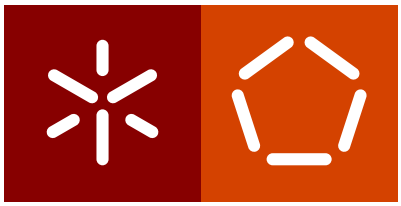
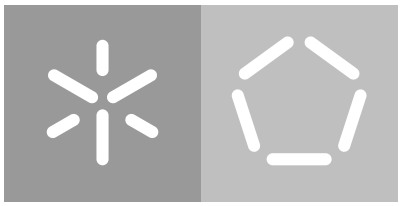


Universidade do Minho
Escola de Engenharia
Departamento de Informática

João Sequeira

Development of an automated pipeline for meta-omics data analysis

February, 2017



Universidade do Minho
Escola de Engenharia
Departamento de Informática

João Sequeira

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

February, 2017

ABSTRACT

Knowing what lies around us has been a goal for many decades now, and the new advances in sequencing technology and in meta-omics approaches have permitted to start answering main questions of microbiology - what is there, and what is it doing? In the last few years, some pipelines have been developed for handling metagenomics (MG), and more recently also metatranscriptomics (MT) and metaproteomics (MP) data. To date, however, there is no developed bioinformatics pipeline to integrate, in an automated way, MG, MT and MP analyses. Furthermore, the existing alternatives are usually not user friendly.

A pipeline for integrative MG, MT and MP data analysis will be developed in a user friendly web design. This pipeline aims to retrieve comprehensive comparative gene/protein expression results obtained from different biological samples. The user will be able to download the data at the end of each step, and final graphical representations will be provided. The pipeline will be constructed with tools tested and validated for meta-omics data analysis.

RESUMO

O objectivo dos microbiólogos, 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. No entanto, até à data, ainda não foi desenvolvida nenhuma pipeline bioinformática que integre, duma forma automatizada, análise metagenómica, metatranscriptómica e metaproteómica. Além disso, as alternativas existentes não costumam ser facilmente manipuláveis por utilizadores sem experiência em informática.

Uma pipeline para análise integrativa de dados de metagenómica, metatranscriptómica e metaproteómica será desenvolvida num formato web com uma interface fácil de usar. Esta pipeline tem como objetivo obter resultados comparativos de expressão génica/proteica entre várias amostras biológicas diferentes. O utilizador poderá descarregar resultados ao fim de cada fase, e serão disponibilizadas representações gráficas finais. Esta pipeline será desenvolvida com ferramentas testadas e validadas para análise de dados de meta-ómica.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

1

INTRODUCTION

NGS technologies have evolved rapidly during the last years, making possible the generation of a large amount of sequencing data obtained from a large variety of organisms. 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 exist, and although the majority are in house developed tools, usually non-available for the scientific community, others are free but usually require deep knowledge on bioinformatics, which most biologists and engineers don't have, or are not flexible because they were created to address specific questions.

More recently, open-source pipelines have been developed which have several advantages, but also drawbacks. One of the major limitations of both MP and MT, but also MG analysis, is the difficulty in choosing the reference database for the identification of genes and proteins. Using MG data to identify transcripts and proteins greatly increases the identification rates, which can increase even more if *de novo* options are considered instead of relying only in reference-based approaches.

The *Integrated Meta-omic Pipeline (IMP)* pipeline (Narayanasamy et al., 2016) considers these aspects, integrating MG with MT datasets in reference alignments or *de novo* genome assembly, the latter being especially important when a significant part of the microbial community is poorly described, thus having no genomic information in databases, which may compromise their functional characterization.

Functional Mapping and Analysis Pipeline (FMAP) (Kim et al., 2016) is the first publicly available implementation of differential quantification and analysis of MT data, with refer-

ence to MG data, making use of ShotgunFunctionalizerR (Kristiansson et al., 2009), an R package, for the differential analysis of RNA expression. Even though FMAP incorporates different analysis for MG and MT, it does lack the implementation of an important feature present in ShotgunFunctionalizerR, the differential analysis between samples.

The MetaProteomeAnalyzer (Muth et al., 2015) is, to date, the only example of an available open-source software for the analysis of MP data. It uses an user friendly interface and has the advantage of identifying a larger number of proteins by merging the results obtained with different database search algorithms, namely OMSSA, Crux, InsPect and X!Tandem, and to reduce the protein redundancy by grouping homologous proteins from different microorganisms in meta-proteins. This last advantage can be seen as a disadvantage when the user wants to assign the proteins identified to specific species.

Even though many computational solutions have already been developed for the study of MG data, only very recently tools for MT and MP analyses have been developed. Furthermore, there is no pipeline developed for an integrated study of MG, MT and MP data in a single workflow. Therefore, a user-friendly pipeline for comparative meta-omics analysis which integrates data from metagenomics with data from metatrancriptomics and metaproteomics is missing.

1.1 OBJECTIVES AND PLAN

The present work proposes the development and implementation of an easy-to-use bioinformatics pipeline for integrated and comparative analysis of metagenomics and metatranscriptomics and metagenomics and metaproteomics data. The objective will be to use open-source pipelines, identify pitfalls and integrate novel tools for a comparative and comprehensive analysis of the data output.

To attain the main aim of the work, a number of technical / scientific tasks have been identified, namely:

1. To review relevant available tools in the field, to identify their advantages and disadvantages and to write a state-of-the-art.
2. To test selected tools/pipelines with simulated and real datasets, obtained from anaerobic digesters from the host group.
3. To create or adapt additional tools which can improve data analysis and comparison, and define and construct workflows for different types of analyses.
4. To test the constructed pipeline with simulated and real datasets, and compare with the results from third party software.
5. To write the thesis and other publications.

The timeline proposed for this work is provided in Figure 1.

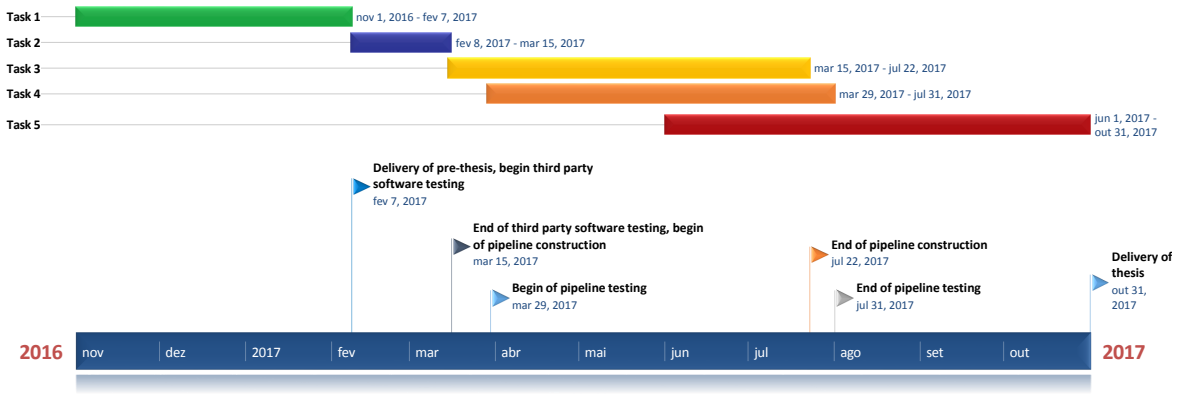


Figure 1: Schedule of the tasks proposed in this work.

2

STATE OF THE ART

2.1 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, and used radioactive chemicals. This technique went on as the main solution to DNA sequencing until the end of the Human Genome Project, although it was expensive and slow. At the time, sequencing could not represent a standard of genomic studies, which instead had to rely on genome-wide association studies using SNP-arrays for the comparison of different genomes.

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 RNA subunit to complete genomes of single organisms or even the entire population, in such low times and with costs that were considered scientific fiction 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 this 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 100*Base 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 8-10Gb 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 through the years, until 2010, when Illumina launched the HiSeq 2000, which had the biggest output per run up to 600 Gb, which it obtains 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, presenting itself more indicated for focused 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, 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 100$ 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 metagenomic studies in which Whole Genome Sequencing is used, since it involves large quantities of genomic material to be 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 - thus it is not an NGS technique. 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 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: NGS summary (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 XL+	Synthesis	Pyrophosphate detection	Single end	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

Many techniques currently exist, with different weaknesses and strengths (see Table 1), but 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 *Bioinformatics tools for Next Generation Sequencing*

NGS methods can produce large ammounts of data, a fact that increases in importance when dealing with meta-omics studies. There are optimized informatics tools that handle such large amount 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, as to allow for easier workflows.

Quality control

In the first *in silico* step of a post-NGS bioinformatics workflow, the datasets of reads produced by sequencing usually undergo a quality check. FASTQ format is the 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 (**?**). In addition, FASTQ comes in several 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.

Preprocessing

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.

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 it is available in databases 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 genome 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 assemblies by further merging contigs into less, 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, or MetaQUAST (Mikheenko et al., 2016) could be used as alternative.

Bining

For population studies, bining 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.

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 in them, first by identifying *Open Reading Frame (ORF)*s in the sequences 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) is the tool that allows for access to InterPro domain information.

Statistical analysis

Many analyses may come from meta-omics studies, and even more varied are 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

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 citepondov2011interactive 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 , while MG-RAST (Glass et al., 2010) provides pie charts representative of several different communities' profiles and heatmaps with differential multisample analysis. MEGAN 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). !!!nao sei o que citar para o MEGAN, nao ha nada que me diga isto que veio do site deles: http://ab.inf.uni-tuebingen.de/software/megan6/

2.2 METAGENOMICS

2.2.1 *Techniques and applications*

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). An answer has come in the form of MG, the partial or complete sequencing of any genome present in a sample - either it be viruses, bacteria or human beings - which in turn serves as a reference to a well defined ecosystem that lacks variance in its biology and chemistry (Illumina Proprietary; 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.

As a first step, 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. After identifying the genes, annotation steps identify the possible transcripts and proteins expressed by the microorganisms.

Another approach is Marker Gene MG, which in the bacterial case has been used, for example, for 16S rRNA gene studies, which is used as a phylogenetic marker. 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 et al., 2012a) populations. But, identifying the microbial taxonomy and genomic potential in a sample is not enough, the next step towards understanding is to know what those identified species are doing there (Tringe and Rubin, 2005; Thomas et al., 2012).

2.2.2 *Metagenomics pipelines*

The first tools used for MG studies were developed for single genome approaches. Because of that, they were not suitable for assembling MG data since they included genomes from several 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. Most MG pipelines don't integrate MT approaches, however, which is an important handicap

when trying to figure out what pathways are being most expressed at the time of sampling (Dudhagara et al., 2015; Nayfact et al., 2016).

The pipelines are considerably diverse (see Table 2): some have fully integrated workflows for MG, from the quality assessment and assembly of sequencing raw data, to the annotation of genes. Installation is easier with some pipelines than with others, and some interfaces are easier to work with; some pipelines are less computationally intensive than others. Some pipelines integrate shotgun MG analysis only, others are exclusive to 16S rRNA analysis, while others possess the flexibility to work with both types of analyses. 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).

Table 2: Main steps of MG data analysis integrated in each pipeline. From Ladoukakis et al. (2014).

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

Ladoukakis et al. (2014) compared seven shotgun MG pipelines - CloVR-metagenomics, Galaxy platform (metagenomics), IMG/M, MetAMOS, MG-RAST, RAMMCAP and SmashCommunity - and two, namely MetAmos and SmashCommunity, were considered the most robust, versatile solutions, by the integration of all steps of MG bioinformatics analysis 2 and by the quality of the results, although not easy to operate by less experienced users. MG-RAST and IMG/M were the suggestions for the functional analysis for already assembled data, with the easier user interface and a database setup designed for the dissemination of results to the scientific community.

Dudhagara et al. (2015) 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 - and MG-RAST and IMG/M were considered the best, having considerably bigger databases and a wide range of annotation tools, followed by CDS Metagenomics and METAVIR. This consideration correlates with the use of these tools through the years (Figure 2). Again, all have specific strengths and weaknesses.

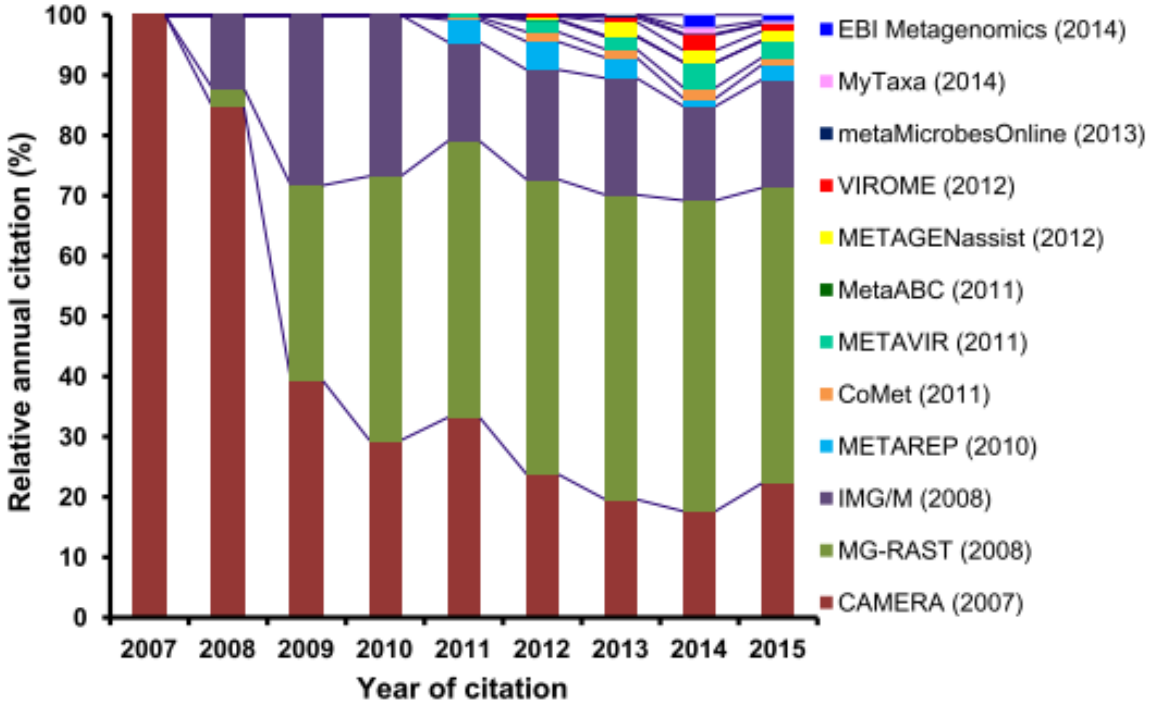


Figure 2: Utilization of web resources of MG analysis throughout the years. From Dudhagara et al. (2015).

Concerning 16S rRNA studies, QIIME is regarded as the most reliable 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

2.3.1 *Techniques and applications*

To know what pathways the microorganisms are utilising is related to what genes are being expressed more intensively, which points to the study of RNA. When collectively studying the RNA 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 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, thus allowing for less complex studies that generate more focused and useful information.

MT distinguishes itself from MG not only in its differential analysis application, but also by the different nature of RNA: for studies aiming at the study of mRNA, a depletion of rRNA is necessary, at the wet- and dry-lab level; 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. Nevertheless, 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 et al., 2012b), kimchi (Jung et al., 2013), bovine rumen (Poulsen et al., 2013) and deep-sea populations (Baker et al., 2013).

2.3.2 *Metatranscriptomics pipelines*

MT pipelines 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, obtaining information on the metabolic pathways expressed and also taxonomic annotation from mRNA analysis results (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 cDNA libraries for paired-ends 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 100bp or paired-end, however, which is not always easy to fulfil (Westreich et al., 2016).

2.4 METAGENOMICS AND METATRANSCRIPTOMICS PIPELINES

There has already been some work involving multiomics approaches, but most of these bioinformatics solutions have not been made public in an automated, well defined way. As such, their results are not easily reproducible, and this is one of the most important aspects of a fully automated pipeline for analysis of omics data. 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.

2.4.1 *IMP*

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 was the first 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, a comparative analysis of different datasets was not considered in this pipeline.

The preprocessing step of the workflow features trimming and quality filtering (by *Trimmomatic*) and ribosomal RNA filtering (by *SortMeRNA 2.0*). After that, 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).

2.4.2 *FMAP*

FMAP is the only implementation to date to integrate both MG and MT analysis in separate ways, coherent with the different nature of both data. The differential abundance of mRNA is not implemented in any other reviewed tool, which is a difficulty to overcome when trying to find the behaviour of a specific community, but besides attempting to solve such

problem, 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 the 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).

2.5 METAPROTEOMICS

2.5.1 *Methods and applications*

MP is the quantification of some or all proteins present in a medium. Even though MT may give information on which genes are being expressed at a given time, MP goes beyond MT in that it identifies proteins and quantifies their expression. Several experiments show that there is often a large discrepancy between RNA transcription and translation, between RNA levels encoding a certain protein and the effective quantity of such protein. Besides, in several ecosystems, many proteins produced by the organisms present in such ecossystems are associated with organic matter (notably, humic acids) and minerals so they can continue their activity even in the absence of the organims that produced them. So, an organism may not exist anymore in a certain medium, and its proteins still be present and active in the

ecossystem. Thus, if a MP work is focused towards classifying organisms taxonomically, it should consider the intracellular proteins (Bastida et al., 2009).

The process is divided in four steps:

1. Extraction - very important to ensure that proteins are not damaged, must consider pH, temperature, proteases and ions;
2. Purification;
3. Separation - through electrophoresis or liquid chromatography, among others, the extracted proteins are separated by their mass, polarity or charge;
4. Identification - mass spectrometry is the most usual technique, and may be done 'bottom-up' or 'top-down'.

Besides organism classification (which already is not a simple task), MP offers a wide range of applications through a perspective not possible before. To know the enzymatic activity the microbial community is to know the biogeochemical potential of the ecosystem it lies in, for example, as pollutant degradator. In the same perspective, knowing the ecosystem recovery potential of a microbial community may allow for programmes of ecosystem recovery, like for example, a specific combination of microbes to repair a certain condition of the ecosystem, or an enzyme or set of enzymes to repair a specific presence (or lack of it) of an important substance. Traditional studies concerning ecosystems quality have focused on ureases, proteases, glucosidases, phosphatases and xylanases, which are enzymes involved in fundamental but general processes, and as such cannot inform about the specific pathways involved in the microbiome balance (Bastida et al., 2009).

Besides its great potential as a tool for understanding microbiology, several challenges have hindered the expansion of MP, and every step of the process has its difficulties. Some medium, like for example the soil, are a poor source of proteins, since many are at very low concentrations - with no option to amplify the molecules obtained like in DNA or RNA sequencing - and protein distribution is usually not even, their location depending on the medium matrix. In the case of extracellular proteins, many are protected by connection with organic molecules, notably humic acids, which interfere with the separation of such proteins, and identification of proteins present in low concentrations. Finally, there is still not enough information in databases for identifying all the proteins expressed in one sample (Bastida et al., 2009; Blackburn and Martens, 2016). The integration of annotated MG and MT data, by reference or *de novo* strategies, may improve the identification of proteins expressed by the same microbial community.

2.5.2 *Bioinformatics tools for Mass Spectrometry*

In MP, the workflow is very different: proteins are extracted and separated prior to LC (Light Chromatography)-*Mass Spectrometry (MS)* techniques of purification and ionization of the peptides fragmented by trypsin to produce the spectra specific for the corresponding protein fragments. The only steps involving informatics are in the interpretation of the spectra data and the functional/taxonomic annotation of the proteins identified among the proteic fragments.

Extraction of organized information from MS data is carried on by parsers, dependent of the search engines used for the identification of proteins by database-reference methods - the parsers MascotDatFile (Helsens et al., 2007), OMSSA Parser (Barsnes et al., 2009) and XTandem Parser (Muth et al., 2010) for the MASCOT (Cottrell and London, 1999), OMSSA (Geer et al., 2004) and X!Tandem (Fenyö and Beavis, 2003) search engines, respectively. If the proteins being identified are not present in the databases - a common situation in MP - *de novo* algorithms must be used, like for example, PepNovo (Frank and Pevzner, 2005), PEAKS (Ma et al., 2003) or Sequit (Demine and Walden, 2004), to assemble the identified fragments into proteins.

Finally, proteins are annotated by cross referencing their sequences with the information present in several databases of different nature, like KEGG, COG, KOG, InterPro and *Universal Protein Resource (UniProt)*, in a similar way to MG and MT and by using the same tools.

2.5.3 *Metaproteomics pipelines*

A universal protein analysis protocol is a hard task, due to the variety of structure, location and function of proteins. It is firmly established that MP benefits greatly from MG data, either if it originates from the same biological sample or even from metagenomes from different samples collected from different environments. A MP pipeline must attempt to integrate all these data in a system approach for the construction of functional models that can predict behaviour of a microbiome in certain conditions (Siggins et al., 2012).

To date, the only attempt at an automated integration of MG and MP data release to public has been the MetaProteomeAnalyzer, a powerful tool developed mainly for MP data analysis. There are four main steps in the *Meta Proteome Analyzer (MPA)* workflow:

1. Search Engine Comparison - the use of four different database search algorithms - X!Tandem, OMSSA, Crux, and InsPect - with the addition of the MASCOT software, for the identification and annotation of peptides, results in much more distinct hits when the databases are combined compared to single database search, and several hits through different databases may also allow for validation of questionable hits. A

custom MG database may also be used to include protein sequences derived from MG sequencing

2. Metaprotein Generation - a set of rules, concerning taxonomy and similarity of peptides and proteins as a whole for the clustering of redundant proteins, allows for a large reduction of the dataset (44-50% reduction in number of proteins), clustering similar proteins into metaproteins and allowing for the conservation of important statistical data

3. Integration of Meta-Information from External Resources - the generated metaproteins are annotated with additional metainformation at several levels: taxonomy from *National Center for Biotechnology Information (NCBI)*, enzyme information from Enzyme Comission, metabolic pathways from KEGG and all-around protein information from UniProt

4. Graph Database Driven Query System - the implementation of the open source graph database Neo4j allows querying using the Cypher language, which makes for more personalized results analysis

The MPA processing of data ends in two types of output - an MPA project file for visualization of results and a CSV file that allows for result analysis by third party software and more levels of integration, like metabolomics.

2.6 THE DATABASES FOR ANNOTATION

2.6.1 *UniProt*

A consortium made of the collaboration between the CDS, the *Protein Information Resource (PIR)* and the *Swiss Institute of Bioinformatics (SIB)*, 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, aggregating 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 *Conversed Domain 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 Domain 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 protei domain 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 has seen 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 the Basic Local Alignment Search Tool (BLAST) (Pruitt et al., 2007; O'Leary et al., 2016).

2.7 DATASETS FOR PIPELINE TESTING AND VALIDATION

This pipeline will be applied to the analysis of real datasets from anaerobic digestion systems. For MG/MT, the data was retrieved from continuous bioreactors, two treating ethanol based wastewater and two with a mixture of volatile fatty acids inoculated with the same inocula. The main goal will be to identify the pathways most utilized and the most active microorganisms and compare these functional and taxonomic information obtained from the different operational conditions.

For MG/MP, the data has been extracted from anaerobic batch reactors converting long-chain fatty acids (LCFA) to methane. The objective will be to compare the microbial communities developed in the anaerobic microcosms incubated with saturated- and unsaturated-LCFA and to identify the ones responsible for specific degradation steps. Palmitate and estearate were the saturated-LCFA, and oleate was the unsaturated-LCFA.

3

DEVELOPMENT

3.1 PROPOSED PIPELINE ARCHITECTURE

This work aims to develop a bioionformatics pipeline for meta-omics studies. The main steps of MG, MT and MP analysis are presented in figure 3. The informatic steps to be implemented are inside gray areas, along with popular bioinformatics tools. The pipeline will be designed integrating MG data with MT or MP data. The purposed backbone of the workflow is given in the following section.

3.1.1 *Integration of metagenomic with metatranscriptomic data*

MG and MT require similar software. For the preprocessing steps, the workflow of IMP seems stronger than that of FMAP, with less house built tools - quality control of reads is done with *FASTQC*, removal of low quality and human (when justified) sequences with *Trimmomatic*, glsrrna depletion (when justified) with *SortMeRNA*.

For assembly, the *de novo* co-assembly of IMP is a very interesting choice, making the most of MG and MT data in the assembly - therefore the integration of *MEGAHIT* and *IDBA-UD* seems a logical step.

After the assembly, *VizBin* may be used for binning the contigs, *MetaQUAST* to verify the quality of the process and BEDTools for calculating depth of coverage. After such analysis, *Prokka* and *MetaQUAST* could be implemented for functional and taxonomic annotation, respectively, of the contigs assembled *de novo*.

For the final analysis, integrating *KronaTools* will allow for KEGG-based functional Krona plots and ShotgunFunctionalizeR will allow for differential quantification of gene expression, in one or between several samples.

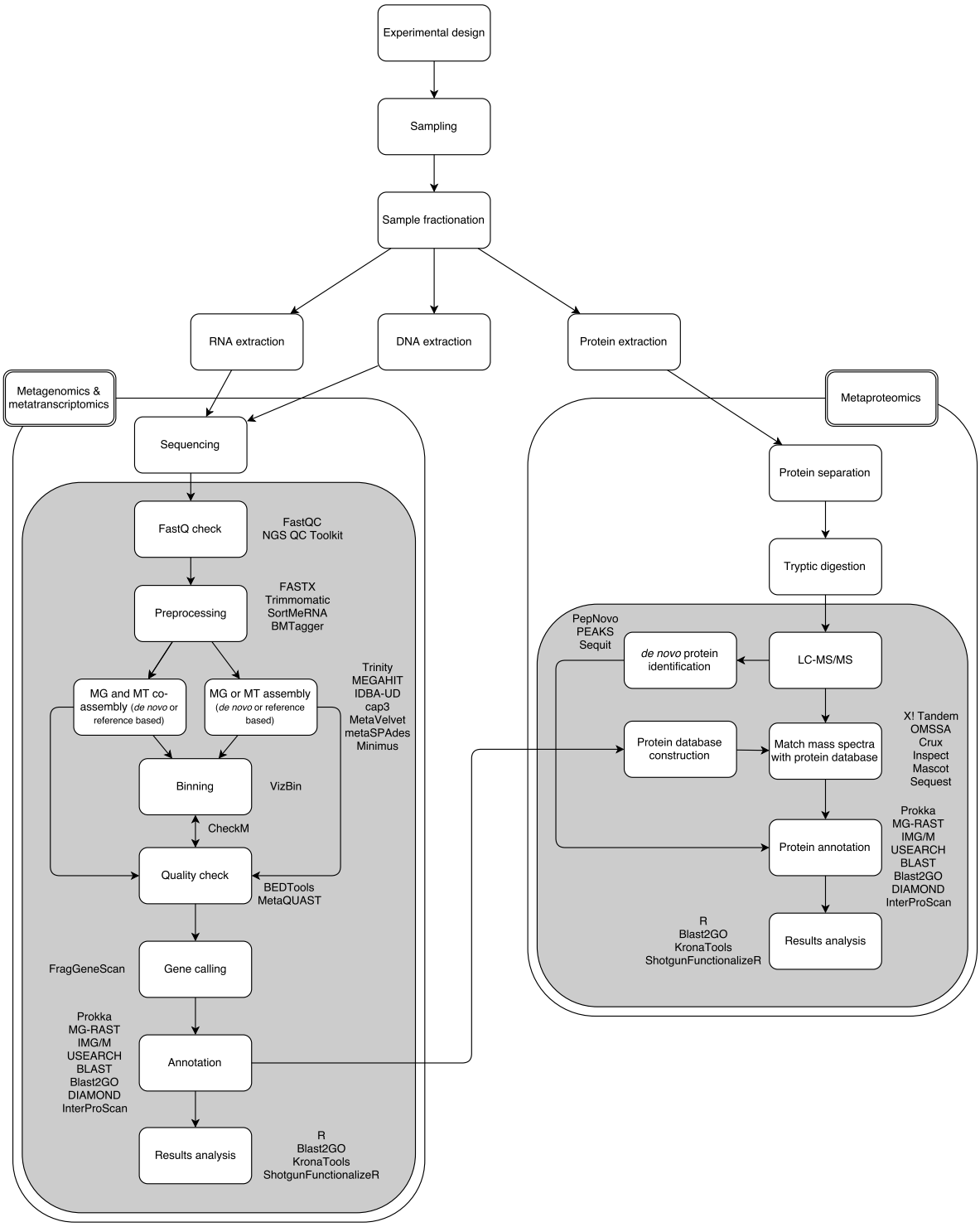


Figure 3: Meta-Omics pipeline workflow, from the laboratory to the bioinformatics processing steps. The tools and software possibilities are represented next to each step of the workflow.

3.1.2 *Integration of metagenomic with metaproteomic data*

The MPA workflow will be the basis for the development of the metaproteomics part of the pipeline. Including the four search engines *X!Tandem*, *OMSSA*, *Crux*, *Inspect*, plus the

Mascot software for the interpretation of MS spectra will produce more protein annotations, validated with the overlap of the search engines' results. But since a MP study deals with many unknown proteins, a *de novo* assembler should be used, such as PepNovo (Frank and Pevzner, 2005), PEAKS (Ma et al., 2003) or SEQUIT (Demine and Walden, 2004), to increase protein identification from microorganisms whose genomic information is not in databases. If MG data is available it will be used to construct the protein database for protein identification. Annotation and results analysis will be performed similarly to what was described for MG and MT analysis.

3.1.3 *General aspects*

The integration of the pipeline will consist in a input/output workflow between the several tools of the pipeline. After each task, the user will have the option of continuing to the next step, or downloading the output file from the previous step. The implementation in a web based environment will be constructed. Web based platforms will be analyzed for selecting the best option of an user interface.

BIBLIOGRAPHY

Vanessa Aguiar-pulido, Wenrui Huang, Victoria Suarez-ulloa, Trevor Cickovski, Kalai Mathee, and Giri Narasimhan. Approaches for Microbiome Analysis. 12:5–16, 2016. ISSN 1176-9343. doi: 10.4137/EBO.S36436.TYPE.

Eyal Akiva, Shoshana Brown, Daniel E. Almonacid, Alan E. Barber, Ashley F. Custer, Michael A. Hicks, Conrad C. Huang, Florian Lauck, Susan T. Mashiyama, Elaine C. Meng, David Mischel, John H. Morris, Sunil Ojha, Alexandra M. Schnoes, Doug Stryke, Jeffrey M. Yunes, Thomas E. Ferrin, Gemma L. Holliday, and Patricia C. Babbitt. The Structure-Function Linkage Database. *Nucleic Acids Research*, 42(D1):521–530, 2014. ISSN 03051048. doi: 10.1093/nar/gkt1130.

Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

Simon Andrews et al. Fastqc: a quality control tool for high throughput sequence data, 2010.

T. K. Attwood. The PRINTS database: A resource for identification of protein families. *Briefings in Bioinformatics*, 3(3):252–263, jan 2002. ISSN 1467-5463. doi: 10.1093/bib/3.3.252. URL http://bib.oxfordjournals.org/content/3/3/252.short.

El Mustapha Bahassi and Peter J. Stambrook. Next-generation sequencing technologies: Breaking the sound barrier of human genetics. *Mutagenesis*, 29(5):303–310, 2014. ISSN 14643804. doi: 10.1093/mutage/geu031.

Brett J Baker, Cody S Sheik, Chris a Taylor, Sunit Jain, Ashwini Bhasi, James D Cavalcoli, and Gregory J Dick. Community transcriptomic assembly reveals microbes that contribute to deep-sea carbon and nitrogen cycling. *The ISME journal*, 7(10):1962–73, 2013. ISSN 1751-7370. doi: 10.1038/ismej.2013.85. URL http://www.ncbi.nlm.nih.gov/pubmed/23702516.

Harald Barsnes, Steffen Huber, Albert Sickmann, Ingvar Eidhammer, and Lennart Martens. Omssa parser: An open-source library to parse and extract data from omssa ms/ms search results. *Proteomics*, 9(14):3772–3774, 2009.

F. Bastida, J. L. Moreno, C. Nicolás, T. Hernández, and C. García. Soil metaproteomics: A review of an emerging environmental science. Significance, methodology and per-

spectives. *European Journal of Soil Science*, 60(6):845–859, 2009. ISSN 13510754. doi: 10.1111/j.1365-2389.2009.01184.x.

Alex Bateman, Lachlan Coin, Richard Durbin, Robert D Finn, Volker Hollich, Sam Grif, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik L L Sonnhammer, David J Studholme, Corin Yeats, and Sean R Eddy. The Pfam protein families database. *Nucleic Acids Res*, 32(Database issue):138D–41, 2004. ISSN 1362-4962. doi: 10.1093/nar/gkh121. URL http://www.ncbi.nlm.nih.gov/pubmed/14681378?ordinalpos=5{&}itool=EntrezSystem2.PEntrez.Pubmed.Pubmed{_}ResultsPanel.Pubmed{_}RVDocSum.

Alex Bateman, Maria Jesus Martin, Claire O'Donovan, Michele Magrane, Rolf Apweiler, Emanuele Alpi, Ricardo Antunes, Joanna Arganiska, Benoit Bely, Mark Bingley, Carlos Bonilla, Ramona Britto, Borisas Bursteinas, Gayatri Chavali, Elena Cibrian-Uhalte, Alan Da Silva, Maurizio De Giorgi, Tunca Dogan, Francesco Fazzini, Paul Gane, Leyla Garcia Castro, Penelope Garmiri, Emma Hatton-Ellis, Reija Hieta, Rachael Huntley, Duncan Legge, Wudong Liu, Jie Luo, Alistair Macdougall, Prudence Mutowo, Andrew Nightingale, Sandra Orchard, Klemens Pichler, Diego Poggioli, Sangya Pundir, Luis Pureza, Guoying Qi, Steven Rosanoff, Rabie Saidi, Tony Sawford, Aleksandra Shypitsyna, Edward Turner, Vladimir Volynkin, Tony Wardell, Xavier Watkins, Hermann Zellner, Andrew Cowley, Luis Figueira, Weizhong Li, Hamish McWilliam, Rodrigo Lopez, Ioannis Xenarios, Lydie Bougueleret, Alan Bridge, Sylvain Poux, Nicole Redaschi, Lucila Aimo, Ghislaine Argoud-Puy, Andrea Auchincloss, Kristian Axelsen, Parit Bansal, Delphine Baratin, Marie Claude Blatter, Brigitte Boeckmann, Jerven Bolleman, Emmanuel Boutet, Lionel Breuza, Cristina Casal-Casas, Edouard De Castro, Elisabeth Coudert, Beatrice Cuche, Mikael Doche, Dolnide Dornevil, Severine Duvaud, Anne Estreicher, Livia Famiglietti, Marc Feuermann, Elisabeth Gasteiger, Sebastien Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz-Gumowski, Ursula Hinz, Chantal Hulo, Florence Jungo, Guillaume Keller, Vicente Lara, Philippe Lemercier, Damien Lieberherr, Thierry Lombardot, Xavier Martin, Patrick Masson, Anne Morgat, Teresa Neto, Nevila Nouspikel, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Monica Pozzato, Manuela Pruess, Catherine Rivoire, Bernd Roechert, Michel Schneider, Christian Sigrist, Karin Sonesson, Sylvie Staehli, Andre Stutz, Shyamala Sundaram, Michael Tognolli, Laure Verbregue, Anne Lise Veuthey, Cathy H. Wu, Cecilia N. Arighi, Leslie Arminski, Chuming Chen, Yongxing Chen, John S. Garavelli, Hongzhan Huang, Kati Laiho, Peter McGarvey, Darren A. Natale, Baris E. Suzek, C. R. Vinayaka, Qinghua Wang, Yuqi Wang, Lai Su Yeh, Meher Shruti Yerramalla, and Jian Zhang. UniProt: A hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212, 2015. ISSN 13624962. doi: 10.1093/nar/gku989.

Courtney A. Benson, Richard W. Bizzoco, David A. Lipson, and Scott T. Kelley. Microbial diversity in nonsulfur, sulfur and iron geothermal steam vents. *FEMS Microbiology Ecology*,

76(1):74–88, 2011. ISSN 01686496. doi: 10.1111/j.1574-6941.2011.01047.x.

Shirley Bikel, Alejandra Valdez-Lara, Fernanda Cornejo-Granados, Karina Rico, Samuel Canizales-Quinteros, Xavier Soberón, Luis Del Pozo-Yauner, and Adrián Ochoa-Leyva. 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, 2015. ISSN 20010370. doi: 10.1016/j.csbj.2015.06.001. URL http://dx.doi.org/10.1016/j.csbj.2015.06.001.

Jonathan M. Blackburn and Lennart Martens. The challenge of metaproteomic analysis in human samples. *Expert Review of Proteomics*, 13(2):135–138, 2016. ISSN 1478-9450. doi: 10.1586/14789450.2016.1135058. URL http://www.tandfonline.com/doi/full/10.1586/14789450.2016.1135058.

Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, page btu170, 2014.

Mya Breitbart, Ian Hewson, Ben Felts, Joseph M Mahaffy, James Nulton, Peter Salamon, and Forest Rohwer. 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, 2003. ISSN 0021-9193. doi: 10.1128/JB.185.20.6220.

Daniel W A Buchan, Adrian J Shepherd, David Lee, Frances M G Pearl, Stuart C G Rison, Janet M Thornton, and Christine A Orengo. Gene3D: structural assignment for whole genes and genomes using the CATH domain structure database. *Genome research*, 12(3):503–14, mar 2002. ISSN 1088-9051. doi: 10.1101/gr.213802. URL http://genome.cshlp.org/content/12/3/503.full.

Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using diamond. *Nature methods*, 12(1):59–60, 2015.

H P J Buermans and J T Den Dunnen. Next generation sequencing technology: Advances and applications. *Biochimica et biophysica acta*, 1842(10):1932–1941, 2014. ISSN 0006-3002. doi: 10.1016/j.bbadis.2014.06.015. URL http://www.ncbi.nlm.nih.gov/pubmed/24995601.

Evelyn Camon, Michele Magrane, Daniel Barrell, Vivian Lee, Emily Dimmer, John Maslen, David Binns, Nicola Harte, Rodrigo Lopez, and Rolf Apweiler. The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. *Nucleic acids research*, 32(suppl 1):D262–D266, 2004.

Lilia C. Carvalhais, Paul G. Dennis, Gene W. Tyson, and Peer M. Schenk. Application of metatranscriptomics to soil environments. *Journal of Microbiological Methods*, 91(2):246–251, 2012. ISSN 01677012. doi: 10.1016/j.mimet.2012.08.011.

Albi Celaj, Janet Markle, Jayne Danska, and John Parkinson. Comparison of assembly algorithms for improving rate of metatranscriptomic functional annotation. *Microbiome*, 2(1):39, 2014. ISSN 2049-2618. doi: 10.1186/2049-2618-2-39. URL http://www.ncbi.nlm.nih.gov/pubmed/25411636.

Ryan Chamberlain and Jennifer Schommer. Using docker to support reproducible research. *DOI: http://dx. doi. org/10.6084/m9. figshare*, 1101910, 2014.

Peter JA Cock, Christopher J Fields, Naohisa Goto, Michael L Heuer, and Peter M Rice. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic acids research*, 38(6):1767–1771, 2010.

Ana Conesa, Stefan Götz, Juan Miguel García-Gómez, Javier Terol, Manuel Talón, and Montserrat Robles. Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676, 2005.

F Corpet. The ProDom database of protein domain families. *Nucleic Acids Research*, 26 (1):323–326, jan 1998. ISSN 13624962. doi: 10.1093/nar/26.1.323. URL http://nar.oxfordjournals.org/content/26/1/323.abstract.

John S Cottrell and U London. Probability-based protein identification by searching sequence databases using mass spectrometry data. *electrophoresis*, 20(18):3551–3567, 1999.

Coralie Damon, Fr??d??ric Lehembre, Christine Oger-Desfeux, Patricia Luis, Jacques Ranger, Laurence Fraissinet-Tachet, and Roland Marmeisse. Metatranscriptomics reveals the diversity of genes expressed by eukaryotes in forest soils. *PLoS ONE*, 7(1), 2012. ISSN 19326203. doi: 10.1371/journal.pone.0028967.

Rodion Demine and Peter Walden. Sequit: software for de novo peptide sequencing by matrix-assisted laser desorption/ionization post-source decay mass spectrometry. *Rapid communications in mass spectrometry*, 18(8):907–913, 2004.

Pravin Dudhagara, Sunil Bhavsar, Chintan Bhagat, Anjana Ghelani, Shreyas Bhatt, and Rajesh Patel. Web Resources for Metagenomics Studies. *Genomics, Proteomics and Bioinformatics*, 13(5):296–303, 2015. ISSN 22103244. doi: 10.1016/j.gpb.2015.10.003. URL http://dx.doi.org/10.1016/j.gpb.2015.10.003.

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

Olof Emanuelsson, Soren Brunak, Gunnar von Heijne, and Henrik Nielsen. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protocols*, 2(4):953–971, apr 2007. ISSN 1754-2189. URL http://dx.doi.org/10.1038/nprot.2007.131.

J.C. et al. Venter. Environmental Genome Shotgun Sequencing of the. *Science*, 1093857(2004): 304, 2004. ISSN 0036-8075. doi: 10.1126/science.1093857.

David Fenyö and Ronald C Beavis. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Analytical chemistry*, 75(4):768–774, 2003.

Robert D Finn, Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, Erik L L Sonnhammer, John Tate, and Marco Punta. Pfam: the protein families database. *Nucleic acids research*, 42(Database issue):D222–30, jan 2014. ISSN 1362-4962. doi: 10.1093/nar/gkt1223. URL http://nar.oxfordjournals.org/content/42/D1/D222.long.

Robert D Finn, Teresa K Attwood, Patricia C Babbitt, Alex Bateman, Peer Bork, Alan J Bridge, Hsin-Yu Chang, Zsuzsanna Dosztányi, Sara El-Gebali, Matthew Fraser, Julian Gough, David Haft, Gemma L Holliday, Hongzhan Huang, Xiaosong Huang, Ivica Letunic, Rodrigo Lopez, Shennan Lu, Aron Marchler-Bauer, Huaiyu Mi, Jaina Mistry, Darren A Natale, Marco Necci, Gift Nuka, Christine A Orengo, Youngmi Park, Sebastien Pesseat, Damiano Piovesan, Simon C Potter, Neil D Rawlings, Nicole Redaschi, Lorna Richardson, Catherine Rivoire, Amaia Sangrador-Vegas, Christian Sigrist, Ian Sillitoe, Ben Smithers, Silvano Squizzato, Granger Sutton, Narmada Thanki, Paul D Thomas, Silvio C E Tosatto, Cathy H Wu, Ioannis Xenarios, Lai-Su Yeh, Siew-Yit Young, and Alex L Mitchell. InterPro in 2017-beyond protein family and domain annotations. *Nucleic acids research*, 45(November 2016):gkw1107, 2016. ISSN 1759-6653. doi: 10.1093/nar/gkw1107. URL http://www.ncbi.nlm.nih.gov/pubmed/27899635.

Ari Frank and Pavel Pevzner. Pepnovo: de novo peptide sequencing via probabilistic network modeling. *Analytical chemistry*, 77(4):964–973, 2005.

Antonio García-Moyano, Elena González-Toril, Ángeles Aguilera, and Ricardo Amils. 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, 2012. ISSN 01686496. doi: 10.1111/j.1574-6941.2012.01346.x.

Lewis Y Geer, Sanford P Markey, Jeffrey A Kowalak, Lukas Wagner, Ming Xu, Dawn M Maynard, Xiaoyu Yang, Wenyao Shi, and Stephen H Bryant. Open mass spectrometry search algorithm. *Journal of proteome research*, 3(5):958–964, 2004.

Elizabeth M Glass, Jared Wilkening, Andreas Wilke, Dionysios Antonopoulos, and Folker Meyer. Using the metagenomics rast server (mg-rast) for analyzing shotgun metagenomes. *Cold Spring Harbor Protocols*, 2010(1):pdb–prot5368, 2010.

Marcin Gołębiewski, Edyta Deja-Sikora, Marcin Cichosz, Andrzej Tretyn, and Borys Wróbel. 16S rDNA pyrosequencing analysis of bacterial community in heavy metals polluted soils. *Microbial ecology*, 67(3):635–47, 2014. ISSN 1432-184X. doi: 10.1007/s00248-013-0344-7. URL http://www.ncbi.nlm.nih.gov/pubmed/24402360{%}5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3962847.

J. Gough. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Research*, 30 (1):268–272, jan 2002. ISSN 13624962. doi: 10.1093/nar/30.1.268. URL http://nar.oxfordjournals.org/content/30/1/268.full.

D. H. Haft. The TIGRFAMs database of protein families. *Nucleic Acids Research*, 31 (1):371–373, jan 2003. ISSN 13624962. doi: 10.1093/nar/gkg128. URL http://nar.oxfordjournals.org/content/31/1/371.full.

Kenny Helsens, Lennart Martens, Joël Vandekerckhove, and Kris Gevaert. Mascotdatfile: An open-source library to fully parse and analyse mascot ms/ms search results. *Proteomics*, 7(3):364–366, 2007.

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

Sarah Hunter, Rolf Apweiler, Teresa K. Attwood, Amos Bairoch, Alex Bateman, David Binns, Peer Bork, Ujjwal Das, Louise Daugherty, Lauranne Duquenne, Robert D. Finn, Julian Gough, Daniel Haft, Nicolas Hulo, Daniel Kahn, Elizabeth Kelly, Aur??lie Laugraud, Ivica Letunic, David Lonsdale, Rodrigo Lopez, Martin Madera, John Maslen, Craig Mcanulla, Jennifer McDowall, Jaina Mistry, Alex Mitchell, Nicola Mulder, Darren Natale, Christine Orengo, Antony F. Quinn, Jeremy D. Selengut, Christian J A Sigrist, Manjula Thimma, Paul D. Thomas, Franck Valentin, Derek Wilson, Cathy H. Wu, and Corin Yeats. InterPro: The integrative protein signature database. *Nucleic Acids Research*, 37(SUPPL. 1):211–215, 2009. ISSN 03051048. doi: 10.1093/nar/gkn785.

Illumina. An Introduction to Next-Generation Sequencing Technology. *Illumina.com*, (illumina):1 – 16, 2015. doi: http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf.

Illumina Proprietary. Microbes and Metagenomics in Human Health.

Shun'ichi Ishii, Shino Suzuki, Aaron Tenney, Trina M. Norden-Krichmar, Kenneth H. Nealson, and Orianna Bretschger. Microbial metabolic networks in a complex electrogenic biofilm recovered from a stimulus-induced metatranscriptomics approach. *Scientific Reports*, 5(October):14840, 2015. ISSN 2045-2322. doi: 10.1038/srep14840. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4595844{&}tool=pmcentrez{&}rendertype=abstract.

Philip Jones, David Binns, Hsin Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F. Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew Yit Yong, Rodrigo Lopez, and Sarah Hunter. InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240, 2014. ISSN 14602059. doi: 10.1093/bioinformatics/btu031.

Ji Young Jung, Se Hee Lee, Hyun Mi Jin, Yoonsoo Hahn, Eugene L. Madsen, and Che Ok Jeon. Metatranscriptomic analysis of lactic acid bacterial gene expression during kimchi fermentation. *International Journal of Food Microbiology*, 163(2-3):171–179, 2013. ISSN 01681605. doi: 10.1016/j.ijfoodmicro.2013.02.022. URL http://dx.doi.org/10.1016/j.ijfoodmicro.2013.02.022.

Lukas Kall, Anders Krogh, and Erik L L Sonnhammer. A combined transmembrane topology and signal peptide prediction method. *Journal of molecular biology*, 338(5):1027–1036, may 2004. ISSN 0022-2836 (Print). doi: 10.1016/j.jmb.2004.03.016.

Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes, jan 2000. ISSN 0305-1048 (Print).

Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44 (D1):D457–D462, 2016. ISSN 13624962. doi: 10.1093/nar/gkv1070.

Jiwoong Kim, Min Soo Kim, Andrew Y. Koh, Yang Xie, Xiaowei Zhan, J Qin, R Li, J Raes, M Arumugam, KS Burgdorf, C Manichanh, T Nielsen, N Pons, F Levenez, T Yamada, F Meyer, D Paarmann, M D'Souza, R Olson, EM Glass, M Kubal, T Paczian, A Rodriguez, R Stevens, A Wilke, S Abubucker, N Segata, J Goll, AM Schubert, J Izard, BL Cantarel, B Rodriguez-Mueller, J Zucker, M Thiagarajan, B Henrissat, DH Huson, N Weber, G Yi, SH Sze, MR Thon, E Kristiansson, P Hugenholtz, D Dalevi, K Rotmistrovsky, R Agarwala, RC Edgar, B Buchfink, C Xie, DH Huson, C UniProt, JN Paulson, OC Stine, HC Bravo, M Pop, S Okuda, AC Yoshizawa, JA Papin, J Stelling, ND Price, S Klamt, S Schuster, BO Palsson, P Khatri, M Sirota, AJ Butte, W Huang, L Li, JR Myers, GT Marth, PJ Turnbaugh, RE Ley, M Hamady, CM Fraser-Liggett, R Knight, JI Gordon, EF DeLong, CM Preston, T Mincer, V Rich, SJ Hallam, NU Frigaard, A Martinez, MB Sullivan, R Edwards, BR Brito, P Belda-Ferre, LD Alcaraz, R Cabrera-Rubio, H Romero,

A Simon-Soro, M Pignatelli, A Mira, AR Erickson, BL Cantarel, R Lamendella, Y Darzi, EF Mongodin, C Pan, M Shah, J Halfvarson, C Tysk, B Henrissat, T Tobe, N Nakanishi, and N Sugimoto. FMAP: Functional Mapping and Analysis Pipeline for metagenomics and metatranscriptomics studies. *BMC Bioinformatics*, 17(1):420, 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-1278-0. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1278-0.

Evguenia Kopylova, Laurent No, and Hlne Touzet. Sortmerna: fast and accurate filtering of ribosomal rnas in metatranscriptomic data. *Bioinformatics*, 28(24):3211, 2012. doi: 10.1093/bioinformatics/bts611. URL +http://dx.doi.org/10.1093/bioinformatics/bts611.

Erik Kristiansson, Philip Hugenholtz, and Daniel Dalevi. Shotgunfunctionalizer: an r-package for functional comparison of metagenomes. *Bioinformatics*, 25(20):2737–2738, 2009.

A Krogh, B Larsson, G von Heijne, and E L Sonnhammer. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology*, 305(3):567–80, jan 2001. ISSN 0022-2836. doi: 10.1006/jmbi.2000.4315. URL http://www.ncbi.nlm.nih.gov/pubmed/11152613.

Cedric C Laczny, Tomasz Sternal, Valentin Plugaru, Piotr Gawron, Arash Atashpendar, Houry Hera Margossian, Sergio Coronado, Laurens Van der Maaten, Nikos Vlassis, and Paul Wilmes. Vizbin-an application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome*, 3(1):1, 2015.

Efthymios Ladoukakis, Fragiskos N Kolisis, and Aristotelis A Chatziioannou. Integrative workflows for metagenomic analysis. *Frontiers in cell and developmental biology*, 2(November):70, 2014. ISSN 2296-634X. doi: 10.3389/fcell.2014.00070. URL http://journal.frontiersin.org/article/10.3389/fcell.2014.00070/abstract.

Ivica Letunic, Richard R Copley, Steffen Schmidt, Francesca D Ciccarelli, Tobias Doerks, Jörg Schultz, Chris P Ponting, and Peer Bork. SMART 4.0: towards genomic data integration. *Nucleic acids research*, 32(Database issue):D142–4, 2004. ISSN 1362-4962. doi: 10.1093/nar/gkh088. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=308822{&}tool=pmcentrez{&}rendertype=abstract.

Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*, page btv033, 2015.

Tania Lima, Andrea H Auchincloss, Elisabeth Coudert, Guillaume Keller, Karine Michoud, Catherine Rivoire, Virginie Bulliard, Edouard de Castro, Corinne Lachaize, Delphine

Baratin, Isabelle Phan, Lydie Bougueleret, and Amos Bairoch. HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot, jan 2009. ISSN 0305-1048 (Print).

Lin Liu, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*, 2012, 2012. ISSN 11107243. doi: 10.1155/2012/251364.

A Lupas, M Van Dyke, and J Stock. Predicting coiled coils from protein sequences. *Science (New York, N.Y.)*, 252(5009):1162–4, may 1991. ISSN 0036-8075. doi: 10.1126/science.252.5009.1162. URL http://www.ncbi.nlm.nih.gov/pubmed/2031185.

Bin Ma, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby, and Gilles Lajoie. Peaks: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry*, 17 (20):2337–2342, 2003.

Aron Marchler-Bauer, Myra K. Derbyshire, Noreen R. Gonzales, Shennan Lu, Farideh Chitsaz, Lewis Y. Geer, Renata C. Geer, Jane He, Marc Gwadz, David I. Hurwitz, Christopher J. Lanczycki, Fu Lu, Gabriele H. Marchler, James S. Song, Narmada Thanki, Zhouxi Wang, Roxanne A. Yamashita, Dachuan Zhang, Chanjuan Zheng, and Stephen H. Bryant. CDD: NCBI's conserved domain database. *Nucleic Acids Research*, 43(D1):D222–D226, 2015. ISSN 13624962. doi: 10.1093/nar/gku1221.

Victor M Markowitz, Natalia N Ivanova, Ernest Szeto, Krishna Palaniappan, Ken Chu, Daniel Dalevi, I-Min A Chen, Yuri Grechkin, Inna Dubchak, Iain Anderson, et al. Img/m: a data management and analysis system for metagenomes. *Nucleic acids research*, 36(suppl 1):D534–D538, 2008.

Xavier Martinez, Marta Pozuelo, Victoria Pascal, David Campos, Ivo Gut, Marta Gut, Fernando Azpiroz, Francisco Guarner, and Chaysavanh Manichanh. MetaTrans: an open-source pipeline for metatranscriptomics. *Scientific reports*, 6:26447, 2016. ISSN 2045-2322. doi: 10.1038/srep26447. URL http://www.ncbi.nlm.nih.gov/pubmed/27211518.

Folker Meyer, Daniel Paarmann, Mark D'Souza, Robert Olson, Elizabeth M Glass, Michael Kubal, Tobias Paczian, A Rodriguez, Rick Stevens, Andreas Wilke, et al. The metagenomics rast server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*, 9(1):386, 2008.

Alla Mikheenko, Vladislav Saveliev, and Alexey Gurevich. Metaquast: evaluation of metagenome assemblies. *Bioinformatics*, 32(7):1088–1090, 2016.

Thilo Muth, Marc Vaudel, Harald Barsnes, Lennart Martens, and Albert Sickmann. XTandem Parser: An open-source library to parse and analyse X!Tandem MS/MS search results. *Proteomics*, 10(7):1522–1524, 2010. ISSN 16159853. doi: 10.1002/pmic.200900759.

Thilo Muth, Alexander Behne, Robert Heyer, Fabian Kohrs, Dirk Benndorf, Marcus Hoffmann, Miro Lehtevä, Udo Reichl, Lennart Martens, and Erdmann Rapp. The MetaProteomeAnalyzer: A powerful open-source software suite for metaproteomics data analysis and interpretation. *Journal of Proteome Research*, 14(3):1557–1565, 2015. ISSN 15353907. doi: 10.1021/pr501246w.

Toshiaki Namiki, Tsuyoshi Hachiya, Hideaki Tanaka, and Yasubumi Sakakibara. Metavelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic acids research*, 40(20):e155–e155, 2012.

Shaman Narayanasamy, Yohan Jarosz, Emilie E.L. Muller, Cédric C. Laczny, Malte Herold, Anne Kaysen, Anna Heintz-Buschart, Nicolás Pinel, Patrick May, and Paul Wilmes. IMP: a pipeline for reproducible metagenomic and metatranscriptomic analyses. *bioRxiv*, (7): 039263, 2016. doi: 10.1101/039263. URL http://biorxiv.org/lookup/doi/10.1101/039263.

Stephen Nayfact, Beltran Rodriguez-Mueller, Nandita Garud, and Katherine Pollard. An integrated metagenomics pipeline for strain profiling reveals novel patterns of transmission and global biogeography of bacteria. *bioRxiv*, 53(9):1689–1699, 2016. ISSN 1098-6596. doi: 10.1017/CBO9781107415324.004.

Sergey Nurk, Dmitry Meleshko, Anton Korobeynikov, and Pavel Pevzner. metaspades: a new versatile de novo metagenomics assembler. *arXiv preprint arXiv:1604.03071*, 2016.

Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M. Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S. Joardar, Vamsi K. Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M. McGarvey, Michael R. Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H. Rangwala, Daniel Rausch, Lillian D. Riddick, Conrad Schoch, Andrei Shkeda, Susan S. Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E. Tully, Anjana R. Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J. Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D. Murphy, and Kim D. Pruitt. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745, 2016. ISSN 13624962. doi: 10.1093/nar/gkv1189.

Anastasis Oulas, Christina Pavloudi, Paraskevi Polymenakou, Georgios A. Pavlopoulos, Nikolas Papanikolaou, Georgios Kotoulas, Christos Arvanitidis, and Ioannis Iliopoulos. Metagenomics: Tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinformatics and Biology Insights*, 9:75–88, 2015. ISSN 11779322. doi: 10.4137/BBI.S12462.

An Overview and Publications Featuring Illumina. Metagenomics Research Review. *Illumina*, page 38, 2012.

Ravi K. Patel and Mukesh Jain. Ngs qc toolkit: A toolkit for quality control of next generation sequencing data. *PLOS ONE*, 7(2):1–7, 02 2012. doi: 10.1371/journal.pone.0030619. URL http://dx.doi.org/10.1371%2Fjournal.pone.0030619.

David A. Pearce, Kevin K. Newsham, Michael A S Thorne, Leo Calvo-Bado, Martin Krsek, Paris Laskaris, Andy Hodson, and Elizabeth M. Wellington. Metagenomic analysis of a southern maritime Antarctic soil. *Frontiers in Microbiology*, 3(DEC), 2012. ISSN 1664302X. doi: 10.3389/fmicb.2012.00403.

Yu Peng, Henry CM Leung, Siu-Ming Yiu, and Francis YL Chin. Idba-ud: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11):1420–1428, 2012.

Erica Plummer and Jimmy Twin. 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, 2015. ISSN 0974276X. doi: 10.4172/jpb.1000381. URL http://www.omicsonline.org/open-access/a-comparison-of-three-bioinformatics-pipelines-for-the-analysis-ofpreterm-gut-microbi php?aid=65142{%}5Cnhttp://www.omicsonline.org/open-access/a-comparison-of-three-b.

Morten Poulsen, Clarissa Schwab, Bent Borg Jensen, Ricarda M Engberg, Anja Spang, Nuria Canibe, Ole Højberg, Gabriel Milinovich, Lena Fragner, Christa Schleper, Wolfram Weckwerth, Peter Lund, Andreas Schramm, and Tim Urich. Methylotrophic methanogenic Thermoplasmata implicated in reduced methane emissions from bovine rumen. *Nature communications*, 4:1428, 2013. ISSN 2041-1723. doi: 10.1038/ncomms2432. URL http://www.ncbi.nlm.nih.gov/pubmed/23385573.

Sean Powell, Damian Szklarczyk, Kalliopi Trachana, Alexander Roth, Michael Kuhn, Jean Muller, Roland Arnold, Thomas Rattei, Ivica Letunic, Tobias Doerks, et al. eggnog v3. 0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic acids research*, 40(D1):D284–D289, 2012.

Kim D. Pruitt, Tatiana Tatusova, and Donna R. Maglott. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 35(SUPPL. 1):501–504, 2007. ISSN 03051048. doi: 10.1093/nar/gkl842.

Michael Quail, Miriam E Smith, Paul Coupland, Thomas D Otto, Simon R Harris, Thomas R Connor, Anna Bertoni, Harold P Swerdlow, Yong Gu, JM Rothberg, W Hinz, TM Rearick, J Schultz, W Mileski, M Davey, JH Leamon, K Johnson, MJ Milgrew, M Edwards, J Eid, A Fehr, J Gray, K Luong, J Lyle, G Otto, P Peluso, D Rank, P Baybayan, B Bettman, DR Bentley, S Balasubramanian, HP Swerdlow, GP Smith, J Milton, CG Brown, KP Hall, DJ Evers, CL Barnes, HR Bignell, I Kozarewa, Z Ning, MA Quail, MJ Sanders, M Berriman, DJ Turner, MA Quail, TD Otto, Y Gu, SR Harris, TF Skelly, JA McQuillan, HP Swerdlow, SO Oyola, F Syed, H Grunenwald, N Caruccio, HYK Lam, MJ Clark, R Chen, R Chen, G Natsoulis, M O'Huallachain, FE Dewey, L Habegger, T Carver, SR Harris, M Berriman, J Parkhill, JA McQuillan, N Ponsting, Z Ning, TD Otto, M Sanders, M Berriman, C Newbold, K Nakamura, T Oshima, T Morimoto, S Ikeda, H Yoshikawa, Y Shiwa, S Ishikawa, MC Linak, A Hirai, H Takahashi, BA Diep, SR Gill, RF Chang, TH Phan, JH Chen, MG Davidson, F Lin, J Lin, HA Carleton, EF Mongodin, EA Achidi, MJ Gardner, N Hall, E Fung, O White, M Berriman, RW Hyman, JM Carlton, A Pain, KE Nelson, S Bowman, M Choi, UI Scholl, W Ji, T Liu, IR Tikhonova, P Zumbo, A Nayir, A Bakkaloglu, S Ozen, S Sanjad, TA Down, VK Rakyan, DJ Turner, P Flicek, H Li, E Kulesha, S Graf, N Johnson, J Herrero, EM Tomazou, PG Giresi, J Kim, RM McDaniell, VR Iyer, JD Lieb, DS Johnson, A Mortazavi, RM Myers, B Wold, GC Langridge, MD Phan, DJ Turner, TT Perkins, L Parts, J Haase, I Charles, DJ Maskell, SE Peters, G Dougan, DD Licatalosi, A Mele, JJ Fak, J Ule, M Kayikci, SW Chi, TA Clark, AC Schweitzer, JE Blume, X Wang, L Mamanova, RM Andrews, KD James, EM Sheridan, PD Ellis, CF Langford, TW Ost, JE Collins, DJ Turner, S Myllykangas, JD Buenrostro, G Natsoulis, JM Bell, HP Ji, NY Shao, HY Hu, Z Yan, Y Xu, H Hu, C Menzel, N Li, W Chen, P Khaitovich, Z Wang, M Gerstein, M Snyder, S Gnerre, I Maccallum, D Przybylski, FJ Ribeiro, JN Burton, BJ Walker, T Sharpe, G Hall, TP Shea, S Sykes, JZ Levin, M Yassour, X Adiconis, C Nusbaum, DA Thompson, N Friedman, A Gnirke, A Regev, A Adey, Asan, X Xun, JO Kitzman, EH Turner, B Stackhouse, AP MacKenzie, NC Caruccio, X Zhang, BA Flusberg, DR Webster, JH Lee, KJ Travers, EC Olivares, TA Clark, J Korlach, SW Turner, TG Holden, JA Lindsay, C Corton, MA Quail, JD Cockfield, S Pathak, R Batra, J Parkhill, SD Bentley, JD Edgeworth, H Li, R Durbin, H Li, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, R Durbin, SV Angiuoli, and SL Salzberg. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics*, 13(1):341, 2012. ISSN 1471-2164. doi: 10.1186/1471-2164-13-341. URL http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-13-341.

E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez. InterProScan: Protein domains identifier. *Nucleic Acids Research*, 33(SUPPL. 2):116–120, 2005. ISSN 03051048. doi: 10.1093/nar/gki442.

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

Kirill Rotmistrovsky and Richa Agarwala. Bmtagger: Best match tagger for removing human reads from metagenomics datasets. 2011.

J Schultz, R R Copley, T Doerks, C P Ponting, and P Bork. SMART: a web-based tool for the study of genetically mobile domains. *Nucleic acids research*, 28(1):231–4, jan 2000. ISSN 0305-1048. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102444{&}tool=pmcentrez{&}rendertype=abstract.

Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, page btu153, 2014.

Alma Siggins, Eoin Gunnigle, and Florence Abram. Exploring mixed microbial community functioning: Recent advances in metaproteomics. *FEMS Microbiology Ecology*, 80(2):265–280, 2012. ISSN 01686496. doi: 10.1111/j.1574-6941.2011.01284.x.

C. J. A. Sigrist. PROSITE: A documented database using patterns and profiles as motif descriptors. *Briefings in Bioinformatics*, 3(3):265–274, jan 2002. ISSN 1467-5463. doi: 10.1093/bib/3.3.265. URL http://bib.oxfordjournals.org/content/3/3/265.short?rss=1{&}ssource=mfc.

Daniel D Sommer, Arthur L Delcher, Steven L Salzberg, and Mihai Pop. Minimus: a fast, lightweight genome assembler. *BMC bioinformatics*, 8(1):64, 2007.

Heike Stevens and Osvaldo Ulloa. Bacterial diversity in the oxygen minimum zone of the eastern tropical South Pacific. *Environmental Microbiology*, 10(5):1244–1259, 2008. ISSN 14622912. doi: 10.1111/j.1462-2920.2007.01539.x.

Boonfei Tan, S Jane Fowler, Nidal Abu Laban, Xiaoli Dong, Christoph W Sensen, Julia Foght, and Lisa M Gieg. Comparative analysis of metagenomes from three methanogenic hydrocarbon-degrading enrichment cultures with 41 environmental samples. *The ISME Journal*, 9(9):2028–2045, 2015. ISSN 1751-7362. doi: 10.1038/ismej.2015.22. URL http://dx.doi.org/10.1038/ismej.2015.22.

Roman L Tatusov, Natalie D Fedorova, John D Jackson, Aviva R Jacobs, Boris Kiryutin, Eugene V Koonin, Dmitri M Krylov, Raja Mazumder, Sergei L Mekhedov, Anastasia N Nikolskaya, B Sridhar Rao, Sergei Smirnov, Alexander V Sverdlov, Sona Vasudevan,

Yuri I Wolf, Jodie J Yin, and Darren A Natale. The COG database: an updated version includes eukaryotes. *BMC bioinformatics*, 4:41, 2003. ISSN 1471-2105. doi: 10.1186/1471-2105-4-41. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=222959{&}tool=pmcentrez{&}rendertype=abstract.

Paul D Thomas, Anish Kejariwal, Michael J Campbell, Huaiyu Mi, Karen Diemer, Nan Guo, Istvan Ladunga, Betty Ulitsky-Lazareva, Anushya Muruganujan, Steven Rabkin, Jody A Vandergriff, and Olivier Doremieux. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification, jan 2003. ISSN 0305-1048 (Print).

Torsten Thomas, Jack Gilbert, and Folker Meyer. Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation*, 2(1):3, 2012. ISSN 2042-5783. doi: 10.1186/2042-5783-2-3. URL http://www.microbialinformaticsj.com/content/2/1/3.

Todd J Treangen, Sergey Koren, Daniel D Sommer, Bo Liu, Irina Astrovskaya, Brian Ondov, Aaron E Darling, Adam M Phillippy, and Mihai Pop. Metamos: a modular and open source metagenomic assembly and analysis pipeline. *Genome biology*, 14(1):R2, 2013.

Susannah Green Tringe and Edward M Rubin. Metagenomics: DNA sequencing of environmental samples. *Nature reviews. Genetics*, 6(11):805–14, 2005. ISSN 1471-0056. doi: 10.1038/nrg1709. URL http://www.ncbi.nlm.nih.gov/pubmed/16304596.

Gene W Tyson, Jarrod Chapman, Philip Hugenholtz, Eric E Allen, Rachna J Ram, Paul M Richardson, Victor V Solovyev, Edward M Rubin, Daniel S Rokhsar, and Jillian F Banfield. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43, 2004. ISSN 0028-0836. doi: 10.1038/nature02340.

UniProt. The Universal Protein Resource. 2010(November 2007):190–195, 2010. ISSN 0305-1048. doi: 10.1093/nar/gkl929. URL http://www.uniprot.org.

Tim Urich, Anders Lanzén, Runar Stokke, Rolf B. Pedersen, Christoph Bayer, Ingunn H. Thorseth, Christa Schleper, Ida H. Steen, and Lise Øvreas. Microbial community structure and functioning in marine sediments associated with diffuse hydrothermal venting assessed by integrated meta-omics. *Environmental Microbiology*, 16(9):2699–2710, 2014. ISSN 14622920. doi: 10.1111/1462-2920.12283.

Erwin L. Van Dijk, Hélène Auger, Yan Jaszczyszyn, and Claude Thermes. Ten years of next-generation sequencing technology. *Trends in Genetics*, pages 1–9, 2014. ISSN 01689525. doi: 10.1016/j.tig.2014.07.001.

Samuel T Westreich, Ian Korf, David A Mills, and Danielle G Lemay. SAMSA: A comprehensive metatranscriptome analysis pipeline. *bioRxiv*, page 046201, 2016. ISSN 1471-2105. doi: 10.1101/046201. URL http://biorxiv.org/lookup/doi/10.1101/046201.

Cathy H Wu, Anastasia Nikolskaya, Hongzhan Huang, Lai-Su L Yeh, Darren A Natale, C R Vinayaka, Zhang-Zhi Hu, Raja Mazumder, Sandeep Kumar, Panagiotis Kourtesis, Robert S Ledley, Baris E Suzek, Leslie Arminski, Yongxing Chen, Jian Zhang, Jorge Louie Cardenas, Sehee Chung, Jorge Castro-Alvear, Georgi Dinkov, and Winona C Barker. PIRSF: family classification system at the Protein Information Resource. *Nucleic acids research*, 32(Database issue):D112–4, jan 2004. ISSN 1362-4962. doi: 10.1093/nar/gkh097. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=308831{&}tool=pmcentrez{&}rendertype=abstract.

Jinbo Xiong, Yongqin Liu, Xiangui Lin, Huayong Zhang, Jun Zeng, Juzhi Hou, Yongping Yang, Tandong Yao, Rob Knight, and Haiyan Chu. Geographic distance and pH drive bacterial distribution in alkaline lake sediments across Tibetan Plateau. *Environmental Microbiology*, 14(9):2457–2466, 2012a. ISSN 14622912. doi: 10.1111/j.1462-2920.2012.02799.x.

Xuejian Xiong, Daniel N. Frank, Charles E. Robertson, Stacy S. Hung, Janet Markle, Angelo J. Canty, Kathy D. McCoy, Andrew J. Macpherson, Philippe Poussier, Jayne S. Danska, and John Parkinson. Generation and analysis of a mouse intestinal metatranscriptome through illumina based rna-sequencing. *PLOS ONE*, 7(4):1–15, 04 2012b. doi: 10.1371/journal.pone.0036009. URL http://dx.doi.org/10.1371%2Fjournal.pone.0036009.

E M Zdobnov and R Apweiler. InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics (Oxford, England)*, 17(9):847–848, 2001. ISSN 1367-4803. doi: 10.1093/bioinformatics/17.9.847. URL http://bioinformatics.oxfordjournals.org/content/17/9/847.short.

Lucia Žifčáková, Tomáš Větrovskỳ, Adina Howe, and Petr Baldrian. Microbial activity in forest soil reflects the changes in ecosystem properties between summer and winter. *Environmental microbiology*, 18(1):288–301, 2016.