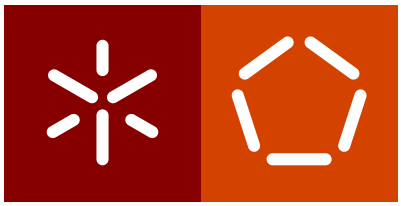
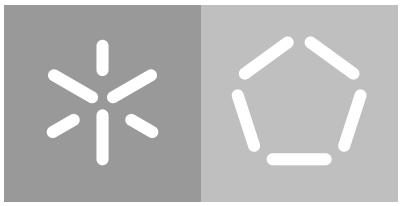


Universidade do Minho
Escola de Engenharia
Departamento de Informática

João Carlos Sequeira da Costa

Development of an automated pipeline for meta-omics data analysis

October, 2017



Universidade do Minho
Escola de Engenharia
Departamento de Informática

João Carlos Sequeira da Costa

Development of an automated pipeline for meta-omics data analysis

Master dissertation
Master Degree in Computer Science

Dissertation supervised by
Andreia Filipa Ferreira Salvador
Miguel Francisco Almeida Pereira Rocha

October, 2017

AGRADECIMENTOS

Ao fazer esta tese, pude contar com o apoio de várias pessoas sem as quais este trabalho seria apenas uma sombra do que se tornou. Este é o meu obrigado a quem em mim acreditou.

À Andreia, pois sem uma boa orientadora muito dificilmente se faz uma boa tese. Tanta correção foi sempre bem aplicada, se não pela minha falta de conhecimento, pelo menos pela minha teimosa desatenção. A uma orientadora exigente até à última mensagem, que nunca deixou de acreditar em mim, o meu obrigado.

Ao professor Miguel, mentor disponível para todos os momentos em que se requeria a sua atenção. A um excelente professor, o meu obrigado.

Ao Vítor, que quando eu estava a um passo da solução, ajudou-me a resolver problema atrás de problema. A um bom amigo, o meu obrigado.

À minha Catarina, parceira nas alturas entusiasmantes, em que tudo corria bem, e nas complicadas, em que tudo era desafio. A ti, que aturaste tudo o que em mim havia a aturar, e que também nunca perdeste a tua confiança em mim, o meu obrigado.

Porque quando o trabalho aperta a família costuma ser a primeira a ficar à distância, aos meus pais, que para tudo se prontificaram para ver o seu filho dar o último passo rumo a mestre. A quem devo tudo, o meu obrigado.

E a Deus. Porque me mostras o caminho.

ABSTRACT

Knowing what lies around us has been a goal for many decades now, and the new advances in sequencing technologies and in meta-omics approaches have permitted to start answering some of the main questions of microbiology - what is there, and what is it doing? The exponential growth of omics studies has been answered by the development of some bioinformatic tools capable of handling Metagenomics (MG) analysis, with a scarce few integrating such analysis with Metatranscriptomics (MT) or Metaproteomics (MP) studies. Furthermore, the existing tools for meta-omics analysis are usually not user friendly, usually limited to command-line usage.

Because of the variety in meta-omics approaches, a standard workflow is not possible, but some routines exist, which may be implemented in a single tool, thereby facilitating the work of laboratory professionals. In the framework of this master thesis, a pipeline for integrative MG and MT data analysis was developed. This pipeline aims to retrieve comprehensive comparative gene/transcript expression results obtained from different biological samples. The user can access the data at the end of each step and summaries containing several parameters of evaluation of the previous step, and final graphical representations, like Krona plots and Differential Expression (DE) heatmaps. Several quality reports are also generated. The pipeline was constructed with tools tested and validated for meta-omics data analysis. Selected tools include FastQC, Trimmomatic and SortMeRNA for preprocessing, MetaSPAdes and Megahit for assembly, MetaQUAST and Bowtie2 for reporting on the quality of the assembly, FragGeneScan and DIAMOND for annotation and DeSEQ2 for DE analysis.

Firstly, the tools were tested separately and then integrated in several python wrappers to construct the software Meta-Omics Software for Community Analysis (MOSCA). MOSCA performs preprocessing of MG and MT reads, assembly of the reads, annotation of the assembled contigs, and a final data analysis.

Real datasets were used to test the capabilities of the tool. Since different types of files can be obtained along the workflow, it is possible to perform further analyses to obtain additional information and/or additional data representations, such as metabolic pathway mapping.

RESUMO

O objectivo da microbiologia, e em particular daqueles que se dedicam ao estudo de comunidades microbianas, é descobrir o que compõe as comunidades, e a função de cada microrganismo no seio da comunidade. Graças aos avanços nas técnicas de sequenciação, em particular no desenvolvimento de tecnologias de *Next Generation Sequencing*, surgiram abordagens de meta-ómicas que têm vindo a ajudar a responder a estas questões. Várias ferramentas foram desenvolvidas para lidar com estas questões, nomeadamente lidando com dados de Metagenómica (MG), e algumas poucas integrando esse tipo de análise com estudos de Metatranscriptómica (MT) e Metaproteómica (MP). Além da escassez de ferramentas bioinformáticas, as que já existem não costumam ser facilmente manipuláveis por utilizadores com pouca experiência em informática, e estão frequentemente limitadas a uso por linha de comando.

Um formato geral para uma ferramenta de análise meta-ómica não é possível devido à grande variedade de aplicações. No entanto, certas aplicações possuem certas rotinas, que são passíveis de serem implementadas numa ferramenta, facilitando assim o trabalho dos profissionais de laboratório. Nesta tese, uma pipeline integrada para análise de dados de MG e MT foi desenvolvida, pretendendo determinar a expressão de genes/transcriptos entre diferentes amostras biológicas. O utilizador tem disponíveis os resultados de cada passo, sumários com vários parâmetros para avaliação do procedimento, e representações gráficas como gráficos Krona e heatmaps de expressão diferencial. Vários relatórios sobre a qualidade dos resultados obtidos também são gerados. A ferramenta foi construída baseada em ferramentas e procedimentos testados e validados com análise de dados de meta-ómica. Essas ferramentas são FastQC, Trimmomatic e SortMeRNA para pré-processamento, Megahit e MetaSPAdes para assemblagem, MetaQUAST e Bowtie2 para controlo da qualidade dos contigs obtidos na assemblagem, FragGeneScan e DIAMOND para anotação e DeSEQ2 para análise de expressão diferencial.

As ferramentas foram testadas uma a uma, e depois integradas em diferentes wrappers de python para compôr a Meta-Omics Software for Community Analysis (MOSCA). A MOSCA executa pré-processamento de reads de MG e MT, assemblagem das reads, anotação dos contigs assemblados, e uma análise de dados final

Foram usados dados reais para testar as capacidades da MOSCA. Como podem ser obtidos diferentes tipos de ficheiros ao longo da execução da MOSCA, é possível levar a cabo análises posteriores para obter informação adicional e/ou representações de dados adicionais, como mapeamento de vias metabólicas.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

API Application Programming Interface.

BLAST Basic Local Alignment Search Tool.
bp Base Pair.

CDD Conserved Domains Database.
cDNA complementary DNA.
CDS Coding Sequence.

DE Differential Expression.
DNA Deoxyribonucleic acid.
dsDNA double stranded DNA.

EBI European Bioinformatics Institute.
EMBL-EBI European Bioinformatics Institute.

FMAP Functional Mapping and Analysis Pipeline.

Gb Gygabase.
GC Guanine/Cytosine.
GUI Graphical User Interface.

HMM Hidden Markov Models.
HPLC High Performance Liquid Chromatography.

IMP Integrated Meta-omic Pipeline.

Kb Kylobase.
KEGG Kyoto Encyclopedia of Genes and Genomes.
KFU KEGG Filtered UniProt.
KO KEGG Orthologous.

LCFA long-chain fatty acids.

Mb Megabase.
MG Metagenomics.
MM Meta-metabolomics.
MOSCA Meta-Omics Software for Community Analysis.

MP Metaproteomics.
MPA Meta Proteome Analyzer.
mRNA Messenger RNA.
MS Mass Spectrometry.
MT Metatranscriptomics.

NCBI National Center for Biotechnology Information.
NGS Next Generation Sequencing.
NR Non redundant.

ORF Open Reading Frame.
OTU Operational Taxonomic Unit.

PCR Polymerase Chain Reaction.
PE Paired end.
PIR Protein Information Resource.

RNA Ribonucleic acid.
RPK Reads per kilobase.
RPKM Reads Per Kilobase per Million.
rRNA Ribossomal RNA.

SE Single end.
SIB Swiss Institute of Bioinformatics.
SMRT Single Molecule Real Time.
SNP Single nucleotide polymorphism.
SOLiD Sequencing by Oligo Ligation Detection.

Tb Terabase.
TPM Transcripts per million.

UniProt Universal Protein Resource.

ACRONYMS

API Application Programming Interface.

BLAST Basic Local Alignment Search Tool.
bp Base Pair.

CDD Conserved Domains Database.
cDNA complementary DNA.
CDS Coding Sequence.

DE Differential Expression.
DNA Deoxyribonucleic acid.
dsDNA double stranded DNA.

EBI European Bioinformatics Institute.
EMBL-EBI European Bioinformatics Institute.

FMAP Functional Mapping and Analysis Pipeline.

Gb Gygabase.
GC Guanine/Cytosine.
GUI Graphical User Interface.

HMM Hidden Markov Models.
HPLC High Performance Liquid Chromatography.

IMP Integrated Meta-omic Pipeline.

Kb Kylobase.
KEGG Kyoto Encyclopedia of Genes and Genomes.
KFU KEGG Filtered UniProt.
KO KEGG Orthologous.

LCFA long-chain fatty acids.

Mb Megabase.
MG Metagenomics.
MM Meta-metabolomics.
MOSCA Meta-Omics Software for Community Analysis.

MP Metaproteomics.
MPA Meta Proteome Analyzer.
mRNA Messenger RNA.
MS Mass Spectrometry.
MT Metatranscriptomics.

NCBI National Center for Biotechnology Information.
NGS Next Generation Sequencing.
NR Non redundant.

ORF Open Reading Frame.
OTU Operational Taxonomic Unit.

PCR Polymerase Chain Reaction.
PE Paired end.
PIR Protein Information Resource.

RNA Ribonucleic acid.
RPK Reads per kilobase.
RPKM Reads Per Kilobase per Million.
rRNA Ribossomal RNA.

SE Single end.
SIB Swiss Institute of Bioinformatics.
SMRT Single Molecule Real Time.
SNP Single nucleotide polymorphism.
SOLiD Sequencing by Oligo Ligation Detection.

Tb Terabase.
TPM Transcripts per million.

UniProt Universal Protein Resource.

1

INTRODUCTION

1.1 CONTEXT AND MOTIVATION

Next Generation Sequencing (NGS) technologies have evolved rapidly during the last years, generating a large amount of sequencing data obtained from a large variety of organisms. Metagenomics (MG), Metatranscriptomics (MT) and Metaproteomics (MP) refer to the study of the genome, transcriptome and proteome of more than one organism occurring in a biological sample, respectively. In nature, microorganisms rarely occur isolated and their metabolism depends on relationships that can be established with the environment (other microorganisms, plants and animals, organic and inorganic materials).

The microbiology of several biotechnological processes is still poorly described due to the difficulty on studying complex microbial communities. This knowledge is crucial for the optimization of biotechnological processes which depend on the activity and interaction of highly diverse microbial communities.

The bioinformatics tools utilized for the study of an isolated organism are usually not suitable for the analysis of diverse communities. Tools for studying complex microbial systems are usually in house developed tools, non-available for the scientific community. There are also freely available tools but usually require deep knowledge on bioinformatics, which most biologists and engineers don't have. Also, some existing tools are not flexible because they were created to address specific questions and to deal with specific data formats.

More recently, open-source pipelines for analyzing meta-omics data have been developed, and several advantages, but also drawbacks can be pointed out. One of the major limitations of both MP and MT, but also MG analysis, is the difficulty in choosing the reference database for mapping and identifying the expressed genes and proteins. Using MG data as a reference to identify transcripts and proteins can greatly increase the identification rates (Heyer et al., 2017). In order to further increase the number of genes/proteins identified, *de novo* options, such as assembling without a reference, or aligning nucleotidic reads against more general databases, can be considered, which do not rely only in reference-based approaches and are particularly relevant in the analysis of complex microbial communities which are poorly characterized. Integrated Meta-omic Pipeline (IMP) (Narayanasamy et al.,

2016) and Functional Mapping and Analysis Pipeline (FMAP) (Kim et al., 2016) already consider some of these aspects. These and other pipelines were reviewed and compared with the pipeline developed in the framework of this thesis, Meta-Omics Software for Community Analysis (MOSCA).

For the various tasks of meta-omics analysis, several tools were tested in their direct, command line interface, and then integrated in python wrappers. Several options exist for every task at hand, but only one tool was chosen for every task, with the exception of the assemblers, because of the reported difference between MetaSPAdes and Megahit results (Vollmers et al., 2017), and the databases for annotation, since different biological databases may contain profoundly different information available for the same sequences. The end goal is always to provide files in useful formats for posterior handling, and the output files from each step are made available.

Testing the tool with Deoxyribonucleic acid (DNA) and Ribonucleic acid (RNA) samples collected from anaerobic bioreactors, containing both MG and MT information, allowed for a comprehensive application of the entirety of the workflow.

1.2 OBJECTIVES AND PLAN

The objective of this thesis was to develop an easy-to-use bioinformatics pipeline for integrated MG and MT data analysis, allowing the comparative analysis between different samples.

To attain this global objective, the following specific aims were defined:

1. To review state-of-the-art pipelines for meta-omic analysis, identifying their advantages and disadvantages, and the tools integrated in their workflows.

2. To create a mostly automated pipeline by only requiring the user input at the beginning to give general information on the type of the data, the place where the data is stored, the assembler to use in the assembly step and the database for the annotation step.

3. To integrate and adapt selected tools which can improve data analysis and comparison, and define and construct workflows for different types of analyses.

4. To test selected tools with real datasets, obtained from anaerobic digesters from the host research group.

1.3 THESIS ORGANIZATION

This thesis presents the construction of MOSCA, and the reasoning behind its build up and necessity. Chapter 2 exposes the fields of MG and MT, their possibilities and challenges. It exposes the major steps in the bioinformatic workflow necessary to obtain meaningful information from the datasets produced in the laboratory, and presents several tools designed for tackling the challenges of the several phases of meta-omics analysis. Several pipelines which integrate these tools into a single software are also explored and the implemented tools and approaches identified.

Chapter 3 explains how MOSCA was designed, the reasoning behind the choice of the tools and how they were tested and integrated. The entire picture of tools, functionalities and input/output files is presented in this chapter.

In chapter 4, the results obtained with real datasets, for testing the pipeline, are shown.

Chapter 5 closes the thesis with the final remarks concerning MOSCA in its present form, and future prospects.

2

STATE OF THE ART

2.1 OVERVIEW OF NEXT GENERATION SEQUENCING TECHNOLOGIES

In 1977, two DNA sequencing techniques were developed, by Frederick Sanger and Walter Gilbert, based on the chain-termination method and chemical modification of DNA respectively. The first generation of sequencing techniques was dominated by Sanger's approach, even though it presented laborious work, used radioactive chemicals, was expensive and slow. This technique went on as the main solution to DNA sequencing until the end of the Human Genome Project.

Even though sequencing is the best tool for comprehensive analysis of several genomic questions, only now is it begining to represent a standard of genomic studies, in part due to the hurdles of performing whole-genome sequencing for every sample and individual. Instead, scientists have had to rely on genome-wide association studies using Single nucleotide polymorphism (SNP)-arrays for the comparison of different genomes (Buermans and Den Dunnen, 2014).

The development of NGS allowed for cheaper and quicker sequencing, distinguishing it from the Sanger approach by enabling massively parallel analysis, and large datasets have emerged as a result. Today, it is possible to sample a certain microbiome and sequence all the genetic material, from the smallest Ribossomal RNA (rRNA) subunit to complete genomes of single organisms or even entire populations, in such low times and with costs that were considered impossible in the recent past (Liu et al., 2012; Buermans and Den Dunnen, 2014; Bahassi and Stambrook, 2014; Van Dijk et al., 2014).

Several NGS techniques have been developed, with a focus on increasing the read length - longer reads are easier to assemble and allow for a better detection of sequencing errors - and the output per run, and decreasing the run time (Quail et al., 2012).

2.1.1 *Roche 454*

454 GS FLX was the first NGS technology viable enough for the market, released in 2005 by 454, following the Human Genome Project, and is based on pyrosequencing. After 454 was

purchased by Roche, two more platforms were developed, namely the FLX+ and the GS Junior systems, which are improvements over the same technology. The advantage of these systems is the larger read length and fast workflow - allowing for 10 hour sequencing - but their smaller outputs have put the competition ahead, and Roche is putting their activity to an end (Liu et al., 2012; Buermans and Den Dunnen, 2014).

2.1.2 *Life technologies*

Sequencing by Oligo Ligation Detection (SOLiD) was developed by Agencourt in 2006, and purchased by Applied Biosystems the same year. The technology of two-base sequencing based on ligation sequencing is capable of generating paired end reads, and the fact that each base is interrogated twice by octomer ligations increases the read accuracy to 99.99%, which represents a main advantage of this method, even though the output is only half the maximum of the competition (Liu et al., 2012; Buermans and Den Dunnen, 2014).

Ion Torrent, developed on the 350 nmCMOS technology, makes use of an oil-water emulsion - thus giving the name emulsion Polymerase Chain Reaction (PCR) to the process - to partition small reaction vesicles, each with ideally one reaction sphere, a library molecule and all the reagents needed for the process, where the PCR reaction will take place. The large output of this method is hindered by some problems of emulsion PCR, besides the normal complications of PCR: it is hard to get one library molecule per reaction vesicle - only about 1/3 of vesicles will have that 1 molecule to 1 vesicle ratio and extraction of the spheres from the emulsion is still not efficient. Nucleotide incorporation is detected through quantification of proton presence in the medium, calculated through pH change of the medium. The lack of an imaging step (unlike Roche's luciferases' reaction or Illumina's fluorescent imaging) leads to a significant decrease in run time. Continuous development of the Ion torrent technology increased the output level from 10Mb up to 1 Gygabase (Gb), and the average read length from 100Base Pair (bp) up to 400bp (Buermans and Den Dunnen, 2014).

Ion Proton Systems were first applied on the Proton-I chips, which were developed on the 110 nmCMOS technology, allowing for a decrease in sphere and censor wells diameter and an increase in number of wells to 165 million per chip and in output up to 8-10 Gb per run. In 2015, the Proton-II chips, which possess double the number of wells because of the corresponding decrease in the sizes of the vesicles, were released, with the announced larger output, but not the decrease in run time.

2.1.3 *Illumina*

Solexa, which developed the Genome Analyzer in 2006, was purchased by Illumina, which improved the technology of sequencing by synthesis used by Genome Analyzer. On 2010, Illumina launched the HiSeq 2000, which had the biggest output per run up to 600 Gb, that can be obtained in 8 days, but with an error rate of almost 2%. Nevertheless, it became the cheapest solution per base when compared with 454 and SOLiD. The second sequencer developed by Illumina was the MiSeq, which has shorter run times and outputs, being more indicated for amplicon sequencing and bacterial genome sequencing (Liu et al., 2012; Buermans and Den Dunnen, 2014).

In 2014, Illumina released two new sequencing models, the NextSeq 500 and the HiSeq X Ten. NextSeq 500 was devised as a smaller, more flexible version of HiSeq 2000, allowing for two modes of operation, both with much less time per run but also with substantially less output, and with two modes of data output, one with up to 60 Gb and another with up to 120 Gb of data output. HiSeq X Ten represents a truly revolutionary breakthrough by seeking the 1000$ genome goal - the challenge of creating a technology capable of sequencing a human genome with less than 1000$ of cost. Introduction of patterned flowcells has allowed for much compacted clusters, and with the application of this method to HiSeq X Ten. This Illumina technology is now regarded as the best option for MG studies in which genomes are sequenced (Illumina, 2015).

2.1.4 *Pacific Biosciences*

The single molecule real-time sequencing technique used by Pacific Biosciences' RSII distinguishes itself from previous sequencing techniques in the fact that it is sensitive enough to detect the incorporation of a single, fluorescently labeled nucleotide, so this method has no need for amplification steps. Although library preparation follows the common workflow of the other methods, it does have some major perks: the adapters have a hairpin structure (Single Molecule Real Time (SMRT) loop adapters), which leads to the double stranded DNA (dsDNA) to become circular after ligation, and the quantity of DNA required for building the library is high, possibly limiting high for some studies such as ChIP-Seq or single-cell genomics.

During sequencing, a molecule may be read several times, depending on a combination of insert size and read length. RSII has no cycles of nucleotide incorporation alternated with imaging or staining, relying instead on a real-time approach, recording the incorporation of the nucleotides at 75 frames per second with use of a powerful optical system, with each kind of nucleotide having its own label. Finally, the enzyme used is a modified version

of phi29, which has no Guanine/Cytosine (GC) bias, high read length, low error rate and strand displacement properties, coming at a cost of decreased 3-5 exonuclease activity.

Even though this process generates a large amount of sequencing errors (10-15%, mostly comprised of insertions/deletions), these errors are randomly distributed across the sequenced molecule, as opposed to the other techniques in which the error rate increases towards the end of the sequences. This allows for an alignment of multiple reads for the same areas of the sequenced molecule to remove most of those errors, and by use of the PacBio Quiver software the error rate may decrease to as low as 0.001%. The long read data, absence of GC bias and insight into the kinetic state of the polymerase during sequencing directs the use of this technique to approaches involved in the study of small genomes, since RSII produces a small output (Buermans and Den Dunnen, 2014).

Table 1 summarizes the main principles of different sequencing technologies and the expected output.

A big progress has went off ever since the days in which 454 pyrosequencer was the only reliable NGS technology, and several technologies exist and continue to emerge in an effort to present the cheapest, quickest, and most reliable technology of NGS. Now that the technology to produce the datasets is available, the big challenge is, and will continue to be in the next years, to store all that data and devise computational solutions for the organization and analysis of such data (Illumina, 2015; Buermans and Den Dunnen, 2014; Bahassi and Stambrook, 2014; Van Dijk et al., 2014).

2.1.5 *Paired-end vs single-end*

With the present higher sequencing capacity, it has become usual to sequence in a paired way, in which the same DNA fragment is sequenced from both ends. This translates in two files with the same number of reads, where the first sequence of the forward file corresponds to the same DNA from which comes the first sequence of the reverse file. All steps in the preprocessing phase must output two files with the same number of reads, or one with the reads from both files interleaved. These sequences may overlap or not, but represent additional clues to the original sequence. By handling paired end data the right way, this is, treating both files as fragments of the same sequences, removal of undesired sequences will be more accurate, and so will the assembling.

2.2 METAGENOMICS

Less than 2% of bacteria can be cultured in laboratory, which immediately raises the question of how can we study the remaining 98% that make up the backbone of most of Earth's ecosystems (Illumina Proprietary, n.d.). An answer has come in the form of MG, which

Table 1: Summary of the main characteristics of NGS technologies (Buermans and Den Dunnen, 2014).

Company	Technology	Sequence by	Detection	Run types	Run time	Read length (bp)	# reads per run	Output per run
Roche	GS FLX Titanium	Synthesis	Pyrophosphate detection	Single and	23h	700	1 million	700 Mb
	GS Junior System	Synthesis	Pyrophosphate detection	Single end	10h	400	0.1 million	40 Mb
LifeTechnologies	Ion torrent	Synthesis	Proton release	Single end	4h	200-400	4 million	1.5-2 Gb
	Proton-I	Synthesis	Proton release	Single end	4h	125	60-80 million	8-10 Gb
	Proton-II	Synthesis	Proton release	Single end	8h	100		24 Gb
	Abi/solid	Ligation	Fluorescence detection of di-base probes	Single & paired-end	10 days	75 + 35	2.7 billion	300 Gb
Illumina/solexa	HiSeq2000/2500	Synthesis	Fluorescence; reversible terminators	Single & paired-end	12 days	2 x 100	3 billion	600 Gb
	MiSeq	Synthesis	Fluorescence; reversible terminators	Single & paired-end	65h	2 x 300	25 million	15 Gb
	NextSeq 500	Synthesis	Fluorescence; reversible terminators	Single & paired-end	16h	2x150	400 million	100 Gb
	HiSeq X Ten	Synthesis	Fluorescence; reversible terminators	Single & paired-end	5 days	2x150	6 billion	1.8 Tb
Pacific biosciences	RSII	Single molecule synthesis	Fluorescence; terminally phospholinked	Single end	2 days	50% of reads > 10kb	0.8 million	5 Gb
Helicos	Heliscope	Single molecule synthesis	Fluorescence; virtual terminator	Single end	10 days	30	500 million	15 Gb

involves the partial or complete sequencing of several genomes present in a DNA sample - either from viruses, bacteria or human beings (Illumina Proprietary, n.d.; Oulas et al., 2015; Tringe and Rubin, 2005; Thomas et al., 2012).

MG comes in two ways, shotgun MG and marker gene MG. Shotgun MG is the complete sequencing of all the genomic material in a sample, including protein coding genes, operons and other information present in the genome, with the sub-goal of identifying what organisms compose the community of a certain sample, but having as its main objective identifying the "community potential" of the sample, the collection of genes present and that may be expressed under certain conditions.

In a MG workflow, after sequencing, shorter reads are assembled into larger contigs by reference-based or by *de novo* assembly. One strategy or even both may be used, depending on the dataset in question, the existence of a reference library and the specifications of the research project. Genes will be then annotated and the possible transcripts and proteins expressed will be identified.

Another approach is Marker Gene MG, where in the taxonomic prokaryotic studies, 16S rRNA is usually the target gene when the aim is to get phylogenetic information, since this gene is common to all prokaryotes. Both approaches present a set of complex challenges, which have been tackled over the years with more advanced and specific informatics solutions (Oulas et al., 2015; Overview and Illumina, 2012). Among them, are:

1. PCR noise and errors - single base pair errors, replicate sequence artifacts, PCR chimeras.

2. Deep sequencing - some genes are not abundant, and may not show up on sequencing results, thus underestimating the diversity of the sample.

3. Mosaicism - horizontal transfer may result in incorrect identification of a genome.

4. Intragenomic Heterogeneity - more specific to 16S rRNA gene studies, since bacteria may have several copies of these genes with significant variations for the same genome.

Having opened the possibility of studying new ecosystems, MG allows their study in native conditions, without the need of culturing in the laboratory. The study of many important ecosystems, like the soil and the human microbiome, benefited greatly from this approach, that has already expanded the knowledge concerning the different composition of microbial life in every corner of the biosphere.

MG shotgun studies have already been applied to human feces (Breitbart et al., 2003; Žifčáková et al., 2016), mines (Tyson et al., 2004), Sargasso Sea (et al. Venter, 2004), oil sands tailings ponds (Tan et al., 2015), hydrocarbon-contaminated aquifers (Tan et al., 2015), marine sediments (Urich et al., 2014), soil (Žifčáková et al., 2016), sewage (Žifčáková et al.,

2016), sewage, swine wastewater sample, treated wastewater, river water and drinking water (Žifčáková et al., 2016), among others.

On the other hand, rRNA 16S marker gene MG has, for example, been applied to soil (Pearce et al., 2012), (Gołębiewski et al., 2014), (Damon et al., 2012), geothermal steamvents (Benson et al., 2011), extremely acidic waters (García-Moyano et al., 2012), oxygen minimum zone of the eastern tropical South Pacific (Stevens and Ulloa, 2008) and Tibetan Plateau (Xiong, Liu, Lin, Zhang, Zeng, Hou, Yang, Yao, Knight and Chu, 2012) populations, and is now a routine approach in most laboratories studying microbial communities. But, identifying the microbial taxonomy and genomic potential in a sample might not be enough, the next step forward would be to understand what those identified species are doing there, and this can be achieved by using shotgun MG (Tringe and Rubin, 2005; Thomas et al., 2012).

2.2.1 *Metagenomics pipelines*

Initially, tools developed for single genome approaches were applied in MG, but they were not suitable for dealing with MG data since genomic information is retrieved from several different organisms (Oulas et al., 2015). Today is a different reality, there are already several pipelines developed for the study of MG, many of them available online.

The pipelines are considerably diverse (see Table 2): some have fully integrated workflows for MG, including quality assessment of sequencing raw data, the assembly and the annotation of genes (Arumugam et al., 2010; Treangen et al., 2013). The difficulty in the installation and in the utilization differs depending on the chosen pipeline, and therefore some will require more computational knowledge from the user than others.

Some pipelines were designed to perform shotgun MG analysis only, others are exclusive to 16S rRNA gene analysis (QIIME, Mothur (Plummer and Twin, 2015)), while others possess the flexibility to work with both types of analyses (MG-RAST (Wilke et al., 2016), EBI metagenomics (Hunter et al., 2014)). As in other areas of informatics, a trade-off is made with these pipelines: more powerful computational solutions with more tasks integrated demand more technical knowledge from the user, while more focused tools have a smoother learning curve with intuitive user interfaces, many even available in the web (Ladoukakis et al., 2014; Oulas et al., 2015).

Ladoukakis et al. (2014) compared seven shotgun MG pipelines - CloVR-metagenomics, Galaxy platform (metagenomics), IMG/M, MetAMOS, MG-RAST, RAMMCAP and SmashCommunity (Table 2). From those, MetAmos and SmashCommunity were considered the most robust and versatile solutions, because they integrate all steps of MG bioinformatics analysis and showed the best quality of the results, although they are not the easiest to operate by less experienced users.

Table 2: Main steps of MG data analysis integrated in common pipelines. From (Ladoukakis et al., 2014).

Pipeline \ Tasks	Quality control	Assembly	Gene detection	Functional annotation	Taxonomic analysis	Comparative analysis	Data management
CloVR-metagenomics	✗	✗	✓	✓	✓	✓	✓
Galaxy platform*	✓	✗	✗	✗	✓	✓	✗
IMG/M	✗	✗	✓	✓	✓	✓	✓
MetAMOS	✓	✓	✓	✓	✓	✓	✓
MG-RAST	✗	✗	✓	✓	✓	✓	✓
RAMMCAP	✗	✗	✓	✓	✓	✓	✗
SmashCommunity	✓	✓	✓	✓	✓	✓	✓

Dudhagara et al. (2015) also reviewed 12 MG pipelines available accessible through the web - MG-RAST, IMG/M, METAREP, CoMet, METAGENassist, MetaABC, MyTaxa, metaMicrobesOnline, Coding Sequence (CDS) Metagenomics, CAMERA, METAVIR and VIROME. These studies consider MG-RAST and IMG/M the best options for the functional analysis for already assembled data, mostly because they presented an easy interface, a database setup designed for dissemination of results to the scientific community and also because of their considerably bigger databases and a wide range of annotation tools.

Concerning 16S rRNA studies, QIIME is the most used pipeline for the corresponding taxonomic interpretation (Oulas et al., 2015), being the fastest and producing similar results to mothur and MG-RAST (Plummer and Twin, 2015).

2.3 METATRANSCRIPTOMICS

To know what pathways the microorganisms are utilizing can be translated in knowing what genes are being expressed more intensively, which leads to the study of Messenger RNA (mRNA). When collectively studying the transcriptome of several organisms in the same ecossystem, it is called MT.

MT analysis integrated with MG have recently been developed, since the differential analysis of the abundance of the mRNA sequences reveals important pathways used by the community and indicate for instance which organisms are vital for maintaining the balance of microbiomes, or through what mechanisms do microorganisms survive extreme conditions.

Indicating the real activity of a microbial population, MT functions as a complement to MG, which only informs about what pathways can be used - the genomic potential of the population (Bikel et al., 2015; Aguiar-pulido et al., 2016). MT presents simpler datasets than MG, since mRNA confines to the coding regions of the genome and not all genes are

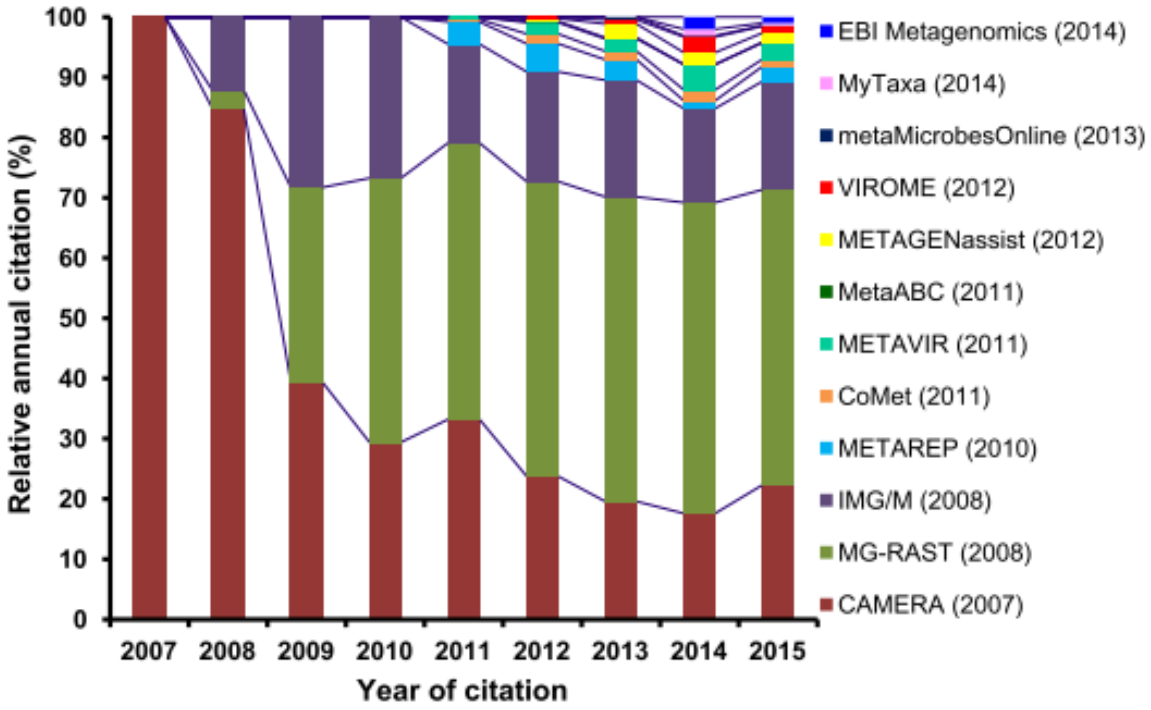


Figure 1: Utilization of web resources of *MG* analysis throughout the years. From Dudhagara et al. (2015).

transcribed under a certain condition, thus allowing for less complex datasets that generate more focused and useful functional information.

MT distinguishes itself from MG in some steps of its analysis:

- For studies aiming at the study of mRNA, a depletion of rRNA is necessary, at the wet- and dry-lab level (Narayanasamy et al., 2016).
- mRNA is very unstable, which might be a factor for ruining the sample before sequencing
- It is harder to distinguish between bacterial and other living beings RNA, which might be a serious problem in human microbiome studies.
- Overrepresented sequences that would be discarded in MG as undesired are normal in a MT dataset, and therefore to be kept.
- The size of RNA fragments is usually smaller than that of DNA, so each RNA sequence is usually sequenced much more times than DNA sequences, thus allowing for more reliable results of sequencing.

With the development of kits of enrichment of bacterial RNA and several techniques of rRNA depletion, several and diverse studies of MT have already been developed (Aguiar-

pulido et al., 2016). Several works have been developed in MT functional and differential analysis, such as the ones applied to soil (Carvalhais et al., 2012), stimulus-induced biofilms (Ishii et al., 2015), mouse intestine (Xiong, Frank, Robertson, Hung, Markle, Canty, McCoy, Macpherson, Poussier, Danska and Parkinson, 2012), kimchi (Jung et al., 2013), bovine rumen (Poulsen et al., 2013) and deep-sea populations (Baker et al., 2013), among others.

2.3.1 *Metatranscriptomics pipelines*

MT pipelines follow basically the same steps and have to deal with most of the problems of their MG counterparts, in addition to the overwhelming presence of rRNA in comparison to mRNA. In four steps - preprocessing of reads, annotation of contigs, aggregation of the annotated contigs, and analysis of results - MT pipelines obtain information about the entire transcriptome, and therefore on the metabolic pathways expressed and also on taxonomic composition of complex microbial communities (Westreich et al., 2016).

Until recently, there was no pipeline fully integrating the steps of assembly and annotation for MT data. A comparison of four assemblers - Trinity, Oases, Metavelvet and IDBA-MT - revealed Trinity as the most reliable in assembling more reads to contigs with annotation value (Celaj et al., 2014). In 2016, the first pipelines to integrate all the steps of metatranscriptomics were released: MetaTrans and SAMSA. MetaTrans performs both taxonomic - making use of 16S rRNA - and gene expression analysis of RNA-Seq, after quality-control assessment and rRNA removal. It uses complementary DNA (cDNA) libraries for PE sequencing, and maps them against functional databases (Martinez et al., 2016). SAMSA identifies the more prominent species and the functional differences between MT datasets, which allows for multisample comparison. It does require reads to be longer than 100 bp or paired-end, however, which is not always easy to fulfill (Westreich et al., 2016).

2.4 Integrated analysis of metatranscriptomics coupled to metagenomics

To be able to figure out which pathways are being most expressed in a given sample, it is necessary to integrate MG and MT data analysis (Dudhagara et al., 2015; Nayfact et al., 2016). However, there are only a few publicly available pipelines developed for multiomics approaches. Very recently, the first publicly available pipelines integrating MG and MT analysis have been presented, and the two available to date will be described in detail below.

IMP and FMAP are two available pipelines developed to analyze simultaneously MG and MT data for a better interpretation of MT studies. IMP is an open-source pipeline designed for the preprocessing, assembly and analysis of MG and MT data (Narayanasamy

et al., 2016). The two main features that distinguish it from other pipelines are its iterative co-assembly of MG and MT reads and its containerization in docker - allowing for reproducibility of its results. Docker (Chamberlain and Schommer, 2014) is a virtual machine whose environment may be built to reproduce the computational environment of the researcher at the time of his work, in an easy to build, accessible way. IMP is the only pipeline to integrate MG and MT reads by iterative co-assembly, which greatly increased the identification of protein coding genes when compared to results obtained based on assembly of MG reads only. However, Differential Expression (DE), the analysis of diferential expression of genes, was not considered in this pipeline.

The preprocessing steps include trimming and quality filtering (by *Trimmomatic*) and rRNA filtering (by *SortMeRNA 2.0*). Assembly is achieved through read mapping (by *bwa*), extracting unmapped reads (by *samtools* and *BEDtools*) and filtering host sequences. After assembly, *VizBin* is used for binning, and in the analysis step, *Prokka* is used for the annotation and *VizBin* for a detailed and interactive look at the results. Written in Bash, Make and Python, docker and python are the only requirements previous to the installation of the IMP for using the tool, allowing for a easier installation and use of the pipeline (Narayanasamy et al., 2016).

FMAP is the only implementation to date that handles both MG and MT analysis while also performing DE analysis with support from the MG information, which is vital to understand the functional behaviour of a community. Besides that, FMAP also implements more typical MG features, like sequence alignment and determination of gene families presence.

In the preprocessing step of the workflow, the usual removal of low-quality and human sequences is done through *BMTagger*, and the alignment of the remaining reads goes through *USEARCH* or *DIAMOND*, with a KEGG Filtered UniProt (KFU) reference cluster as database - enriched in bacteria, fungi and archaea sequences, for more robust and informative results. The result of this alignment may be extracted for annotation with another software.

After assembly, gene abundance is determined by raw count or Reads Per Kilobase per Million (RPKM), and for differential quantification, *metagenomeSeq*, using raw count, and Kruskal-Wallis and quasi-Poisson, both using RPKM, are the tools selected for that step. The three show variable quality of results depending on the situation, and so the three are implemented in FMAP.

Enriched operons analysis is based on the differential gene abundance, considering differential abundant the operon corresponding to the gene differentially abundant. The definition of operon in FMAP is tied to the KEGG Orthologous (KO) concept, where each KO corresponds to a molecular-level function. In FMAP, the differential abundance of operons also means the differential abundance of its corresponding pathway, and another output of

FMAP is an input to Kyoto Encyclopedia of Genes and Genomes (KEGG) online pathway map tool, where it is possible to easily visualize which pathways are more represented in the sample, among all pathways available in KEGG (Kim et al., 2016).

A limitation of FMAP is that it was not designed for comparison of data between different samples, although having ShotgunFunctionalizeR in its workflow, an R package specific for functional comparison of metagenomes of different samples (Kristiansson et al., 2009).

A typical workflow for MG/MT studies, including the main steps and tools for a complete bioinformatics data analysis, is presented in figure 2.

2.5 Steps and tools for MG/MT data analysis

NGS methods can produce large amounts of data, which are necessary in meta-omics studies. There are optimized informatics tools that handle such large amounts of data and that were designed for the different steps of meta-omics data analysis, namely quality control, preprocessing, assembly, bining, annotation and statistical and visual analysis. Several of these tools have been integrated in pipelines, to allow for easier workflows (Table 3). A brief overview of the main steps and most common tools utilized in each step of MG and/or MT data analysis will be given.

2.5.1 *Steps and tools for Preprocessing*

The preprocessing mainly consists on the removal of undesired sequences from NGS datasets. All pipelines include a preprocessing phase, but some integrate only a small number of steps, like MetAMOS, while others offer more complex preprocessings, like IMP and MetaTrans (Table 3).

After the quality of the sequencing reads has been inspected, preprocessing prepares the dataset for the further steps, by removing undesirable sequences. Trimming is to remove the sequences of less interest from the datasets, either because they are too short, they are from species not in the scope of the study or there is too much doubt about the consensus sequence, the latter being quantified by the scores relative to each position of the read. Trimmomatic (Bolger et al., 2014) is one of the solutions for removing low quality and short sequences and BMTagger (Rotmistrovsky and Agarwala, 2011) is capable of identifying and removing human sequences. When the work involves study of mRNA, a depletion of rRNA sequences is necessary. SortMeRNA (Kopylova et al., 2012) identifies the rRNA sequences and removes them from the dataset.

Quality check of FASTQ reads In the first *in silico* step of a post-NGS bioinformatics workflow, the datasets of reads produced by sequencing usually undergo quality check.

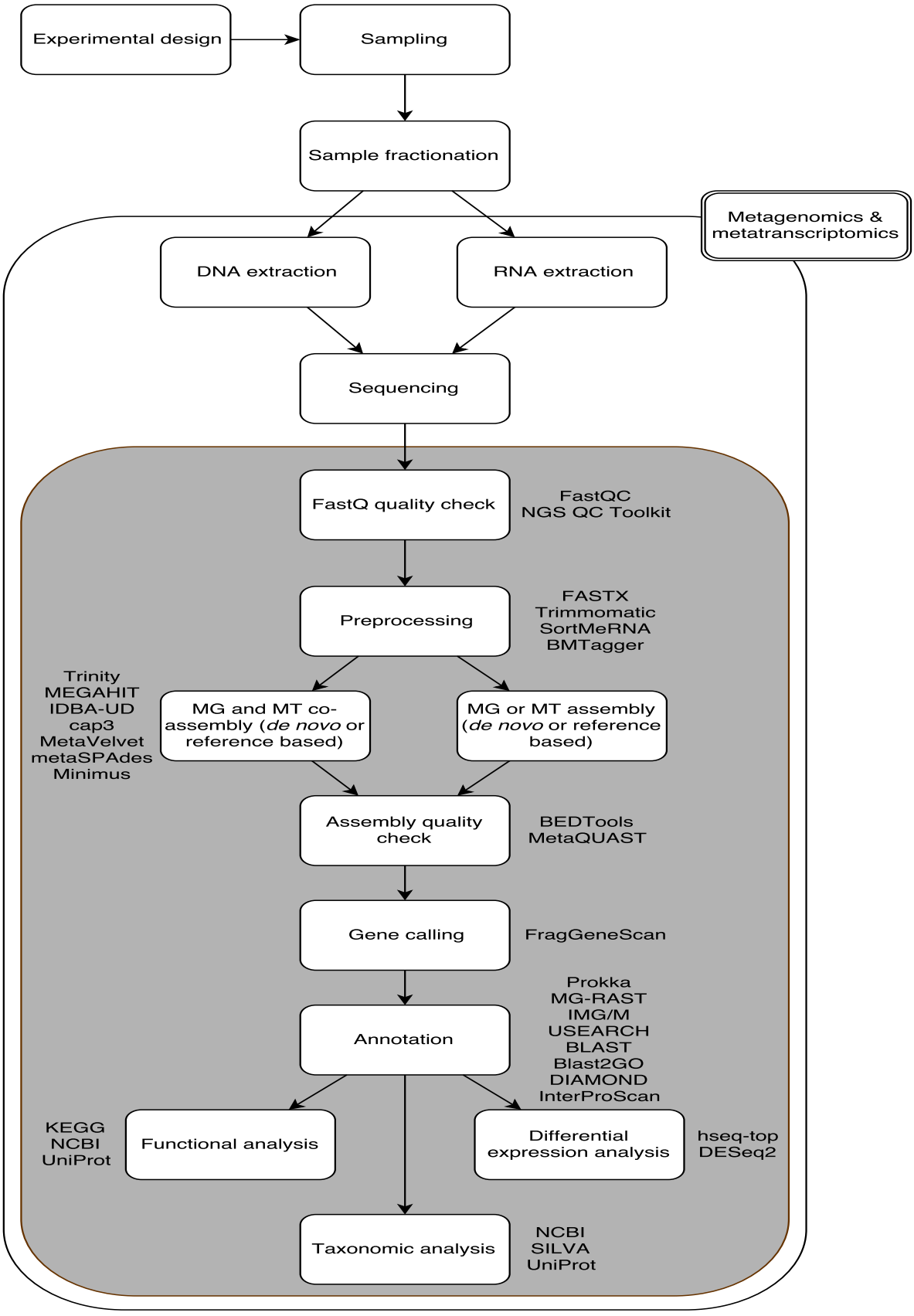


Figure 2: Typical Meta-Omics pipeline workflow, from the laboratory to the bioinformatics processing steps. Inside the squares are the major steps. The tools and software possibilities are represented next to each step of the workflow.

Table 3: Comparison of different steps and tools present in some MG and MT pipelines.

	Steps \ Pipelines	IMP	SAMSA	MetaTrans	FMAP	MOCAT2	MG-RAST	metagenomics	MetAMOS
Preprocessing	Quality check	FastQC	None	FastQC	None	SolexaQA	DRISEE, kmer profiles, nucleotide histograms	None	FastQC
	Removal of low quality sequences/regions	Trimmomatic	Trimmomatic	Kraken tools	BMTagger	FastX, SolexaQA, custom scripts	SolexaQA	Trimmomatic, Biopython	fastx_toolkit
	Adapter removal	Not specified	Trimmomatic	Kraken tools	None	Usearch	None	Biopython	None
	rRNA removal	SortMeRNA	None	SortMeRNA	None	None	BLAT	rRNASelector	None
	Host sequences removal	bwa	None	None	BMTagger	SOAPAligner2	Bowtie	None	None
	Alignment of paired end reads	None	FLASh	Kraken tools	None	None	None	SeqPrep	None
Assembly	Assembly	IDBA-UD, MEGAHIT	None	None	None	SOAPdenovo	None	None	SOAPdenovo, Newbler, Velvet, Velvet-SC, MetaVelvet, Meta-IDBA, CABOG and Minimus
	Quality control of assembly	MetaQUAST	None	None	None	SOAPdenovo, BWA	None	None	bowtie, bowtie2, Bambus2, several plots
Annotation	Functional annotation	Prokka	MG-RAST	SOAP2 (with in house scripts)	USEARCH, DIAMOND	fetchMG, DIAMOND, perl script	sBLAT	InterProScan	BLAST, MetaPhyler, PHMMER, PhyloSift, PhymmBL, FCP
	Taxonomic annotation	MetaQUAST	MG-RAST	SOAP2 (with in house scripts)	None	fetchMG, mOTU-LG, specI	sBLAT	Qiime	BLAST, MetaPhyler, PHMMER, PhyloSift, PhymmBL, FCP
	Depth of coverage calculation	BEDtools	None	None	None	None	Not specified	None	Repeatoire
	Variant calling	Samtools mpileup, Freebayes, Platypus	None	None	None	None	None	None	None
	Binning	VizBin	None	None	None	None	Not specified	None	None
	Gene calling	MetaQUAST	MG-RAST (FragGeneScan)	FragGeneScan	None	Prodigal, MetaGeneMark	FragGeneScan	rRNAselector, FragGeneScan	MetaGeneMark, FragGeneScan, Glimmer-MG
Results analysis	Taxonomic analysis	Barplots	Barplots	Plots	None	None	phyloseq, rank abundance, rarefaction	Pie chart, barplot, Krona	Krona, heatmaps
	Abundance of pathways	Heatmap, Krona	Barplots	DESeq2, iPath2	KEGG, ShotgunFunctionalizeR	None	heatmap, Krona	None	Krona
	Variants analysis	VizBin	None	None	None	None	None	None	Sequence alignment
	Species sequences detection	VizBin	None	None	None	None	None	None	None
	Depth of coverage	VizBin	None	None	None	None	None	None	R
	General sequences information	None	None	None	None	None	Flowchart	None	Plots

FASTQ format is an usual output file format of NGS, but a quality control tool should also have the capacity to analyze SAM, a compressed version of FASTQ, and BAM files, the binary version of SAM (Jones et al., 2012). In addition, two scoring formats, PHRED-33 and PHRED-64, originate two different FASTQ versions depending on the NGS technology used

(Cock et al., 2010). Finally, a quality check should output statistical and graphical analyses of several subjects about the quality of the datasets being studied. FastQC (Andrews et al., 2010) and NGS QC Toolkit (Patel and Jain, 2012) are two solutions for this task, with FastQC being the most used and validated.

FastQC (Andrews et al., 2010) is currently the most popular tool at checking the quality of NGS data, but other options exist, like DRISEE for calculating sequencing errors and custom application of histograms for measuring presence of bases in every position of the dataset, and SolexaQA (Cox et al., 2010), which posses several informative graphical results aswell. FastQC retrieves a quality report showing the important characteristics of NGS data in boxplots, line plots and including a colored code indicating if there is any quality problem and at which level. This analysis is fast, easy to use and to interpret, which has led FastQC to become the most used quality control tool. The strength of FastQC is reflected in the fact that several pipelines used FastQC to generate the initial quality control report, and in further steps during the preprocessing stage.

UNDESIRED SEQUENCES REMOVAL After the quality of the sequencing reads has been inspected, preprocessing prepares the dataset for the further steps, by removing undesirable sequences. Trimming is to remove the sequences of less interest from the datasets, either because they are too short, from species not in the scope of the study, there is not enough confidence about the consensus sequence, quantified by the scores relative to each position of the read, or there are artificial sequences from the experimental work of obtaining the nucleotidic sequences.

Trimmomatic (Bolger et al., 2014) is becoming more and more popular as a toolbox for tailoring NGS datasets in many different ways, most concerned with data quality, making use of the quality related information contained in FastQ files. Different tools allow for a versatile response to data problems, which is translated into a powerful and versatile solution for quality and artificial sequences trimming. These are some of the reasons that explain its incorporation in many pipelines, mainly for removing artificial sequences and for cropping or removing entire sequences. SolexaQA and FastX (Gordon and Hannon, 2010) are examples of alternative toolboxes available for the bioinformatics community, and can be found integrated in some pipelines as well.

RRNA AND HOST SEQUENCES REMOVAL Because most of the pipelines are focused on MG data analysis, which contains much less rRNA sequences than MT data, removal of rRNA sequences is not widespread throughout their preprocessing workflows. For the identification and removal of rRNA sequences in NGS datasets, the choice of both software and reference database is highly important. SILVA databases (Quast et al., 2012) have become the golden standard among rRNA databases, containing all types of rRNA sequences

in a single resource: all domains - prokaryotic, archaea and eukaryotic - and all rRNA subunits - 16S, 23S, 18S and 28S. Different tools can be used to map the datasets to databases, which use different approaches. For example, prior to alignment of sequence reads to databases, SortMeRNA (Kopylova et al., 2012) and BLAT (Kent, 2002) generate an index for each database, while rRNASelector (Lee et al., 2011) uses HMMER and Hidden Markov Models (HMM).

Host sequences removal is a very important step, since even when not working with samples from human gut, it still must be assured that the samples are voided of human sequences for submission in certain web servers for annotation and posterior analysis. Like rRNA depletion, it is, however, not implemented in many of the state-of-the-art pipelines for MG and MT data analysis. The most important question is, again, what database to use. Examples of tools for removing human and other hosts derived sequences are BMTagger (Rotmistrovsky and Agarwala, 2011), SOAPaligner (Gu et al., 2013) and Bowtie (Langmead and Salzberg, 2012).

2.5.2 *Bioinformatic tools for Assembly*

Aligning the reads into contigs that represent as closely as possible the sequences present in the original sample is a task already approached with different strategies to make use of the most resources possible. If there are available examples of the target organisms genomes or of closely related species, then a database reference based assembly may be the best solution, which is the strategy of Minimus (Sommer et al., 2007), from the MetAMOS pipeline (Treangen et al., 2013). If there is no closely related genomes available, *de novo* assembly is a solution, for which there is Trinity (Celaj et al., 2014) for MT data assembling, and MetaVelvet (Namiki et al., 2012), metaSPAdes (Nurk et al., 2016), MEGAHIT (Li et al., 2015) and IDBA-UD (Peng et al., 2012) for MG. cap3 (Huang and Madan, 1999) can also be used as a standalone assembler or as a complement to other programs by further merging contigs into bigger ones.

After an assembly, it is important to evaluate the results concerning genes and species detection. BEDTools (Quinlan and Hall, 2010) possesses a suite of tools for that task, while MetaQUAST (Mikheenko et al., 2016) could be used as alternative, by presenting several classic metrics such total assembly size and N50. By reconstructing the assembly back to the original contigs it is possible to have an ideia of the ammount of reads used in the assembly, allowing to better understand how much of the original information is present in the contigs. Such is achieved by aligning the reads to the contigs, using Bowtie2 (Langmead and Salzberg, 2012) or BWA (Li and Durbin, 2009), for example.

Despite its proven usefulness, assembling is not included in most pipelines (Table 3). IMP, MOCAT2 and MetAMOS present examples of implementation of an assembling rou-

tine, with MetAMOS possessing by far the largest collection of assembling options, with eight different tools to choose from, and three types of assembly, single genome/isolate, metagenomic and single cell assembly. Some of its assemblers are not capable of performing all types of assembly, however, and the tool has been specifically designed towards metagenomic analysis (it was built from AMOS, a genome assembly framework).

IMP incorporates two assemblers and the option to co-assemble MG with MT data, in an iterative procedure - after a first assembly, the first contigs and the remaining unused reads serve as input for a second assembly. Additional rounds of assembly could be implemented, but do not show a significant increase in number of contigs after the second assembly.

MOCAT2 (Kultima et al., 2012) assembles the reads into contigs and scaftigs (scaftigs are contigs extended using paired-end information).

QUALITY CONTROL OF ASSEMBLY MetAMOS also implements by far the most complex evaluation and correction of assembling, which starts by mapping the original reads used for the assembling back to the contigs determined by the assemblers (using Bowtie and Bowtie2), and using such mapping to estimate depth of coverage, filter contigs with no reads mapped to them, create links between contigs for scaffolding and re-estimate fragment lengths. Still in MetAMOS, Repeatoire annotates repetitive contigs, allowing for identifying under-collapsed sequence output from the assembling step. Some later applications of other tools also facilitate some annotation steps, and are discussed in the annotation section bellow.

IMP makes use of MetaQUAST to calculate several metrics concerning the contigs produced, such as N50 and number of contigs. Even though this are classic metrics used in the evaluation of assemblies, it lacks an approach such as that of bowtie to deconstruct the assemblies and relate the original reads with the contigs. MOCAT2 makes use of SOAPdenovo and BWA to correct for indels and chimeric regions.

MOCAT2 (Kultima et al., 2012) uses BWA to correct base errors and short indels in the assembly, by aligning the reads to the generated scaftigs, and SOAPaligner2 to resolve chimeric regions, by aligning the several contigs.

2.5.3 *Bioinformatic tools for Annotation*

After obtaining a list of partial and entire genomes predicted to be present in the sample of study, it is a common step to identify what genes are present, first by identifying ORFs in the sequences, which can be done with FragGeneScan (Rho et al., 2010) and then by submitting each ORF detected to a search software that searches the databases for the closest entries to the sequence, either by homology or pre-determined features. Tools developed for this task are Prokka (Seemann, 2014), IMG/M (Markowitz et al., 2008), BLAST (Altschul et al.,

1990), DIAMOND (Buchfink et al., 2015), USEARCH (Edgar, 2010), and Blast2GO (Conesa et al., 2005), the latter through use of Gene Ontology. MG-RAST (Glass et al., 2010) is also widely used for this purpose, while integrating additional features such as phylogenetic and functional classifications of metagenomes (Meyer et al., 2008), and InterProScan (Jones et al., 2014), the tool that allows for access to InterPro domain information.

All pipelines infer the origin and function of the original reads or of the contigs produced by the assembly steps, in their annotation phase (Table 3). Such determination usually starts by identifying the ORFs, in a process known as gene calling. After that, functional and taxonomic annotation is generally performed by aligning the obtained sequences to protein databases. Binning help in separating the total transcriptome into smaller, taxon specific clusters, necessary for understanding the individual importance of such taxon. Variant calling allows to identify small differences in similar sequences, although it may sacrifice some speed of the pipeline because of the sheer size of data usually processed.

GENE CALLING The identification of ORFs approach is very dependent on the specific characteristics of the analyzed contigs. One of the most used tools is FragGeneScan (Table 3), which incorporates a machine learning algorithm, and even though it works well when challenged with short reads, it is versatile enough to be used with longer reads from both MG and genomic datasets, and by combining models of sequencing error and codon usages in its HMM, it performs very well with error prone reads. Other tools, such as MetaGeneMark and Glimmer-MG, have also been specifically developed for handling gene calling in MG samples. rRNAselector identifies regions encoding for rRNA in MG and MT datasets.

FUNCTIONAL AND TAXONOMIC ANNOTATION Functional and taxonomic annotation is usually tackled by aligning the contigs against databases containing the specific sequences expected to be found on the datasets. In meta-omics studies, however, such databases must be more general, for it is usually not known the totality of species present in the environments from where the samples came. More functional databases, such as UniProt and KEGG, are used to identify the functions of the proteins found in the datasets, which allows for mapping such functions to metabolic pathways, and discover which are the most important ones. Taxonomy may be determined from the alignments in the functional annotation, or it may be determined by aligning the 16S rRNA or DNA sequences to 16S databases, such as SILVA, Greengenes and National Center for Biotechnology Information (NCBI).

BLAST has been the classic tool used for such alignments, but DIAMOND applies an algorithm tailored for short reads, much faster in such situations when compared with the other solutions. USEARCH and VSEARCH are alternatives similar to DIAMOND, faster than BLAST and tailored for MG. Specifically for taxonomic annotation, MetaQUAST incor-

porates an algorithm for searching the SILVA database automatically, when assessing the quality metrics of the contigs (Table 3).

DEPTH OF COVERAGE Depth of coverage is measured by how many times a certain sequence appears in the dataset. Average depth of coverage is that concept applied to the full dataset. A higher coverage confirms the veracity of the sequences (less randomness from sequencing and assembling errors), and allows to detect SNPs when present on those sequences. MG data usually exhibits less coverage, however, which might compromise organisms identification. Coverage is also used to measure gene expression when applied to MT datasets, since original higher abundance of material related to a certain gene is usually the reason for a higher coverage, which in such cases might be magnitudes of times more variable than in MG datasets. BEDtools and Repeatoire are two options for calculating depth of coverage, but it is not a much implemented step in the compared pipelines (Table 3).

VARIANT CALLING Currently, IMP is the only pipeline to integrate the process of identifying and putting in relevance small variations between very similar contigs, since because of the sheer scale of the data, identifying such strain specific characteristics is not a priority. Nevertheless, it has its place when studying multiple strains of the same organism. In IMP, samtoolsś mpileup tool is used to provide a summary of the depth of coverage for each base pair on the sequences aligned to the assembled contigs, Freebays detects such variations but on the original reads, without aligning them back to the assembly, and Platypus performs both types of variant detection.

BINNING For population studies, binning has proven to be a helpful step in assigning an Operational Taxonomic Unit (OTU) to the contigs originated in assembly, which are aggregated in clusters in this step and classified with the OTU for more comprehensive population level information. VizBin (Laczny et al., 2015) accomplishes this, while also producing visual representations of the results, and CheckM (Parks et al., 2015) makes available a suite of tools for evaluating the binning quality.

Binning may serve two purposes: the resulting contigs may be organized into clusters (bins) to diminish data size by only considering one or a few representative sequences from each bin in future analysis; and the contigs may be assigned to a specific taxa, where the assembly serves as the binning process itself. Ideally, there will be one contig by different genome, but in MG it is usually hard to avoid co-assemblies and misassemblies (Kunin et al., 2008). Not just the pure sequence alignment but some nucleotide features related to processes directly involved with DNA, like its repair, and codon usage, might be used for the binning as well.

The only two pipelines to integrate binning have applied it for the first function, in an effort to simplify the datasets (Table 3). VizBin, for example, provides interactive 2D maps of points representing the contigs, where it may be possible to discern clusters that might represent several closely related contigs. This maps allow the user to isolate this clusters and reduce them to one single contig.

2.5.4 *Bioinformatic tools for Statistical analysis*

Many analyses may come from meta-omics studies, and these are even more varied when talking about multi-omics approaches: determining GC content of the genomes, main pathways active, MT/MG ratios, and much more. There are many R packages, such as ShotgunFunctionalizer (Kristiansson et al., 2009), that tackle several of these challenges, as does Blast2GO (Conesa et al., 2005).

VISUALIZATION OF THE RESULTS As seeing the analysis results in a comprehensive, intuitive way is usually easier and more useful than to read results in text, many tools already incorporate several types of graphics, many even interactive, for a more helpful approach to presenting results. As an example, VizBin (Laczny et al., 2015) provides several graphical solutions to perceive the binning results, and Krona tools (Ondov et al., 2011) is an example of interactive graphic results where a user has access to several layers of the same information. Blast2GO (Conesa et al., 2005) makes use of colour changes for a better understanding of the annotation process and of graph representations for highlighting the most important GO, while MG-RAST (Glass et al., 2010) provides pie charts representative of several different communities' profiles and heatmaps with differential multisample analysis. MEGAN (Huson et al., 2007) is a pipeline designed for handling the latter steps in the analysis of microbiome data, integrating tools such as DIAMOND for annotation, but making available a large suite of tools for visual analysis, involving, for example, Voronoi tree maps, principal coordinates analysis and interaction with InterPro2GO (Camon et al., 2004), eggNOG (Powell et al., 2012) and KEGG (Kanehisa and Goto, 2000).

2.6 THE DATABASES FOR ANNOTATION

2.6.1 *UniProt*

A consortium made of the collaboration between the European Bioinformatics Institute (EMBL-EBI), the Protein Information Resource (PIR) and the Swiss Institute of Bioinformatics (SIB), Universal Protein Resource (UniProt) is composed of four approaches to the storage of protein information: the UniProt Knowledgebase (UniProtKB), the UniProt Archive

(UniParc), the UniProt Reference Clusters (UniRef) and the UniProt Metagenomic and Environmental Sequences (UniMES).

UniProtKB is the main point of acess to UniProt, and is divided in two, distinct parts - UniProtKB SwissProt, where the increment in information is more supervised, with information extracted from the literature or from computational scrutiny, and UniProtKB TrEMBL, with more automated and less reviewed information.

UniParc is a repository of past information not only concerning entries of UniProt, but also of several other databases, in a comprehensive, aggregated way. UniRef speeds database query by merging the information contained in UniProtKB into clusters according to the percentage of identity - 50%, 90% and 100% of identity lies between the clusters of UniRef50, UniRef90 and UniRef100 respectively, and UniMES contains data concerning a number of metagenomic studies not available in UniProtKB (UniProt, 2010; Bateman et al., 2015). The massive increase in sequencing initiatives has been followed by a massive increase in UniProt data, and the UniProt interface has seen many changes to facilitate the user survey through such a big database (Bateman et al., 2015).

2.6.2 *KEGG*

Starting on 1995 as a repository of information derived from the Human Genome Program, KEGG has grown to become one of the most relevant databases concerning functional information of organisms and analysis of pathways (Kanehisa and Goto, 2000). Starting based on three databases, it is now divided in eighteen: KEGG PATHWAY, KEGG BRITE and KEGG MODULE are designed for systems information, for understanding life at the system level; KEGG ORTHOLOGY, KEGG GENOME and KEGG GENES organize information retrieved from NCBI's RefSeq and GenBank, aggregating it into KOs, for an easier access to better organized and diverse Genomic Information, with links to other databases such as NCBI; KEGG COMPOUND, KEGG GLYCAN, KEGG REACTION, KEGG RPAIR, KEGG RCLASS, and KEGG ENZYME compose the Chemical Information of KEGG, with information detailing each metabolite and enzyme present in the pathways of KEGG; KEGG DISEASE, KEGG DRUG, KEGG DGROUP, KEGG ENVIRON, JAPIC and DailyMed provide Health Information, concerning diseases and drugs. KEGG presents itself as database of information organized by, besides the normal formats, its pathways and orthologies that allow for a more visual and direct search (Kanehisa and Goto, 2000; Kanehisa et al., 2016).

2.6.3 *Conserved Domains Database*

Starting as a mirror for Pfam (Bateman et al., 2004), a collection of protein families and domains, Simple Modular Architecture Research Tool (SMART) (Letunic et al., 2004), a tool

for annotation of protein domains, and Clusters of Orthologous Groups (COG) (Tatusov et al., 2003), a database for clustering of genes for generation of taxonomic information, the Conserved Domains Database (CDD) has grown to become an important repository of protein domain information, with an ever increasing amount of information concerning domain models and description. With resources such as superfamily clustering and domain annotation with attention to common domain architectures, CDD has gone beyond a simple aggregation of databases to become a tool in itself for the analysis of sequences in respect to their structure.

2.6.4 *InterPro*

InterPro is an European Bioinformatics Institute (EBI) database that collects protein domains information from several databases: HAMAP (Lima et al., 2009), PANTHER (Thomas et al., 2003), PfamA (Finn et al., 2014), PIRSF (Wu et al., 2004), ProDom (Corpet, 1998), PRINTS (Attwood, 2002), Prosite-Profiles (Sigrist, 2002), SMART (Schultz et al., 2000), TIGRFAM (Haft, 2003) and Prosite-Patterns (Sigrist, 2002) for information about protein families, domains, functional sites and repeats regions, Gene3d (Buchan et al., 2002) and SUPERFAMILY (Gough, 2002) for structural information and Coils (Lupas et al., 1991), Phobius (Kall et al., 2004), SignalP (Emanuelsson et al., 2007) and TMHMM (Krogh et al., 2001) for additional features information (Hunter et al., 2009; Finn et al., 2016).

Recently, InterPro added to its consortium the databases SFLD (Akiva et al., 2014), structure/function relational information, and CDD (Marchler-Bauer et al., 2015), which is, like InterPro, supported by a consortium of databases, but from which only the CDD entries are extracted (since the other databases are already integrated in InterPro) (Finn et al., 2016). This diverse consortium of database partners results in varied data, that along the years has been made more uniform by InterPro, in an effort to facilitate access and reference. InterPro also has taken a direction towards studying how the signatures from this databases are connected, thus generating new, more in-depth, information (Hunter et al., 2009).

InterProScan is the service provided by EBI that extends the funcionality of InterPro by facilitating study of proteic and nucleic sequences against InterPro data (Zdobnov and Apweiler, 2001; Quevillon et al., 2005). Besides an interactive interface, it allows programmatic access through REST and SOAP protocols in several programmatic languages (Jones et al., 2014). Through InterProScan, it is possible to access most information kept in InterPro from all its member databases, which is returned in several formats, namely HTML, which contains the graphic interpretation of the results, organized by repeated matches (because some databases will match for the same domain, with the same InterPro identifier), GFF, which contains information regarding the location of the domain, the database of origin, the GO terms associated to that particular domain, the E-value associated with the match, among

other information, and XML, which contains all the information contained in the GFF file, with more detailed information, like, for example, the models associated with that domain and the location of the domain concerning Hidden Markov Models and environment data. It also returns other kinds of information, like log files (Quevillon et al., 2005).

2.6.5 *NCBI's RefSeq*

RefSeq is a non-redundant database of genomic, transcriptomic and proteomic sequences built and maintained by the NCBI. It organizes information from many sources, both from NCBI, like CDD and GenBank (the redundant version of RefSeq, with several entries for the same sequences), and from databases of other organizations, like the Saccharomyces Genome Database (SGD) and The Institute for Genomic Research (TIGR), collecting the information regarding each entity into a single entry. The information from RefSeq is available directly from the site itself, from other NCBI's resources, or from tools like Basic Local Alignment Search Tool (BLAST) (Pruitt et al., 2007; O'Leary et al., 2016).

3

DEVELOPMENT

MOSCA was developed by constructing wrappers to integrate bioinformatic tools for meta-omics analysis (Table **??**). Firstly by testing with variable arguments, and afterwards by developing python scripts to replicate the behavior of the tools, MOSCA was developed for full automation of metagenomic and metatranscriptomic analysis of single- or paired-end data.

3.1 PIPELINE ARCHITECTURE AND IMPLEMENTATION IN MOSCA

3.1.1 *Preprocessing*

The preprocessing of MOSCA aims to remove all undesired sequences from the raw data, to obtain a dataset with only meaningful, good quality reads and free of rRNA. It uses FastQC for several quality evaluations of the data during the preprocessing workflow, to evaluate the quality of the raw data, before trimming by Trimmomatic, and after SortMeRNA. The main steps of the preprocessing, including the tools and the name of the python scripts, are represented in Figure 3.

The preprocessing of MOSCA includes the following steps:

1. Quality control of raw NGS data by running FastQC (version 0.11.4) - provides information about the quality of the reads, and allows the identification of adapters and bias during sequencing;

2. Adapters removal - supplying the right file of adapters, Trimmomatic (version 0.32) excels at removing adapter presence in the data, with its "ILLUMINACLIP" tool. If the data was obtained by Illumina sequencing the adapter files can be obtained together with the Trimmomatic distribution available in GitHub [1]. MOSCA will automatically search for adapter sequences in the data, identify which Illumina adapters were used and then remove them from the data. The presence of adapters results in the detec-

1 https://github.com/timflutre/trimmomatic

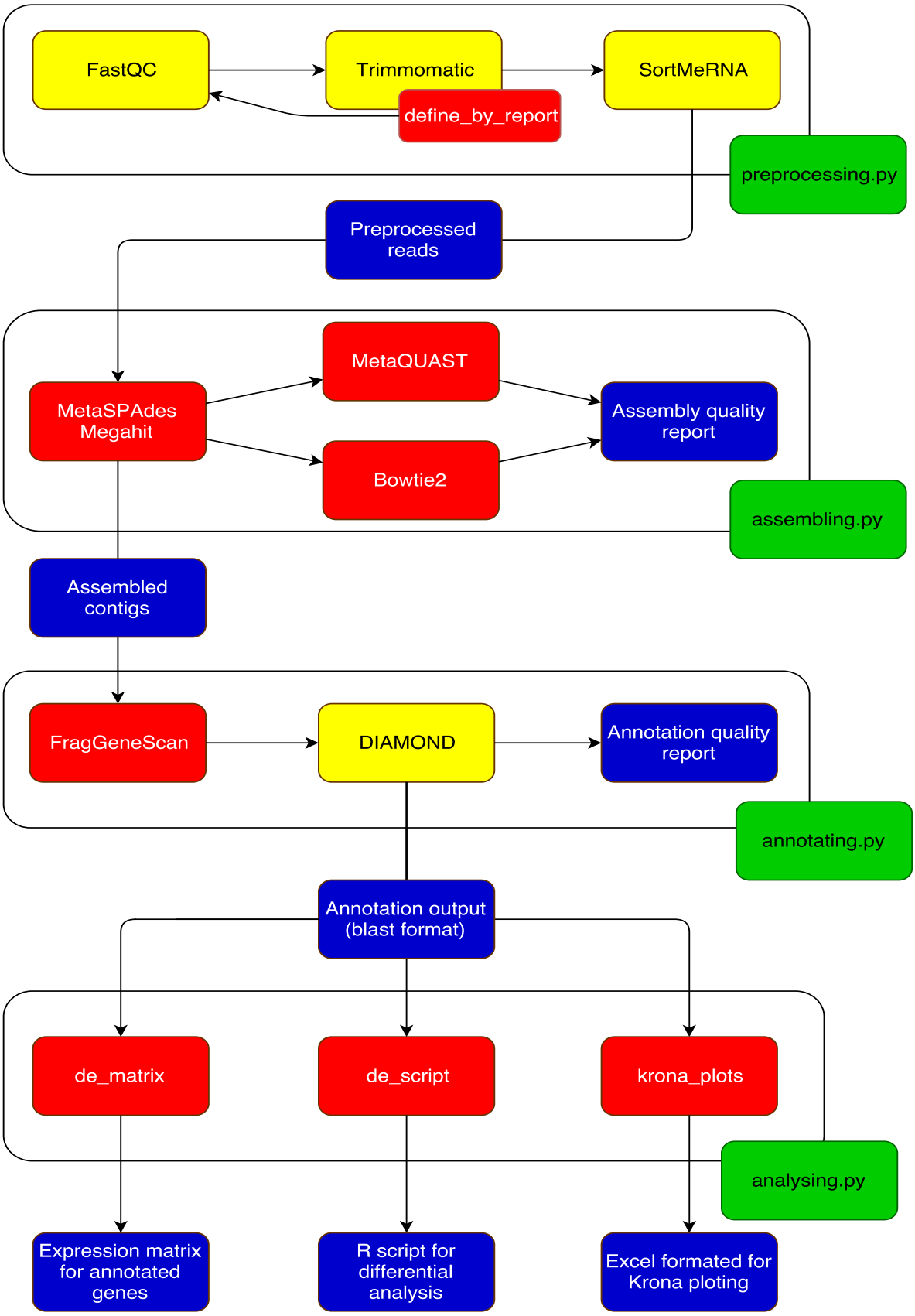


Figure 3: The four scripts (green) integrating the four steps of meta-omics analysis, by incorporating wrappers for some tools in the form of classes (yellow) and functions (red). Some of the functions integrate additional functionalities, like the ones present in the analysis phase. Output files (blue) connect the various steps of the pipeline.

tion of overrepresented sequences by FastQC. After the adapter removal step, those overrepresented sequences corresponding to the adapters should disappear;

3. Low quality sequences removal - MOSCA reads the FastQC report obtained after the adapters removal step and solves the problems related to the "warning" and "fail" flags associated with bad data, by using several tools available in the Trimmomatic toolbox;

4. Removal of rRNA - rRNA sequences are filtered by SortMeRNA (version 2.0) using SILVA database as the reference database;

5. Quality assessment post-preprocessing after all preprocessing steps, reads are again submitted to FastQC quality check. Ideally, the final FastQC report should not have any warning or failed flags.

A general script calls all the classes that compose the collection of wrappers for these four tools (FastQC, Trimmomatic and SortMeRNA) and the additional functions to Trimmomatic, namely the determination of adapters with ILLUMINACLIP and FastQC and the quality trimming adusted in the arguments for CROP and HEADCROP with reference to the FastQC report.

The preprocessing.py script makes use of the classes developed as wrappers of FastQC, Trimmomatic and SortMeRNA for the preprocessing steps (figure 3). Here, the data with a different nature namely, files containing MG, MT or 16S rRNA gene data are distinctively manipulated in some steps. It also distinguishes between single- or paired-end reads (all data manipulation tools integrated in MOSCA support single- and paired-end modes).

FastQC provides a report in tabular and visual (HTML) formats, with information concerning several aspects of NGS data quality, divided in several analysis modules: Basic statistics, Per base sequence quality, Per tile sequence quality, Per sequence quality scores, Per base sequence content, Per sequence GC content, Per base N content, Sequence length distribution, Sequence duplication levels, Overrepresented sequences, Adapter content and Kmer content.

For each analysis module a flag is given, meaning that the result is entirely normal (green), slightly abnormal (yellow) or very unusual (red), concerning the degree of deviation from an excellent report.

FastQC results are used to define the parameters for adapter removal and quality trimming with Trimmomatic. In the adapter removal phase, FastQCś tabular report is parsed, and analyzed - if overrepresented sequences are detected, and are related to adapters, ILLUMINACLIP will be tested with all the adapter files. The adapter files used will be those that have "PE" in the name in the case of a PE analysis, or "SE" in the case of a SE analysis. After each ILLUMINACLIP trimming, a FastQC report will be generated over the resulting file, and if the file no longer has adapter presence, then it will be used for the next steps.

ILLUMINACLIP is used with values of 2, 30 and 10 for seed mismatches, palindrome clip threshold and simple clip threshold, respectively. The adapter files are downloaded from Trimmomatic's github project [2], if no adapter files are available. All this is implemented through the function "remove_adapters".

Trimmomatic is a toolkit packed with several functionalities for NGS data trimming, designed specifically to work with Illumina reads. Five of its tools are implemented in MOSCA:

1. ILLUMINACLIP - Trimmomatic distribution contains several Illumina files (Table 4) containing the platform's artificial sequences that can be formed during the sequencing workflow. These artificial sequences files can be assessed through the Trimmomatics GitHub project [3] and have not been updated since 2015. However, MOSCA will ask if the user wants to check for updated files, and if so it will automatically download the files from the given location.

2. HEADCROP The FastQC module "Per Base Sequence Content" gives the proportion of each base position. In a random library, little differences between the different bases (A, T, G and C) are expected and, thus, the different lines obtained should be as parallel as possible. If not, it means there is an abnormal presence of a certain nucleotide at a certain position in the data. Usually, abnormal situations occur at the beginning of the sequences, and so HEADCROP will remove those first bases at the beginning, to the point where the difference between A, T, G and C is lower than 10% . there is no longer a significant bias towards any nucleotide.

3. CROP - Sequencing quality usually decreases at the end of the reads. MOSCA uses the CROP tool to remove all the nucleotides from the position where the quality of the reads falls below a certain threshold (lower quartile below 10 or median less than 25), based on the "Per Base Sequence Quality" module results from the FastQC report.

4. AVGQUAL - in an effort to increase the robustness of preprocessing NGS datasets, a minimum average quality is imposed, so only sequences above a defined threshold follow the next steps of analysis

5. MINLEN - MOSCA keeps reads with a minimum read length of 100 nucleotides. Reads with less than 100 nucleotides are removed using the MINLEN tool from Trimmomatic.

The six files with Illumina artificial sequences were considered for the preprocessing 4. This sequences are available from Trimmomatic distribution.

2 https://github.com/timflutre/trimmomatic/tree/master/adapters
3 https://github.com/timflutre/trimmomatic

Table 4: The six artificial sequences files available from Trimmomatic distribution, two for SE and four for PE mode, and the corresponding sequencing kits from Illumina. The TruSeq3-PE-2.fa file contains the same sequences as the TruSeq3-PE.fa, but containing their reverse complements, it allows for palindrome clipping.

Adapter files available	Sequencing technology
NexteraPE-PE	Nextera DNA kits
TruSeq2-SE, TruSeq2-PE	TruSeq RNA Library Prep Kit v2
TruSeq3-SE, TruSeq3-PE, TruSeq3-PE-2	TruSeq PE Cluster Kit v3-cBot-HS

After adapter removal by ILLUMINACLIP, MOSCA calls the function "define_by_report", making use of several different Trimmomatićs tools for different quality trimming tasks. With CROP, MOSCA removes the bases of all reads from the position after which the average quality in that position falls below the defined threshold for a warning flag in the FastQC report - if, in the boxplot representing the quality distribution in that position, the lower quartile falls below 10 or the median is less than 25. With HEADCROP, MOSCA removes bases until the last position at which there is a significant difference in the quantity of certain bases (if the difference between A and T or G and C is over 10%). MOSCA removes reads with an average quality under 20 (using AVGQUAL), and finally removes all sequences with less than 100 nucleotides in length (using "MINLEN"). Because every FastQC report is related to one file, even when handling paired-end data, MOSCA will use Trimmomatic in single-end mode for cropping the less desirable extremities of the reads, but will not remove any read, as to conserve the paired-end nature of the data. When removing reads with the AVGQUAL tool, it will again use Trimmomatic in paired-end mode.

SortMeRNA is used with the SILVA database, separating the sequences aligned to the database from the sequences not aligned. Because SortMeRNA accepts only one input file at a time, when working in paired-end mode two bash scripts [4] must be used, one for merging the two files into one with interlaced reads, and the other to separate the reads into two files, after rRNA removal. Paired-in and paired-out arguments must be used to specify to SortMeRNA that the input file is paired-end data. The non aligned reads file is used for the next steps.

3.1.2 *Assembly*

MOSCA allows to choose between two assemblers, MetaSPAdes (version 3.9.0) and Megahit (version 1.1.1). They were chosen because of their superior performance in previous studies, and because even though they are similar tools, they implement several distinctive solutions to problems inherent to assembly. Megahit, for example, employs "succint de

4 https://github.com/biocore/sortmerna/tree/master/scripts

Bruijn graphs" to reduce memory requirements and discards singleton k-mers, while also implementing "mercy-k-mer" strategy, thus obtaining more relevant data, while MetaSPAdes uses complete read information together with preassembled reads at every step of its workflow, and in contrast to most assemblers, it directly incorporates paired end information in the graphs by building it with k-bimers, instead of using ofter construction of the Bruijn graph for simplification steps (Vollmers et al., 2017).

In their default mode, MetaSPAdes and Megahit iterate through several kmers for the same assembling procedure, generating several contig files, from which the most reliable contigs are chosen. Even though they both support reference-based assembling, for now MOSCA only operates on *de novo* assembling, since it aims to analyze data collected from complex microbial communities containing several unknown microorganisms without complete genomic information available.

After the assembly, MetaQUAST (version 4.5) is used as a tool for quality control of the assembly. It produces reports for several classical metrics on the final contigs, like N50, number of contigs for several sizes, and number of misassemblies. Bowtie2 (version 2.2.6) is also used to complete the assembly report by giving the percentage of preprocessed reads that can be aligned to the obtained contigs, i.e., the percentage of reads used in the assembly.

The assembling.py script takes into account the existence of two options for assembler - MetaSPAdes and Megahit -, generating the commands for both tools, and integrating MetaQUAST and Bowtie2 for posterior check on the quality of the assembly, returning a final report with the analysis performed by these two tools (Figure 3).

Both MetaSPAdes and Megahit run as default to make a multi-kmer assembly:

1. MetaSPAdes iterates over k-mer sizes 21, 33 and 51, and outputs several files for every k-mer and for the more reliable contigs,:
   a) The contigs in FASTA format (contigs.fasta)
   b) The scaffolds (scaffolds.fasta)
   c) The assembly graph (assembly_graphp.fastg)
   d) The scaffold paths in the assembly graph (scaffolds.paths)
   e) The contigs before read resolution (before_rr.fasta)
2. Megahit iterates through the k-mer values 21, 29, 39, 59, 79, 99, 119 and 141, and besides the final contigs file, for each contig value it outputs five files:
   a) The contigs in FASTA format (.contigs.fa)
   b) The contigs with local low coverage *unitigs* removed (.addi.fa)
   c) The locally assembled contigs for that specific kmer (.local.fa)

d) The stand-alone contigs for that specific kmer (.final.contigs.fa)

e) A file related with bubble representation of contigs (.bubble_seq.fa)

MOSCA uses the final contigs files (i.e., contigs.fa or contigs.fasta) for posterior analysis obtained from the default set of parameters. These files are used as input for generating the quality report.

The quality control is based on the report from MetaQUAST, with one final line added from the alignment of Bowtie2. MetaQUAST and Bowtie2 together report for a number of metrics in MOSCA:

1. Number of contigs, total and for several intervals (over 10000bp, for example)
2. Duplication ratio - total number of bases aligned to the assembly divided by the total number of aligned bases in the reference genome
3. Genome fraction (%) - percentage of aligned bases in the reference genome
4. N50, N75, L50 and L75
5. Reads aligned (%)

3.1.3 *Annotation*

In MOSCA, FragGeneScan (version 1.15) is used for gene calling in the annotation step.

A function was developed to build the FragGeneScan command and perform gene calling, with the contigs files (.fa file from Megahit or .fasta file from Metaspades) as input. Because the contigs were used instead of reads, the arguments were set as following: complete as ”1” and train as ”complete”.

FragGeneScan outputs three files:

1. The list of the coordinates of putative genes (.out)
2. A FASTA of nucleotide sequences corresponding to the putative genes (.ffn)
3. A FASTA of aminoacid sequences corresponding to the putative genes (.fna)

DIAMOND (version 0.9.9) is used to align the assembled contigs with the sequences present in reference databases in FASTA format, and is run in default conditions. It was chosen because it is much faster than the existing alternatives, such as BLAST. MOSCA determines the number of ORFs identified and the number of ORFs that could be annotated by using the selected FASTA database.

MOSCA uses the FragGeneScan aminoacid sequences file as input for DIAMOND. DIAMOND uses ”blastp” for the identification of aminoacid sequences, and requires a database

(.dmnd) as input for aligning the sequences. Compressed or uncompressed FASTA databases can be converted to .dmnd format with the "makeblastdb" tool from DIAMOND.

Blastp may output several alignments for each of the sequences, from the most to the less confident, and MOSCA only considers the best hit, with the smaller e-value. The annotated reads are then outputed in a .blast file, and the unaligned reads in a .fasta file.

3.1.4 *Data analysis*

MOSCA integrates UniProt's database in its workflow, allowing to use UniProt's ID mapping service for obtaining relevant data in specific formats: a tab-separated report with taxonomic (superkingdom, phylum, class, order, family, genus and species) and systems (EC number, pathways and protein names) information, and the GFF file with information about the features present in the sequences. Because UniProt has a maximum limit of 2Mb of information for submission, the list of IDs must be broken into chunks, and one chunk must be submitted at a time. This taxonomy and functional information is used to create CSV files formatted for generating Krona plots.

For MT experiments, MOSCA produces an expression matrix from the values of coverage of the annotated genes, and generates a script for DE using DeSEQ2 (version 1.18.1). Heatmaps are generated giving information on the degree of similarity between samples, and on the differential gene expression.

The analysing.py script receives as input the blast result from DIAMOND (aligned.blast file), and converts it into a pandas DataFrame object, for which it has integrated three functionalities (Figure 3):

1. krona_plots - If the annotation was performed with reference to the UniProt database, MOSCA will use the Application Programming Interface (API) provided by UniProt to retrieve relevant taxonomic and systems information and output it as CSVs. Quantification of annotated ORFs for each taxonomic and systems category determines the area for the Krona plots.

2. de_matrix - For MT studies, de_matrix builds an expression matrix from the coverage values calculated by the assemblers, and retrieves the sequences IDs from DIAMOND's report, thus building a matrix that is outputted for the DE analysis in R. The values on this matrix are normalized by Transcripts per million (TPM).

3. de_script - For MT studies, de_script builds the R script for DE analysis with DESeq2, to be used over the built matrix.

The TPM normalization that takes place in the build up of the DE matrix follows three steps:

1. For each annotated gene, the coverage value is divided by the length of the corresponding gene, in the length column of DIAMOND's report. This gives Reads per kilobase (RPK).

2. The sum of all RPK values in the sample is calculated, and divided by 1000000. This is the "per million" scaling factor.

3. The RPK values are divided by the "per million" scaling factor, giving TPM.

This analysis outputs two types of graphics:

1. From the taxonomic and systems information, two CSV files are outputted in a format that can be used directly for krona ploting, using pandas *convert_to_csv* function.

2. From the R package DeSEQ2, heatmaps are outputed concerning multisample similarity comparison and DE of the most significantly expressed genes.

3.1.5 *Implementation details*

MOSCA was developed in python 3.6.0, and tested in a Ubuntu xenial 16.04.2 desktop version. The source code of MOSCA and more information concerning the pipeline can be found on github [5].

The following libraries have been used in the development for distinct purposes.

SUBPROCESS Implements command line functionality through python scripting, used for the input/output workflows between the several tools integrated [6]. In MOSCA, usually the command is first generated as a string, in the form that it would be inserted in the command line, and then subprocess runs the command.

PANDAS Implements data structures and analysis [7]. In MOSCA, it is used for storing the result of parsing FastQC's and DIAMOND's reports and UniProt's information files into DataFrames.

REQUESTS A python library for performing HTTP requests [8]. Used for performing the HTTP request steps in the analysis of results from annotation.

5 https://github.com/iquasere/MOSCA
6 https://docs.python.org/2/library/subprocess.html
7 http://pandas.pydata.org
8 http://docs.python-requests.org

4

PIPELINE TESTING

4.1 DATASETS FOR PIPELINE TESTING

MOSCA was tested with real datasets obtained from lab scale anaerobic digestion bioreactors. For MG/MT, the data was retrieved from four continuous bioreactors inoculated with the same inoculumtreating synthetic wastewater containing ethanol or a mixture of volatile fatty acids. The main goals were to identify the pathways most utilized and the most active microorganisms and compare these functional and taxonomic information obtained from the different operational conditions. In this work, the MG samples from this study were named "DNA1", "DNA2", "DNA3" and "DNA4", while the corresponding MT samples were named "RNA1", "RNA2", "RNA3" and "RNA4".

For MG/MP, the data was extracted from anaerobic batch reactors converting different long-chain fatty acids (LCFA) to methane. In this work, only the MG samples, named "DNA5", "DNA6", "DNA7" and "DNA8", were tested with MOSCA.

Three datasets retrieved from amplicon sequencing experiments targeting the 16S rRNA gene, also obtained from anaerobic bioreactors, were used, as well, for testing some of the tools implemented in MOSCA. These samples were designated by "RRNA1", "RRNA2" and "RRNA3".

All FastQ files were obtained in PE format, and so all the tools used until the assembly, inclusively, were used in PE format. PE sequencing produces two different files per sample, one named "forward" and the other "reverse", since one has the sequences from the forward strand and the other the ones from the reverse strand.

4.2 RESULTS

4.2.1 *Preprocessing*

Initial data quality assessment

Before testing the artificial sequences removal capacity of ILLUMINACLIP, eight metagenomic FastQ files (DNA1 to DNA8) were submitted to FastQC evaluation.

It was possible to obtain high quality FastQ files at the end of the trimming with Trimmomatic as it is shown in table 5. All the original files presented "warn" or "fail" based on FASTQC analysis for overrepresented sequences. In general, after the first step of trimming, the ILLUMINACLIP, the adapters were removed and the warnings resolved. Per base sequence quality (PBSQ) passed the quality check only after the trimming based on FASTQC report (that gives the information on the quality of the sequences). This trimming step used the CROP tool to remove the bases at the end of the sequences that presented low quality. These results were obtained by setting the following parameters:

1. ILLUMINACLIP - seed mismatches of 2, palindrome clip threshold of 30 and simple clip threshold of 10
2. AVGQUAL - minimum quality of 20
3. MAXINFO - target length of 40, strictness of 0.5

The ILLUMINACLIP was capable, when using the right adapter file, of removing completely the presence of overrepresented sequences, and even reducing "fail" and "warn" flags in the Kmer Content to smoother reports, showing that those irregularities were associated with the presence of artificial sequences in the datasets.

FastQ files obtained from 16S metagenomics showed much different reports, due to the different nature of the data. "Per base sequence content", "Overrepresented sequences" and "Sequence Duplication Levels" all reported severe bias, but these biases were due to the conserved nature of rRNA and to the fact that these files were obtained from amplicon sequencing targeting the 16S rRNA gene (table 6).

Trimming by FastQC report

In the case of MG data (DNA1 to DNA8), for all reverse files, a warning was issued by FastQC concerning the general quality of the reads towards the last positions, in a per position report, and towards the first positions, when considering relative nucleotide abundance. After using the CROP tool adjusted to the first position from which the quality is under the defined threshold and using the HEADCROP tool for the position after which there is a normal relative abundance of each nucleotide, MOSCA was capable of improving

Table 5: Quality evaluation results using FastQC on MG FastQ files (DNA1 to DNA8) before and after trimming with Trimmomatic. Green color means "pass", yellow color means "warn" and red color means "fail".

	DNA1	Adapters	PBSQ	PTSQ	PSQS	PBSC	PSGCC	PBNC	SLD	SDL	OS	AC	KC
forward	Initial	None											
	After adapter removal	NexteraPE-PE											
		TruSeq2-PE											
		TruSeq3-PE											
		TruSeq3-PE-2											
	After trimming by FastQC report	TruSeq2-PE											
	After trimming by FastQC report	TruSeq3-PE-2											
	After AVGQUAL	TruSeq3-PE-2											
	After MAXINFO	TruSeq3-PE-2											
reverse	Initial	None											
	After adapter removal	NexteraPE-PE											
		TruSeq2-PE											
		TruSeq3-PE											
		TruSeq3-PE-2											
	After trimming by FastQC report	TruSeq2-PE											
	After trimming by FastQC report	TruSeq3-PE-2											
	After AVGQUAL	TruSeq3-PE-2											
	After MAXINFO	TruSeq3-PE-2											

	DNA2	Adaptors	PBSQ	PTSQ	PSQS	PBSC	PSGCC	PBNC	SLD	SDL	OS	AC	KC
forward	Initial	None											
	After adapter removal	NexteraPE-PE											
		TruSeq2-PE											
		TruSeq3-PE											
		TruSeq3-PE-2											
	After trimming by FastQC report	TruSeq2-PE											
	After trimming by FastQC report	TruSeq3-PE-2											
	After AVGQUAL	TruSeq3-PE-2											
	After MAXINFO	TruSeq3-PE-2											
reverse	Initial	None											
	After adapter removal	NexteraPE-PE											
		TruSeq2-PE											
		TruSeq3-PE											
		TruSeq3-PE-2											
	After trimming by FastQC report	TruSeq2-PE											
	After trimming by FastQC report	TruSeq3-PE-2											
	After AVGQUAL	TruSeq3-PE-2											
	After MAXINFO	TruSeq3-PE-2											

	DNA3	Adaptors	PBSQ	PTSQ	PSQS	PBSC	PSGCC	PBNC	SLD	SDL	OS	AC	KC
forward	Initial	None											
	After adapter removal	NexteraPE-PE											
		TruSeq2-PE											
		TruSeq3-PE											
		TruSeq3-PE-2											
	After trimming by FastQC report	TruSeq2-PE											
	After trimming by FastQC report	TruSeq3-PE-2											
	After AVGQUAL	TruSeq3-PE-2											
	After MAXINFO	TruSeq3-PE-2											
reverse	Initial	None											
	After adapter removal	NexteraPE-PE											
		TruSeq2-PE											
		TruSeq3-PE											
		TruSeq3-PE-2											
	After trimming by FastQC report	TruSeq2-PE											
	After trimming by FastQC report	TruSeq3-PE-2											
	After AVGQUAL	TruSeq3-PE-2											
	After MAXINFO	TruSeq3-PE-2											

	DNA4	Adaptors	PBSQ	PTSQ	PSQS	PBSC	PSGCC	PBNC	SLD	SDL	OS	AC	KC
forward	Initial	None											
	After adapter removal	NexteraPE-PE											
		TruSeq2-PE											
		TruSeq3-PE											
		TruSeq3-PE-2											
	After trimming by FastQC report	TruSeq2-PE											
	After trimming by FastQC report	TruSeq3-PE-2											
	After AVGQUAL	TruSeq3-PE-2											
	After MAXINFO	TruSeq3-PE-2											
reverse	Initial	None											
	After adapter removal	NexteraPE-PE											
		TruSeq2-PE											
		TruSeq3-PE											
		TruSeq3-PE-2											
	After trimming by FastQC report	TruSeq2-PE											
	After trimming by FastQC report	TruSeq3-PE-2											
	After AVGQUAL	TruSeq3-PE-2											
	After MAXINFO	TruSeq3-PE-2											

PBSQ: Per Base Sequence Quality; PTSQ: Per Tile Sequence Quality; PSQS: Per Sequence Quality Scores; PBSC: Per Base Sequence Content (PBSC); PSGCC: Per Sequence GC Content; PBNC: Per Base N Content; SLD: Sequence Length Distribution; SDL: Sequence Duplication Levels; OS: Overrepresented Sequences; AC: Adapter Content; KC: Kmer Content.

Table 5: Continued.

Table 6: Quality evaluation results using FastQC on metagenomics 16S rRNA genes FastQ files (RRNA 1 to 3) before and after trimming with Trimmomatic. Green color means "pass", yellow color means "warn" and red color means "fail".

	RRNA1	PBSQ	PTSQ	PSQS	PBSC	PSGCC	PBNC	SLD	SDL	OS	AC	KC
forward	Initial											
	After trimming by FastQC report											
	After AVGQUAL											
	After MAXINFO											
reverse	Initial											
	After trimming by FastQC report											
	After AVGQUAL = 20											
	After MAXINFO: target length = 38; strictness = 0.5											

	RRNA2	PBSQ	PTSQ	PSQS	PBSC	PSGCC	PBNC	SLD	SDL	OS	AC	KC
forward	Initial											
	After trimming by FastQC report											
	After AVGQUAL											
	After MAXINFO											
reverse	Initial											
	After trimming by FastQC report											
	After AVGQUAL = 20											
	After MAXINFO: target length = 38; strictness = 0.5											

	RRNA3	PBSQ	PTSQ	PSQS	PBSC	PSGCC	PBNC	SLD	SDL	OS	AC	KC
forward	Initial											
	After trimming by FastQC report											
	After AVGQUAL											
	After MAXINFO											
reverse	Initial											
	After trimming by FastQC report											
	After AVGQUAL = 20											
	After MAXINFO: target length = 38; strictness = 0.5											

the quality of the data, since this two simple steps solved the irregularities reported by FastQC on the corresponding modules (Table 5 and 6).

In the analysis of 16S rRNA files no sequences were identified as artificial by FastQC, so ILLUMINACLIP was not used. FastQC reported a strong bias at several positions of the data, which is justified by the conserved nature of 16S sequences. Therefore, the HEADCROP functionality was not used for the trimming, but because of decline in quality at the end of sequences, CROP was used, with results similar to those of DNA samples.

rRNA removal by SortMeRNA

Table 7: Number of reads in the datasets throughout preprocessing, number of contigs after assembly, number of ORFs identified in the contigs and number of genes annotated with reference to the UniProt database.

Sample	Initial number of reads	Number of reads after Trimmomatic	Number of reads after SortMeRNA	Percentage of reads removed by preprocessing	Contigs	Identified ORFs	Annotations by UniProt
DNA1	620978	572664	566324	9	45170	43461	34193
DNA2	40368346	38425046	38025476	6	691967	836439	636232
DNA3	24162338	22498888	22301428	8	577041	688246	523773
DNA4	16545148	15114390	14961130	10	479468	561966	427102
DNA5	8366918	8009440	7990204	5	343350	484310	383144
DNA6	7759654	7250992	7230616	7	368139	518502	411381
DNA7	6923546	6441510	6424360	7	337091	471541	377357
DNA8	7563918	7002928	6986178	8	341131	473630	373088
RNA1	49925424	49154422	48917218	2	1168700	1292267	1174612
RNA2	70304382	68570414	68085172	3	1290096	1453985	1077019
RNA3	460622226	423112624	420629878	9	12398739	13493827	9376991
RNA4	29898224	29683410	29408190	2	980083	1003412	731554
RRNA1	76678	75624	10	100	-	-	-
RRNA2	71004	69914	2	100	-	-	-
RRNA3	64696	63800	2	100	-	-	-

rRNA sequences were aligned to those present in SILVA and RFAM databases in order to detect and remove the 16S rRNA sequences by using Sortmerna (Table 7). Two different kinds of dataset were tested: three samples, forward and reverse, of rRNA amplicon sequencing, where ideally it would be expected total identification of reads as rRNA; eight samples of DNA /MG and four of RNA (MT), with forward and reverse information, where only a low percentage should be identified as rRNA. SortMeRNA was very efficient in identifying the rRNA from the 16S samples, with 99.99% of sequences identified as rRNA. Less than 1% of the reads were removed, on average, from the DNA samples (Table 7), thus proving the sensibility and specificity of SortMeRNA.

4.2.2 *Assembly*

MEGAHIT and MetaSPAdes were run using the predefined/default arguments (see section 3.1.2), and both used a multi-kmer approach for assembling the MG reads. The percentage

of reads used for the assembly was estimated, by using Bowtie2 to align the original reads with the contigs produced, and several metrics were calculated by MetaQUAST (Tables 8 and 9).

At least 70% of reads were aligned with Bowtie2 to the contigs, with the exception of DNA1, which may be related with the small size of the dataset, when compared to the others. With MetaSPAdes, the results were even better, with more reads aligned, more contigs assembled and higher N50s. For this reason, MetaSPAdes contigs were used in the next steps.

Table 8: Several metrics concerning the quality of the contigs produced by MEGAHIT, obtained by MetaQUAST and Bowtie2.

	DNA1	DNA2	DNA3	DNA4	DNA5	DNA6	DNA7	DNA8
# N's per 100 kbp	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
# contigs	2146	157032	122795	98115	158962	167679	141461	149486
# contigs (>= 0 bp)	7221	291686	233529	187446	269259	301973	258605	249642
# contigs (>= 1000 bp)	248	68066	51095	39442	54757	58359	48473	48551
# contigs (>= 5000 bp)	3	7803	5311	3658	3675	4427	4032	3271
# contigs (>= 10000 bp)	0	2678	1792	1323	1072	1250	1221	966
# contigs (>= 25000 bp)	0	510	411	326	191	195	206	172
# contigs (>= 50000 bp)	0	151	128	87	52	47	35	39
# indels per 100 kbp	-	35.42	35.73	33.98	20.97	21.81	19.20	17.99
# local misassemblies	-	131	138	106	45	88	76	42
# misassembled contigs	-	154	163	165	88	197	165	86
# misassemblies	-	165	179	188	95	208	175	87
# mismatches per 100 kbp	-	1305.98	1437.66	1208.13	607.84	1138.29	898.12	572.64
# unaligned contigs	-	153265 + 510 part	117791 + 560 part	95436 + 358 part	157539 + 165 part	160773 + 460 part	133578 + 380 part	147797 + 209 part
# unaligned mis. contigs	-	35	46	32	16	34	23	18
Duplication ratio	-	1.196	1.171	1.121	1.118	1.099	1.068	1.143
Genome fraction (%)	-	90.597	73.359	88.469	91.090	51.230	46.366	91.833
L50	781	20152	16236	13882	30186	30700	24667	28952
L75	1409	62421	50397	41944	78759	81888	68102	75390
Largest alignment	-	33829	40875	37591	21582	18605	31325	23252
Largest contig	5527	336120	280276	228421	165310	114545	157288	100744
Misassembled contigs length	-	679002	650413	999161	256679	323518	278508	202033
N50	689	2559	2345	2070	1465	1515	1553	1380
N75	580	1073	1010	955	798	811	815	772
Reference length	-	7052336	10393517	7052336	3026645	15897113	20428964	3026645
Total aligned length	-	7399530	8682213	6768965	2936758	8646921	9773162	3035355
Total length	1579645	268718963	200790882	152000468	202801485	218444940	187250445	184564422
Total length (>= 0 bp)	3466605	319349253	242532181	185720540	245103504	268861328	231427448	223102892
Total length (>= 1000 bp)	363644	207382110	151295229	111556300	130819175	142770228	122804017	115075522
Total length (>= 5000 bp)	16316	91480253	64660508	45267832	38691783	45047082	42296213	34302308
Total length (>= 10000 bp)	0	56469975	40848293	29490106	21055547	23611474	23125246	18684516
Total length (>= 25000 bp)	0	24600947	20480318	14804515	8380598	8311228	8375220	7081355
Total length (>= 50000 bp)	0	12600805	10879697	6669939	3746286	3215821	2630751	2580626
Unaligned length	-	261224533	192001856	145144590	199827248	209653731	177344794	181490122
Reads aligned (%)	18.32	86.57	78.71	74.82	77.61	73,85	71.80	ND

4.2.3 *Annotation*

FragGeneScan was used to detect the ORFs in the contigs generated by MetaSPAdes. DIAMOND was used to align the ORFs identified by FragGeneScan to sequences in the UniProt database. Approximately 78% of the identified ORFs could be annotated by aligning to UniProt. For each read, only the most confident alignment was considered, and for each alignment, the following information was retrieved: the ID of the sequence that the ORF aligned to, the percentage of identity between the ORF and the sequence it aligned to, the length of the sequence from UniProt, the number of mismatch between the two sequences,

Table 9: Several metrics concerning the quality of the contigs produced by MetaSPAdes, obtained by MetaQUAST and Bowtie2.

	DNA1	DNA2	DNA4	DNA5	DNA6	DNA7	DNA8
# N's per 100 kbp	0.00	0.00	0.00	0.00	0.00	0.00	0.00
# contigs	182550	210149	140184	159824	159790	135597	156378
# contigs (>= 0 bp)	1168700	1290096	980083	343350	368139	337091	341131
# contigs (>= 1000 bp)	64992	74975	48810	51261	52488	43660	46762
# contigs (>= 5000 bp)	7851	9699	4087	3965	4926	4499	3542
# contigs (>= 10000 bp)	3280	4011	1144	1621	1816	1742	1450
# contigs (>= 25000 bp)	838	1098	134	402	407	384	391
# contigs (>= 50000 bp)	236	371	17	102	133	113	88
# indels per 100 kbp	-	31.94	29.43	24.57	23.36	19.67	21.10
# local misassemblies	-	76	43	42	65	61	38
# misassembled contigs	-	70	86	51	95	78	46
# misassemblies	-	77	94	61	100	86	51
# mismatches per 100 kbp	-	1517.50	1024.77	671.42	1182.16	961.56	633.50
# unaligned contigs	-	206866 + 254 part	137980 + 176 part	159277 + 77 part	156158 + 212 part	131224 + 156 part	155782 + 82 part
# unaligned mis. contigs	-	8	4	1	15	10	3
Duplication ratio	-	1.031	1.031	1.028	1.027	1.033	1.017
GC (%)	57.22						
Genome fraction (%)	-	80.308	84.658	91.175	61.668	53.814	92.372
L50	19613	20136	24872	25948	23404	18499	26936
L75	71814	77949	66825	76189	73016	60783	77044
Largest alignment	-	35305	34682	37368	35338	46332	36506
Largest contig	214986	288163	78157	186121	236216	204583	177738
Misassembled contigs length	-	423619	365638	590078	462438	476383	465994
N50	2400	2773	1561	1520	1691	1752	1377
N75	932	971	811	799	826	827	765
Reference length	-	9750659	7052336	3026645	11543203	15991827	3026645
Total aligned length	-	7894382	5995405	2709102	7118092	7276048	2715734
Total length	291091142	351428560	182413605	211244375	221525430	190949820	196885390
Total length (>= 0 bp)	561431491	652215162	403952307	281184416	299312233	266695067	268474837
Total length (>= 1000 bp)	211736669	260642340	120659137	136291872	147571203	127716732	121427952
Total length (>= 5000 bp)	106308986	137656146	39238486	52672544	61794102	57479439	46555829
Total length (>= 10000 bp)	74673068	98521955	19264480	36679249	40602672	38526161	32291625
Total length (>= 25000 bp)	37981991	54720396	4768964	18190171	19967990	18399808	16236780
Total length (>= 50000 bp)	17526138	29567663	1010645	7997489	10825618	8990568	6035448
Unaligned length	-	343480921	176350197	208513646	214341744	183623017	194153830
Reads aligned	40.38	89.26	81.63	80.83	ND	76.13	79.63

the number of gap openings in the alignment, the first and last position of the ORF and of the sequence it aligned to, the e-value and the score of the alignment.

4.2.4 *Data analysis*

The UniProt ids were retrieved from the alignment files produced by DIAMOND, and submitted to the UniProt mapping service. Taxonomic and systems information was obtained, and the presence of the different taxa and subsystems was quantified, and represented in Krona plots (figures 4 and 5, respectively).

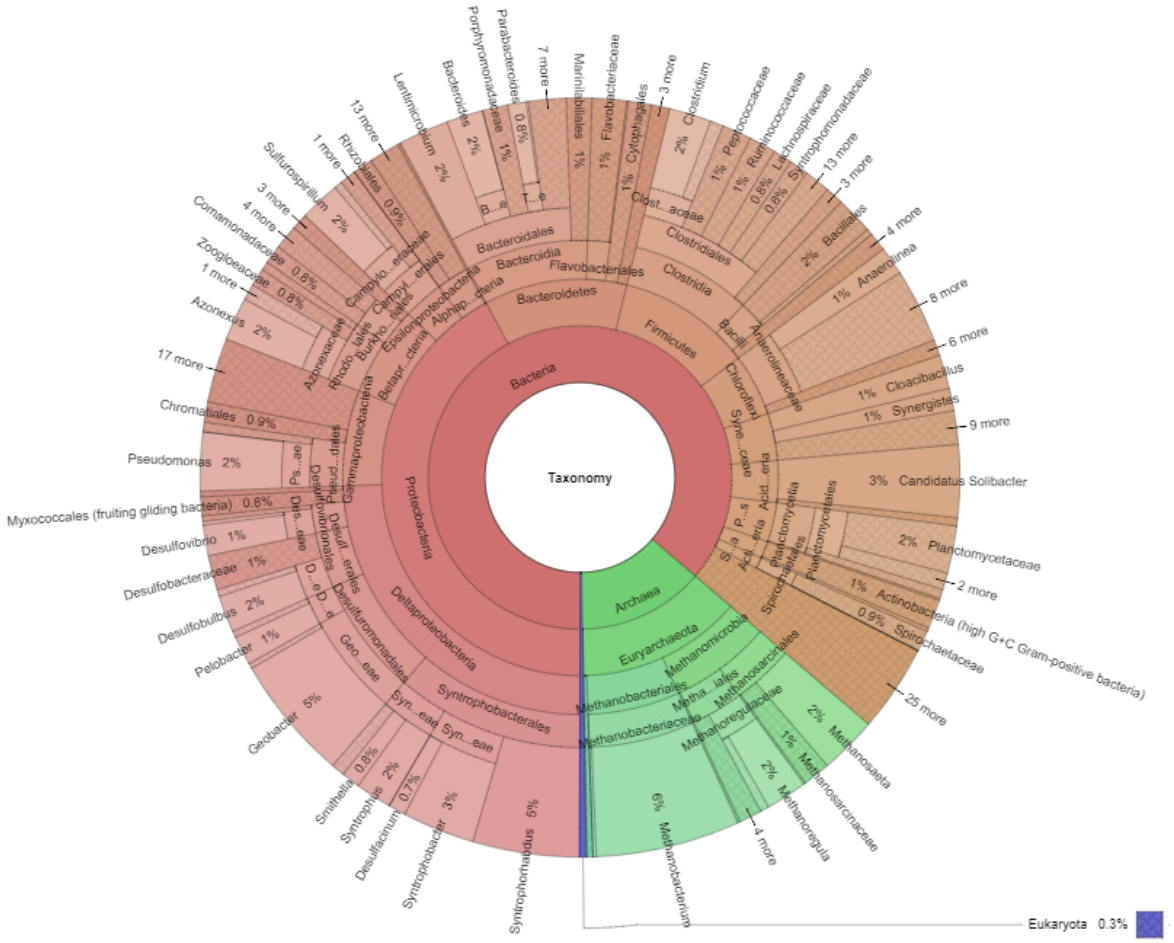


Figure 4: Taxonomic identification of the species present in sample DNA2.

DE analysis was performed for three samples of RNA. Based on the coverage values obtained from assembling, a matrix was built, integrating information about every annotated gene in the assembled contigs. The values of that matrix were normalized with TPM (Conesa et al., 2016). DE analysis was performed in R using DeSEQ2 package and heatmaps

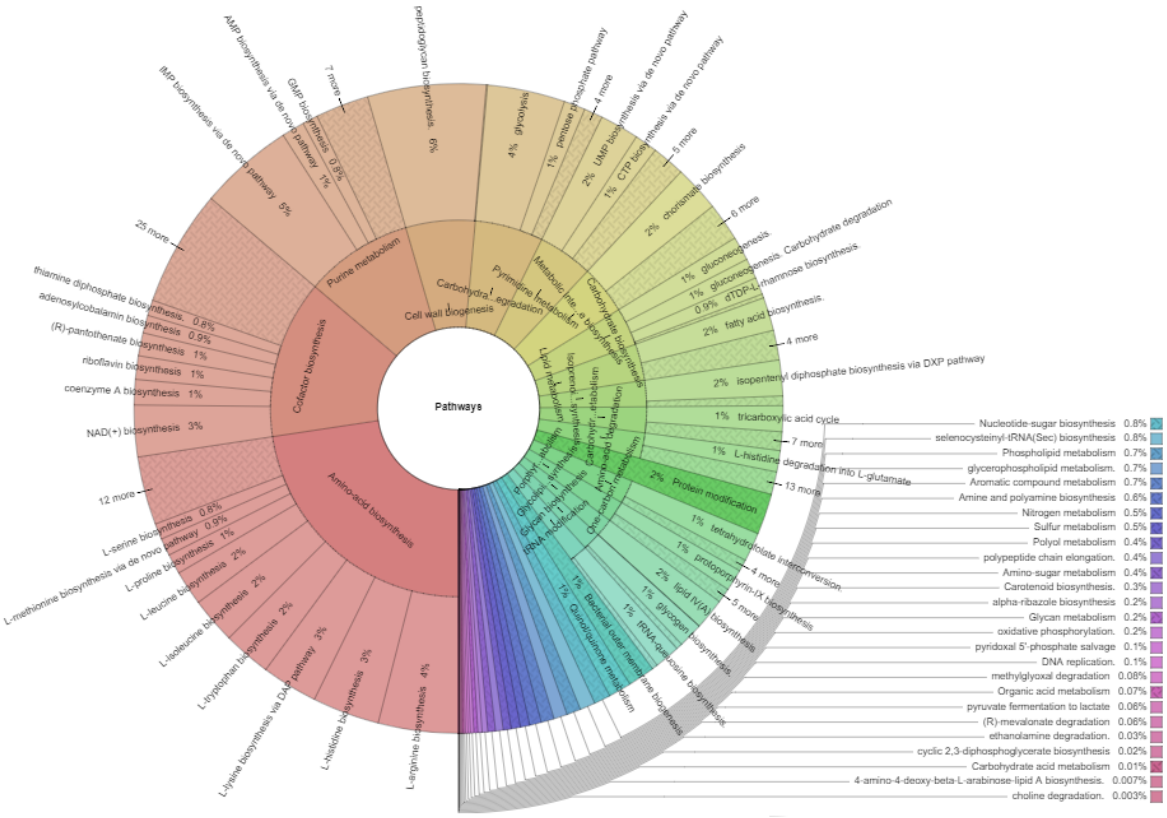


Figure 5: Assignment of genes to pathways for the DNA2 sample.

were obtained for most differentially expressed genes (Figure 6) and for denoting the differences between samples (Figure 7).

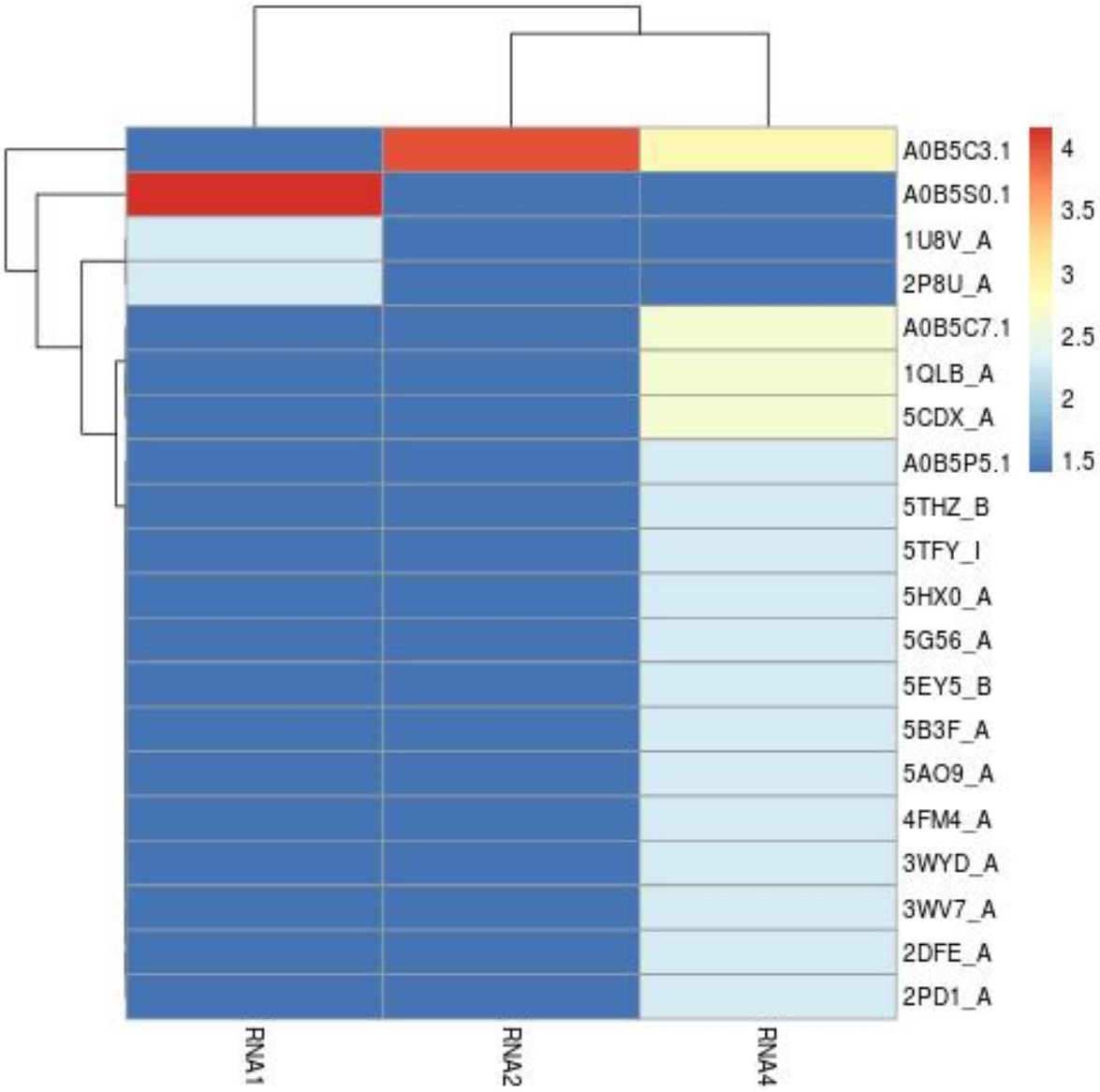


Figure 6: Example of heatmap representing the most expressed genes in the three samples (RNA1, RNA2 and RNA4), and evidencing the differences in expression of the genes by a colour gradient.

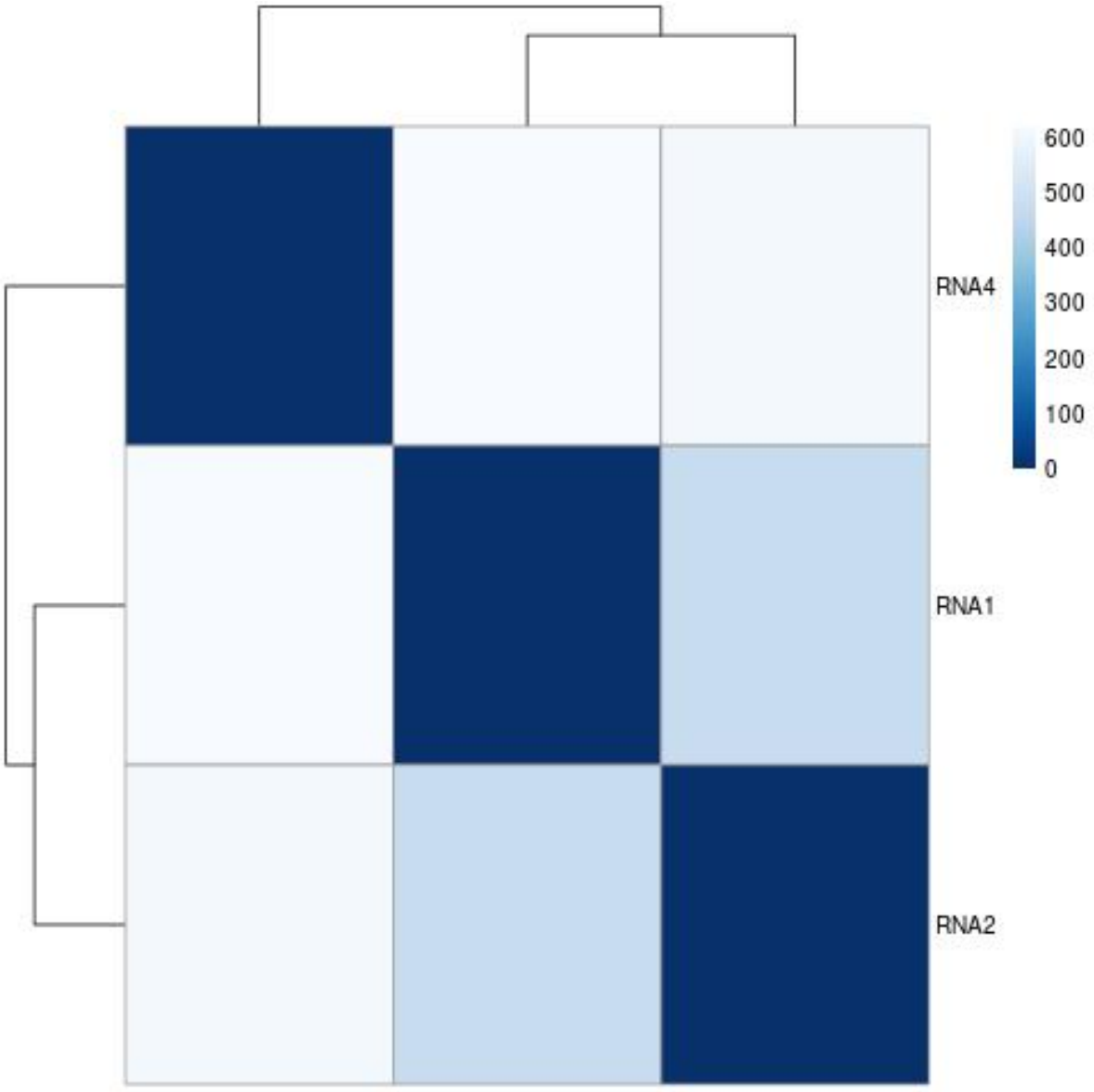


Figure 7: Example of heatmap denoting the distance between the three samples (RNA1, RNA2 and RNA4), illustrated in a colour gradient, with clustering of the distance values.

5

CONCLUSION

MOSCA was developed as a Python command line tool for full automation of MG and MT data analysis. MOSCA starts by applying successfully several pre-processing steps to raw input reads, then assemblies the reads into contigs and annotates the resulting sequences into taxonomic and biosystems information. For every main step there is a quality report, either obtained through custom scripts inside MOSCA or through the functionalities already present in other tools.

A major improvement is that MOSCA adapts the datasets trimming by adjusting Trimmomatic's arguments based on FastQC's reports. For MT experiments, after the annotation step, MOSCA builds a normalized matrix and generates a script that allows for DE analysis in R, thus allowing for multi-sample comparison of transcriptomes. MOSCA is, therefore, the third tool to integrate MG and MT analysis, after IMP and FMAP.

The results obtained are promising but there is space for several improvements. In the future, MOSCA can be easily expanded to include MP and/or Meta-metabolomics (MM) data analysis as well. Also, having a Graphical User Interface (GUI) would be useful to turn this automated pipeline into a more user friendly tool for integrated meta-omics data analysis. Even though MOSCA is developed to maximize automation, more user input could allow for human augmented data analysis - VizBin is a good example, where the user handles the data in an interactively, visual way, for binning the contigs into subsets.

BIBLIOGRAPHY

Aguiar-pulido, V., Huang, W., Suarez-ulloa, V., Cickovski, T., Mathee, K. and Narasimhan, G. (2016), 'Approaches for Microbiome Analysis', **12**, 5–16.

Akiva, E., Brown, S., Almonacid, D. E., Barber, A. E., Custer, A. F., Hicks, M. A., Huang, C. C., Lauck, F., Mashiyama, S. T., Meng, E. C., Mischel, D., Morris, J. H., Ojha, S., Schnoes, A. M., Stryke, D., Yunes, J. M., Ferrin, T. E., Holliday, G. L. and Babbitt, P. C. (2014), 'The Structure-Function Linkage Database', *Nucleic Acids Research* **42**(D1), 521–530.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990), 'Basic local alignment search tool', *Journal of molecular biology* **215**(3), 403–410.

Andrews, S. et al. (2010), 'Fastqc: a quality control tool for high throughput sequence data'.

Arumugam, M., Harrington, E. D., Foerstner, K. U., Raes, J. and Bork, P. (2010), 'Smashcommunity: a metagenomic annotation and analysis tool', *Bioinformatics* **26**(23), 2977–2978.

Attwood, T. K. (2002), 'The PRINTS database: A resource for identification of protein families', *Briefings in Bioinformatics* **3**(3), 252–263.

Bahassi, E. M. and Stambrook, P. J. (2014), 'Next-generation sequencing technologies: Breaking the sound barrier of human genetics', *Mutagenesis* **29**(5), 303–310.

Baker, B. J., Sheik, C. S., Taylor, C. a., Jain, S., Bhasi, A., Cavalcoli, J. D. and Dick, G. J. (2013), 'Community transcriptomic assembly reveals microbes that contribute to deep-sea carbon and nitrogen cycling.', *The ISME journal* **7**(10), 1962–73.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Grif, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C. and Eddy, S. R. (2004), 'The Pfam protein families database.', *Nucleic Acids Res* **32**(Database issue), 138D–41.

Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Apweiler, R., Alpi, E., Antunes, R., Arganiska, J., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas, B., Chavali, G., Cibrian-Uhalte, E., Da Silva, A., De Giorgi, M., Dogan, T., Fazzini, F., Gane, P., Castro, L. G., Garmiri, P., Hatton-Ellis, E., Hieta, R., Huntley, R., Legge, D., Liu, W., Luo, J., Macdougall, A., Mutowo, P., Nightingale, A., Orchard, S., Pichler, K., Poggioli, D., Pundir, S., Pureza, L., Qi, G., Rosanoff, S., Saidi, R., Sawford, T., Shypitsyna, A., Turner, E., Volynkin, V., Wardell, T., Watkins, X., Zellner, H., Cowley, A., Figueira, L., Li, W., McWilliam,

H., Lopez, R., Xenarios, I., Bougueleret, L., Bridge, A., Poux, S., Redaschi, N., Aimo, L., Argoud-Puy, G., Auchincloss, A., Axelsen, K., Bansal, P., Baratin, D., Blatter, M. C., Boeckmann, B., Bolleman, J., Boutet, E., Breuza, L., Casal-Casas, C., De Castro, E., Coudert, E., Cuche, B., Doche, M., Dornevil, D., Duvaud, S., Estreicher, A., Famiglietti, L., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Jungo, F., Keller, G., Lara, V., Lemercier, P., Lieberherr, D., Lombardot, T., Martin, X., Masson, P., Morgat, A., Neto, T., Nouspikel, N., Paesano, S., Pedruzzi, I., Pilbout, S., Pozzato, M., Pruess, M., Rivoire, C., Roechert, B., Schneider, M., Sigrist, C., Sonesson, K., Staehli, S., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Veuthey, A. L., Wu, C. H., Arighi, C. N., Arminski, L., Chen, C., Chen, Y., Garavelli, J. S., Huang, H., Laiho, K., McGarvey, P., Natale, D. A., Suzek, B. E., Vinayaka, C. R., Wang, Q., Wang, Y., Yeh, L. S., Yerramalla, M. S. and Zhang, J. (2015), 'UniProt: A hub for protein information', *Nucleic Acids Research* **43**(D1), D204–D212.

Benson, C. A., Bizzoco, R. W., Lipson, D. A. and Kelley, S. T. (2011), 'Microbial diversity in nonsulfur, sulfur and iron geothermal steam vents', *FEMS Microbiology Ecology* **76**(1), 74–88.

Bikel, S., Valdez-Lara, A., Cornejo-Granados, F., Rico, K., Canizales-Quinteros, S., Soberón, X., Del Pozo-Yauner, L. and Ochoa-Leyva, A. (2015), 'Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: Towards a systems-level understanding of human microbiome', *Computational and Structural Biotechnology Journal* **13**, 390–401.

Bolger, A. M., Lohse, M. and Usadel, B. (2014), 'Trimmomatic: a flexible trimmer for illumina sequence data', *Bioinformatics* p. btu170.

Breitbart, M., Hewson, I., Felts, B., Mahaffy, J. M., Nulton, J., Salamon, P. and Rohwer, F. (2003), 'Metagenomic Analyses of an Uncultured Viral Community from Human Feces Metagenomic Analyses of an Uncultured Viral Community from Human Feces Downloaded from http://jb.asm.org/ on December 8 , 2013 by National Institute of Technology and Evaluation', *Journal of Bacteriology* **185**(20), 6220–6223.

Buchan, D. W. A., Shepherd, A. J., Lee, D., Pearl, F. M. G., Rison, S. C. G., Thornton, J. M. and Orengo, C. A. (2002), 'Gene3D: structural assignment for whole genes and genomes using the CATH domain structure database.', *Genome research* **12**(3), 503–14.

Buchfink, B., Xie, C. and Huson, D. H. (2015), 'Fast and sensitive protein alignment using diamond', *Nature methods* **12**(1), 59–60.

Buermans, H. P. J. and Den Dunnen, J. T. (2014), 'Next generation sequencing technology: Advances and applications.', *Biochimica et biophysica acta* **1842**(10), 1932–1941.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R. and Apweiler, R. (2004), 'The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology', *Nucleic acids research* **32**(suppl 1), D262–D266.

Carvalhais, L. C., Dennis, P. G., Tyson, G. W. and Schenk, P. M. (2012), 'Application of metatranscriptomics to soil environments', *Journal of Microbiological Methods* **91**(2), 246–251.

Celaj, A., Markle, J., Danska, J. and Parkinson, J. (2014), 'Comparison of assembly algorithms for improving rate of metatranscriptomic functional annotation.', *Microbiome* **2**(1), 39.

Chamberlain, R. and Schommer, J. (2014), 'Using docker to support reproducible research', *DOI: http://dx. doi. org/10.6084/m9. figshare* **1101910**.

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L. and Rice, P. M. (2010), 'The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants', *Nucleic acids research* **38**(6), 1767–1771.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M. and Robles, M. (2005), 'Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research', *Bioinformatics* **21**(18), 3674–3676.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X. et al. (2016), 'A survey of best practices for rna-seq data analysis', *Genome biology* **17**(1), 13.

Corpet, F. (1998), 'The ProDom database of protein domain families', *Nucleic Acids Research* **26**(1), 323–326.

Cox, M. P., Peterson, D. A. and Biggs, P. J. (2010), 'Solexaqa: At-a-glance quality assessment of illumina second-generation sequencing data', *BMC bioinformatics* **11**(1), 485.

Damon, C., Lehembre, F., Oger-Desfeux, C., Luis, P., Ranger, J., Fraissinet-Tachet, L. and Marmeisse, R. (2012), 'Metatranscriptomics reveals the diversity of genes expressed by eukaryotes in forest soils', *PLoS ONE* **7**(1).

Dudhagara, P., Bhavsar, S., Bhagat, C., Ghelani, A., Bhatt, S. and Patel, R. (2015), 'Web Resources for Metagenomics Studies', *Genomics, Proteomics and Bioinformatics* **13**(5), 296–303.

Edgar, R. C. (2010), 'Search and clustering orders of magnitude faster than blast', *Bioinformatics* **26**(19), 2460–2461.

Emanuelsson, O., Brunak, S., von Heijne, G. and Nielsen, H. (2007), 'Locating proteins in the cell using TargetP, SignalP and related tools', *Nat. Protocols* **2**(4), 953–971.

et al. Venter, J. (2004), 'Environmental Genome Shotgun Sequencing of the', *Science* **1093857**(2004), 304.

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., Chang, H.-Y., Dosztányi, Z., El-Gebali, S., Fraser, M., Gough, J., Haft, D., Holliday, G. L., Huang, H., Huang, X., Letunic, I., Lopez, R., Lu, S., Marchler-Bauer, A., Mi, H., Mistry, J., Natale, D. A., Necci, M., Nuka, G., Orengo, C. A., Park, Y., Pesseat, S., Piovesan, D., Potter, S. C., Rawlings, N. D., Redaschi, N., Richardson, L., Rivoire, C., Sangrador-Vegas, A., Sigrist, C., Sillitoe, I., Smithers, B., Squizzato, S., Sutton, G., Thanki, N., Thomas, P. D., Tosatto, S. C. E., Wu, C. H., Xenarios, I., Yeh, L.-S., Young, S.-Y. and Mitchell, A. L. (2016), 'InterPro in 2017-beyond protein family and domain annotations.', *Nucleic acids research* **45**(November 2016), gkw1107.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J. and Punta, M. (2014), 'Pfam: the protein families database.', *Nucleic acids research* **42**(Database issue), D222–30.

García-Moyano, A., González-Toril, E., Aguilera, Á. and Amils, R. (2012), 'Comparative microbial ecology study of the sediments and the water column of the R??o Tinto, an extreme acidic environment', *FEMS Microbiology Ecology* **81**(2), 303–314.

Glass, E. M., Wilkening, J., Wilke, A., Antonopoulos, D. and Meyer, F. (2010), 'Using the metagenomics rast server (mg-rast) for analyzing shotgun metagenomes', *Cold Spring Harbor Protocols* **2010**(1), pdb–prot5368.

Gołębiewski, M., Deja-Sikora, E., Cichosz, M., Tretyn, A. and Wróbel, B. (2014), '16S rDNA pyrosequencing analysis of bacterial community in heavy metals polluted soils.', *Microbial ecology* **67**(3), 635–47.

Gordon, A. and Hannon, G. (2010), 'Fastx-toolkit', *FASTQ/A short-reads preprocessing tools (unpublished) http://hannonlab. cshl. edu/fastx_toolkit* .

Gough, J. (2002), 'SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments', *Nucleic Acids Research* **30**(1), 268–272.

Gu, S., Fang, L. and Xu, X. (2013), 'Using soapaligner for short reads alignment', *Current Protocols in Bioinformatics* pp. 11–11.

Haft, D. H. (2003), 'The TIGRFAMs database of protein families', *Nucleic Acids Research* **31**(1), 371–373.

Heyer, R., Schallert, K., Zoun, R., Becher, B., Saake, G. and Benndorf, D. (2017), 'Challenges and perspectives of metaproteomic data analysis', *Journal of biotechnology* **261**, 24–36.

Huang, X. and Madan, A. (1999), 'Cap3: A dna sequence assembly program', *Genome research* **9**(9), 868–877.

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., Mcanulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A. F., Selengut, J. D., Sigrist, C. J. A., Thimma, M., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H. and Yeats, C. (2009), 'InterPro: The integrative protein signature database', *Nucleic Acids Research* **37**(SUPPL. 1), 211–215.

Hunter, S., Corbett, M., Denise, H., Fraser, M., Gonzalez-Beltran, A., Hunter, C., Jones, P., Leinonen, R., McAnulla, C., Maguire, E., Maslen, J., Mitchell, A., Nuka, G., Oisel, A., Pesseat, S., Radhakrishnan, R., Rocca-Serra, P., Scheremetjew, M., Sterk, P., Vaughan, D., Cochrane, G., Field, D. and Sansone, S. A. (2014), 'EBI metagenomics - A new resource for the analysis and archiving of metagenomic data', *Nucleic Acids Research* **42**(D1), 1–7.

Huson, D. H., Auch, A. F., Qi, J. and Schuster, S. C. (2007), 'Megan analysis of metagenomic data', *Genome research* **17**(3), 377–386.

Illumina (2015), 'An Introduction to Next-Generation Sequencing Technology', *Illumina.com* (illumina), 1 – 16.

Illumina Proprietary (n.d.), 'Microbes and Metagenomics in Human Health'.

Ishii, S., Suzuki, S., Tenney, A., Norden-Krichmar, T. M., Nealson, K. H. and Bretschger, O. (2015), 'Microbial metabolic networks in a complex electrogenic biofilm recovered from a stimulus-induced metatranscriptomics approach', *Scientific Reports* **5**(October), 14840.

Jones, D. C., Ruzzo, W. L., Peng, X. and Katze, M. G. (2012), 'Compression of next-generation sequencing reads aided by highly efficient de novo assembly', *Nucleic acids research* p. gks754.

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S. Y., Lopez, R. and Hunter, S. (2014), 'InterProScan 5: Genome-scale protein function classification', *Bioinformatics* **30**(9), 1236–1240.

Jung, J. Y., Lee, S. H., Jin, H. M., Hahn, Y., Madsen, E. L. and Jeon, C. O. (2013), 'Metatranscriptomic analysis of lactic acid bacterial gene expression during kimchi fermentation', *International Journal of Food Microbiology* **163**(2-3), 171–179.

Kall, L., Krogh, A. and Sonnhammer, E. L. L. (2004), 'A combined transmembrane topology and signal peptide prediction method.', *Journal of molecular biology* **338**(5), 1027–1036.

Kanehisa, M. and Goto, S. (2000), 'KEGG: Kyoto Encyclopedia of Genes and Genomes'.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M. (2016), 'KEGG as a reference resource for gene and protein annotation', *Nucleic Acids Research* **44**(D1), D457–D462.

Kent, W. J. (2002), 'Blatthe blast-like alignment tool', *Genome research* **12**(4), 656–664.

Kim, J., Kim, M. S., Koh, A. Y., Xie, Y., Zhan, X., Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Abubucker, S., Segata, N., Goll, J., Schubert, A., Izard, J., Cantarel, B., Rodriguez-Mueller, B., Zucker, J., Thiagarajan, M., Henrissat, B., Huson, D., Weber, N., Yi, G., Sze, S., Thon, M., Kristiansson, E., Hugenholtz, P., Dalevi, D., Rotmistrovsky, K., Agarwala, R., Edgar, R., Buchfink, B., Xie, C., Huson, D., UniProt, C., Paulson, J., Stine, O., Bravo, H., Pop, M., Okuda, S., Yoshizawa, A., Papin, J., Stelling, J., Price, N., Klamt, S., Schuster, S., Palsson, B., Khatri, P., Sirota, M., Butte, A., Huang, W., Li, L., Myers, J., Marth, G., Turnbaugh, P., Ley, R., Hamady, M., Fraser-Liggett, C., Knight, R., Gordon, J., DeLong, E., Preston, C., Mincer, T., Rich, V., Hallam, S., Frigaard, N., Martinez, A., Sullivan, M., Edwards, R., Brito, B., Belda-Ferre, P., Alcaraz, L., Cabrera-Rubio, R., Romero, H., Simon-Soro, A., Pignatelli, M., Mira, A., Erickson, A., Cantarel, B., Lamendella, R., Darzi, Y., Mongodin, E., Pan, C., Shah, M., Halfvarson, J., Tysk, C., Henrissat, B., Tobe, T., Nakanishi, N. and Sugimoto, N. (2016), 'FMAP: Functional Mapping and Analysis Pipeline for metagenomics and metatranscriptomics studies', *BMC Bioinformatics* **17**(1), 420.

Kopylova, E., No, L. and Touzet, H. (2012), 'Sortmerna: fast and accurate filtering of ribosomal rnas in metatranscriptomic data', *Bioinformatics* **28**(24), 3211.

Kristiansson, E., Hugenholtz, P. and Dalevi, D. (2009), 'Shotgunfunctionalizer: an r-package for functional comparison of metagenomes', *Bioinformatics* **25**(20), 2737–2738.

Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E. L. (2001), 'Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.', *Journal of molecular biology* **305**(3), 567–80.

Kultima, J. R., Sunagawa, S., Li, J., Chen, W., Chen, H., Mende, D. R., Arumugam, M., Pan, Q., Liu, B., Qin, J., Wang, J. and Bork, P. (2012), 'MOCAT : A Metagenomics Assembly and Gene Prediction Toolkit', **7**(10), 1–6.

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K. and Hugenholtz, P. (2008), 'A bioinformatician's guide to metagenomics.', *Microbiology and molecular biology reviews : MMBR* **72**(4), 557–78, Table of Contents.

Laczny, C. C., Sternal, T., Plugaru, V., Gawron, P., Atashpendar, A., Margossian, H. H., Coronado, S., Van der Maaten, L., Vlassis, N. and Wilmes, P. (2015), 'Vizbin-an application for reference-independent visualization and human-augmented binning of metagenomic data', *Microbiome* **3**(1), 1.

Ladoukakis, E., Kolisis, F. N. and Chatziioannou, A. A. (2014), 'Integrative workflows for metagenomic analysis.', *Frontiers in cell and developmental biology* **2**(November), 70.

Langmead, B. and Salzberg, S. L. (2012), 'Fast gapped-read alignment with bowtie 2', *Nature methods* **9**(4), 357–359.

Lee, J.-H., Yi, H. and Chun, J. (2011), 'rrnaselector: a computer program for selecting ribosomal rna encoding sequences from metagenomic and metatranscriptomic shotgun libraries', *The Journal of Microbiology* **49**(4), 689–691.

Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J., Ponting, C. P. and Bork, P. (2004), 'SMART 4.0: towards genomic data integration.', *Nucleic acids research* **32**(Database issue), D142–4.

Li, D., Liu, C.-M., Luo, R., Sadakane, K. and Lam, T.-W. (2015), 'Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph', *Bioinformatics* p. btv033.

Li, H. and Durbin, R. (2009), 'Fast and accurate short read alignment with burrowswheeler transform', *Bioinformatics* **25**(14), 1754.

Lima, T., Auchincloss, A. H., Coudert, E., Keller, G., Michoud, K., Rivoire, C., Bulliard, V., de Castro, E., Lachaize, C., Baratin, D., Phan, I., Bougueleret, L. and Bairoch, A. (2009), 'HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot'.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. and Law, M. (2012), 'Comparison of next-generation sequencing systems', *Journal of Biomedicine and Biotechnology* **2012**.

Lupas, A., Van Dyke, M. and Stock, J. (1991), 'Predicting coiled coils from protein sequences.', *Science (New York, N.Y.)* **252**(5009), 1162–4.

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., Geer, R. C., He, J., Gwadz, M., Hurwitz, D. I., Lanczycki, C. J., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Wang, Z., Yamashita, R. A., Zhang, D., Zheng, C. and Bryant, S. H. (2015), 'CDD: NCBI's conserved domain database', *Nucleic Acids Research* **43**(D1), D222–D226.

Markowitz, V. M., Ivanova, N. N., Szeto, E., Palaniappan, K., Chu, K., Dalevi, D., Chen, I.-M. A., Grechkin, Y., Dubchak, I., Anderson, I. et al. (2008), 'Img/m: a data management and analysis system for metagenomes', *Nucleic acids research* **36**(suppl 1), D534–D538.

Martinez, X., Pozuelo, M., Pascal, V., Campos, D., Gut, I., Gut, M., Azpiroz, F., Guarner, F. and Manichanh, C. (2016), 'MetaTrans: an open-source pipeline for metatranscriptomics.', *Scientific reports* **6**, 26447.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A. et al. (2008), 'The metagenomics rast server–a public resource for the automatic phylogenetic and functional analysis of metagenomes', *BMC bioinformatics* **9**(1), 386.

Mikheenko, A., Saveliev, V. and Gurevich, A. (2016), 'Metaquast: evaluation of metagenome assemblies', *Bioinformatics* **32**(7), 1088–1090.

Namiki, T., Hachiya, T., Tanaka, H. and Sakakibara, Y. (2012), 'Metavelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads', *Nucleic acids research* **40**(20), e155–e155.

Narayanasamy, S., Jarosz, Y., Muller, E. E., Laczny, C. C., Herold, M., Kaysen, A., Heintz-Buschart, A., Pinel, N., May, P. and Wilmes, P. (2016), 'IMP: a pipeline for reproducible metagenomic and metatranscriptomic analyses', *bioRxiv* (7), 039263.

Nayfact, S., Rodriguez-Mueller, B., Garud, N. and Pollard, K. (2016), 'An integrated metagenomics pipeline for strain profiling reveals novel patterns of transmission and global biogeography of bacteria', *bioRxiv* **53**(9), 1689–1699.

Nurk, S., Meleshko, D., Korobeynikov, A. and Pevzner, P. (2016), 'metaspades: a new versatile de novo metagenomics assembler', *arXiv preprint arXiv:1604.03071* .

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D. and Pruitt, K. D. (2016), 'Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation', *Nucleic Acids Research* **44**(D1), D733–D745.

Ondov, B. D., Bergman, N. H. and Phillippy, A. M. (2011), 'Interactive metagenomic visualization in a web browser', *BMC bioinformatics* **12**(1), 1.

Oulas, A., Pavloudi, C., Polymenakou, P., Pavlopoulos, G. A., Papanikolaou, N., Kotoulas, G., Arvanitidis, C. and Iliopoulos, I. (2015), 'Metagenomics: Tools and insights for analyzing next-generation sequencing data derived from biodiversity studies', *Bioinformatics and Biology Insights* **9**, 75–88.

Overview, A. and Illumina, P. F. (2012), 'Metagenomics Research Review', *Illumina* p. 38.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. and Tyson, G. W. (2015), 'Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes', *Genome research* **25**(7), 1043–1055.

Patel, R. K. and Jain, M. (2012), 'Ngs qc toolkit: A toolkit for quality control of next generation sequencing data', *PLOS ONE* **7**(2), 1–7.

Pearce, D. A., Newsham, K. K., Thorne, M. A. S., Calvo-Bado, L., Krsek, M., Laskaris, P., Hodson, A. and Wellington, E. M. (2012), 'Metagenomic analysis of a southern maritime Antarctic soil', *Frontiers in Microbiology* **3**(DEC).

Peng, Y., Leung, H. C., Yiu, S.-M. and Chin, F. Y. (2012), 'Idba-ud: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth', *Bioinformatics* **28**(11), 1420–1428.

Plummer, E. and Twin, J. (2015), 'A Comparison of Three Bioinformatics Pipelines for the Analysis of Preterm Gut Microbiota using 16S rRNA Gene Sequencing Data', *Journal of Proteomics & Bioinformatics* **8**(12), 283–291.

Poulsen, M., Schwab, C., Jensen, B. B., Engberg, R. M., Spang, A., Canibe, N., Højberg, O., Milinovich, G., Fragner, L., Schleper, C., Weckwerth, W., Lund, P., Schramm, A. and Urich, T. (2013), 'Methylotrophic methanogenic Thermoplasmata implicated in reduced methane emissions from bovine rumen.', *Nature communications* **4**, 1428.

Powell, S., Szklarczyk, D., Trachana, K., Roth, A., Kuhn, M., Muller, J., Arnold, R., Rattei, T., Letunic, I., Doerks, T. et al. (2012), 'eggnog v3. 0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges', *Nucleic acids research* **40**(D1), D284–D289.

Pruitt, K. D., Tatusova, T. and Maglott, D. R. (2007), 'NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins', *Nucleic Acids Research* **35**(SUPPL. 1), 501–504.

Quail, M., Smith, M. E., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., Gu, Y., Rothberg, J., Hinz, W., Rearick, T., Schultz, J., Mileski, W.,

Davey, M., Leamon, J., Johnson, K., Milgrew, M., Edwards, M., Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bentley, D., Balasubramanian, S., Swerdlow, H., Smith, G., Milton, J., Brown, C., Hall, K., Evers, D., Barnes, C., Bignell, H., Kozarewa, I., Ning, Z., Quail, M., Sanders, M., Berriman, M., Turner, D., Quail, M., Otto, T., Gu, Y., Harris, S., Skelly, T., McQuillan, J., Swerdlow, H., Oyola, S., Syed, F., Grunenwald, H., Caruccio, N., Lam, H., Clark, M., Chen, R., Chen, R., Natsoulis, G., O'Huallachain, M., Dewey, F., Habegger, L., Carver, T., Harris, S., Berriman, M., Parkhill, J., McQuillan, J., Ponsting, N., Ning, Z., Otto, T., Sanders, M., Berriman, M., Newbold, C., Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M., Hirai, A., Takahashi, H., Diep, B., Gill, S., Chang, R., Phan, T., Chen, J., Davidson, M., Lin, F., Lin, J., Carleton, H., Mongodin, E., Achidi, E., Gardner, M., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R., Carlton, J., Pain, A., Nelson, K., Bowman, S., Choi, M., Scholl, U., Ji, W., Liu, T., Tikhonova, I., Zumbo, P., Nayir, A., Bakkaloglu, A., Ozen, S., Sanjad, S., Down, T., Rakyan, V., Turner, D., Flicek, P., Li, H., Kulesha, E., Graf, S., Johnson, N., Herrero, J., Tomazou, E., Giresi, P., Kim, J., McDaniell, R., Iyer, V., Lieb, J., Johnson, D., Mortazavi, A., Myers, R., Wold, B., Langridge, G., Phan, M., Turner, D., Perkins, T., Parts, L., Haase, J., Charles, I., Maskell, D., Peters, S., Dougan, G., Licatalosi, D., Mele, A., Fak, J., Ule, J., Kayikci, M., Chi, S., Clark, T., Schweitzer, A., Blume, J., Wang, X., Mamanova, L., Andrews, R., James, K., Sheridan, E., Ellis, P., Langford, C., Ost, T., Collins, J., Turner, D., Myllykangas, S., Buenrostro, J., Natsoulis, G., Bell, J., Ji, H., Shao, N., Hu, H., Yan, Z., Xu, Y., Hu, H., Menzel, C., Li, N., Chen, W., Khaitovich, P., Wang, Z., Gerstein, M., Snyder, M., Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F., Burton, J., Walker, B., Sharpe, T., Hall, G., Shea, T., Sykes, S., Levin, J., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D., Friedman, N., Gnirke, A., Regev, A., Adey, A., Asan, Xun, X., Kitzman, J., Turner, E., Stackhouse, B., MacKenzie, A., Caruccio, N., Zhang, X., Flusberg, B., Webster, D., Lee, J., Travers, K., Olivares, E., Clark, T., Korlach, J., Turner, S., Holden, T., Lindsay, J., Corton, C., Quail, M., Cockfield, J., Pathak, S., Batra, R., Parkhill, J., Bentley, S., Edgeworth, J., Li, H., Durbin, R., Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Angiuoli, S. and Salzberg, S. (2012), 'A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers', *BMC Genomics* **13**(1), 341.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glöckner, F. O. (2012), 'The silva ribosomal rna gene database project: improved data processing and web-based tools', *Nucleic acids research* **41**(D1), D590–D596.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R. and Lopez, R. (2005), 'InterProScan: Protein domains identifier', *Nucleic Acids Research* **33**(SUPPL. 2), 116–120.

Quinlan, A. R. and Hall, I. M. (2010), 'Bedtools: a flexible suite of utilities for comparing genomic features', *Bioinformatics* **26**(6), 841–842.

Rho, M., Tang, H. and Ye, Y. (2010), 'Fraggenescan: predicting genes in short and error-prone reads', *Nucleic acids research* **38**(20), e191–e191.

Rotmistrovsky, K. and Agarwala, R. (2011), 'Bmtagger: Best match tagger for removing human reads from metagenomics datasets'.

Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P. and Bork, P. (2000), 'SMART: a web-based tool for the study of genetically mobile domains.', *Nucleic acids research* **28**(1), 231–4.

Seemann, T. (2014), 'Prokka: rapid prokaryotic genome annotation', *Bioinformatics* p. btu153.

Sigrist, C. J. A. (2002), 'PROSITE: A documented database using patterns and profiles as motif descriptors', *Briefings in Bioinformatics* **3**(3), 265–274.

Sommer, D. D., Delcher, A. L., Salzberg, S. L. and Pop, M. (2007), 'Minimus: a fast, lightweight genome assembler', *BMC bioinformatics* **8**(1), 64.

Stevens, H. and Ulloa, O. (2008), 'Bacterial diversity in the oxygen minimum zone of the eastern tropical South Pacific', *Environmental Microbiology* **10**(5), 1244–1259.

Tan, B., Jane Fowler, S., Laban, N. A., Dong, X., Sensen, C. W., Foght, J. and Gieg, L. M. (2015), 'Comparative analysis of metagenomes from three methanogenic hydrocarbon-degrading enrichment cultures with 41 environmental samples', *The ISME Journal* **9**(9), 2028–2045.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J. and Natale, D. A. (2003), 'The COG database: an updated version includes eukaryotes.', *BMC bioinformatics* **4**, 41.

Thomas, P. D., Kejariwal, A., Campbell, M. J., Mi, H., Diemer, K., Guo, N., Ladunga, I., Ulitsky-Lazareva, B., Muruganujan, A., Rabkin, S., Vandergriff, J. A. and Doremieux, O. (2003), 'PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification'.

Thomas, T., Gilbert, J. and Meyer, F. (2012), 'Metagenomics - a guide from sampling to data analysis', *Microbial Informatics and Experimentation* **2**(1), 3.

Treangen, T. J., Koren, S., Sommer, D. D., Liu, B., Astrovskaya, I., Ondov, B., Darling, A. E., Phillippy, A. M. and Pop, M. (2013), 'Metamos: a modular and open source metagenomic assembly and analysis pipeline', *Genome biology* **14**(1), R2.

Tringe, S. G. and Rubin, E. M. (2005), 'Metagenomics: DNA sequencing of environmental samples.', *Nature reviews. Genetics* **6**(11), 805–14.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S. and Banfield, J. F. (2004), 'Community structure and metabolism through reconstruction of microbial genomes from the environment.', *Nature* **428**(6978), 37–43.

UniProt (2010), 'The Universal Protein Resource', **2010**(November 2007), 190–195.

Urich, T., Lanzén, A., Stokke, R., Pedersen, R. B., Bayer, C., Thorseth, I. H., Schleper, C., Steen, I. H. and Øvreas, L. (2014), 'Microbial community structure and functioning in marine sediments associated with diffuse hydrothermal venting assessed by integrated meta-omics', *Environmental Microbiology* **16**(9), 2699–2710.

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y. and Thermes, C. (2014), 'Ten years of next-generation sequencing technology', *Trends in Genetics* pp. 1–9.

Vollmers, J., Wiegand, S. and Kaster, A.-k. (2017), *Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist ' s Perspective - Not Only Size Matters !*

Westreich, S. T., Korf, I., Mills, D. A. and Lemay, D. G. (2016), 'SAMSA: A comprehensive metatranscriptome analysis pipeline', *bioRxiv* p. 046201.

Wilke, A., Bischof, J., Gerlach, W., Glass, E., Harrison, T., Keegan, K. P., Paczian, T., Trimble, W. L., Bagchi, S., Grama, A., Chaterji, S. and Meyer, F. (2016), 'The MG-RAST metagenomics database and portal in 2015', *Nucleic Acids Research* **44**(D1), D590–D594.

Wu, C. H., Nikolskaya, A., Huang, H., Yeh, L.-S. L., Natale, D. A., Vinayaka, C. R., Hu, Z.-Z., Mazumder, R., Kumar, S., Kourtesis, P., Ledley, R. S., Suzek, B. E., Arminski, L., Chen, Y., Zhang, J., Cardenas, J. L., Chung, S., Castro-Alvear, J., Dinkov, G. and Barker, W. C. (2004), 'PIRSF: family classification system at the Protein Information Resource.', *Nucleic acids research* **32**(Database issue), D112–4.

Xiong, J., Liu, Y., Lin, X., Zhang, H., Zeng, J., Hou, J., Yang, Y., Yao, T., Knight, R. and Chu, H. (2012), 'Geographic distance and pH drive bacterial distribution in alkaline lake sediments across Tibetan Plateau', *Environmental Microbiology* **14**(9), 2457–2466.

Xiong, X., Frank, D. N., Robertson, C. E., Hung, S. S., Markle, J., Canty, A. J., McCoy, K. D., Macpherson, A. J., Poussier, P., Danska, J. S. and Parkinson, J. (2012), 'Generation and analysis of a mouse intestinal metatranscriptome through illumina based rna-sequencing', *PLOS ONE* **7**(4), 1–15.

Zdobnov, E. M. and Apweiler, R. (2001), 'InterProScan–an integration platform for the signature-recognition methods in InterPro.', *Bioinformatics (Oxford, England)* **17**(9), 847–848.

Žifčáková, L., Větrovskỳ, T., Howe, A. and Baldrian, P. (2016), 'Microbial activity in forest soil reflects the changes in ecosystem properties between summer and winter', *Environmental microbiology* **18**(1), 288–301.