UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

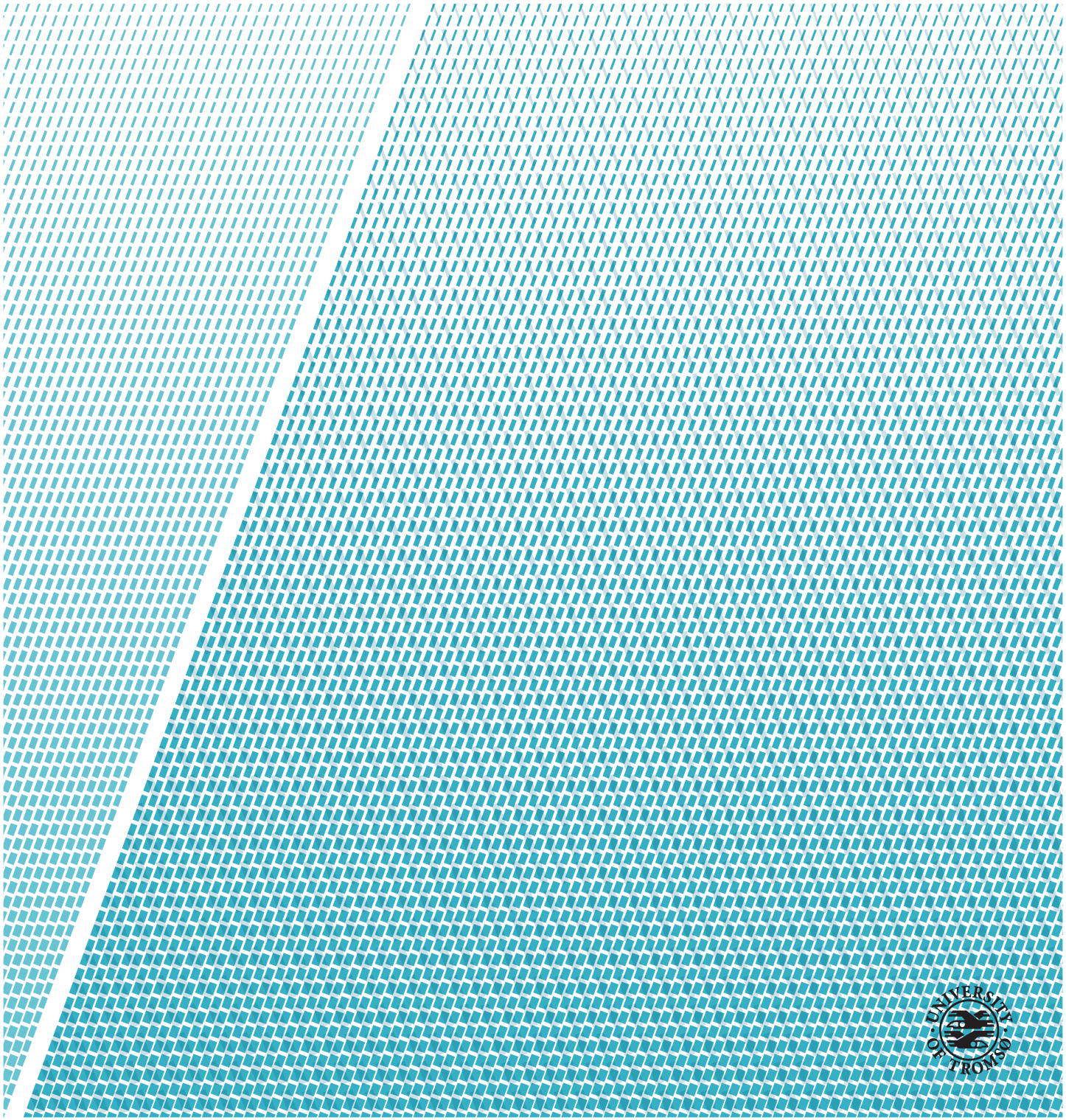
Department of Chemistry, Faculty of Science and Technology

META-pipe – Distributed Pipeline Analysis of Marine Metagenomic Sequence Data

—

Espen Mikal Robertsen

A dissertation for the degree of Philosophiae Doctor



Acknowledgements

The presented study was carried out at the Faculty of Science and Technology, Department of Chemistry, UiT The Arctic University of Norway, Norway, from November 2012 to February 2017. Financial support for this study was provided by UiT The Arctic University of Norway.

Firstly, I would like to express my sincere gratitude to my supervisors Professor Nils-Peder Willassen, Lars Ailo Bongo and Professor Peik Haugen for the continuous support in my Ph.D study, and for their patience, motivation and guidance. I would like to thank all the additional people involved in the development of META-pipe: Edvard Pedersen, Inge Alexander Raknes, Tim Kahlke, Erik Hjerde, Aleksandr Agafonov and Giacomo Tartari. I am grateful for all your contributions and hard work. Special thanks to the "computer guys" for slowly but surely introducing me to the field of computer science with top-shelf expertise at only an arms length. Invaluable.

I would like to thank all the people at Norstruct for a great work environment, trips and social activities, as well as my lunch crew for all the alternative topics we have discussed over the years. I would like to thank the department, technical staff and other co-workers for their help and support. You know who you are.

Many thanks go to my family for their support and patience through my PhD period, especially through the strenuous effort of writing this thesis and the sustained suffering. At one point I even considered having a brain transplant. But i changed my mind...

Lastly, to everyone else, i would like to thank you for attending my defense, which is probably why you are reading this. If you do not understand a word of what I am saying, don't worry. I left you some extra space on several pages in this thesis you can use for doodling or drawing. This way you can look super serious as you were taking notes, while still not paying any attention at all.

Tromsø, February 2017

Espen Mikal Robertsen

Abstract

With the accelerated advances in sequencing technology the last decade, the field of metagenomics has progressed immensely. Sampling and sequencing of metagenomic data is now prevalent, and publicly available data sets from mundane soil and water environments to exotic niche habitats such as geothermal hot springs are readily available through sequence data repositories such as the European Nucleotide Archive. Meanwhile, the computational resource requirements for a complete and comprehensive analysis of metagenomic data have escalated dramatically, due to a tremendous increase in data set sizes. To analyze and make sense of these samples, researchers can choose to employ public resources for metagenomic analysis. However, most of the available public resources provide generic analyses and are not suited for applications such as bioprospecting or samples from complex habitats such as the marine domain.

In this thesis, we introduce a metagenomic analysis pipeline coined META-pipe. With META-pipe, we aim to supply a public analysis resource catered for the marine domain, with an emphasis on analysis of full-length genes. META-pipe offers pre-processing, assembly, taxonomic classification and functional analysis of metagenomic sequence data. The pipeline has gone through several iterations, both in terms of functionality and implementation. In **Paper 1** we describe the initial version of META-pipe, including biological functionality, implementation details and integration with identity provider services, distributed storage, distributed computation and the Galaxy workflow manager. We evaluate the performance of META-pipe through two separate use cases, as presented in **Paper 2** and **Paper 3**. These use cases demonstrate the usability of META-pipe and gave us an opportunity to refine and enhance the pipeline through evaluation of biological results and computational performance characteristics. In summary, this dissertation gives an overview of common strategies for metagenomic analysis in a pipeline context. It discusses the development of META-pipe through refinement and presents the current version. The pipeline is now a deliverable to the ELIXIR infrastructure, hence future versions of META-pipe will continue to improve and expand both in functionality and public usage, providing a sustainable resource for metagenomic analysis in years to come.

Abbreviations

ANN	artificial neural net
BLAST	basic local alignment search tool
bp	base pair
CCS	circular consensus sequencing
DDBJ	DNA Data Bank of Japan
DNA	deoxyribonucleic acid
EMP	EBI Metagenomics Portal
ENA	the European Nucleotide Archive
EnvO	the Environment Ontology
FEIDE	Felles Elektronisk IDEntitet
GO	Gene Ontology
GOS	Global Ocean Sampling Expedition
GSC	the Genomic Standards Consortium
HGP	the Human Genome Project
HMP	the Human Microbiome Project
KEGG	Kyoto Encyclopedia of Genes and Genomes
LCA	Lowest Common Ancestor
MetaHIT	Metagenomics of the Human Intestinal Tract
MG-RAST	The metagenomics RAST server
MIxS	Minimum Information about any (x) Sequence
NCBI	National Center for Biotechnology Information
NeLS	Norwegian e-infrastructure for Life Science
NGS	next-generation sequencing
OTU	operational taxonomic unit
PCR	polymerase chain reaction
RDP	Ribosomal Database Project
RNA	ribonucleic acid
rRNA	ribosomal RNA
SDS	sodium dodecyl sulfate
WGS	whole genome sequencing

Contents

List of Figures

List of Tables

Part I

Thesis

1 Introduction

1.1 Metagenomics

Studies of microbial communities can be traced back as far as 1676, when Antonie van Leeuwenhoek, coined "The Father of microbiology" first described micro organisms in oral cavities [1]. In the following 300 years, studies of microbial communities were mostly based on phenotypic traits, growth experiments and selection. In the late 19th century, efforts were made by Robert Koch to count and visualize microorganisms using cultivation, and he is still renowned for his achievements in identifying the specific causative agents of tuberculosis, cholera, and anthrax [2]. Later, significant improvements in microscopy and staining techniques such as Gram-staining [3] became available and slowly propelled the field of microbiology forward. At the time, it was conspicuous that there was a discrepancy between the amount of organisms identifiable through the use of microscopes and the amount actually procured in culture. With the ideas and work of Sergei Winogradsky, it soon became evident that most microorganisms need special environmental growth conditions to thrive [4]. Cultivation-based methods can only isolate a fraction of the microbial diversity present within a given environmental sample [5, 6]. After Carl Woese proposed the use of ribosomal RNA genes as marker genes for taxonomic classification in 1977 [7], and Sanger automated sequencing was introduced the same year [8], environmental profiling of microorganism diversity based on rRNA genes became the defacto standard for classification of microorgansims. Suddenly, the concept of microbial ecology, a study of microorganisms and their environmental roles and habitats gave hints towards a previously hidden

diversity of microscopic life.

In the last two decades, the field metagenomics, stemming from microbiology, ecology and genomics has slowly emerged and proven its importance. As it is multidisciplinary, metagenomics as a term is prone to varying definitions, but briefly metagenomics is the study of genetic material sampled directly from the environment. The birth of metagenomics as a field is most commonly referenced by the publication of a 1998 article Jo Handelsman et al. [9], where they cloned environmental DNA and explored the biosynthetic machinery of the collective genomes of soil microflora. The term "metagenomics" was introduced for the very first time in this publication and references the idea that a collection of genes sequenced from the environment can be analyzed using similar principles as when analyzing a single genome. Later on, shotgun sequencing, which is arguably one of the corner stone technologies in terms of rapid progression of the field was introduced and utilized [10, 11]. The introduction of the metagenomic approach, coupled with this leap in sequencing technology revolutionized microbial research and offered scientist a lens to view the microbial world in a completely new way.

Today, metagenomics offers vast possibilities of sample analyses (A simplified typical overview is depicted in Figure 1.1, which is described in detail in sections 1.2 and 1.3). Although the development of novel approaches and methodology in metagenomics has accelerated dramatically since its introduction, most research can still be summed up in three questions, "Who is there?", "What are they doing" and "Who is doing what?". The first question addresses taxonomic classification, the identification and quantification of organisms in a sample. Traditionally, environmental profiling of sample diversity was solely based on rRNA genes due to its high degree of conservation between species, but with the advances in sequencing technology, novel approaches to answer this question has emerged. Taxonomic classification of species in a metagenomic sample can now be inferred not only from marker genes based on rRNA, but also using clade-specific marker genes, binning of sequences or mapping to references through assembly [12]. This is possible as experimental design and sequencing strategies in metagenomic projects are slowly moving away from the standard amplicon approach, to shotgun sequencing of all available DNA within a sample (often referred to as whole genome sequencing for single genomes, or WGS). With shotgun sequencing, access to the complete functional gene composition of whole microbial communities is granted, paving way for a myriad of novel functional analysis methods which aims to answer the question, "What are they doing?". Through assembly, fragments of genetic sequences are reconstructed to contigs, continuous stretches of genomic DNA from represented species in the sample. This allows for prediction of full-length genes and non-coding features, operon analyses, pathway analyses, protein family diversities and countless other types of functional annotation. The third question, "Who is doing what?", involves combining taxonomic and functional analyses, which can give evidence of genomic linkages between function and phylogeny,

and evolutionary profiles of community function through biomarker discovery [13, 14]. Additionally, conclusions of grander scales can be made by comparison to publicly available samples and their metadata from repositories like the European Nucleotide Archive (ENA) [15]. Alternative related approaches to standard metagenomic methodology has started to surface recently. Examples include metatranscriptomics (transcriptomics on metagenomic data), the analysis of the total expression of genes in a community [16]. Also, metaproteomics (proteomics on metagenomic data) has been utilized to determine the relative abundances of proteins in a metagenomic sample [17]. Lastly, metagenomics in it self provides an invaluable resource through bioprospecting, where the aim is to discover novel enzymes or other bioactive compounds which can have huge impacts in biotechnological applications [18].

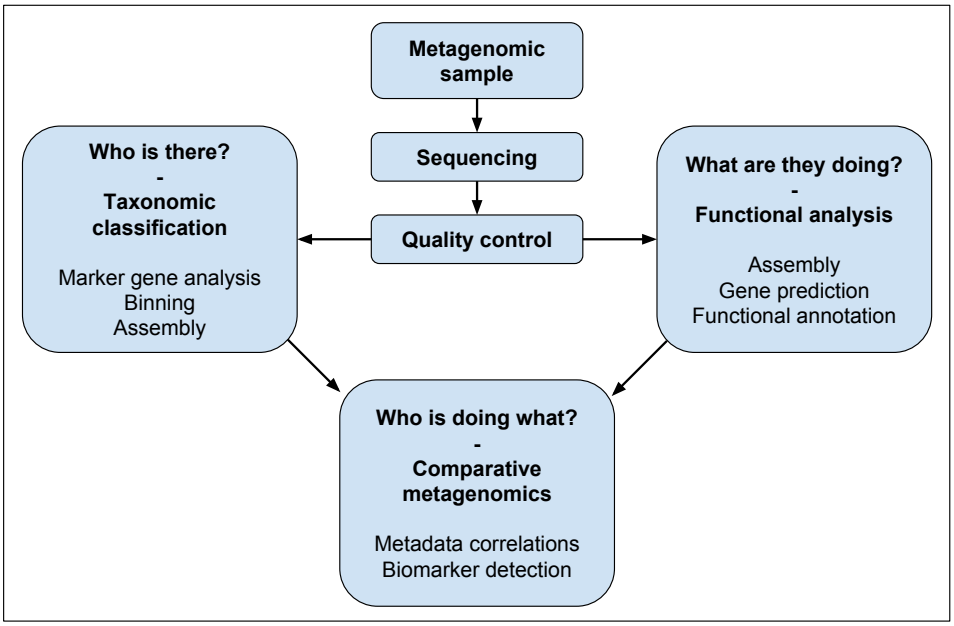


Figure 1.1: The common steps involved in a typical metagenomic analysis workflow.

1.1.1 Applications of metagenomics

Most of the activity in the field of metagenomics has so far been done in a research context, however this research is obviously a driver for novel applications of metagenomics. In medicine, projects such as The Human Microbiome Project (HMP) [19] has revealed that the microflora present in human gut and intestines has a huge impact on health, both directly through dysbiosis [20] and as a fingerprint of other diseases or afflictions [21]. Recently, inflammatory bowel diseases such as Crohn's and ulcerative colitis, cirrhosis of the liver and

colorectal cancer has been shown to be predictable using supervised machine learning [22], which illustrates how metagenomics has enormous potential in diagnosis. However, so far most practical applications of metagenomics in medical diagnosis consists of identifying known pathogenic organisms through sequencing of gut microflora. The process of bioprospecting, where one tries to discover and possibly commercialize novel bioactive compounds from biological resources such as metagenomic samples, has also provided some new antibiotics, including beta-lactamases [23], Fasamycin and Violacein [24]. Other compounds screened from metagenomic samples are enzymes such as cellulose [25, 26] and xylanase [27, 28], involved in the conversion of biomass into biofuel by subsequent fermentation into ethanol. Such biofuels has been adopted by a wide range of vehicles in public transportation in recent years. Another area of application is bioremediation, where metagenomic approaches are used in treatment of oil spills. Using chemical surfactants, petroleum hydrocarbons are made soluble by emulsification and can be easily degraded by microbes. However, these chemical surfactants have been proven to be toxic to the environment [29]. An environmentally friendly alternative with low toxicity and high biodegradability are biosurfactants [30]. Efforts have been made to develop screening methods for biosurfactant producing microorganisms from metagenomic samples [31]. By employing these methods, novel genes involved in biosurfactant production can be identified and hence accelerate the development of bioremediation technologies.

1.1.2 Large-scale projects and outcomes

Even though metagenomics as a field is arguably still in its birth fase, funding within the field has started to increase and large-scale projects has started to emerge. Such large-scale projects not only provide novel insights and breakthrough discoveries on their own, but also help to steer research activity in a field with seemingly endless possibilities. Publicly available data, software resources, tools and standards developed and released in tandem with such grand collaborations are also highly beneficial for medium and small-scale projects. Because of the multidisciplinary nature of metagenomics and its apparent grand challenges like the immense complexity of microbial communities and geographical scale of sampling, these large-scale projects are especially important to support valid generalizations and "proof of concepts", which are not possible to achieve from small single-investigator projects. An early example from the field of genomics that emphasizes this concept is the Human Genome Project (HGP)[32] which launched in 1990. In 2003, they closed the gaps in the sequenced human genome and released a high-quality publicly available sequenced genome, along with freely available tools for researchers to analyse their data. Since 2006, when the first next-generation sequencers was commercialized, terrabase-scale metagenomic sequencing projects have emerged. An

illustrative, but non-exhaustive list includes the projects:

- Global Ocean Sampling Expedition (GOS) [11, 33, 34]
- Metagenomics of the Human Intestinal Tract (MetaHIT) [35]
- The Human Microbiome Project (HMP) [19]
- TARA Oceans [36]
- Malaspina [37]
- MetaSoil [38]
- The JGI Great Prairie Grand Challenge pilot study [39]

Projects such as MetaHIT and HMP have greatly accelerated science towards understanding the gut microbiome in relation to human health. In HMP, the aim is to study the complexity of human-associated microbial communities using not only metagenomic approaches, but also transcriptomic, proteomic and metabolomic approaches. This way, multiple levels of data will provide insight into how the microbiome and the human host interact to support health or to trigger disease. HMP consists of six different initiatives and is associated with over 500 publications, as well as providing a myriad of tools, methods and reference databases for the scientific community. The project MetaHIT has a similar objective: to establish associations between the genes of the human intestinal microbiota and health and disease. It involves 13 partners from 8 different countries and lasted from 2008 until 2012 with a funding estimate set to 21 million euros. With a focus on obesity and inflammatory bowel diseases, several publications and other important resources such as reference genomes, novel methodology [40] and innovative tools [41] are all outcomes attributed to this large-scale project during its 4 year life span. Large-scale circumnavigation projects such as Tara Oceans, Malaspina and GOS aims to asses genetic diversity in marine microbial communities, and provide invaluable resources in terms of publicly available sequence data. The Tara Oceans project also has a green agenda, focusing their efforts on understanding human effects on the environment, such as impact of plastic debris in the environment and effects of global warming. Soil directed projects such as MetaSoil and the Great Prairie Grand Challenge pilot study aim to determine the impact of land management (such as tillage and fertilization) on soil microbial communities, including cycling of carbon and nitrogen.

1.1.3 Novel challenges in metagenomic analysis

In recent years, the amount of sequencing data produced in the field of metagenomics has increased exponentially. This fact introduces a new and somewhat unexpected set of challenges. Firstly, with sequencing machines yielding up to terrabyte size datasets per run, storage and archiving of data has become increasingly expensive. Sequencing yield has now surpassed Kryder's Law [42],

a postulate that hard disk space doubles annually. In fact, the cost of sequencing a base is now cheaper than storing a byte on a hard disk [43]. Surprisingly, storing metagenomic samples in freezers and sequencing them when needed might become a more feasible economical solution at some point. Secondly, the increase in data size demands an equal increase in computation resources. Traditionally, tools and software for analyzing both genomic and metagenomic data were run on laptops and workstations, single machines capable of handling the amount of data with ease. Today, with some exceptions, a complete metagenomic analysis is typically run on cluster computers with significantly more resources in terms of computation and memory, such as EMP [44] and MG-RAST [45] (described in detail in sections 1.4.1 and 1.4.2, respectively). Thirdly, tools and libraries developed for analysis and handling of sequencdata are rapidly made obsolete due to increasing data amounts [46]. As an example, the libraries Bioperl [47] and Biopython [48] are no longer able to handle the tremendous amounts of sequences generated by next-generation sequencing technologies, and have been replaced by libraries such as HTSeq [49]. Assembly tools such as MEGAHIT need hundreds of gigabytes of memory to assembly a single sample [50], amounts of memory which are not common in high-end laptops and workstations. Even non-redundant databases such as Uniprot are growing exponentially [51], making utilization of such resources more time consuming in terms of computation, but also more comprehensive in an analysis context. Recently, a focus on making resources available in the cloud to provide a more flexible solution to some of these problems have been embraced by the community and providers of computation-as-a-service platforms have emerged [43].

1.1.4 Specific challenges in marine metagenomics

The most essential factor in attaining a comprehensive analysis of a metagenomic sample is the quality and composition of reference data. Through large-scale projects such as HMP and MetaHIT, microorganisms from habitats such as the human gut and intestinal tract are readily represented in reference databases. However, due to the complexity, diversity and general neglection of the marine domain, marine metagenomic reference data is severely insufficient. This causes serious data biases in existing generic reference databases, effectively over-representing well studied organisms and generating a skewed representation of the database. The fact that less than 0.1% of all microbes in the oceans today has been discovered highlights the severity of this issue [52]. In fact, no substantial reference databases explicitly for the marine domain of metagenomics currently exist [53], making specific annotation and analysis of marine samples a serious challenge. Additionally, due to the diversity and complexity of the marine environment, assembly of marine metagenomic samples are also especially difficult. This is discussed in detail in section 1.2.2.

1.2 Production of metagenomic sequence data

Before any bioinformatic analysis can be done on in a metagenomic project, several task must be performed in advance in order to attain sequence data from metagenomic samples. Producing sequence data typically involves experimental design, sample processing, sequencing and notation of metadata. This section describes these steps in detail in a successive manner.

1.2.1 Sample collection, preparation and metadata

In a metagenomics project, the first step is experimental design (Figure 1.2). Most importantly, this task should be guided by the research question at hand, so that ideal sequencing technologies, libraries and protocols for the given project are utilized [54]. Also, any technical, operational or cost based restrictions should be readily avoided, so that the statistical significance of the analyzed results are not undermined in any way [55]. Tools such as Metastats [56] have been developed to focus on this particular aspect of comparative analysis. Sample collection is the second step in a metagenomics project. The sample can come from anywhere there is microbial life, which has lately expanded in terms of extreme environments, such as low oxygen [57], alkalinity [58], acidity [59] and extreme temperature [60, 61]. Ideally, the sample should contain DNA representing the isolated microbial community as a whole, meaning the complexity, abundance and diversity of organisms should be properly reflected through subsequent analysis of a sample. However, sample processing is crucial and introduces potential biases. Depending on the origin of the sample, different protocols are used to ensure a favorable yield of DNA [62, 63]. DNA extraction from metagenomic samples generally consists of three steps, fractionation, lysis and purification [64]. If the sample is host associated, sample fractionation can be used to remove potential host DNA within the sample, especially if the host genome is very large and potentially overwhelming in terms of DNA yield [65]. In some cases, fractionation of certain parts of an environmental sample is necessary to study distinct taxonomical divisions, like separating bacteria from communities with macroalgae [66]. To isolate the DNA within a sample, different types of cell lysis methods can be utilized and combined, which are generally divided into thermal, mechanical, chemical and enzymatic methods. In thermal and mechanical lysis, the physical force generated from for example bead-beating or sonification destroys cell walls and shear DNA into fragments, which is ideal for library construction. With chemical and enzymatic lysis, more subtle ways of DNA isolation are achieved, like dissolving cell membranes with sodium dodecyl sulfate (SDS) or sample digestion with various enzymes. Lastly, purification of the sample is needed to remove any contaminants which might interfere with any subsequent steps like quantification of DNA, enymatic reactions or sequencing [64]. In soil samples, humic acid is

a particular nuisance with similar physiochemical properties as DNA, causing problems in subsequent steps of library preparation [67]. In the end, acquiring the total amount of DNA from a metagenomic sample is not possible because of the extreme microbial diversity and low abundance of certain organisms. Up to fifty percent losses of DNA should be accounted for in this stage alone [68].

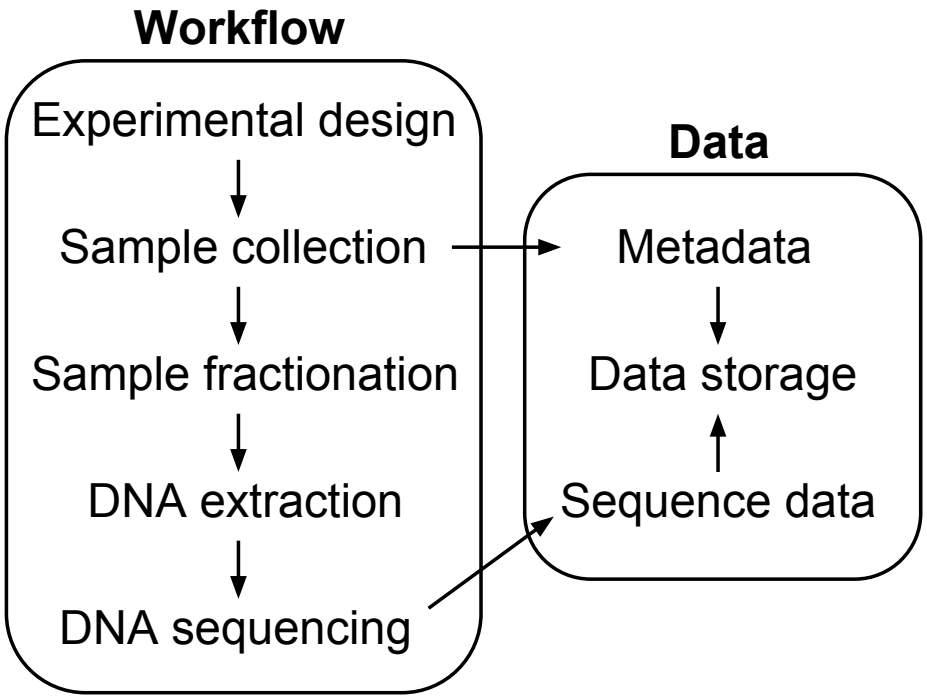


Figure 1.2: The main steps in production of sequence data from a metagenomic sample

With recovering samples from the environment, additional information in the form of metadata and contextual data is also acquired, which in earlier years of metagenomics was not very well taken care of. Metadata are the descriptors of what, how, when and where your sample was taken from, while contextual data describes the environmental conditions. Today, it is common practice to archive recovered samples in repositories like ENA [15], National Center for Biotechnology Information (NCBI) [69] and DNA data bank of Japan (DDBJ) [70], which provides not only a permanent storage of sequence data but also rich metadata information submitted by the user. Controlled vocabularies like EnvO (The Environment Ontology) [71] and MIxS [72] governed by The Genomic Standards Consortium (GSC) have been introduced to handle the description of metagenomic samples in a uniform way. This way, the research community as a whole can benefit from publicly available data through metastudies and comparative analyses.

1.2.2 Sequencing

In order to do any kind of bioinformatic analysis on a metagenomic sample, it needs to be sequenced to produce sequence data from genetic material present in the sample. In 1977, Applied biosystems commercialized the first DNA sequencing method, coined Sanger sequencing [8, 73], a sequencing method based on polymerase chain reaction (PCR) yielding minimal amounts of sequence information. Today, next generation sequencing technology from companies such as 454 Life Sciences and Illumina are extensively applied to metagenomic samples, and can generate terrabytes of information from one sequencing run. This information is produced in the form of reads, fragments of DNA from species present in a metagenomic sample. Even though sequencing technology has made exceptional advances in recent years, the available technologies thus far are not perfect. Sequencers based on 454 pyrosequencing technology typically struggle with sequencing errors, especially DNA homopolymers [74] and have a relatively low output of under 1 gigabyte of sequence information. However, 454 sequencers generate relatively long reads (up to 1000 base pairs), which are advantageous in assembly and subsequent functional annotation. In comparison, Illumina based sequencers offer a substantially higher output of up to 1 terrabyte per run and lower sequencing costs, but with shorter reads (typically in the range of 100 to 300 base pairs). These reads have shown a tendency to have have high errors rates at the tail end [75]. However, all Illumina systems are capable of paired-end sequencing, which yields two reads per DNA insert, one from the forward and one from the reverse template strand with a known distance between them. This strategy is particularly useful for handling DNA with genomic rearrangements and repetitive sequence elements, as the positional information between the two reads helps in alignment to a reference or extending contig lengths in de novo assembly [76, 77, 78]. Newer, less extensively tested technologies with ground breaking properties have also started to emerge lately. Pacific Biosciences offers a sequencing technology coined "circular consensus sequencing" (CCS), which can achieve read lengths of several thousand base pairs. This is especially useful in assembly, annotation and functional assignment, as well as extending contig lengths in hybrid assemblies [79]. Using nanopore technology, Oxford Nanopore is developing portable solutions such as the MinION, capable of detection and analysis of pathogens in-field [80]. They are even developing a sequencer called the SmidgION, which operates as an in-field accessory to a smart-phone.

In addition to sequencing technologies there are also different variations in approaches, depending on the application or research question at hand. In metagenomics, two of the most widely used approaches are amplicon sequencing and shotgun sequencing (WGS) (Shown in Figure 1.3). With amplicon sequencing, the goal is usually to uncover the species present in a metagenomic sample through a taxonomic classification. The sequencing target is the marker gene 16S rRNA, which has become the de facto standard for taxonomic anal-

ysis of prokaryotic diversity due to its inherent conservation between species. Parts of the 16S rRNA gene is sequenced using primers covering one or more of the variable regions within the gene using PCR. Using this method, only sequences stemming from the 16S rRNA gene of all organisms present within a sample will be sequenced. Reducing the target template to only this locus produces a deeper coverage than for example shotgun sequencing, meaning it is generally less costly and samples can be multiplexed (multiple samples per sequencing run, separated using barcode sequences). However, this method has some proven biases. The unspecific primers used to amplify the 16S rRNA gene might not adhere to all present strains with equal affinity, resulting in a distorted representation of the actual diversity of the sample. Primer affinity can be checked using tools such as TestPrime [81].

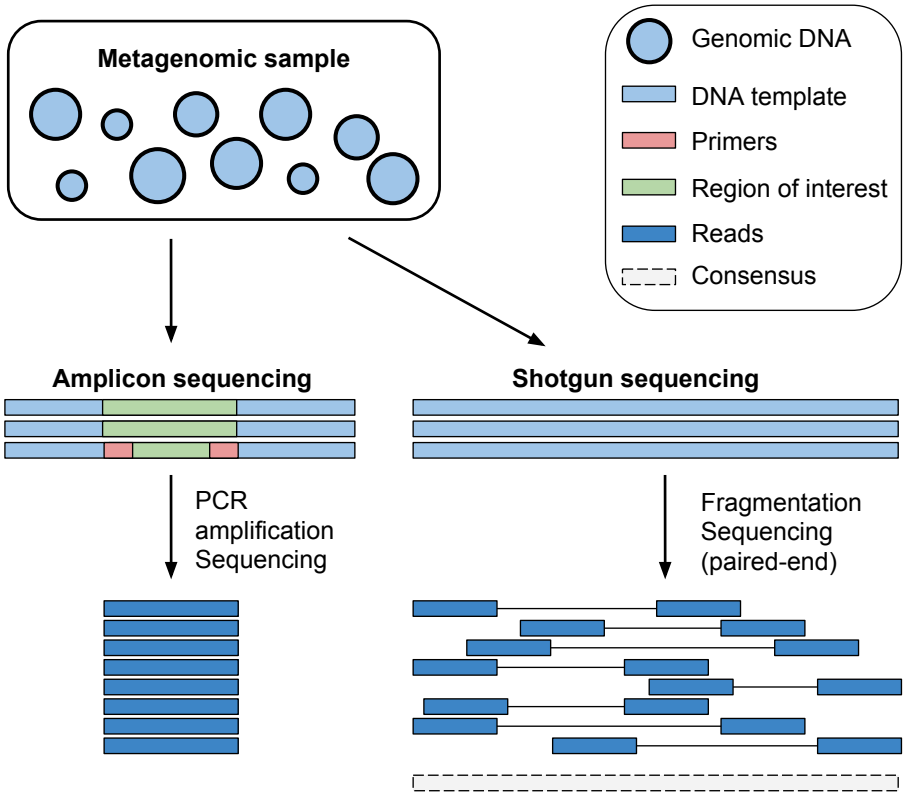


Figure 1.3: A simplified comparison of the two sequencing apporaches: Amplicon sequencing and shotgun sequencing. Amplicon sequencing targets a particular region of interest, usually part of the 16S rRNA gene for prokaryotic taxonomy analysis. With shotgun sequencing, random fragments of DNA from all species are produced, which is built into longer contigs (consensus)

With shotgun sequencing, random fragments of DNA from all species in a sample are produced [82, 83]. This offers not only a means for analyzing

sample diversity, but also the sequences of coding genes and other forms of non-coding DNA, which yields additional functional knowledge about the microbial community as a whole. Since this approach targets all DNA present in a sample, and not just a specific gene or locus such as in amplicon sequencing, a larger volume of data is needed to achieve a viable coverage of the sample. This is a particular problem in metagenomic assembly, where reads are built or reconstructed into longer segments of DNA called contigs. The diversity and complexity of a metagenomic sample implies that not all genomes present will be represented by reads, making this reconstruction especially challenging [84]. A naive way to tackle this problem is to use a sequencer with especially high output capacity, but since the distribution of abundance in the sample remains the same, normalization of reads should be employed afterwards [85]. Paired end libraries can also be used to help facilitate the joining of contigs more easily, as the distance between read pairs are known (as shown in Figure 1.3). However, even though sequencing technology has progressed tremendously in recent years, most sequencers can still only scratch the surface of the actual DNA available in a complex metagenomic sample using a shotgun sequencing approach.

1.3 Pipeline analysis of metagenomic sequence data

An efficient way to solve some of the novel challenges in recent metagenomics projects is to use pipelines backed by substantial computational resources. These pipelines represent automatic or semi-automatic work flows that process a sample from raw data to a complete taxonomical and functional analysis of a metagenomic sample. This section will give an overview of the most common steps involved in a metagenomic analysis pipeline in a successive order, as well as describing popular tools involved in each step. The focus of this step-by-step overview is shotgun sequencing specific, however some steps are applicable to amplicon data as well, such as sections 1.3.1 and 1.3.3.

1.3.1 Quality control

An important first step before starting any analysis is to assess the output quality of the data from a sequencing run. Removal or trimming of low quality reads is vital to produce analysis results with minimal biases stemming from sequencing errors. Depending on the sequencing technology used, particular biases or patterns of errors intrinsic to the sequencing technology used need to be considered and evaluated carefully [86]. Omitting this step will have nega-

tive effects in characterizing the microbial community present in the sample and reconstructing genomic DNA in the process of assembly [86, 87]. As an assembly is often the basis for a functional analysis of a sample, an insufficient assembly causes an incomplete assessment of functionality. Quality filtering is also important from a computational perspective. Too much low quality input in assembly increases processing time and memory requirements [88].

Tool name	*Type*	*Description*	*Reference*
AmpliconNoise	Denoising	Reduction of errors from amplicon sequence data	[89]
BBTools	QC-filtering	Toolkit for sequence data. Available at http://jgi.doe.gov/data-and-tools/bbtools/	-
DeconSeq	Host contamination	Removal of host sequencing reads from host-associated samples	[90]
FastQC	QC-filtering, evaluation	Quality filtering with a graphical user interface for easy assessment	[91]
Fastx-Toolkit	QC-filtering	Collection of commandline tools for short read preprocessing	[92]
PRINSEQ	QC-filtering	Preprocessing of genomic and metagenomic sequence data	[93]
Trimmomatic	QC-filtering	Trimming of Illumina sequence data	[94]

Table 1.1: A list of common software used in evaluation and filtering of raw sequencing data

To evaluate and trim reads from next-generation sequencing (NGS) data, several programs are available (mentioned alternatives are referenced and listed in Table 1.1). Collectively, they provide calculated statistics such as number of reads, over-representation of reads, length, quality profiles and more. Programs such as FastQC, Fastx-Toolkit, PRINSEQ and Trimmomatic are generic QC-tools offering filtration, trimming and removal of low quality sequencing reads. Some of them can also remove platform-specific artifacts, like adapter sequences. In an automatic pipeline context, setting generic parameters for such tools can be challenging, as no sample has the exact same quality characteristics. Evaluation of QC results should ideally be manually examined to find the optimal trade off between average sequence quality and discarded sequence data. However, this requires specific user-competence and represents a manual intervention not ideal in an automatic pipeline. Other more situational types of quality control programs are also available. DeconSeq can remove host contamination using a reference sequence, which is often necessary in host-associated samples with reference sequences available. The software package BBTools includes normalization to optimize distribution of sequencing reads, which can decrease memory and computation resources needed in subsequent analyses drastically, depending on the diversity and complexity of the sample.

With amplicon sequencing, the same generic QC-filtering applies, however additional processing is often required. Depending on the sequencing technology used, the sequence data might need a certain extent of denoising to reduce intrinsic errors stemming from sequencing errors. This can be done with software such as AmpliconNoise. However, the effects of this process in terms of sample richness, diversity and evenness depending on which algorithms are used can vary greatly [95].

1.3.2 Assembly

In an assembly, QC-filtered reads are built into long stretches of DNA called contigs, exploiting the inherent overlap of reads stemming from the sequencing process. This is done to gain access to full length genes and operons, which can provide valuable functional information about the community as a whole. When assembling single genomes, the dataset consists of only a single organism, which is a task that has been thoroughly studied and effective algorithms have been developed. However, assembly of metagenomic data is not as trivial. This section introduces common strategies and tools available in metagenomic assembly (referenced and listed in Table 1.2). Some of the introduced tools are designed for genomic assembly, but can be used for metagenomic assembly with special care.

In a metagenomics project, assembly is especially difficult due to the diversity and abundance of organisms in the sample. Firstly, a metagenomic sample represents a distribution of abundance between organisms, meaning abundant organisms will be represented with sufficient sequence data coverage, while less abundant organisms are effectively impossible to assemble. Secondly, some species may contain homologous genes or other sequencing artifacts representing a consensus sequence between them, which typically causes spurious and chimeric contigs [104]. Additionally, since the abundance and diversity of organisms in a metagenomic sample are vast compared to a genomic sample, the memory footprint using metagenomic assemblers can reach hundreds of giga bytes, an amount not suited for a common workstation or laptop. In fact, sequence assembly has been proven to be NP-hard, a class of computational problems which can not be solved in polynomial time [105]. Several approaches to tackle these challenges have been employed, such as binning by sequencing depth, effectively categorizing reads by abundance as seen with Meta-IDBA and MetaVelvet. However, assembly yield is inevitably bound to sample coverage, complexity and abundance of organisms.

When assembling metagenomic data, two distinct routes can be taken: De novo assembly and reference based assembly. Reference based assembly involves mapping metagenomic reads to a collection of known references, hereby achieving a more precise assembly and species or genus specific taxonomic information on contigs in the process. However, this approach needs an extensive

Tool name	*Type*	*Description*	*Reference*
Artemis	Visualization	Sequence annotation and visualization tool	[96]
Celera	Assembler	Consensus and variant detection using whole genome sequencing datasets	[97]
MEGAHIT	Assembler	Fast and memory efficient de novo assembly of metagenomic data	[50]
Meta-IDBA	Assembler	De novo assembly of metagenomic data using partitioning and creation of consensus sequences	[98]
MetAMOS	Assembly (pipeline)	Metagenomic assembly and analysis pipline	[99]
MetaQUAST	Quality evaluation	Evaluation tool for metagenomics assemblies	[100]
MetaVelvet	Assembler	De novo assembly of metagenomic data using coverage decomposed graphs	[101]
MIRA	Assembler	Multi-pass sequence assembler and mapper	[101]
Ray Meta	Assembler	Scalable de novo metagenomic assembly	[102]
TIGR	Assembler	Greedy assembler	[103]

Table 1.2: A list of common tools used in assembly

reference database tailored for the specific habitat of the sample at hand. If the reference database is insufficient, or the sample is from an especially complex habitat, any reads belonging to unrepresented references will not be assembled. The algorithms employed are generally faster and more memory-efficient, making this approach viable for standard computers. Examples of tools able to perform this type of assembly are MIRA and MetAMOS.

With de novo assembly, no reference sequences are used, and overlapping reads are built into contigs without any supporting knowledge. The algorithms utilized in this process can be divided into three distinct types, greedy assemblers, overlap assemblers and De Bruijn-graph assemblers. These types have different strengths in terms of memory usage, processing time and precision. Greedy assemblers are generally simple implementations which iteratively merges contigs through maximum overlap, and is effective when assembling data with no repeats. An example of such an assembler is TIGR. Overlap assemblers uses a pairwise overlap approach, which tackles error prone reads well, but is not optimal with high coverage datasets as the pairwise computation becomes strenuous. Noteworthy overlap assemblers include Celera, which was used to reconstruct the human genome [106] and the Arachne assembler. De Bruijn graph assemblers are generally considered state-of-the-art and uses kmers, fragments of input reads of a set length to construct graphs. Contigs are

reconstructed by analyzing "walks", routes through the constructed graph based on kmer count which decides which contigs to keep and discard. Assemblers that utilize De Bruijn graphs for metagenomic assembly include Meta-IDBA, MEGAHIT and MetaVelvet. However, if a dataset is very complex or contains a high proportion of sequencing errors, this graph will grow out of proportion and require extensive amounts of memory. One way around this caveat is to distribute graph construction, a functionality provided with the metagenome assembler Ray Meta. This way memory requirements and computation time can be mitigated between multiple computers in a cluster environment.

When an assembly of metagenomic data is completed, it is often necessary to evaluate the performance of the applied tool and parameters set. Most assemblers offer simple statistics such as N50 (weighted median contig size), total contigs, largest contigs, percentage of bases assembled etc., however this is not sufficient information to properly validate an assembly. Some of these values are also repeatedly misunderstood in a metagenomic context. As an example, the N50 measurement is a rather meaningless value as one does not know the correct size of the combined genomes. To assess an assembly in a more rigorous manner, tools such MetaQUAST can be utilized. MetaQUAST aligns assembled contigs to reference genomes and outputs a detailed overview of coverage and mis-assemblies relative to the provided references. This way, if a sufficient reference dataset exists, a thorough assessment of the assembly can be achieved. Furthermore, assemblies can be quantified and evaluated using visualization tools such as Artemis. This sequence viewer allows for in-depth analysis of coverage information in a per base context. This allows for identification of specific gaps or assembly errors, but is tedious for longer contig segments and should be utilized only for specific loci of interest.

1.3.3 Taxonomic classification

When performing a taxonomic classification, the aim is to characterize and quantify the microbial community. This is vital to uncover the richness and abundance of organisms present, and answers the question "Who is there?" for a given sample. Depending on the type dataset analyzed, this assessment can be quantified using three distinct approaches, namely marker gene analysis, binning or assembly (Figure 1.4). These approaches are not mutually exclusive, and combinations of approaches are implemented in various publicly available tools able to perform this task. In this section, an individual explanation of these approaches as well as common tools and databases utilized is described. Any tools or databases mentioned in this section are summarized and referenced in Table 1.3

The traditional and most extensively adopted way of assessing taxonomic diversity is using marker genes. This strategy involves comparing sequenced reads against databases with taxonomically informative marker genes to identify ho-

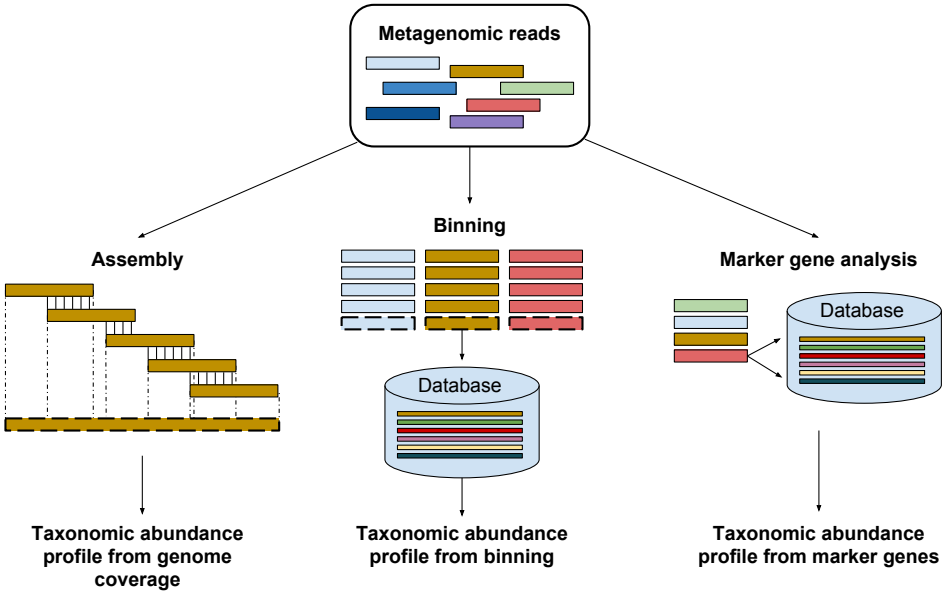


Figure 1.4: An overview of taxonomic classification approaches

Tool name	*Type*	*Description*	*Reference*
LCAClassifier	Marker genes	Taxonomic classificatoin using the lowest common ancestor algorithm	[107]
Greengenes	Database	16S rRNA gene database	[108]
MetaPhlAn	Marker genes	Taxonomic classification of microbial communities using clade-specific markers	[109]
PhyloPithia	Binning, Phylogeny	Phylogenetic classification of DNA fragments	[110]
Phymm	Binning, Phylogeny	Phylogenetic classification using Markov models	[111, 112]
QIIME	Binning, Marker genes	Pipeline for microbiome analysis of metagenomic data	[113]
SortMeRNA	rRNA prediction	Prediction, mapping and OTU picking of rRNA sequences	[114]
RDP	Database	Bacterial, archaeal and fungal rRNA sequence database	[115]
rRNASelector	rRNA prediction	Prediction of rRNA sequences in metagenomic data	[116]
Silva	Database	Small and large subunit rRNA sequence database	[117]

Table 1.3: A list of common tools and databases used in taxonomic classification

mologous matches. Most commonly, marker genes are represented by rRNA sequences due to their inherent conservation between species. For prokaryotic

assignment, the 16S rRNA subunit is commonly used, however many databases include 18S rRNA for eukaryotic assignment, as well as their large subunit counterparts (23S/28S). State-of-the-art databases commonly used for homologous comparison of rRNA includes Silva, Greengenes and the Ribosomal Database Project (RDP). For amplicon datasets this strategy is relatively straight forward as they consist of only rRNA sequences, but rRNA sequences can also be predicted and extracted in whole genome sequencing datasets using software such as rRNASelector and SortMerRNA. The resulting set of extracted rRNA sequences can be analyzed in a similar manner, albeit with special care as they are more fragmented due to the random nature of whole genome sequencing reads. The most common method for taxonomic assignment is the Lowest Common Ancestor (LCA) method. With this method, a read with multiple homologous database hits is assigned to the taxa which is the lowest common ancestor to the acquired hits (descendants) in a hierarchical graph context, given a set of stringent parameters. This way the method is relatively accurate, but lacks resolution at strain and family-level taxa [107], such as can be seen in LCAClassifier. Marker genes can also be represented by clade-specific (genes only common to a monophylitic group of taxa) protein coding genes, such as with MetaPhlAn, but this requires whole genome sequencing datasets with protein coding genes. Common for most tools using this approach is an effective and computationally efficient classification, as databases are relatively small. However, it assumes that the fraction of sequences with homologous hits to marker genes reflects the total diversity within the sample. Depending on the coverage of the database used, and the environment the sample represents, this might not hold true [118].

With binning the aim is to assign sequences into groups, either by shared characteristics (such as homology or GC-content) or by comparison to reference data. The binning approach is often a precursor to other approaches, such as marker gene analysis or assembly reference comparison, effectively sorting sequences before taxonomic assignment. Sorting sequences this way provides a number of benefits. Firstly, it reduces the complexity of input data, so that subsequent analyses are generally less computationally expensive and can be executed on individual bins or sets. Secondly, it provides the ability to discover novel strains in metagenomic data otherwise difficult to identify due to lack of reference data [119]. A popular tool utilizing this technique is QIIME. Sequenced reads are binned into OTUs (operational taxonomic units) based on identity, representing provisional groups of unknown taxa which are subsequently taxonomically assigned using a reference database. The tools Phymm and PhyloPithia both use compositional binning (oligonucleotide frequency and length) to produce an overview of phylogenetic linages and discover novel unknown organisms, respectively. However, the binning process also introduces some caveats. As reads are effectively represented as bins, annotation or classification of a bin does not necessarily reflect the true annotation or classification of an individual read, depending on the specific parameters used in the binning

process.

A taxonomic classification is also obtainable using assembled sequences (Described in section 1.3.2). Contigs produced from assembly can be quantified by tracking coverage, meaning to count reads aligning to each specific contig. This way, annotated contigs representing individual strains can be quantified, effectively producing a profile of taxonomic diversity and abundance. Tools such as MetaVelvet and Meta-IDBA (listed in table 1.2) generate sub graphs in their effort to separate the microbial community into groups, which can be treated as a representation of genome-specific divisions. Caveats discussed in section 1.3.2 also applies in taxonomic classification. Any spurious or chimeric contigs produced in assembly will naturally effect the precision of taxonomic classification using this approach. Also, as an assembly is only viable for relatively abundant organisms in a sample, low coverage strains will not be identified.

1.3.4 Gene prediction

Following assembly, predicting genes or other features from genomic contigs is usually a precursor to a functional analysis. These reconstructed stretches of DNA will most likely contain genes which can be predicted and extracted using metagenomic gene prediction tools. This step is imperative as a set of coding genes from a metagenomic sample naturally reflects the profile of its collective biological functions. In this section, common tools and strategies used in metagenomic gene prediction will be introduced. Any gene prediction tools covered in this section are listed in Table 1.4.

Assembly is not necessarily a prerequisite for gene prediction; genes can be predicted directly from raw reads as can be achieved with FragGeneScan. This tool also incorporates sequencing platform specific error models, increasing the accuracy of genes predicted from raw reads. However this will produce mostly fragmented genes depending on the sequencing technology and length of reads, which is not ideal in a functional analysis context [135]. A set of fragmented genes will produce an overview of functionality based on fragments, but is not sufficient if the aim is to mine for novel full-length proteins or enzymes in a bioprospecting context. Longer contigs (upwards of 500 bp) will have higher chances of containing non-truncated full length genes, hence the quality of assembly is an important factor in this step. Gene prediction in longer fragments of DNA is generally easier to implement, and fewer genes are missed by gene prediction tools [136]. A number of tools specially developed to provide de novo gene prediction in metagenomic sequence data has been introduced, including MetaGeneAnnotator, MetaGeneMark and Orphelia. Collectively, they all use models that are trained using sequence properties such as GC-content, codon usage and length to optimize prediction and discriminate coding an non-coding stretches from a DNA template. As these tools do not rely on any reference databases or alignment algorithms, they are also relatively fast com-

Name	*Type*	*Application*	*Reference*
BLAST	Annotation tool	Basic Local Alignment Search Tool	[120]
FragGeneScan	Gene prediction	Gene prediction in fragmented short reads	[121]
FROMP	Gene prediction	Fragment recuitment using metabolic pathways	[122]
GO	Ontology	A comprehensive, computational model of biological systems	
HMMer	Algorithm	Hidden Markov Model search and alignment tool	[123]
InterPro	Database	Functional analysis of proteins and protein families	[124]
InterProScan	Annotation tool	Annotation tool that integrates with Interpro	[125]
KEGG	Ontology	Database resource for annotation of functions and utilities in biological systems	[126]
MetaGene Annotator	Gene prediction	Gene-finding program for prokaryotic and phage sequences	[127]
MetaGeneMark	Gene prediction	Gene prediction exploiting oligonucleotide frequencies and nucleotide composition	[110]
MetaPath	Annotation tool	identification of differentially abundant metabolic pathways in metagenomic datasets	[128]
NCBI	Databases	The National Center for Biotechnology Information	[129]
Orphelia	Gene prediction	Metagenomic gene prediction tool	[130]
Pfam	Database	Collection of protein families, represented by sequence alignments	[131, 132]
PRIAM	Annotation tool	Enzyme-specific profiles for metabolic pathway prediction	[133]
Uniprot	Database	Universal protein resource catalog	[134, 51]

Table 1.4: A list of common tools and databases used in gene prediction and functional analysis

pared to other approaches. Other approaches include identifying genes through alignment to reference sequences or binning (fragment recruiment) which was utilized in the analysis of metagenomic data from the Global Ocean Sampling Expedition [33] and in the metabolic pathway profiling tool FROMP.

1.3.5 Functional analysis

Using a set of protein coding genes predicted by a metagenomic gene prediction tool, the functional diversity can be quantified by functional annotation

(Figure 1.5). This annotation is the basis for a functional analysis, which answers the question "What are they doing?" for a given sample. A lot of resources and algorithms are available to annotate metagenomic sequence data and enrich the description of each specific gene predicted in a metagenomic sample (listed and referenced in table 1.4 and described in the next paragraphs). Since this step relies heavily on alignment algorithms and relatively big databases, it represents a considerable computational effort in terms of processing and storage. A complete functional analysis is best performed on a distributed computer cluster or cloud environment in order to finish in a reasonable amount of time.

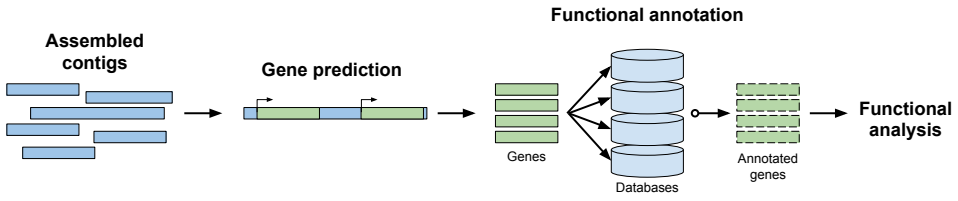


Figure 1.5: Functional analysis workflow

Databases used in functional annotation generally come in two types: sequence databases and HMM (Hidden Markov Model) databases. Sequence databases consist of multiple sequence entries which provide specific hits for a query sequence that is closely related to sequences in the database. Databases suitable for metagenomic data are mostly non-redundant, however the definition of redundancy varies between different database providers. As an example, Uniprot (Universal Protein Resource) consists of three different databases, UniprotKB ("KnowledgeBase", which integrates TrEMBL and Swiss-prot), Uniparc (Archive) and UniRef (Reference clusters). UniprotKB contains one record for all 100% identical full-length sequences in one species, while UniRef contains one record for all 100% identical sequences regardless of species. These variations of Uniprot databases are used extensively in metagenomic analysis as they represent a comprehensive resource in terms of protein annotation. However, the growth of uniprot is exponential [51] and consists of over 70 million entries as of 2017, which poses a problem when used for sequence similarity searches. As these databases are typically queried by BLAST-programs (Basic Local Alignment Search Tool), all query sequences are compared to all sequences in the database, which represents a substantial and exponentially growing computational effort as datasets and databases increase in size. A similar major resource for sequence analysis is NCBI, providing not only sequence databases, but also additional resources such as literature search engines and software. In a metagenomic functional analysis context, the most relevant database is Protein, which consists of sequences from several external sources and provides biological structure and function determination. Other relevant databases under the NCBI umbrella relevant to metagenomic analy-

sis are RefSeq for reference sequence analysis and GenBank which provides an extensive nucleotide archive incorporating genes, genomes, protein and transcripts from several sources.

HMM databases consists of profiles built from seeds, aligned homologous sequences that represent a related entity, for example a protein family (Pfam) or a class of enzymes (PRIAM). Compared to sequence databases, HMM databases generally identifies more distantly related relationships, as the profile is a probabilistic model built from seed sequences from divergent sources. This type of database is commonly queried by the tool HMMer, which also offers database formatting from plain sequence data. One of the most extensively used HMMer-based databases in a metagenomic context is Pfam. With Pfam, protein sequences from predicted genes are classified into protein families, which represents groups of evolutionarily related protein sequences. Pfam is also included in InterPro, a collection of 14 databases incorporated into a single searchable resource. InterPro provides functional annotation and classification using these integrated databases, and equips the specialized search tool InterProScan for easy integration with InterPro.

When predicted genes are annotated, further mapping to descriptive ontologies which can summarize and clarify the annotation in a comprehensive manner should be performed. Many of these ontologies are available, most commonly used are the Kyoto Encyclopedia of Genes and Genomes (KEGG) for metabolic pathway analysis and Gene Ontology (GO) mappings to describe functional relationships within a sample. Several tools are available to reconstruct metabolic pathways from metagenomic data, such as MetaPath. MetaPath uses statistic methods and prior pathway knowledge to identify differentially abundant pathways present in a sample. The functional annotation tool PRIAM can also map enzyme annotations from predicted genes to KEGG identifiers automatically. A characterization of the overall metabolic pathway within a sample is important to fully understand its complete enzymatic capabilities and synergies between species. The more functional descriptive ontology GO aims to define concepts of gene function through classes, such as metabolic function, biological process and cellular components. Various slimmed versions of this ontology is available to reduce the vocabulary and simplify the functional description, such as metagenomic slim for metagenomic data. Together, metabolic and functional ontologies like these serve as a basis for a functional analysis of a metagenomic sample.

1.3.6 Comparative analysis and visualization

One of the central challenges in the study of metagenomic data is making sense of differences between samples from different microbial communities. A comparative analysis involves finding genes, organisms, pathways and other elements that consistently explain these differences coined biomarkers, and

present or visualize them in a comprehensive manner. In taxonomic classification, the terms alpha, beta and gamma diversity are used to explain the observed differences between taxonomic profiles of samples. These terms were introduced by Whittaker in 1960 [137], who proposed the idea that gamma diversity, the total species diversity of an ecosystem was defined by alpha and beta diversity. Alpha diversity refers to the mean diversity of microorganisms at a specific site or habitat (local species pool), while beta diversity defines the differentiation among these habitats. These terms are often used in tandem with tools for comparing taxonomic classifications, such as QIIME (listed in table 1.3). The biomarker discovery tool LEfSe [14] uses genomic features such as pathways, genes and taxonomic information to characterize the differences between samples on both taxonomic and functional levels. However, in a pipeline context, rigorous comparative analysis between samples has long been neglected, with a focus solely on overviews of taxonomy and function, but the newer versions of both EMP and MG-RAST pipelines has started incorporating tools for comparison (described in sections 1.4.1 and 1.4.2). A number of tools and exist to visualize metagenomic data. The interactive visualization tool KronaTools [138] uses HTML5 and JavaScript to create dynamic pie charts viewable in a web context, and supports a number of bioinformatics related data formats out of the box. This tool is used extensively in recent publications [139, 140, 141] as well as in the web user interface of EMP for displaying taxonomic classifications, as it offers snapshots of publication-ready SVG-files out of the box as seen in Figure 1.4. Elvis [142] is another interactive visualization tool for metagenomic assemblies. It offers the capability to correlate meta data with attributes of assembly, such as GC-content, contig length and relative abundance. With this interactive solution, the quality and attributes of metagenomic assembly can be studied in real time and hypothesis generation and testing can be greatly accelerated.

1.4 Established pipelines

1.4.1 EMP - European Bioinformatics Institute

The EBI Metagenomics Portal (EMP) is an online resource for metagenomics analysis [44]. Through a state-of-the-art pipeline users of this resource can have their metagenomics samples analyzed in exchange for making their analysis results and raw sequence data publicly available. EMP accepts both metagenomics and metatranscriptomics data and have seen rapid growth over the years, representing one of the largest freely available resources for metagenomic analysis today. As of 1st of January 2017, EMP has analyzed 792 publicly available projects totaling over 50.000 samples. The pipeline has undergone several updates and changes through the years, and now includes a more com-

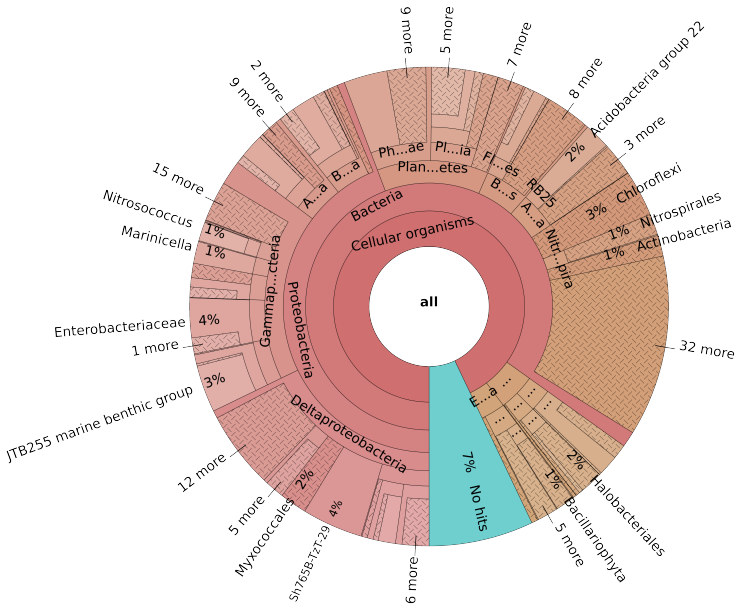


Figure 1.6: Visualization of a taxonomic classification of a marine metagenomic dataset by KronaTools

prehensive and easy to use website with web based tools for functional comparison of samples within a study [143]. EMP is coupled with ENA, which handles metadata and sequence daya submission via the ENA Webin tool. The EMP pipeline offers quality filtering and both taxonomic classification using QIIME and functional analysis using FragGeneScan and InterProScan (described in in detail in section 1.3). When samples are processed, users can access their analyses through the EBI Metagenomics web portal and browse taxonomy, functional annotation and download results. EMP represents a major resource for metagenomic analysis and has expanded considerably in recent years, both in terms of data sets analyzed and pipeline development and capacity.

1.4.2 MG-RAST - Argonne National Laboratory

MG-RAST [45], similarly to EMP, is another major metagenomic data analysis resource. The fully automated pipeline offers processing, analyses, sharing and dissemination of metagenomic data. With over 200.000 publicly available samples, users can download sequence data and corresponding metadata from a rich diversity of biomes and locations around the world. MG-RAST can process shotgun and amplicon metagenomes, as well as metatransciptomes via upload to the site itself, script-based submission or RESTful API. The pipeline was recently updated to version 4.0 [144], and now offers quality control, protein prediction, clustering and similarity-based annotation, effectively producing

both taxonomic profiles and functional analyses of metagenomic data. The web interface allows for comparison using statistical methods based on KEGG and Clusters of Orthologous Groups (COG) on multiple levels of resolution, as well as meta data incorporation in sample comparison. MG-RAST is a bonafide effort to centralize metagenomic resources in one place and standardize analyses.

2

Aims of the study

Main objectives
The main objective of the presented work was to develop a state-of-the-art metagenomic analysis pipeline. As the field of metagenomics has seen an extreme progression through the rapid development of sequencing technology, analysis is no longer possible on standard laptops and workstations. A complete and comprehensive metagenomic analysis requires an extensive amount of computational resources and carefully selected tools running sucessively in the form of a pipeline, where each tool performs a distinct task necessary in a metagenomic analysis workflow.

The pipeline should:

1. Include necessary components for a thorough rigorous analysis, namely preprocessing, assembly, taxonomic classification and functional analysis.
2. Scale to contemporary datasets, efficiently utilizing distributed computer clusters for parallel computation.
3. Be publicly available as a resource for external users.
4. Undergo rigorous performance and quality evaluations through biological use cases using metagenomic sequence data.

3

Included papers

This chapter gives an overview and a short description of the included papers in this thesis and my own contribtuions to each paper.

3.1 Paper 1

Title	META-pipe – Pipeline annotation, analysis and visualization of marine metagenomic sequence data.
Authors	Espen Mikal Robertsen, Tim Kahlke, Inge Alexander Raknes, Edvard Pedersen, Erik Kjærner Semb, Martin Ernstsen, Lars Ailo Bongo, Nils Peder Willassen.
Description	This paper describes the biological context, design and implementation of the initial version of META-Pipe, and an experimental evaluation of the pipeline tools, the biological context and future work
Contribution	I developed the initial version of META-pipe based on the GePan framework, contributed in performance evaluation, intergration and redesign, and wrote the paper.
Publication date	14.04.2016
Publication venue	Manuscript, archived in arXiv.
Citation	[145] Robertsen, E.M., Kahlke, T., Raknes, I.A., Pedersen, E., Semb, E.K., Ernstsen, M., Bongo, L.A., Willassen, N.P.: Meta-pipe - pipeline annotation, analysis and visualization of marine metagenomic sequence data (2016) arXiv:1604.04103

3.2 Paper 2

Title	ELIXIR pilot action: Marine metagenomics – towards a domain specific set of sustainable services
Authors	Espen Mikal Robertsen, Hubert Denise, Alex Mitchell, Robert D. Finn, Lars Ailo Bongo, Nils Peder Willassen
Description	This paper describes the harmonization and interoperability evaluation of META-pipe and the EMG pipeline through comparison of analysis output, as well as gap analysis of available resources for the marine domain.
Contribution	I performed the comparison of biological results, wrote the paper and subsequently refined and enhanced the pipeline based on this evaluation
Publication date	23.01.2017
Publication venue	F1000Research
Citation	[135] Robertsen, E.M, Denise, H., Mitchell, A. et al. ELIXIR pilot action: Marine metagenomics – towards a domain specific set of sustainable services [version 1; referees: awaiting peer review]. F1000Research 2017, 6(ELIXIR):70 (doi: 10.12688/f1000research.10443.1)

3.3 Paper 3

Title	Automatic Contextual Data Curation – Applying Artificial Neural Nets to Taxonomic Classifications of Metagenomes
Authors	Espen Mikal Robertsen, Lars Ailo Bongo, Nils Peder Willassen.
Description	This paper describes a novel application for automatic suggestion of metadata for metagenomic samples in a user submission context
Contribution	I designed the experiment, implemented the necessary code, performed the analysis and wrote the paper.
Publication date	-
Publication venue	Manuscript submitted to BMC Bioinformatics (05.12.2016)
Citation	-

4

Results and Discussion

The presented work is based on the implementation and application of a metagenomic analysis pipeline coined META-pipe. The pipeline is developed at the Center for Bioinformatics at UiT - Arctic University of Norway through collaborative effort from several present and previous members of our team[1]. META-pipe represents a corner stone in deliverables to projects such as ELIXIR [146] and NeLS (Norwegian e-infrastructure for Life Science) [147] and is continually enhanced and maintained to fit the focus of our core interests. The initial period of this project was used to remodel and develop META-pipe from its previous genome centric analysis focus to a bonafide metagenomic pipeline. Subsequently, the pipeline was applied as an analysis tool for metagenomic data in two different use cases. This chapter is divided into two parts. Section 4.1.1 summarizes and discusses the development of META-pipe through all its iterations and discusses functionality, integration with a distributed computer cluster and caveats as presented in **Paper 1** and **Paper 2**. Section 4.2 discusses the results and outcomes of two specific uses cases:

1. A pilot study focused on analysis comparison and interoperability assessment with the EMP pipeline [44, 143] offered as a public analysis resource by EBI Metagenomics as presented in **Paper 2**.

1. Members involved in this collaboration through various contributions are: Tim Kahlke, Inge Alexander Raknes, Edvard Pedersen, Espen Mikal Robertsen, Giacomo Tartari, Aleksandr Agofonof, Erik Kjærner-Semb, Martin Ernstsen, Erik Hjerde, Lars Ailo Bongo and Nils Peder Willassen.

2. The prototype implementation of an automated metadata curation approach using artificial neural nets as presented in **Paper 3**

Both uses cases demonstrate the application of META-pipe on metagenomic data sets and evaluates its functionality in different contexts.

4.1 META-pipe

META-pipe is based on the now obsolete GePan framework [148] for annotation of complete genomes originally developed by Tim Kahlke. The initial motivation behind META-pipe was to implement a pipeline suited for bioprospecting from full-length novel genes. Hence, in contrast to many other publicly available metagenomic pipelines such as EMP and MG-RAST, META-pipe targets its analysis on assembled metagenomic contigs. We have slowly refined the focus of the pipeline towards the marine domain, and marine specific databases are under development to support this endeavor. META-pipe can be accessed through Galaxy [149] (described in section 4.1.3 for users with FEIDE login credentials, and as a standalone web-portal currently in development. The standalone web-portal is part of a complete reimplementation of META-pipe implemented in Scala (unpublished) with Apache Spark [150] as cluster-computing framework (described in section 4.1.4).

4.1.1 Development of META-pipe

Implementation of the first version of META-pipe consisted largely of evaluation and incorporation of new tools suited for metagenomic analysis, as well as addressing scalability challenges stemming from the sheer increase in average data set sizes. As an example, the UniProtKB [51] databases described in **Paper 1** have now outgrown their usability in a metagenomic sequence alignment context, simply because homology searches demand too much computational resources and are not feasible even on a large distributed system. As a result, UniProtKB was exchanged with the smaller UniRef50 clustered version, which offers less detailed functional annotation but only at a fraction of the runtime. Additionally, GePan consisted of some intermediate modules for splitting, handling and exporting of intermediate files which did not scale to 1000-fold increased dataset sizes, and had to be reimplemented to reduce unnecessary computation time overhead (discussed in **Paper 1**). These modules were not geared towards the 1000-fold increase in data size a metagenomic data set represents compared to a genome, and have been replaced. A metagenomic pipeline needs to scale with available computational resources to be able to handle the ever increasing data set sizes the field of metagenomics produces.

However, the hardest challenge to overcome was metagenomic assembly. The initial version utilized Mira [151], a sophisticated assembler meant for genomic data, which translates to poor metagenomic assemblies and extremely high memory usage pr. run (upwards of 120 GB for medium sized data sets). As a result, this tool was replaced with Ray Meta [102] to capitalize on its distributed assembly strategy via use of MPI (Message Passing Interface) between distributed nodes. However, as described in **Paper 1** this tool does not deliver what it promises in terms of scalability, which prompted an evaluation of several other candidates capable of distributed assembly. None of these delivered what they promised, further provoking the need for novel assemblers capable of handling metagenomic data in an efficient and timely manner. We settled on MEGAHIT [50], a non-distributed, but state-of-the-art assembly tool published in 2015 which through our experience produces high yield, high quality contigs using tolerable amounts of memory.

4.1.2 Overview of the current version

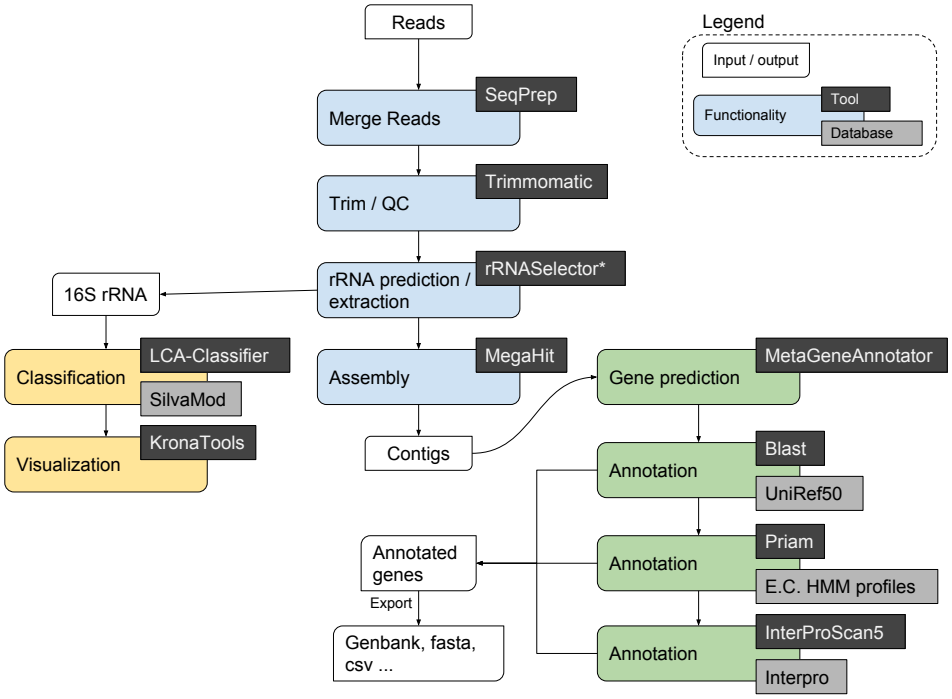


Figure 4.1: An overview of tools and databases currently included in META-pipe

The current version of the pipeline differs slightly from the version presented i **Paper 1**, and now consists of state-of-the-art tools for quality control, assembly, taxonomic classification and functional analysis 4.1. These differences

are the result of continuous evaluation and benchmarking of the performance of tools, both in context of computational resources and biological relevance and quality. The first step is preprocessing, which involves merging, trimming, rRNA prediction and assembly. With Illumina data sets, the first step is to merge overlapping paired-end reads to obtain a better basis for assembly (discussed in detail in section 4.2.1 and **Paper 2**). The trimming and quality check is performed by Trimmomatic [94], to trim 5'-ends and discard low quality reads. An in-house reimplementation of rRNASelector [116] is utilized to predict and trim rRNA fragments used in taxonomic classification, as the original version did not support programmatic submission in a pipeline context. 5S, 16S and 23S rRNA sequences are also removed prior to assembly to reduce the amount of chimeric and spurious contigs caused by these especially conserved sequences (discussed in detail in section 4.2.1 and **Paper 2**). Assembly is performed using MEGAHIT [50].

Predicted 16S rRNA sequences are queried towards the manually curated SilvaMod database (part of LCAClassifier) using megablast. This database is specially tailored for the classification tool LCAClassifier [107], which utilizes the Lowest Common Ancestor (LCA) algorithm to classify rRNA sequences and assign taxonomy. The output assignment is visualized using KronaTools [138], an interactive web-based pie chart that allows in-depth browsing and analysis of the LCAClassifier output. The taxonomic classification module of the pipeline has not been changed since its introduction, as it still performs exceptionally well compared to other alternatives as presented in **Paper 2**. However, we have made a necessary alteration to the functional analysis module. As discussed previously, the UniprotKB database has outgrown its use as a sequence alignment database, and is too costly to use. We changed it with UniRef50, a clustered reference protein sequence database only a fraction of the size of the complete UniprotKB, translating to faster computation times with only marginally less detailed annotation. We feel that this switch was necessary as metagenomic datasets will only continue to grow in the future. The gene prediction tool MetaGeneAnnotator [127], and the annotation tools PRIAM [133] and InterproScan [125] are still present in the current version.

The initial implementation of META-pipe supports embl, tsv and a METAREP-specific output formats, but is now deprecated (discussed in section 4.1.4). The recent reimplementation of META-pipe implemented in Scala (unpublished) outputs only genbank formatted files so far, however more output formats are planned to allow for more flexible post-processing by users. META-pipe in its current form offers no tools for comparison, however an in-house modified version of METAREP [152] offering extended functionality is locally available at our institute. This open sourced tool enables viewing, querying, browsing and comparing of metagenomic data sets, and has a dedicated output format in META-pipe. Our modified version grants access to blast-formatted databases on the backend, which allows effective retrieval and download of sequences of interest.

4.1.3 Galaxy and distributed computer cluster integration

META-pipe has been publicly available through the web-based analysis platform Galaxy for several years now, exclusively for users with Felles Elektronisk IDEntitet (FEIDE) login credentials or users associated with NeLS (Figure 4.2). This platform was chosen as a standard in the NeLS infrastructure and provides integration with external storage, sequencing and computational resources such as the supercomputer Stallo, localized at UiT - The Arctic University of Norway.

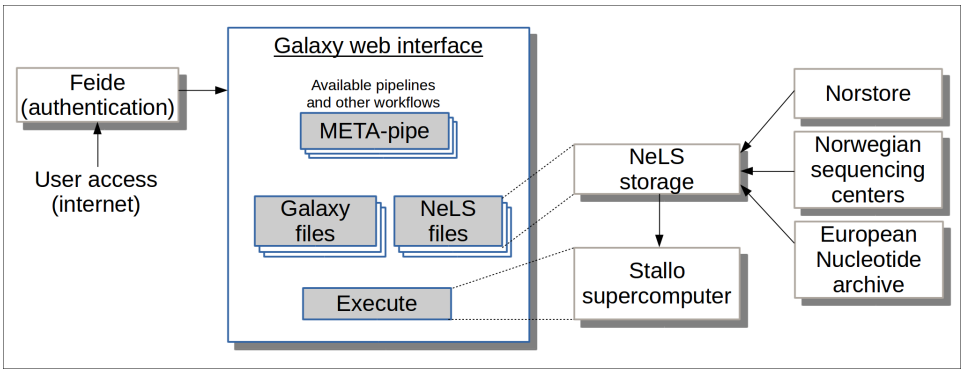


Figure 4.2: META-pipe integration with galaxy and associated storage, computation and sequencing resources as described in **Paper 1**.

However, this integration revealed several challenges and limitations of the initial implementation, and is the main reason we decided to fully re-implement the META-pipe work flow source code. The initial version of META-pipe implemented in Perl was designed for small homogeneous computer clusters where deployment assumed superuser access to the file system, specific queuing systems and direct submission as its own work flow manager. This non-flexible implementation presented several challenges in the integration process with Galaxy (presented in **Paper 1**). Firstly, Galaxy assumes a tool runs and finishes, which is reflected in its user interface by yellow and green boxes, respectively. The initial implementation used a wrapper script to submit bash-scripts to the queue, which did not fit with this scheme, and made it impossible to reflect whether a job was done or not. This caveat was fixed by an additional ad-hoc python script to check the status of jobs in the queue, basically wrapping the wrapper script, and adding unnecessary complexity and instability to the system. Secondly, Galaxy needs the Pulsar service (part of the Galaxy Project) to execute META-pipe tools in parallel on Stallo via a message broker (RabbitMQ [153]), a combination which has added additional instability. Providing the distributed assembly tool Ray Meta through Galaxy was also impossible for similar reasons.

The initial implementation also had some more native limitations and de-

sign flaws, regardless of Galaxy and Stallo integration. The way parallel jobs was submitted, it introduced unnecessary overhead in queue time as batches of tasks had to wait for the slowest member to finish before the next batch could start (depicted in Figure 4.3). The initial assumption was that the tasks within these batches would finish more or less simultaneously due to an equal split of input data, however this does not necessarily hold true. This became even more apparent on Stallo as nodes are shared between multiple users and utilized hardware is possibly non-homogeneous, which can cause straggler nodes (as shown in Figure 10 in **Paper 1**). This issue is addressed in the re-implemented version of META-pipe by submitting Spark workers to the Stallo queue, continuously executing parallel tasks as resources become available. The initial implementation also provided minimal amounts of logging, making debugging and troubleshooting especially time consuming.

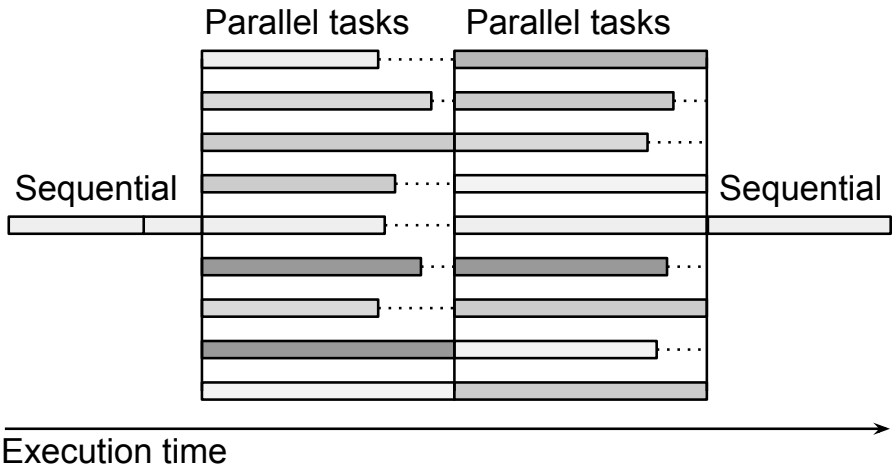


Figure 4.3: Depiction of parallel task implementation overhead. Idle CPU-time marked is in dashed lines.

4.1.4 Future work

The current version of the pipeline is a reimplementation in the functional programming language Scala. In addition to a complete redesign of the underlying framework, this version also provides a standalone web user interface separate from Galaxy, where users can upload and run their analyses. We felt this was necessary to circumvent the mentioned limitations and caveats of the initial version. This allows for added flexibility and less maintenance based on previous experiences, both in terms of the current production system and future integrations and deployment on external computing infrastructures. The initial version of META-pipe has been subject to deployment on the Embassy Cloud

[154] (part of the work performed in **Paper 2**, but not published), which proved a significant headache in terms of intergration. However, the redesigned version has been successfully integrated with the Finnish CSC (IT Center For Science) cPouta cloud infrastructure, using an automatic deployment script based on Ansible Playbooks [155]. To provide much needed resources for the marine domain, an effort to establish databases specific to the marine environment has been initiated through the ELIXIR infrastructure (discussed in **Paper 2**). In this context, META-pipe will provide the analysis of novel samples included in the MarCat database, a marine metagenomic sequence and meta data resource database, including an extensive collection of samples from global-scale projects such as Tara Oceans [36].

4.2 Use cases

4.2.1 Interoperability assessment with the EMP pipeline and pilot studies of marine datasets

Through the ELIXIR infrastructure, we launched a pilot project with the aim to investigate the need to establish dedicated data resources and harmonized metagenomics pipelines for the marine domain. This project enveloped several tasks, but the main focus was to assess and evaluate the EMP-pipeline and META-pipe together, to investigate possible points of interoperability and differences in analysis of data. This was a good opportunity to benchmark some of the included tools in META-pipe and provide an in-depth overview of strengths and weaknesses through comparison with EMP. Firstly, we analysed the biological results from four selected samples, two marine datasets from the Barents sea and two gut samples from Norwegian moose and sea urchin, produced by the two pipelines independently. Through comparison, the main findings can be attributed to two distinct causes, differences in methodology and differences in the general performance of the included tools. While META-pipe performs assembly of sequence data, EMP does analysis on raw reads. This fact causes significant differences in functional analysis. Through an in-depth analysis of functional analysis, we discovered that the functional fingerprint from the pipelines differ both quantitatively and qualitatively. We believe that these differences stem mainly from the fact that an assembly effectively sets a cut-off on the portion of organisms analysed based on relative abundance (Figure 4.4). Organisms with low relative abundance does not meet the coverage criteria for assembly into contigs with our 500 bp minimum contig cut-off, an are thus excluded from the functional analysis in META-pipe. Any genes present in these excluded organisms will not get predicted and annotated, causing a quantitative and qualitative difference in output as seen in Figure 5 in **Paper 2**. Comparing taxonomic classifications, we experienced that the results were

more comparable than the output from the functional analysis. However, the performance of tools vary depending on the origin of samples and the resolution at various taxonomic levels. For marine samples, META-pipe produces a richer classification and also includes eukaryotic organisms through classification of mitochondrial rRNA, but for gut and intestinal samples EMP is seemingly a better option.

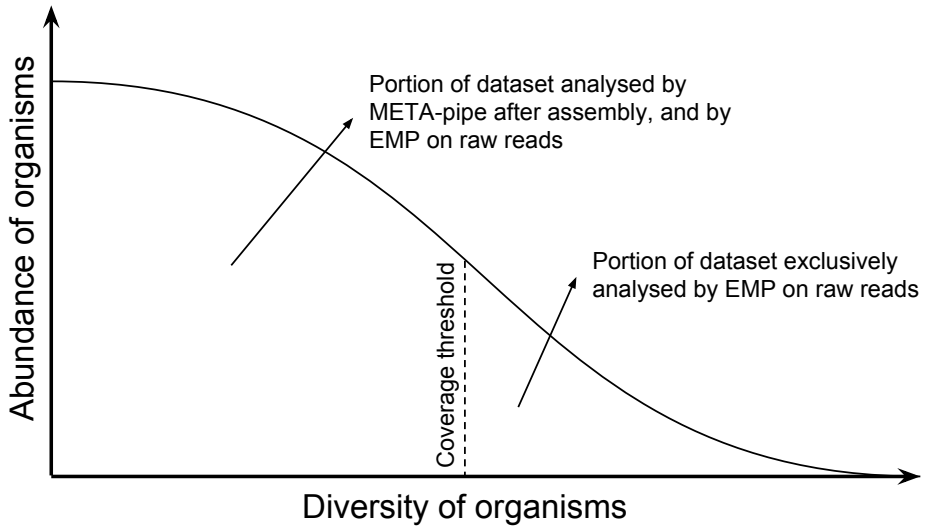


Figure 4.4: Given an unequal abundance distribution of strains in a sample, only strains with sufficient sequence information are assembled, effectively excluding parts of the functional fingerprint of a sample

Secondly, we decided to evaluate the performance of included tools in META-pipe based on the in-depth analysis of the biological results produced. Some of the subsequent changes to the pipeline were directly adopted from the EMP workflow, such as merging of paired-end reads. We believe that since roughly two thirds of reads overlap, merging them will result in a better basis for both assembly and rRNA prediction as input reads will be longer. We have already run some performance evaluations comparing merged and un-merged input to assembly using MEGAHIT (unpublished), showing that merging produces better assemblies with lower computation time and memory footprint. We also decided to adopt Trimmomatic as a quality filtering tool, as it confers to all our needs regarding functionality and performance. The final change to our pipeline based on this evaluation was the introduction of rRNA filtering before assembly. As discussed in **Paper 2** (Table 1), eliminating rRNA reduces mismatches in assembled contigs and produces a more correct basis for gene prediction and functional annotation.

The project also had a focus on gap analysis, an effort to investigate the actions needed to develop sustainable resources for the marine domain. One of the significant outcomes of this analysis was the initiative to establish databases

specific to the marine domain. In evaluation of metagenomic assemblies for the Barents sea samples in this project, we used an in-house marine reference database consisting of 337 manually curated strains. This reference database has now grown and is the first out of three marine specific databases, MarRef, MarDB and MarCat, respectively. These databases are now parts of a deliverable to the ELIXIR infrastructure as an independent project. While MarRef will serve as a complete marine reference database, MarDB will contain all marine strains publicly available and MarCat will serve as a gene catalogue, incorporating genetic information from substantial projects such as Tara Oceans and Global Ocean Sampling Expedition. With manually curated metadata and publicly available sequence data through blast databases, these resources will provide a much needed asset for marine research in years to come.

4.2.2 Automatic metadata curation using machine learning

As experienced through the establishment of the previously mentioned marine specific Mar-databases, manual curation of metadata is a tedious task and requires excessive amounts of man-hours to maintain. In this project (**Paper 3**), we applied artificial neural nets (ANN), a sub-division of machine learning in an effort to automate metadata submission for users. While curation of reference databases need manual attention to ensure the highest possible quality of metadata, user submitted metadata for novel metagenomic samples can be automated to a certain extent. When uploading metagenomic samples to publicly available analysis resources such as EMP and MG-RAST [45, 144], users are required to submit accommodating metadata and contextual data. From a user perspective, submitting metadata and contextual data compliant with for example the MIxS-standards is a dreary assignment, and causes some submitters to leave out vital sample information in the process. In an effort to alleviate this burden, and consequently enrich and quality-assure submitted metadata, we wanted to prototype a suggestive tool for meta data submission. The tool uses an ANN trained with datasets from multiple environments, making the model able to predict the source of origin for a new submitted sample based on its taxonomic classification. This way, we can make a tool that suggests possible metadata for users, which we believe lowers the threshold for submitting fulfilling and consistent information. Naturally, this would imply processing some of the sample before metadata submission as we need its taxonomic classification before being able to utilize this novel application. The standard operating procedure on portals such as EMP and MG-RAST is to submit metadata before processing of samples are begun, however since taxonomic classification requires relatively small amounts of processing, we do not see this as a problem.

Using publicly available datasets from MG-RAST we were able to acquire

88,4 % accuracy in prediction of source environment on novel samples from our prototype model. However, as discussed in **Paper 3**, our prototype has some limitations. The accompanying metadata downloaded from MG-RAST struggles with the same impediment as we ultimately want to fix, namely a lack of complete metadata annotation. As some of the training sets are wrongfully annotated, they will effectively lower the accuracy of our trained model. Also, training sets are not stratified, meaning there is an unequal distribution of training sets from different possible environments, as we wanted to see how this model would perform with no manual curation involved. The reasoning behind this was that once a user submits a novel sample with complete metadata annotation, it will automatically get included as a training set, effectively increasing accuracy as more data is added to the model. These caveats are easily fixed in a potential production system by manually curating and distributing training sets, a task that will require a lot less effort once the Mar-Databases are complete. With processed samples and metadata from the MarCat database (described in section 4.2.1) we will have a manually curated starting point to produce a stratified and completely annotated training set for metadata suggestion in a production system. This application is thought implemented in the META-pipe user portal to aid in metadata submission, offering an easier and less strenuous user experience.

4.3 Concluding remarks

Analysis of metagenomic sequence data represents a continuously escalating challenge in terms of computational resource utilization and implementation of efficient algorithms. With META-pipe, we have implemented an automatic analysis pipeline for metagenomic data with a focus on the marine domain and full-length genes. While other big actors such as EMP and MG-RAST serve as more generic alternatives, processing thousands of diverse samples each year, we opt to specialize our analysis with specific databases and tools suited for marine data sets. As presented in **Paper 1** and discussed in this thesis, development of an analysis resource such as META-pipe requires not only an initial implementation, but testing and refinement through well designed use cases with important biological and computational aspects in mind. The use cases described in **Paper 2** and **Paper 3** not only serve as a demonstration of pipeline functionality, but also suggest possible improvements to the pipeline through rigorous evaluation of biological results. The current version of META-pipe implements several improvements stemming from the work performed in **Paper 2** and the metadata user application discussed in **Paper 3** is due to be implemented in the new version of META-pipe soon. We are confident that the new version will address and fix the mentioned caveats and limitations described in **Paper 1**, and that future versions of META-pipe will continue to

improve and expand, both in functionality and public usage.

Bibliography

[1] S. Hamarneh, "Measuring the Invisible World. The life and works of Antoni van Leeuwenhoek. A. Schierbeek. Abelard-Schuman, New York, 1959. 223 pp. $4," *Science*, vol. 132, no. 3422, pp. 289–290, Jul. 1960. [Online]. Available: http://science.sciencemag.org/content/132/3422/289

[2] S. M. Blevins and M. S. Bronze, "Robert koch and the 'golden age' of bacteriology," vol. 14, no. 9, pp. e744–e751. [Online]. Available: /article/S1201-9712(10)02314-3/abstract

[3] T. J. Beveridge, "Use of the gram stain in microbiology," vol. 76, no. 3, pp. 111–118.

[4] M. McFall-Ngai, "Are biologists in 'future shock'? symbiosis integrates biology across domains," vol. 6, no. 10, pp. 789–792. [Online]. Available: http://www.nature.com/nrmicro/journal/v6/n10/full/nrmicro1982.html

[5] P. Hugenholtz, B. M. Goebel, and N. R. Pace, "Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity," vol. 180, no. 18, pp. 4765–4774. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC107498/

[6] V. Torsvik, J. Goksøyr, and F. L. Daae, "High diversity in DNA of soil bacteria." *Applied and Environmental Microbiology*, vol. 56, no. 3, pp. 782–787, Mar. 1990. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC183421/

[7] C. R. Woese and G. E. Fox, "Phylogenetic structure of the prokaryotic domain: the primary kingdoms," vol. 74, no. 11, pp. 5088–5090.

[8] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," vol. 74, no. 12, pp. 5463–5467.

[9] J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman, "Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products," *Chemistry & Biology*, vol. 5, no. 10, pp. R245–249, Oct. 1998.

[10] G. W. Tyson, J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield, "Community structure and metabolism through reconstruction of microbial genomes from the environment," *Nature*, vol. 428, no.

6978, pp. 37–43, Mar. 2004. [Online]. Available: http://www.nature.com/nature/journal/v428/n6978/full/nature02340.html

[11] J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y.-H. Rogers, and H. O. Smith, "Environmental Genome Shotgun Sequencing of the Sargasso Sea," *Science*, vol. 304, no. 5667, pp. 66–74, Apr. 2004. [Online]. Available: http://science.sciencemag.org/content/304/5667/66

[12] S. Lindgreen, K. L. Adair, and P. P. Gardner, "An evaluation of the accuracy and speed of metagenome analysis tools," *Scientific Reports*, vol. 6, p. 19233, Jan. 2016. [Online]. Available: http://www.nature.com/srep/2016/160118/srep19233/full/srep19233.html

[13] J. M. Walter, D. A. Tschoeke, P. M. Meirelles, L. d. Oliveira, L. Leomil, M. Tenório, R. Valle, P. S. Salomon, C. C. Thompson, and F. L. Thompson, "Taxonomic and Functional Metagenomic Signature of Turfs in the Abrolhos Reef System (Brazil)," *PLOS ONE*, vol. 11, no. 8, p. e0161168, Aug. 2016. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0161168

[14] N. Segata, J. Izard, L. Waldron, D. Gevers, L. Miropolsky, W. S. Garrett, and C. Huttenhower, "Metagenomic biomarker discovery and explanation," *Genome Biology*, vol. 12, no. 6, p. R60, Jun. 2011.

[15] R. Leinonen, R. Akhtar, E. Birney, L. Bower, A. Cerdeno-Tárraga, Y. Cheng, I. Cleland, N. Faruque, N. Goodgame, R. Gibson, G. Hoad, M. Jang, N. Pakseresht, S. Plaister, R. Radhakrishnan, K. Reddy, S. Sobhany, P. Ten Hoopen, R. Vaughan, V. Zalunin, and G. Cochrane, "The European Nucleotide Archive," *Nucleic Acids Research*, vol. 39, no. Database issue, pp. D28–D31, Jan. 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3013801/

[16] M. J. Gosalbes, A. Durbán, M. Pignatelli, J. J. Abellan, N. Jiménez-Hernández, A. E. Pérez-Cobas, A. Latorre, and A. Moya, "Metatranscriptomic approach to analyze the functional human gut microbiota," *PloS One*, vol. 6, no. 3, p. e17447, Mar. 2011.

[17] C. Pan and J. F. Banfield, "Quantitative metaproteomics: functional insights into microbial communities," *Methods in Molecular Biology (Clifton, N.J.)*, vol. 1096, pp. 231–240, 2014.

[18] A. C. Almeida, P. F. Azevedo, R. D. Marques, E. Mello-Peixoto, and L. S. Matsumoto, "Bioprospecting of cave bacteria with antifungal activity," *Planta Medica*, vol. 81, no. S 01, pp. S1–S381, Dec. 2016.

[19] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon, "The Human Microbiome Project," *Nature*, vol. 449, no. 7164, pp. 804–810, Oct. 2007. [Online]. Available: http://www.nature.com/nature/journal/v449/n7164/full/nature06244.html

[20] M. A. Carl, I. M. Ndao, A. C. Springman, S. D. Manning, J. R.

Johnson, B. D. Johnston, C.-A. D. Burnham, E. S. Weinstock, G. M. Weinstock, T. N. Wylie, M. Mitreva, S. Abubucker, Y. Zhou, H. J. Stevens, C. Hall-Moore, S. Julian, N. Shaikh, B. B. Warner, and P. I. Tarr, "Sepsis From the Gut: The Enteric Habitat of Bacteria That Cause Late-Onset Neonatal Bloodstream Infections," *Clinical Infectious Diseases*, p. ciu084, Mar. 2014. [Online]. Available: http://cid.oxfordjournals.org/content/early/2014/03/12/cid.ciu084

[21] J. D. Lewis, E. Z. Chen, R. N. Baldassano, A. R. Otley, A. M. Griffiths, D. Lee, K. Bittinger, A. Bailey, E. S. Friedman, C. Hoffmann, L. Albenberg, R. Sinha, C. Compher, E. Gilroy, L. Nessel, A. Grant, C. Chehoud, H. Li, G. D. Wu, and F. D. Bushman, "Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease," *Cell Host & Microbe*, vol. 18, no. 4, pp. 489–500, Oct. 2015.

[22] E. Pasolli, D. T. Truong, F. Malik, L. Waldron, and N. Segata, "Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights," *PLOS Computational Biology*, vol. 12, no. 7, p. e1004977, Jul. 2016. [Online]. Available: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004977

[23] H. K. Allen, L. A. Moe, J. Rodbumrer, A. Gaarder, and J. Handelsman, "Functional metagenomics reveals diverse beta-lactamases in a remote Alaskan soil," *The ISME journal*, vol. 3, no. 2, pp. 243–251, Feb. 2009.

[24] Z. Feng, D. Chakraborty, S. B. Dewell, B. V. B. Reddy, and S. F. Brady, "Environmental DNA-encoded antibiotics fasamycins A and B inhibit FabF in type II fatty acid biosynthesis," *Journal of the American Chemical Society*, vol. 134, no. 6, pp. 2981–2987, Feb. 2012.

[25] T. M. Alvarez, J. H. Paiva, D. M. Ruiz, J. P. L. F. Cairo, I. O. Pereira, D. A. A. Paixão, R. F. d. Almeida, C. C. C. Tonoli, R. Ruller, C. R. Santos, F. M. Squina, and M. T. Murakami, "Structure and Function of a Novel Cellulase 5 from Sugarcane Soil Metagenome," *PLOS ONE*, vol. 8, no. 12, p. e83635, Dec. 2013. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0083635

[26] H. Nacke, M. Engelhaupt, S. Brady, C. Fischer, J. Tautzt, and R. Daniel, "Identification and characterization of novel cellulolytic and hemicellulolytic genes and enzymes derived from German grassland soil metagenomes," *Biotechnology Letters*, vol. 34, no. 4, pp. 663–675, Apr. 2012.

[27] F. Cheng, J. Sheng, R. Dong, Y. Men, L. Gan, and L. Shen, "Novel xylanase from a holstein cattle rumen metagenomic library and its application in xylooligosaccharide and ferulic Acid production from wheat straw," *Journal of Agricultural and Food Chemistry*, vol. 60, no. 51, pp. 12 516–12 524, Dec. 2012.

[28] Y. S. Jeong, H. B. Na, S. K. Kim, Y. H. Kim, E. J. Kwon, J. Kim, H. D. Yun, J.-K. Lee, and H. Kim, "Characterization of xyn10j, a novel family 10 xylanase from a compost metagenomic library," *Applied Biochemistry*

and Biotechnology, vol. 166, no. 5, pp. 1328–1339, Mar. 2012.

[29] J. Kennedy, N. D. O'Leary, G. S. Kiran, J. P. Morrissey, F. O'Gara, J. Selvin, and A. D. W. Dobson, "Functional metagenomic strategies for the discovery of novel enzymes and biosurfactants with biotechnological applications from marine ecosystems," *Journal of Applied Microbiology*, vol. 111, no. 4, pp. 787–799, Oct. 2011.

[30] M. Pacwa-Płociniczak, G. A. Płaza, Z. Piotrowska-Seget, and S. S. Cameotra, "Environmental applications of biosurfactants: recent advances," *International Journal of Molecular Sciences*, vol. 12, no. 1, pp. 633–654, Jan. 2011.

[31] A. Y. Burch, B. K. Shimada, P. J. Browne, and S. E. Lindow, "Novel High-Throughput Detection Method To Assess Bacterial Surfactant Production," *Applied and Environmental Microbiology*, vol. 76, no. 16, pp. 5363–5372, Aug. 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2918974/

[32] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A.

Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, and International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, Feb. 2001.

[33] D. B. Rusch, A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y.-H. Rogers, L. I. Falcón, V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Nealson, R. Friedman, M. Frazier, and J. C. Venter, "The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific," *PLOS Biology*, vol. 5, no. 3, p. e77, Mar. 2007. [Online]. Available: http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0050077

[34] S. Yooseph, G. Sutton, D. B. Rusch, A. L. Halpern, S. J. Williamson, K. Remington, J. A. Eisen, K. B. Heidelberg, G. Manning, W. Li, L. Jaroszewski, P. Cieplak, C. S. Miller, H. Li, S. T. Mashiyama, M. P. Joachimiak, C. v. Belle, J.-M. Chandonia, D. A. Soergel, Y. Zhai, K. Natarajan, S. Lee, B. J. Raphael, V. Bafna, R. Friedman, S. E. Brenner, A. Godzik, D. Eisenberg, J. E. Dixon, S. S. Taylor, R. L. Strausberg, M. Frazier, and J. C. Venter, "The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families," *PLOS Biology*, vol. 5, no. 3, p. e16, Mar. 2007. [Online]. Available: http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0050016

[35] MetaHIT Consortium, "Metagenomics of the Human Intestinal Tract (MetaHIT) Consortium," 2008. [Online]. Available: http://www.metahit.eu/

[36] E. Karsenti, S. G. Acinas, P. Bork, C. Bowler, C. D. Vargas, J. Raes, M. Sullivan, D. Arendt, F. Benzoni, J.-M. Claverie, M. Follows, G. Gorsky,

P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels-Lewis, U. Krzic, F. Not, H. Ogata, S. Pesant, E. G. Reynaud, C. Sardet, M. E. Sieracki, S. Speich, D. Velayoudon, J. Weissenbach, P. Wincker, and t. T. O. Consortium, "A Holistic Approach to Marine Eco-Systems Biology," *PLOS Biology*, vol. 9, no. 10, p. e1001177, Oct. 2011. [Online]. Available: http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001177

[37] State Agency Superior Council for Scientific Research, "The Malaspina 2010 Global Circumnavigation Expedition." Aug. 2016. [Online]. Available: http://scientific.expedicionmalaspina.es/

[38] T. M. Vogel, "Metagenomic discovery and exploitation of the soil microbial community," 2009. [Online]. Available: http://www.genomenviron.org/Projects/METASOIL.html

[39] J. Jansson, "Great Prairie Soil Metagenome Grand Challenge," 2009. [Online]. Available: http://genome.jgi.doe.gov/GrePraGChallenge_2/GrePraGChallenge_2.info.html

[40] O. Lakhdari, A. Cultrone, J. Tap, K. Gloux, F. Bernard, S. D. Ehrlich, F. Lefevre, J. Dore, and H. M. Blottiere, "Functional metagenomics: a high throughput screening method to decipher microbiota-driven NF-kB modulation in the human gut," *PloS One*, vol. 5, no. 9, Sep. 2010.

[41] M. Arumugam, E. D. Harrington, K. U. Foerstner, J. Raes, and P. Bork, "SmashCommunity: a metagenomic annotation and analysis tool," *Bioinformatics (Oxford, England)*, vol. 26, no. 23, pp. 2977–2978, Dec. 2010.

[42] C. Walter, "Kryder's Law," *Scientific American*, vol. 293, no. 2, pp. 32–33, Aug. 2005. [Online]. Available: http://www.nature.com/scientificamerican/journal/v293/n2/full/scientificamerican0805-32.html

[43] L. D. Stein, "The case for cloud computing in genome informatics," *Genome Biology*, vol. 11, p. 207, 2010. [Online]. Available: http://dx.doi.org/10.1186/gb-2010-11-5-207

[44] S. Hunter, M. Corbett, H. Denise, M. Fraser, A. Gonzalez-Beltran, C. Hunter, P. Jones, R. Leinonen, C. McAnulla, E. Maguire, J. Maslen, A. Mitchell, G. Nuka, A. Oisel, S. Pesseat, R. Radhakrishnan, P. Rocca-Serra, M. Scheremetjew, P. Sterk, D. Vaughan, G. Cochrane, D. Field, and S.-A. Sansone, "EBI metagenomics—a new resource for the analysis and archiving of metagenomic data," *Nucleic Acids Research*, p. gkt961, Oct. 2013. [Online]. Available: http://nar.oxfordjournals.org/content/early/2013/10/26/nar.gkt961

[45] F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. Edwards, "The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes," *BMC Bioinformatics*, vol. 9, p. 386, 2008. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-9-386

[46] M. Baker, "Next-generation sequencing: adjusting to data overload,"

Nature Methods, vol. 7, no. 7, pp. 495–499, Jul. 2010. [Online]. Available: http://www.nature.com/nmeth/journal/v7/n7/full/nmeth0710-495.html

[47] J. E. Stajich, D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fuellen, J. G. Gilbert, I. Korf, H. Lapp, H. Lehväslaiho, C. Matsalla, C. J. Mungall, B. I. Osborne, M. R. Pocock, P. Schattner, M. Senger, L. D. Stein, E. Stupka, M. D. Wilkinson, and E. Birney, "The Bioperl Toolkit: Perl Modules for the Life Sciences," *Genome Research*, vol. 12, no. 10, pp. 1611–1618, Oct. 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC187536/

[48] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon, "Biopython: freely available Python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, Jun. 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2682512/

[49] S. Anders, P. T. Pyl, and W. Huber, "HTSeq – A Python framework to work with high-throughput sequencing data," *Bioinformatics*, p. btu638, Sep. 2014. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/early/2014/09/25/bioinformatics.btu638

[50] D. Li, C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam, "MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph," *Bioinformatics*, p. btv033, Jan. 2015. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/early/2015/01/19/bioinformatics.btv033

[51] T. U. Consortium, "UniProt: a hub for protein information," *Nucleic Acids Research*, p. gku989, Oct. 2014. [Online]. Available: http://nar.oxfordjournals.org/content/early/2014/10/27/nar.gku989

[52] C. Simon and R. Daniel, "Achievements and new knowledge unraveled by metagenomic approaches," *Applied Microbiology and Biotechnology*, vol. 85, no. 2, pp. 265–276, Nov. 2009. [Online]. Available: http://link.springer.com/article/10.1007/s00253-009-2233-z

[53] K. Mineta and T. Gojobori, "Databases of the marine metagenomics," *Gene*, vol. 576, no. 2 Pt 1, pp. 724–728, Feb. 2016.

[54] R. Knight, J. Jansson, D. Field, N. Fierer, N. Desai, J. A. Fuhrman, P. Hugenholtz, D. van der Lelie, F. Meyer, R. Stevens, M. J. Bailey, J. I. Gordon, G. A. Kowalchuk, and J. A. Gilbert, "Unlocking the potential of metagenomics through replicated experimental design," *Nature biotechnology*, vol. 30, no. 6, pp. 513–520, Jun. 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4902277/

[55] J. I. Prosser, "Replicate or lie," *Environmental Microbiology*, vol. 12, no. 7, pp. 1806–1810, Jul. 2010. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1462-2920.2010.02201.x/abstract

[56] J. R. White, N. Nagarajan, and M. Pop, "Statistical Methods for

Detecting Differentially Abundant Features in Clinical Metagenomic Samples," *PLOS Computational Biology*, vol. 5, no. 4, p. e1000352, Apr. 2009. [Online]. Available: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000352

[57] J. A. Bryant, F. J. Stewart, J. M. Eppley, and E. F. DeLong, "Microbial community phylogenetic and trait diversity declines with depth in a marine oxygen minimum zone," *Ecology*, vol. 93, no. 7, pp. 1659–1673, Jul. 2012.

[58] J. Xiong, Y. Liu, X. Lin, H. Zhang, J. Zeng, J. Hou, Y. Yang, T. Yao, R. Knight, and H. Chu, "Geographic distance and pH drive bacterial distribution in alkaline lake sediments across Tibetan Plateau," *Environmental Microbiology*, vol. 14, no. 9, pp. 2457–2466, Sep. 2012.

[59] A. Garcia-Moyano, E. Gonzalez-Toril, A. Aguilera, and R. Amils, "Comparative microbial ecology study of the sediments and the water column of the Río Tinto, an extreme acidic environment," *FEMS microbiology ecology*, vol. 81, no. 2, pp. 303–314, Aug. 2012.

[60] D. A. Pearce, K. K. Newsham, M. A. S. Thorne, L. Calvo-Bado, M. Krsek, P. Laskaris, A. Hodson, and E. M. Wellington, "Metagenomic analysis of a southern maritime antarctic soil," *Frontiers in Microbiology*, vol. 3, p. 403, 2012.

[61] M. A. Bradford, C. A. Davies, S. D. Frey, T. R. Maddox, J. M. Melillo, J. E. Mohan, J. F. Reynolds, K. K. Treseder, and M. D. Wallenstein, "Thermal adaptation of soil microbial respiration to elevated temperature," *Ecology Letters*, vol. 11, no. 12, pp. 1316–1327, Dec. 2008.

[62] T. O. Delmont, P. Robe, I. Clark, P. Simonet, and T. M. Vogel, "Metagenomic comparison of direct and indirect soil DNA extraction approaches," *Journal of Microbiological Methods*, vol. 86, no. 3, pp. 397–400, Sep. 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167701211002351

[63] J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y.-H. Rogers, and H. O. Smith, "Environmental Genome Shotgun Sequencing of the Sargasso Sea," *Science*, vol. 304, no. 5667, pp. 66–74, Apr. 2004. [Online]. Available: http://science.sciencemag.org/content/304/5667/66

[64] A. Felczykowska, A. Krajewska, S. Zieliska, and J. M. Ło, "Sampling, metadata and DNA extraction - important steps in metagenomic studies," *Acta Biochimica Polonica*, vol. 62, no. 1, pp. 151–160, 2015.

[65] T. Thomas, D. Rusch, M. Z. DeMaere, P. Y. Yung, M. Lewis, A. Halpern, K. B. Heidelberg, S. Egan, P. D. Steinberg, and S. Kjelleberg, "Functional genomic signatures of sponge bacteria reveal unique and shared features of symbiosis," *The ISME Journal*, vol. 4, no. 12, pp. 1557–1567, Dec. 2010. [Online]. Available: http:

//www.nature.com/ismej/journal/v4/n12/full/ismej201074a.html

[66] C. Burke, S. Kjelleberg, and T. Thomas, "Selective Extraction of Bacterial DNA from the Surfaces of Macroalgae," *Applied and Environmental Microbiology*, vol. 75, no. 1, pp. 252–256, Jan. 2009. [Online]. Available: http://aem.asm.org/content/75/1/252

[67] Y. Hu, Z. Liu, J. Yan, X. Qi, J. Li, S. Zhong, J. Yu, and Q. Liu, "A developed DNA extraction method for different soil samples," *Journal of Basic Microbiology*, vol. 50, no. 4, pp. 401–407, Aug. 2010.

[68] C. Carrigg, O. Rice, S. Kavanagh, G. Collins, and V. O'Flaherty, "DNA extraction method affects microbial community profiles from soils and sediment," *Applied Microbiology and Biotechnology*, vol. 77, no. 4, pp. 955–964, Dec. 2007. [Online]. Available: http://link.springer.com/article/10.1007/s00253-007-1219-y

[69] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, "GenBank," *Nucleic Acids Research*, vol. 38, no. Database issue, pp. D46–51, Jan. 2010.

[70] E. Kaminuma, J. Mashima, Y. Kodama, T. Gojobori, O. Ogasawara, K. Okubo, T. Takagi, and Y. Nakamura, "DDBJ launches a new archive database with analytical tools for next-generation sequence data," *Nucleic Acids Research*, vol. 38, no. Database issue, pp. D33–D38, Jan. 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2808917/

[71] P. L. Buttigieg, N. Morrison, B. Smith, C. J. Mungall, and S. E. Lewis, "The environment ontology: contextualising biological and biomedical entities," *Journal of Biomedical Semantics*, vol. 4, p. 43, Dec. 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3904460/

[72] R. Kottmann, T. Gray, S. Murphy, L. Kagan, S. Kravitz, T. Lombardot, D. Field, F. O. Glöckner, and Genomic Standards Consortium, "A standard MIGS/MIMS compliant XML Schema: toward the development of the Genomic Contextual Data Markup Language (GCDML)," *Omics: A Journal of Integrative Biology*, vol. 12, no. 2, pp. 115–121, Jun. 2008.

[73] F. Sanger and A. R. Coulson, "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase," *Journal of Molecular Biology*, vol. 94, no. 3, pp. 441–448, May 1975. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0022283675902132

[74] J. M. Rothberg and J. H. Leamon, "The development and impact of 454 sequencing," *Nature Biotechnology*, vol. 26, no. 10, pp. 1117–1124, Oct. 2008.

[75] K. Nakamura, T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa, Y. Shiwa, S. Ishikawa, M. C. Linak, A. Hirai, H. Takahashi, M. Altaf-Ul-Amin, N. Ogasawara, and S. Kanaya, "Sequence-specific error profile of Illumina sequencers," *Nucleic Acids Research*, vol. 39, no. 13, p. e90, Jul. 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/

PMC3141275/

[76] J. O. Korbel, A. E. Urban, J. P. Affourtit, B. Godwin, F. Grubert, J. F. Simons, P. M. Kim, D. Palejev, N. J. Carriero, L. Du, B. E. Taillon, Z. Chen, A. Tanzer, A. C. E. Saunders, J. Chi, F. Yang, N. P. Carter, M. E. Hurles, S. M. Weissman, T. T. Harkins, M. B. Gerstein, M. Egholm, and M. Snyder, "Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome," *Science (New York, N.Y.)*, vol. 318, no. 5849, pp. 420–426, Oct. 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2674581/

[77] M. P. Dempsey, J. Nietfeldt, J. Ravel, S. Hinrichs, R. Crawford, and A. K. Benson, "Paired-End Sequence Mapping Detects Extensive Genomic Rearrangement and Translocation during Divergence of Francisella tularensis subsp. tularensis and Francisella tularensis subsp. holarctica Populations," *Journal of Bacteriology*, vol. 188, no. 16, pp. 5904–5914, Aug. 2006. [Online]. Available: http://jb.asm.org/content/188/16/5904

[78] J. A. Chapman, I. Ho, S. Sunkara, S. Luo, G. P. Schroth, and D. S. Rokhsar, "Meraculous: De Novo Genome Assembly with Short Paired-End Reads," *PLOS ONE*, vol. 6, no. 8, p. e23501, Aug. 2011. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0023501

[79] J. A. Frank, Y. Pan, A. Tooming-Klunderud, V. G. H. Eijsink, A. C. McHardy, A. J. Nederbragt, and P. B. Pope, "Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data," *Scientific Reports*, vol. 6, p. 25373, May 2016. [Online]. Available: http://www.nature.com/srep/2016/160509/srep25373/full/srep25373.html

[80] M. Jain, H. E. Olsen, B. Paten, and M. Akeson, "The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community," *Genome Biology*, vol. 17, p. 239, 2016. [Online]. Available: http://dx.doi.org/10.1186/s13059-016-1103-0

[81] A. Klindworth, E. Pruesse, T. Schweer, J. Peplies, C. Quast, M. Horn, and F. O. Glöckner, "Evaluation of general 16s ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies," *Nucleic Acids Research*, vol. 41, no. 1, pp. e1–e1, Jan. 2013. [Online]. Available: http://nar.oxfordjournals.org/content/41/1/e1

[82] S. Anderson, "Shotgun DNA sequencing using cloned DNase I-generated fragments." *Nucleic Acids Research*, vol. 9, no. 13, pp. 3015–3027, Jul. 1981. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC327328/

[83] R. Staden, "A strategy of DNA sequencing employing computer programs." *Nucleic Acids Research*, vol. 6, no. 7, pp. 2601–2610, Jun. 1979. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC327874/

[84] D. R. Mende, A. S. Waller, S. Sunagawa, A. I. Järvelin, M. M. Chan, M. Arumugam, J. Raes, and P. Bork, "Assessment of metagenomic assembly using simulated next generation sequencing data," *PloS One*, vol. 7,

no. 2, p. e31386, 2012.

[85] P. Chouvarine, L. Wiehlmann, P. M. Losada, D. S. DeLuca, and B. Tümmler, "Filtration and Normalization of Sequencing Read Data in Whole-Metagenome Shotgun Samples," *PLOS ONE*, vol. 11, no. 10, p. e0165015, Oct. 2016. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0165015

[86] C. Luo, D. Tsementzi, N. Kyrpides, T. Read, and K. T. Konstantinidis, "Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample," *PLOS ONE*, vol. 7, no. 2, p. e30087, Feb. 2012. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0030087

[87] N. A. Bokulich, S. Subramanian, J. J. Faith, D. Gevers, J. I. Gordon, R. Knight, D. A. Mills, and J. G. Caporaso, "Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing," *Nature Methods*, vol. 10, no. 1, pp. 57–59, Jan. 2013. [Online]. Available: http://www.nature.com/nmeth/journal/v10/n1/full/nmeth.2276.html

[88] C. D. Fabbro, S. Scalabrin, M. Morgante, and F. M. Giorgi, "An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis," *PLOS ONE*, vol. 8, no. 12, p. e85024, Dec. 2013. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0085024

[89] C. Quince, A. Lanzen, R. J. Davenport, and P. J. Turnbaugh, "Removing noise from pyrosequenced amplicons," *BMC bioinformatics*, vol. 12, p. 38, Jan. 2011.

[90] R. Schmieder and R. Edwards, "Fast identification and removal of sequence contamination from genomic and metagenomic datasets," *PloS One*, vol. 6, no. 3, p. e17288, Mar. 2011.

[91] S. Andrews, "FastQC A Quality Control tool for High Throughput Sequence Data," Aug. 2016. [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[92] H. Lab, "FASTX Toolkit," 2009. [Online]. Available: http://hannonlab.cshl.edu/fastx_toolkit/

[93] R. Schmieder and R. Edwards, "Quality control and preprocessing of metagenomic datasets," *Bioinformatics*, p. btr026, Jan. 2011. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/early/2011/01/27/bioinformatics.btr026

[94] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, Aug. 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103590/

[95] M. G. Bakker, Z. J. Tu, J. M. Bradeen, and L. L. Kinkel, "Implications of pyrosequencing error correction for biological data interpretation," *PloS One*, vol. 7, no. 8, p. e44357, 2012.

[96] K. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P. Rice, M. A. Rajandream, and B. Barrell, "Artemis: sequence visualization and annotation,"

Bioinformatics (Oxford, England), vol. 16, no. 10, pp. 944–945, Oct. 2000.

[97] G. Denisov, B. Walenz, A. L. Halpern, J. Miller, N. Axelrod, S. Levy, and G. Sutton, "Consensus Generation and Variant Detection by Celera Assembler," *Bioinformatics*, vol. 24, no. 8, pp. 1035–1040, Apr. 2008. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btn074

[98] Y. Peng, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin, "Meta-IDBA: a de Novo assembler for metagenomic data," *Bioinformatics (Oxford, England)*, vol. 27, no. 13, pp. i94–101, Jul. 2011.

[99] T. J. Treangen, S. Koren, D. D. Sommer, B. Liu, I. Astrovskaya, B. Ondov, A. E. Darling, A. M. Phillippy, and M. Pop, "MetAMOS: a modular and open source metagenomic assembly and analysis pipeline," *Genome Biology*, vol. 14, no. 1, p. R2, Jan. 2013.

[100] A. Mikheenko, V. Saveliev, and A. Gurevich, "MetaQUAST: evaluation of metagenome assemblies," *Bioinformatics (Oxford, England)*, vol. 32, no. 7, pp. 1088–1090, Apr. 2016.

[101] T. Namiki, T. Hachiya, H. Tanaka, and Y. Sakakibara, "MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads," *Nucleic Acids Research*, vol. 40, no. 20, p. e155, Nov. 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3488206/

[102] S. Boisvert, F. Raymond, E. Godzaridis, F. Laviolette, and J. Corbeil, "Ray Meta: scalable de novo metagenome assembly and profiling," *Genome Biology*, vol. 13, p. R122, 2012. [Online]. Available: http://dx.doi.org/10.1186/gb-2012-13-12-r122

[103] G. G. Sutton, O. White, M. D. Adams, and A. R. Kerlavage, "TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects," *Genome Science and Technology*, vol. 1, no. 1, pp. 9–19, Jan. 1995. [Online]. Available: http://online.liebertpub.com/doi/abs/10.1089/gst.1995.1.9

[104] K. Mavromatis, N. Ivanova, K. Barry, H. Shapiro, E. Goltsman, A. C. McHardy, I. Rigoutsos, A. Salamov, F. Korzeniewski, M. Land, A. Lapidus, I. Grigoriev, P. Richardson, P. Hugenholtz, and N. C. Kyrpides, "Use of simulated data sets to evaluate the fidelity of metagenomic processing methods," vol. 4, no. 6, pp. 495–500. [Online]. Available: http://www.nature.com/nmeth/journal/v4/n6/full/nmeth1043.html

[105] P. Medvedev, K. Georgiou, G. Myers, and M. Brudno, "Computability of Models for Sequence Assembly," in *Algorithms in Bioinformatics*. Springer, Berlin, Heidelberg, Sep. 2007, pp. 289–301, dOI: 10.1007/978-3-540-74126-8_27. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-540-74126-8_27

[106] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D.

Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu, "The sequence of the human genome," *Science (New York, N.Y.)*, vol. 291, no. 5507, pp. 1304–1351, Feb. 2001.

[107] A. Lanzén, S. L. Jørgensen, D. H. Huson, M. Gorfer, S. H. Grindhaug, I. Jonassen, L. Øvreås, and T. Urich, "CREST – Classification Resources for Environmental Sequence Tags," *PLOS ONE*, vol. 7, no. 11, p. e49334,

Nov. 2012. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0049334

[108] T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen, "Greengenes, a chimera-checked 16s rRNA gene database and workbench compatible with ARB," *Applied and Environmental Microbiology*, vol. 72, no. 7, pp. 5069–5072, Jul. 2006.

[109] N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower, "Metagenomic microbial community profiling using unique clade-specific marker genes," *Nature Methods*, vol. 9, no. 8, pp. 811–814, Aug. 2012. [Online]. Available: http://www.nature.com/nmeth/journal/v9/n8/full/nmeth.2066.html

[110] A. C. McHardy, H. G. Martín, A. Tsirigos, P. Hugenholtz, and I. Rigoutsos, "Accurate phylogenetic classification of variable-length DNA fragments," *Nature Methods*, vol. 4, no. 1, pp. 63–72, Jan. 2007.

[111] A. Brady and S. L. Salzberg, "Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models," *Nature Methods*, vol. 6, no. 9, pp. 673–676, Sep. 2009. [Online]. Available: http://www.nature.com/nmeth/journal/v6/n9/full/nmeth.1358.html

[112] A. Brady and S. Salzberg, "PhymmBL expanded: confidence scores, custom databases, parallelization and more," *Nature Methods*, vol. 8, no. 5, pp. 367–367, May 2011. [Online]. Available: http://www.nature.com/nmeth/journal/v8/n5/full/nmeth0511-367.html

[113] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight, "QIIME allows analysis of high-throughput community sequencing data," *Nature Methods*, vol. 7, no. 5, pp. 335–336, May 2010. [Online]. Available: http://www.nature.com/nmeth/journal/v7/n5/full/nmeth.f.303.html

[114] E. Kopylova, L. Noé, and H. Touzet, "SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data," *Bioinformatics (Oxford, England)*, vol. 28, no. 24, pp. 3211–3217, Dec. 2012.

[115] B. L. Maidak, G. J. Olsen, N. Larsen, R. Overbeek, M. J. McCaughey, and C. R. Woese, "The Ribosomal Database Project (RDP)," *Nucleic Acids Research*, vol. 24, no. 1, pp. 82–85, Jan. 1996. [Online]. Available: http://nar.oxfordjournals.org/content/24/1/82

[116] J.-H. Lee, H. Yi, and J. Chun, "rRNASelector: a computer program for selecting ribosomal RNA encoding sequences from metagenomic and metatranscriptomic shotgun libraries," *Journal of Microbiology (Seoul, Korea)*, vol. 49, no. 4, pp. 689–691, Aug. 2011.

[117] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies,

and F. O. Glöckner, "The SILVA ribosomal RNA gene database project: improved data processing and web-based tools," *Nucleic Acids Research*, vol. 41, no. Database issue, pp. D590–596, Jan. 2013.

[118] D. Wu, P. Hugenholtz, K. Mavromatis, R. Pukall, E. Dalin, N. N. Ivanova, V. Kunin, L. Goodwin, M. Wu, B. J. Tindall, S. D. Hooper, A. Pati, A. Lykidis, S. Spring, I. J. Anderson, P. D'haeseleer, A. Zemla, M. Singer, A. Lapidus, M. Nolan, A. Copeland, C. Han, F. Chen, J.-F. Cheng, S. Lucas, C. Kerfeld, E. Lang, S. Gronow, P. Chain, D. Bruce, E. M. Rubin, N. C. Kyrpides, H.-P. Klenk, and J. A. Eisen, "A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea," *Nature*, vol. 462, no. 7276, pp. 1056–1060, Dec. 2009. [Online]. Available: http://www.nature.com/nature/journal/v462/n7276/full/nature08656.html

[119] T. J. Sharpton, "An introduction to the analysis of shotgun metagenomic data," *Frontiers in Plant Science*, vol. 5, 2014. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fpls.2014.00209/abstract

[120] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, Oct. 1990.

[121] M. Rho, H. Tang, and Y. Ye, "FragGeneScan: predicting genes in short and error-prone reads," *Nucleic Acids Research*, vol. 38, no. 20, p. e191, Nov. 2010.

[122] D. K. Desai, H. Schunck, J. W. Löser, and J. LaRoche, "Fragment Recruitment on Metabolic Pathways (FROMP): Comparative metabolic profiling of metagenomes and metatranscriptomes," *Bioinformatics*, p. bts721, Jan. 2013. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/early/2013/01/09/bioinformatics.bts721

[123] S. R. Eddy, "Profile hidden Markov models." *Bioinformatics*, vol. 14, no. 9, pp. 755–763, Jan. 1998. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/14/9/755

[124] A. Mitchell, H.-Y. Chang, L. Daugherty, M. Fraser, S. Hunter, R. Lopez, C. McAnulla, C. McMenamin, G. Nuka, S. Pesseat, A. Sangrador-Vegas, M. Scheremetjew, C. Rato, S.-Y. Yong, A. Bateman, M. Punta, T. K. Attwood, C. J. A. Sigrist, N. Redaschi, C. Rivoire, I. Xenarios, D. Kahn, D. Guyot, P. Bork, I. Letunic, J. Gough, M. Oates, D. Haft, H. Huang, D. A. Natale, C. H. Wu, C. Orengo, I. Sillitoe, H. Mi, P. D. Thomas, and R. D. Finn, "The InterPro protein families database: the classification resource after 15 years," *Nucleic Acids Research*, p. gku1243, Nov. 2014. [Online]. Available: http://nar.oxfordjournals.org/content/early/2014/11/26/nar.gku1243

[125] P. Jones, D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S.-Y. Yong, R. Lopez, and S. Hunter, "InterProScan 5: genome-scale protein function classification," *Bioinformatics (Oxford, England)*, vol. 30, no. 9, pp. 1236–1240, May 2014.

[126] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 27, no. 1, pp. 29–34, Jan. 1999. [Online]. Available: http://dx.doi.org/10.1093/nar/27.1.29

[127] H. Noguchi, T. Taniguchi, and T. Itoh, "MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes," *DNA research: an international journal for rapid publication of reports on genes and genomes*, vol. 15, no. 6, pp. 387–396, Dec. 2008.

[128] B. Liu and M. Pop, "MetaPath: identifying differentially abundant metabolic pathways in metagenomic datasets," *BMC Proceedings*, vol. 5, no. 2, p. S9, 2011. [Online]. Available: http://dx.doi.org/10.1186/1753-6561-5-S2-S9

[129] "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 44, no. Database issue, pp. D7–D19, Jan. 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702911/

[130] K. J. Hoff, T. Lingner, P. Meinicke, and M. Tech, "Orphelia: predicting genes in metagenomic sequencing reads," *Nucleic Acids Research*, vol. 37, no. Web Server issue, pp. W101–105, Jul. 2009.

[131] A. Bateman, E. Birney, R. Durbin, S. R. Eddy, K. L. Howe, and E. L. L. Sonnhammer, "The Pfam Protein Families Database," *Nucleic Acids Research*, vol. 28, no. 1, pp. 263–266, Jan. 2000. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC102420/

[132] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, and A. Bateman, "The Pfam protein families database: towards a more sustainable future," *Nucleic Acids Research*, vol. 44, no. D1, pp. D279–D285, Jan. 2016. [Online]. Available: http://nar.oxfordjournals.org/content/44/D1/D279

[133] C. Claudel-Renard, C. Chevalet, T. Faraut, and D. Kahn, "Enzyme-specific profiles for genome annotation: PRIAM," *Nucleic Acids Research*, vol. 31, no. 22, pp. 6633–6639, Nov. 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC275543/

[134] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L.-S. L. Yeh, "UniProt: the Universal Protein knowledgebase," *Nucleic Acids Research*, vol. 32, no. Database issue, pp. D115–D119, Jan. 2004. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC308865/

[135] E. M. Robertsen, H. Denise, A. Mitchell, R. D. Finn, L. A. Bongo, and N. P. Willassen, "ELIXIR pilot action: Marine metagenomics – towards a domain specific set of sustainable services," *F1000Research*, vol. 6, p. 70, Jan. 2017. [Online]. Available: https://f1000research.com/articles/6-

70/v1

[136] N. Yok and G. Rosen, "Benchmarking of gene prediction programs for metagenomic data," *Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, vol. 2010, pp. 6190–6193, 2010.

[137] R. H. Whittaker, "Vegetation of the Siskiyou Mountains, Oregon and California," *Ecological Monographs*, vol. 30, no. 3, pp. 279–338, Feb. 1960. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.2307/1943563/abstract

[138] B. D. Ondov, N. H. Bergman, and A. M. Phillippy, "Interactive metagenomic visualization in a Web browser," *BMC Bioinformatics*, vol. 12, p. 385, 2011. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-12-385

[139] J. Hubert, M. Kamler, M. Nesvorna, O. Ledvinka, J. Kopecky, and T. Erban, "Comparison of Varroa destructor and Worker Honeybee Microbiota Within Hives Indicates Shared Bacteria," *Microbial Ecology*, vol. 72, no. 2, pp. 448–459, Aug. 2016. [Online]. Available: http://link.springer.com/article/10.1007/s00248-016-0776-y

[140] M. Schubert, L. Ermini, C. D. Sarkissian, H. Jónsson, A. Ginolhac, R. Schaefer, M. D. Martin, R. Fernández, M. Kircher, M. McCue, E. Willerslev, and L. Orlando, "Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX," *Nature Protocols*, vol. 9, no. 5, pp. 1056–1082, May 2014. [Online]. Available: http://www.nature.com/nprot/journal/v9/n5/full/nprot.2014.063.html

[141] M. Vastra, P. Salvin, and C. Roos, "MIC of carbon steel in Amazonian environment: Electrochemical, biological and surface analyses," *International Biodeterioration & Biodegradation*, vol. 112, pp. 98–107, Aug. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0964830516301524

[142] M. Cantor, H. Nordberg, T. Smirnova, M. Hess, S. Tringe, and I. Dubchak, "Elviz – exploration of metagenome assemblies with an interactive visualization tool," *BMC Bioinformatics*, vol. 16, p. 130, 2015. [Online]. Available: http://dx.doi.org/10.1186/s12859-015-0566-4

[143] A. Mitchell, F. Bucchini, G. Cochrane, H. Denise, P. t. Hoopen, M. Fraser, S. Pesseat, S. Potter, M. Scheremetjew, P. Sterk, and R. D. Finn, "EBI metagenomics in 2016 - an expanding and evolving resource for the analysis and archiving of metagenomic data," *Nucleic Acids Research*, p. gkv1195, Nov. 2015. [Online]. Available: http://nar.oxfordjournals.org/content/early/2015/11/17/nar.gkv1195

[144] A. Wilke, J. Bischof, W. Gerlach, E. Glass, T. Harrison, K. P. Keegan, T. Paczian, W. L. Trimble, S. Bagchi, A. Grama, S. Chaterji, and F. Meyer, "The MG-RAST metagenomics database and portal in

2015," *Nucleic Acids Research*, p. gkv1322, Dec. 2015. [Online]. Available: http://nar.oxfordjournals.org/content/early/2015/12/09/nar.gkv1322

[145] E. M. Robertsen, T. Kahlke, I. A. Raknes, E. Pedersen, E. K. Semb, M. Ernstsen, L. A. Bongo, and N. P. Willassen, "META-pipe - Pipeline Annotation, Analysis and Visualization of Marine Metagenomic Sequence Data," *arXiv:1604.04103 [cs]*, Apr. 2016. [Online]. Available: http://arxiv.org/abs/1604.04103

[146] "ELIXIR - A distributed infrastructure for life-science information," 2017. [Online]. Available: https://www.elixir-europe.org/

[147] "Norwegian e-Infrastructure for Life Sciences," 2017. [Online]. Available: https://nels.bioinfo.no/

[148] T. Kahlke, "Analysis of the Vibrionaceae pan-genome," Ph.D. dissertation, UiT - The Arctic University of Norway, 2013. [Online]. Available: http://hdl.handle.net/10037/5248

[149] Penn State University, Johns Hopkins University, "The Galaxy Project," 2017. [Online]. Available: https://galaxyproject.org/

[150] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin, A. Ghodsi, J. Gonzalez, S. Shenker, and I. Stoica, "Apache Spark: A Unified Engine for Big Data Processing," *Commun. ACM*, vol. 59, no. 11, pp. 56–65, Oct. 2016. [Online]. Available: http://doi.acm.org/10.1145/2934664

[151] B. Chevreux, T. Pfisterer, B. Drescher, A. J. Driesel, W. E. G. Müller, T. Wetter, and S. Suhai, "Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs," *Genome Research*, vol. 14, no. 6, pp. 1147–1159, Jun. 2004.

[152] J. Goll, D. B. Rusch, D. M. Tanenbaum, M. Thiagarajan, K. Li, B. A. Methé, and S. Yooseph, "METAREP: JCVI metagenomics reports—an open source tool for high-performance comparative metagenomics," *Bioinformatics*, vol. 26, no. 20, pp. 2631–2632, Oct. 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2951084/

[153] Pivotal Software, "RabbitMQ," 2007. [Online]. Available: https://www.rabbitmq.com/

[154] C. E. Cook, M. T. Bergman, R. D. Finn, G. Cochrane, E. Birney, and R. Apweiler, "The European Bioinformatics Institute in 2016: Data growth and integration," *Nucleic Acids Research*, vol. 44, no. D1, pp. D20–D26, Jan. 2016. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1352

[155] Red Hat, Inc., "Ansible Playbooks," 2016. [Online]. Available: http://docs.ansible.com/ansible/playbooks.html

Part II

Collection of publications