2-1-2013

De novo metagenomic assembly of microbial communities from the lower convective layer of the Red Sea Atlantis II brine environment.

Osama Said Ali

The American University in Cairo
School of Sciences and Engineering

De novo metagenomic assembly of microbial communities from the lower convective layer of the Red Sea Atlantis II brine environment.

A Thesis Submitted to
The Biology Department

in partial fulfillment of the requirements for
the degree of Master of Science

by Osama Said Ali

under the supervision of Dr. Hamza El Dorry
January 2013

The American University in Cairo

De novo metagenomic assembly of microbial communities from the lower convective layer of the Red Sea Atlantis II brine environment

A Thesis Submitted by

Osama Said Ali

To the Biotechnology Graduate Program

Month/ Year

In partial fulfillment of the requirements for
The degree of Master of Science

Has been approved by

Thesis Committee Supervisor/Chair ______________________________________

Affiliation ____

Thesis Committee Reader/Examiner _____________________________________

Affiliation ____

Thesis Committee Reader/Examiner _____________________________________

Affiliation____

Thesis Committee Reader/External Examiner ______________________________

Affiliation ____

__________ ______________ ______________ _____________
Dept. Chair/Director Date Dean Date

DEDICATION

To The Soul of My Father Who Taught me That Knowledge Is The Real Heritage To Be Left For The Future.

To My Beloved Mother Whose Sacrifices Made Me Able To Conquer My Achievements.

To My Life Partner, Ayat, For Her Unlimited Support And Encouragement To Complete This Work.

Last, But Not Least, To My Whole Family For Being True Fans During My Journey.

Acknowledgements

I am particularly grateful to my advisor Dr. Hamza El Dorry for his constant support and advice. Also, I am really grateful for his confidence in my capabilities, the secrete Encouraging word behind the conquer of unexpected difficulties.

Many thanks to dr. rania siam, chair of the biology department for her continuous encouragement and support

i am heartily gratefull to Dr. Ari Ferreira who taught me that annotation is not a process but a vision.

I am grateful to Dr. Ahmed Sayed abdel Aziz and Dr. Mohamed Ghazi for their helpful tranining in the lab

I am greatly thankful to Mr. hazem sharaf, Ms. mariam rizkallah, Mr. Mustafa adel and Mr. amged ouf for their continuous help and support

Finally, I would like to express my gratefulness to my colleagues at the biology department for their kind support

ABSTRACT

The American University in Cairo

De novo metagenomic assembly of microbial communities from the lower convective layer of the Red Sea Atlantis II brine environment.

by Osama Said Ali
under the supervision of Dr. Hamza El Dorry

The lower convective layer of the Red Sea Atlantis II brine pool (ATII-LCL) is an unexplored environment that is characterized by harsh conditions of high temperature (68 ºC), high salinity (26%), high concentration of heavy metals and very low oxygen content. Microbial communities inhabiting this extreme environment are expected to have unique structural and functional adaptations to survive such harsh conditions. These adaptations can be expressed by novel genes or new metabolic pathways.

The recent advances in the next generation sequencing technologies have increased the size of the generated reads (500 bps in 454 pyrosequencing) and lowered the sequencing cost per gigabase. As a result, research efforts became more feasible to reveal the mystery of such an interesting environment and to discover novel proteins that might have a useful biotechnological application.

This study is the first attempt to establish a metagenomic assembled dataset of the environmental sample taken from the ATII-LCL. Three successive runs of 454 random shotgun sequencing were performed producing a large size dataset of 1.5 Gbs and 4.4 million reads. This approach has been used to increase the sequence coverage of metagenomic datasets in order to overcome the high diversity of some microbial communities. *De novo* assembly of the pooled reads from all sequencing runs resulted in a 40,693 contigs with maximum contig size of 350 kb. The comparison of different assembly versions of individual runs showed that we have not yet reached a complete coverage of the genomes contained in the metagenomic sample. Also, this metagenomic dataset has shown a high complexity concerning the community structure due to the absence of a dominant taxonomic classification. The taxonomic classification of the assembled dataset has been distributed between three major bacterial orders, Burkholderiales, Rhizobiales and Pseudomonadales and one Archaeal class Euryarchaeota.

The newly established dataset has been used to annotate an operon for mercury resistance genes. The annotated Mercuric reductase gene (MerA) has been synthesized and expressed in the lab showing a high enzyme activity compared to its terrestrial peers.

TABLE OF CONTENTS

LIST OF FIGURS

LIST OF TABLES

LIST OF ABBREVIATIONS

LCL	Lower Convective Layer
DNA	Deoxyribonucleic acid
BAC	Bacterial Artificial Chromosome
HGT	Horizontal Gene Transfer
GOS	Global Ocean Sampling
AAnP	Aerobic Anoxygenic Photosynthestic bacteria
AT II	Atlantis II
DD	Discovery Deep
CD	Chain Deep
OLC	Overlap / Layout / Consensus
DBG	De Bruijn Graph
BLAST	Basic Local Alignment Search Tool
MAP	Metagenomic Assembly Program
AMD	Acid Mine Drainage
LT	Lake Tyrell
TCA	Tricarboxylic Citric Acid pathway
MGA	Metagene Annotator
ORF	Open Reading Frame
CDs	Coding Sequences
PTP	Pico Titre Plate

1. Literature Review

1.1 Introduction

Earth is dominated by microbial life. It has been estimated that the total number of prokaryotic live cells on earth is $5X10^{30}$, comprising 10^6 to 10^8 separate species[1]. Moreover, microbes are a primary source of nutrients for other life forms and the main recyclers of dead organisms by hydrolyzing them into the essential organic substances. Bacteria and Archaea constitute the major component in many marine habitats especially those with harsh physical and chemical conditions of high temperature, pressure and anoxia. For example, some deep sea vents have high temperature that reach 340 ºC at which bacteria and archaea are the only form of life available[1].

Early genomic studies started with sequencing the genomes of bacteriophages MS2; (3,569 nucleotides) long [2] and Phi-X174 (5,386 nucleotides) long [3]. In 1995, *Haemophilus influenza* was the first bacterial genome to be sequenced (1,830,137 base pairs)[4]. The microbial genomes of about 3,447 bacterial, 1,762 viral and 230 archaeal species are recorded in the GeneBank database release 192.0 till October 2012. However, the relationships among different species within the community and the interactions of the species with the surrounding environment have not been enough explained. This limitation results from the fact that only a very low percentage of microorganisms are cultivable (0.001-0.01% from sea water, 0.25% from fresh water or sediments and 3% from soil)[5]. Therefore, the clonal cultures prepared from any microbial community will reflect a biased image with regard to the abundance of the species in this community. Also, since microorganisms do not live as single species, the cultured clones will not be able to describe the impact of the surrounding environment and other species on the regulation of gene expression and the biological functions.

1.2 Red Sea Atlantis II Brine Pool, Lower Convective Layer

Red Sea formation was suggested to have started 25 million years ago through the continuous separation of both the African and Arabian tectonic plates[6]. The topography of the Red Sea bottom is characterized by an axial rift that extends from the southern area where it is fully-formed to the northern area where it is in the form of scattered deeps[7]. It is believed that some of the deep basins of the rift were filled with high salty brines five million years ago due to the high volcanic activity of the submarine shallow layer. Brine water is characterized by hypersaline content, high temperature and lack of oxygen. Twenty-five brine pools were discovered in the Red Sea to date[7]. The Atlantis II deep is the largest deep-sea brine pool in the Red Sea and is located at 21^{o} 21'N on the axial rift (Figure 1). Atlantis II deep is considered to be hydrothermally active due to the long term observations for the continuous increase in the maximum recorded temperature from 55.9 to 68.2°C over thirty years period[8]. Sediments of Atlantis II brine are cooler in temperature by 1.7 C^{o} than the brine water and they have a high metal content especially, iron, copper, zinc and other heavy metals[8],[9]. Atlantis II deep is divided into three main distinctive layers, upper convective layer (UCL) that has a thickness of 46 meters, lower convective layer(LCL) of 135 meters thickness and a seawater-brine interface layer which is about 14 meters[7]. Atlantis II deep has a maximum depth of about 2194 meters with a sharp increase in both temperature of a maximum $68.2C^{o}$ and salinity of a maximum 25.7% at the lower convective layer[9] (Figure 2). On the other hand, Oxygen content decreases to reach 0 mM at LCL layer which is considered highly anoxic.

Harsh conditions represented by high temperature, hypersalinity and acidic pH of Atlantis II deep brines have attracted recent microbiological studies to explore this extreme environment looking for novel microbial species and identifying new taxonomic groups[10]. Due to the advances in sampling tools and molecular investigation technologies, the study of the biodiversity and functional analysis of the microbial communities inhabiting Red Sea brines became more feasible and opened a door for new discoveries regarding microorganisms that have not been cultured before.

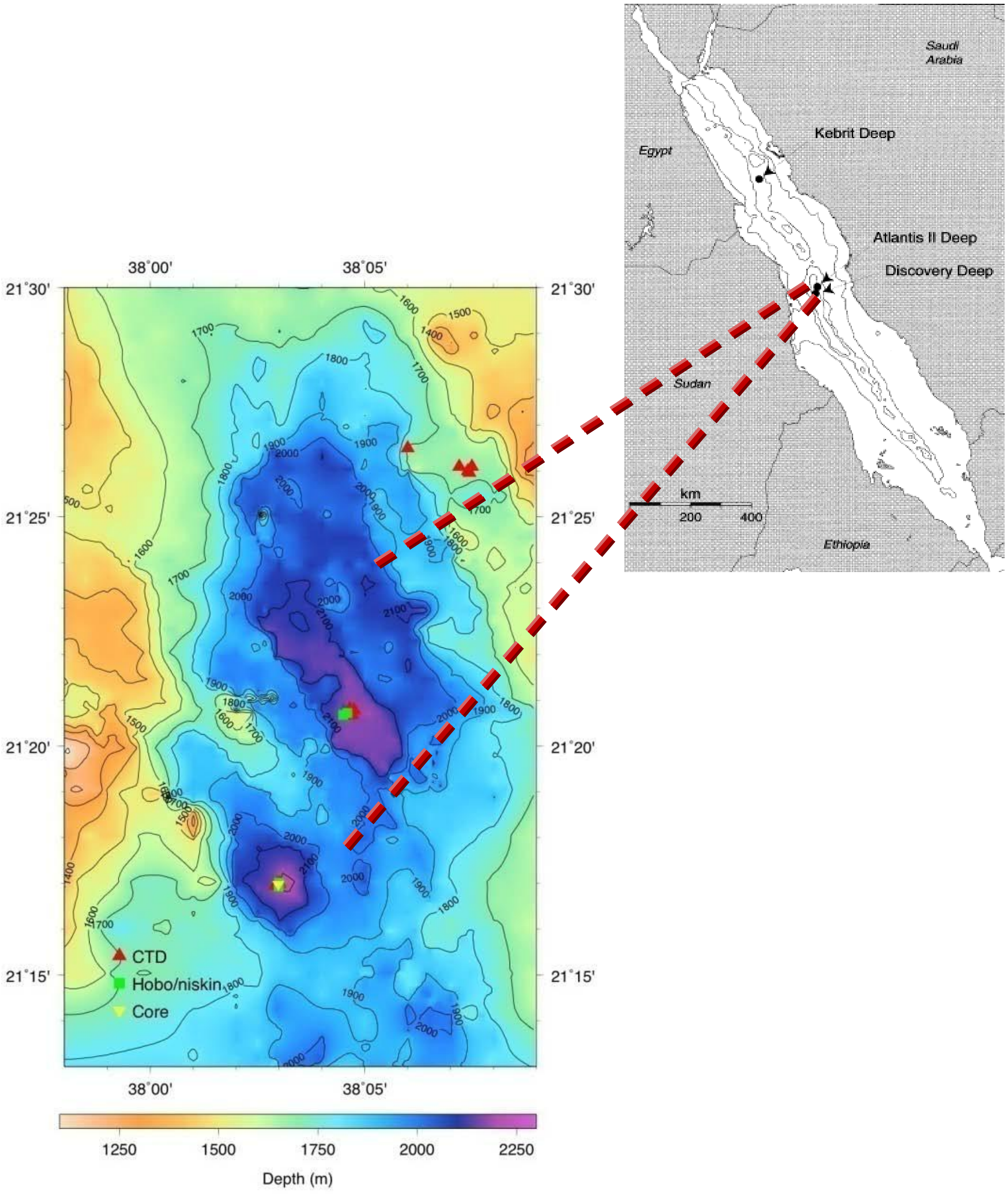
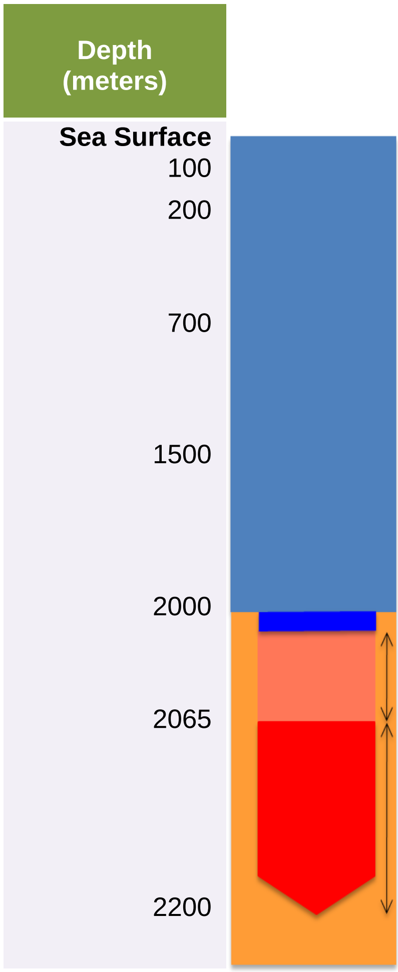


Figure 1: Atlantis II Deep geographical location

The Atlantis II deep is the largest deep-sea brine pool in the Red Sea and is located at 21^{o}21’N 38^{o}04.61' E on the axial rift



	Thickness (meters)	Salinity (%)	Temperature (C^o)	Oxygen (mM)
		4.0	22.0	180
		4.0	22.0	25
		4.0	22.0	100
		4.0	22.0	100
Interface	14	8.7	43	25
UCL	46	9-16	43-57	15
LCL	135	26	68	0

Figure 2: Physical characteristics of the Atlantis II Brine Pool

1.3 Metagenomics

The Metagenomics term has been used to describe a new research field that aims at direct analyzing the uncultured total genomic DNA extracted from microorganisms collected from environmental samples. Over the past decade, metagenomic research succeeded to shed light on understanding the diversity of microbes, their functions and evolution in different habitats such as water[11], soil[12] and the digestive systems of both humans[13] and animals[14].

Metagenomics research can be divided into two major divisions; environmental gene surveys and random shotgun studies of genomes in an environmental sample[15]. The first division comprises targeted studies that use polymerase chain reaction techniques (PCR) to amplify a specific gene and sequence it to determine different orthologs of this gene in the studied community. The second approach, random shotgun metagenomic studies, deals with total DNA extracted from environmental samples. This DNA is then sequenced to create a catalog of genes present within the sample (Figure 3).

Due to the high abundance of prokaryotes in the marine habitat, research efforts were drawn to discover the role of microbial organisms in establishing communities that inhabit the extreme life conditions of deep seas. What follows is an extensive review of the major marine metagenomic studies done in the last decade.

1.3.1 Literature review of marine metagenomic studies

Research work for the marine microbial shotgun metagenomics use different approaches including establishing a metagenomic libraries in adequate vector (fosmid, cosmid, and bacterial artificial chromosome-(BAC) followed by random sequencing of recombined clones. In addition, direct sequencing approach of metagenomic DNA using next-generation sequencing techniques (NGS) to study the community structure and function as well as isolating individual genomes from assembled metagenomic datasets[15].

The isolation of the first marine microbial metagenomic DNA using BAC cloning vector was one of the early studies that provided an insight at marine Archaea[16]. Then, metagenomic fosmid libraries were used to describe the marine archaeal phylum Crenarchaeota from the temperate Pacific Ocean and the Antarctic Ocean[17]. Another fosmid-derived metagenomic study succeeded to isolate proteins from the coastal water of the Antarctic that are adapted to the extreme cold environments[18]. Also, environmental fosmid inserts have been used to indicate the taxonomy of their source genomes in addition to determine gene clusters produced as a result of horizontal gene transfer (HGT)[19].

A major fosmid study at the Hawaii Ocean Time-series (HOT) station in the Pacific ocean produced 64Mbp of DNA sequence[20]. One of the important findings of this study was that the variations in protein sequences of marine microbial communities reflected a vertical stratified distribution pattern of taxonomical groups, functional gene profiles and metabolic pathways. Also, the metadata of this study provided a valuable source of information about a well-studied environment for comparative metagenomic approaches.

The first application of Sanger sequencing technology using small-insert clones was performed on metagenomic DNA from the Saragasso Sea was performed by Venter and colleagues[21]. The discovery of 1800 genomes, 48 unknown bacterial taxonomic groups as well as, 1.2 million novel genes was one of the essential outcomes of this study. Also, it is worth mentioning that new bioinformatic approaches were applied to analyze the resulted dataset composed of one billion base pair. This study was the milestone for the coming metagenomic research since it was considered as a proof of concept for the feasibility for using genomic sequencing techniques in metagenomic samples. However, according to an estimation that one milliliter of seawater contains about one million bacterial cells with an average genome size of two million base pair, only 0.05% of genomic information was sequenced per one milliliter of seawater of the Saragasso Sea project [15]. The last hypothesis demonstrates the huge amount of required data if a full sequencing coverage of all the components of a microbial community would be expected.

The Global Ocean Sampling(GOS) expedition is an another example of an extensive metagenomic sampling efforts that produced a total of 6.3 billion base pairs from 7.7 million sequenced reads[11]. Having these data uploaded in public databases opened the door for applying new bioinformatics tools in many independent research efforts. For example, Yutin et al.(2007) used GOS metagenomic datasets to study abundance and diversity of aerobic anoxygenic photosynthetic bacteria (AAnP) in different oceanic regions[22]. AAnP bacteria is a type of marine bacteria that utilize aerobic Oxygen in respiration and perform photosynthesis to fix CO_2 but they do not produce oxygen (anoxygenic) as a by-product of the process[23]. The study indicated that AAnP groups have different relative abundance in the GOS datasets according to the difference in environmental conditions from open ocean, where AanP bacteria represent 1-5% of the total microbial community, to coastal regions, where they account for more than 10% of the microbes. Furthermore, they showed that the marine AAnP bacteria are essential component of the bacterioplankton assemblages in specific oceanic regions.

The study of Wilhelm et al.(2007)[24] is another research work that benefited from the public dataset of the Saragasso Sea metagenomic project[15]. A genome of marine alpha-proteobacterium SAR11 was isolated form coastal regions of Oregon, USA and used as a query sequence to compare it with the metagenomic dataset of Saragasso Sea project aiming at studying genomic variations of SAR11 genomes found in different oceanic regions. It was concluded that although, natural selection maintains a common core features along all SAR11 genomes (71%) from different marine locations, a significant variation in four hypervariable genomic regions was recorded. This suggested that different marine habitats can affect the selection of genomic content of microbial populations[24].

Since 2008 and due to advances in sequencing technologies that lowered the sequencing cost per gigabase, many marine metagenomic research studies have been performed in different marine habitat such as coastal pelagic ecosystems, marine hydrothermal vents, marine sediments, open ocean, host-associated communities as well as comparative metagenomic pyrosequencing studies[15].

Coastal pelagic ecosystem is the coastal shelf sea that is far from the shore by about 5 kilometers. An example of recent research work of coastal pelagic ecosystem is a major metagenomic study of the western English Channel L4 sampling site[25]. In this study, a novel combination of different techniques was used to study methylotrophic microbial communities which usually exist in low abundance due to lack of one-carbon substrates needed for their feeding. The first technique was isotope probing with^{13}C-labeled methanol to test for organisms that utilize methanol as a substrate. Then, multiple displacement amplification (MDA) was used to enrich the isolated picograms of ^{13}C-DNA into microgram quantities to construct fosmid library of about 10,000 clones. Finally, polymerase chain reaction (PCR) screening of 1500 clones of ^{13}C-DNA fosmid library to find methanol metabolism genes and a shotgun Sanger sequencing of the screened insert were performed. The assembly of the data resulted in formation of 9Kb-contig carrying a cluster of genes involved in methanol metabolism in this marine community. It was demonstrated that the dominant group involved in methanol metabolism, in this community, was closely related to *Methylophaga* genus. The study highlighted the role of using combined research methods to circumvent the challenge of isolating specific genes from low-abundant organisms in a marine microbial community.

Palenik and colleagues performed another enrichment study using different strategy to raise the relative abundance of cells of the cyanobacterial genus *Synechococcus* from a metagenomic sample taken from surface seawater off the coast of California[26]. Targeted cells were enriched by sorting them out from the complex metagenomic sample using flow cytometry technology. DNA was extracted and pyrosequenced using

454 platform producing 370,000 reads. Reads were aligned to model *Synechococcus* reference genomes that were isolated from different marine locations except for two genomes. One of them was isolated from the same environment of the study and the other one was from a similar coastal area. It was indicated that sequence identity is very high with the genes of the two coastal model genomes while a significant difference in identity was noticed with genomes of other locations. This pattern suggested a role for horizontal gene transfer affecting the genetic structure in different environments and also among different strains.

Although the hydrothermal vents represent an interesting marine ecosystem due to their extreme physical and chemical conditions, few metagenomic studies were performed. In a recent study, DNA was extracted from the metagenomic sample of the biofilm layer of the carbonate chimneys of the Lost City Hydrothermal Field on the Mid-Atlantic Ridge[27].The biofilm microbial community was dominated by one phylotype (*Methanosarcinales*) which represented more than 80% of the total community structure. DNA was cloned and randomly end-sequenced to obtain 46,316 shotgun reads of a total size 35 Mbp. It was found that transposases constituted more than 8% of the whole community which is ten times more than any other compared metagenome. Also, although the assembled contigs of transposases showed a very high coverage they were small in size. The previous results indicated that transposases were abundant in the community but they are located on small, extragenomic molecules such as plasmids or viruses. It was concluded that lateral gene transfer in low-complex communities that are dominated by few number of organisms plays an important role in determining the phenotypic diversity for the members of the community.

An example of metagenomic studies of marine sediments is the study by Huang et al. (2009)[28]. Metagenomic samples were isolated from sediments at 1200 m, 1300 m and 2900 m depths from the South China Sea. Environmental DNA was extracted and a fosmid library of 40,000 clones was prepared with an insert size ranging from 24-45 kb. Clone screening resulted in determining one specific clone, that was called, fss61, that was able to alter the phenotype of its *Escherichia coli* host. This alteration was represented by the ability of *E. coli* cells to produce melanin. The selected fosmid clone was fully sequenced using Sanger method and the sequences were analyzed. Sequence analysis identified the open reading frame (ORF) responsible for melanin production. The deduced protein from this ORF was highly similar to 4-hydroxyphenylpyruvate dioxygenase (HPPD) from deep-sea bacteria *Idiomarina loihiensis*. This study showed that the extracted metagenomic DNA can be a source of novel gene discoveries.

The work of Siam and colleagues is one of the few recent comparative taxonomic studies on metagenomic sediment samples of three Red Sea brine pools including Atlantis II(ATII), Discovery Deep (DD) and Chain Deep (CD) as well as sediment samples from an adjacent brine-influenced site (BI)[29]. In this study, the environmental 16S ribosomal RNA genes (16S rDNA) were isolated, amplified and pyrosequenced using 454 sequencing system. The analysis of the resulted datasets showed evidences for a distinctive structural difference in the microbial populations found in both ATII and DD sites compared to the other study sites. The most abundant bacterial and archaeal phyla in ATII and DD were *Proteobacteria, Actinobacteria, Cyanobacteria, Deferribacteres, and Euryarchaeota*. In addition, the 16S rDNA pyrotag analysis made it possible to classify the bacterial and archaeal communities into three major groups; group I which characterized the sulphur-rich ATII site, group II which was dominant in nitrogen-rich DD sample and group III which was found in the rest of sampling sites.

An interesting study was performed in 2008 using a hybrid approach that combined both metagenomic and metatranscriptomic strategies to identify functional and taxonomic diversity in open ocean communities[30]. Metagenomic DNA was extracted from a sample taken from the North pacific. Cyanobacteria and unknown bacterial taxa were the most abundant gene transcripts. Also a significant number of transcripts came from genes located in the hypervariable regions of cyanobacterial genomes, confirming the notion that these genomic variations are essential for habitat differentian. This study showed the power of creating two complementary metagenomic and metatranscriptomic datasets to better understand the open ocean environment.

1.3.2 Challenges facing a metagenomic project

The development in metagenomic studies is driven by the advances in sequencing technologies as well as the improvement in bioinformatics data analysis algorithms. A general framework structure of any metagenomic project can be summarized as sampling, sequencing and data analysis. Each one of these steps have some challenges which affect the expected final outcome of a metagenomic study.

1.3.2.1 Sampling

Sampling, being the first step, has a significant impact on the results of any metagenomic project. The extracted DNA from an environmental sample should be representative to all cells present in the sample and has enough amounts and a high quality yield for further library preparation and sequencing steps[31]. If the metagenomic study targets a certain part of the community, physical fractionation by a series of selective filtrations can be used to concentrate the targeted material and to be sure that no contamination from the non-target parts of the community is available[21]. In case of low DNA yield, DNA amplification method can be used to provide sufficient amounts of genomic DNA for further sequencing step. However, amplification process is associated with some problems as a potential chimeric sequence formation and amplification bias towards the most abundant organism in the sample[32]. Hence, a careful assessment of how many rounds of amplification are needed and the proper amount of the starter DNA required has to be done prior applying any DNA amplification process.

1.3.2.2 Sequencing

Sequencing is the process determining the right order of the four building blocks thymine (T), adenine (A), guanine (G) and cytosine (C) that form any DNA strand of an organism. Metagenomic sequencing aims at studying community composition including taxonomic structure and abundance ratio of different species, functional analysis of genetic profile of community members and intra-species or, intra-population genetic relationships[33] . Sequencing platform has a crucial role in the results obtained from any metagenomic project. Metagenomic random shotgun sequencing can be classified into Sanger sequencing technology[34] and next generation sequencing(NGS) technology.

454/Roche and Illumina/Solexa systems are the most applied next generation sequencing technologies in metagenomic analysis.

The principles of Sanger sequencing is based on chain termination of the replicated DNA fragments using dideoxy derivatives of the four nucleotides (ddNTPs)[34]. Sanger sequencing is characterized by a low error rate and long read length (>700bp). However, disadvantages of Sanger sequencing can be summarized as the high sequencing cost per gigabase (approximately USD 400,000)[31] and the cloning bias against toxic genes for the host cells[35]. Accordingly, Sanger metagenomic sequencing approach suits better the reconstruction of complete genomes from low-diversity environmental samples[36].

On the other hand, 454/Roche system uses emulsion polymerase chain reaction (ePCR) to amplify clones of random DNA fragments that are attached to microscopic beads located in the wells of a picotitre plate. The picotitre plate is subjected to a parallel pyrosequencing process in which the four dNTPs are added sequentially to all the template DNA strands. The incorporation of dNTP molecule in the new strand formation results in a release of a pyrophosphate molecule which is interpreted by the system into light signal. About 1.2 million light signals are emitted from polymerization reactions running on the picotitre plate. The strength of the emitted light signal determines the source nucleotide and the system will translate these signals into their comparable sequence[37]. 454/Roche pyrosequencing is one of the most suitable next generation sequencing choices for metagenomic projects due to the low sequencing cost per a gigabase (about USD 20,000), good average read length between 300 to 600 bps, low amount of genomic DNA needed for the run (few nanograms in single end sequencing) and multiplexing which allows simultaneous sequencing of up to 12 samples in 500Mbp run[31]. However the artificial replicates and homopolymer error are the essential disadvantages of applying 454 system in metagenomic sequencing.

1.3.2.3 Data analysis

The typical goal of analyzing datasets of metagenomic samples is to reconstruct all genomes found in an environment. However, this is not feasible because of the computational complexity of the available analysis solutions. Accordingly, the major approaches applied for the analysis of metagenomic datasets have fundamental limitations that challenge performing a thorough and complete study of environmental samples[33]. The first approach is to assemble reads resulted from metagenomic sequencing and carries out a contig-based taxonomic and functional analysis (Figure 3). Problems of this approach can be summarized as the high computer memory required for assembly due to the large size of metagenomic datasets, variable abundance of the genomes in a community will prevent the assembly of low-abundant organisms and the population heterogeneity will be a source of chimeric contigs. On the other hand, read-based analysis is used as a second method to reconstruct both taxonomical and functional components of a metagenome. However, this approach faces some challenges of large number of reads resulted from NGS data which lead to long analysis time. Also, the size of the reads will add another source for errors. The next part will focus on the detailed overview of research achievements with regard to the assembly approach of metagenomic data analysis since it matches the scope of this study.

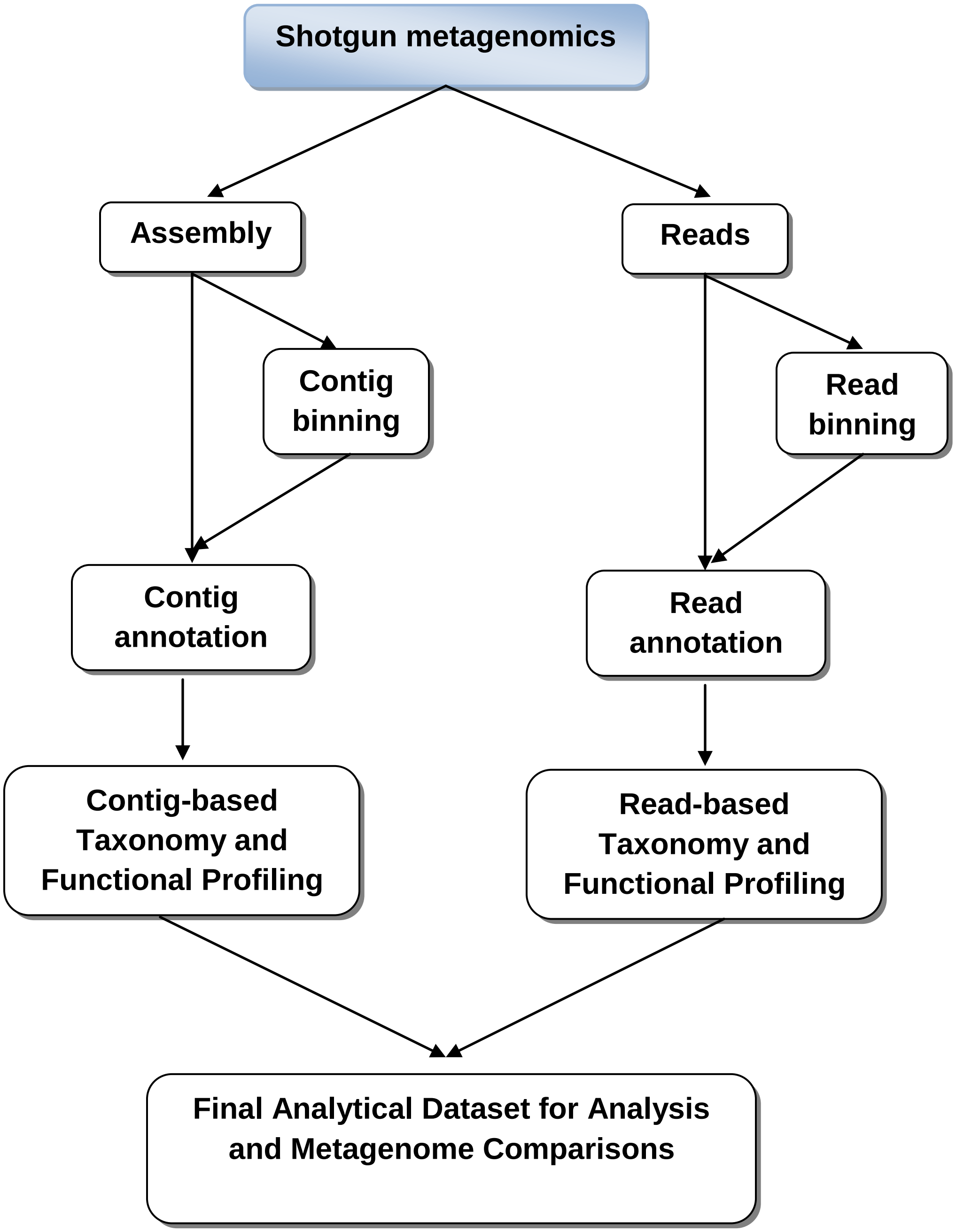

Shotgun metagenomics
Assembly
Reads
Contig binning
Read binning
Contig annotation
Read annotation
Contig-based Taxonomy and Functional Profiling
Read-based Taxonomy and Functional Profiling
Final Analytical Dataset for Analysis and Metagenome Comparisons

Figure 3: Data Analysis flow chart for metagenomic shotgun sequencing

1.3.2.3.1 Assembly of metagenomic datasets

In general, a genomic assembly is a data structure of whole genome shotgun (WGS) reads that are sorted in a hierarchical arrangement so that all reads within a predefined identity percentage are aligned together and form a common contiguous sequence called contig. Contigs are, in turn, grouped into a larger structure called scaffolds which determine the order of the contigs and the size of the gap separating any two successive contigs to form a complete genome sequence [38]. Although there is no a definite measure of the assembly accuracy, N50 parameter can be used to indicate the quality of the assembled data. N50 is defined as the contig size; when contigs are sorted from the largest to the smallest, at which the percentage of the total number of bases contributing in the assembly equals to 50%[38].

The development in algorithms of the assembly software is strongly tied to the advances of sequencing technologies. The Overlap/Layout/Consensus (OLC) and de Bruijn Graph (DBG) are the most widely used assembly algorithms for the next generation sequencing(NGS) assemblers [39]. Contig construction is the main target of both assembly algorithms. However, the main difference between them is the pattern which each algorithm uses to build contigs. OLC algorithm applies pairwise alignment among all reads and construct a graph layout which constitutes the aligned reads as nodes and the overlap links between reads as edges. The final step is the interpretation of the read graph into contigs by calling consensus sequences from multiple sequence alignment of the reads. It is worth mentioning that OLC graph construction is a CPU-intensive process and it needs more efficient computational resources as the size of the sequenced reads dataset becomes larger. Accordingly, OLC algorithm is the preferred approach for lower-coverage and long reads (100-800bp) as in case of Roche/454 sequencing platform.

On the other hand, DBG algorithm does not have any pairwise alignment for the reads but instead it cuts each read into a predetermined number of bases called k-mer. The overlapped k-mers are linked together in a directed layout which has one entrance and one exit for each k-mer node. Since the resulted DBG graph does not use actual reads in its construction, it saves computational resources and allow for larger datasets. This makes DBG algorithm more suitable for high-coverage short-reads (< 100 bp) as in case of Illumina or Solexa sequencing platforms[40]. Examples of assembly software that apply OLC algorithm are Newbler[41] which is the official Roche assembler and distributed by 454 life sciences and Celera[42]. On the other hand, Velvet[43] and ABySS[44] are the examples of assemblers using DBG algorithm. Regardless of the type of the applied algorithm, the main core of assembly steps can be classified into data preprocessing, contig construction, scaffold linkage and gap closure. Assembly software faces many challenges of repeat sequences of the assembled genomic regions, limited read length and sequencing errors which significantly affect the accuracy of the resulted datasets. Moreover in case of metagenomic assembly, a new level of complexity is added to the process due to high genomic diversity of the environmental samples, the abundance variability within populations, the high cost of efficient computational resources that can handle large datasets and finally, lack of specialized metagenomic assemblers that are able to separate the closely related species from microbial communities. However recently, few *De novo* metagenomic assemblers have been introduced as an approach to overcome the complexity issues of metagenomic samples. Meta-velvet[45] and Meta-IDBA[46] are examples of metagenomic assemblers that use DBG algorithm. Also, metagenomic assembly program (MAP)[47] is another application using OLC algorithm for 454 reads.

The assembly of metagenomic datasets aims at studying either the reconstruction of genomes from environmental samples or, creating longer pieces of coding DNA sequences (CDs) for further characterization. In the last case, contigs are not an end product by themselves but they are used as a mean to understand the structure and function of the microbial community [48].

Assembly approaches can be classified into a reference-based and *De novo* assembly. A reference-based assembly is applied when a closely related reference genome to the metagenomic dataset is found. While in the *De novo* method the metagenomic reads are assembled from scratch without having any reference sequence due to the high complexity of the microbial communities especially at the level of species and strains. This is why the *De novo* assembly of a metagenome is the most commonly used approach since it is difficult to find reference genomes for complex environmental samples.

De novo assembly of metagenomic datasets can be a source of novel findings even in a previously well-studied habitats as in case of hypersaline Lake Tyrrell(LT) in Australia [49]. In this study, the deeply-sequenced libraries with both Sanger and 454 pyrosequencing technologies were assembled either independently or, by using different combinations. The phylogenetic analysis of the assembled contigs resulted in the discovery of two new halophilic archaeal lineages that are highly abundant in the surface water of LT. The reconstruction of these two novel uncultured genomes proved the promising capabilities of *De novo* metagenomic assembly.

The construction of microbial community profiles from metagenomes is another interesting application of the *De novo* metagenomic assembly approach. As an example, a recent study succeeded to create a catalogue of the human gut microbial genes by using *De novo* illumina-based metagenomic sequencing, assembly and characterization of 3.3 million non-redundant human intestinal microbial genes [13].

The possibility of getting draft genomes or even complete ones from a metagenomic sample increases when it is dominated by few number of organisms or the target species shows low interspecies variations. Early metagenomic studies on low complexity environment was able to isolate near-complete genomes of two leptospirillum group II and Ferroplasma type II bacteria from the acid mine drainage(AMD) biofilm of Rhichmond mine in California[50].. The metagenomic DNA sample of the AMD microbial community was dominated by few genomically distinctive species that were Leptospirillum groupII 75%, Liptospirillum groupIII 10%, Archaea 10%, Euarkyotes 4% and sulfobacillus spp. 1%. The community structure of AMD metagenome facilitated the assembly process since about 85% of the shotgun reads were assembled into scaffolds of size larger than 2kb. This study was a milestone for further metagenomic studies since it shed light on the complexity of the environmental samples as a real challenge for the recovery of complete genomes. Pelletier *et al. was* the first research group who was able to recover a draft genome of low abundant uncultured anaerobic bacterium ‘ *Candidatus* Cloacamonas acidaminovorans ’ from a metagenomic sample from digester of wastewater treatment plant [51] . This study confirmed the possibility of reconstructing either partial or complete genomes from complex metagenomic sample. As an attempt to overcome the assembly difficulty of interstrain variations in complex environment, Iverson *et al.*[52] applied a massively parallel sequencing approach using Solid technology[53] to have a total of 58.5 gigabases of 50-base mate-paired short reads of the surface seawater metagenome. A *De novo* assembly of the high-coverage mate-paired sequences was performed to isolate a nearly complete genome of an uncultured class of marine group II Euryarchaeota despite they were represented in the sequenced reads by 1.7% only. The recovered genome of group II Euryarchaeota describes a type of motile photo-heterotrophic marine archaea that are specialized in lipids and protein degradation and explains the origin of proteorhodopsin which is considered as a beneficial source of energy for organisms moving long distances searching for food. Similarly, the Cow rumen metagenome is another example of a recent research work that used massive sequencing of DNA extracted from a complex environmental microbial community to characterize the biomass degrading genes and genomes of the cow rumen microbes [54]. The study was able to predict 27,755 carbohydrate-active genes and assemble 15 complete genomes of uncultured bacteria from 268 gigabases of metagenomic DNA. The data sets generated by this study represent a catalogue of both genes and genomes responsible for the degradation of cellulosic biomass.

It was demonstrated that genomes of a distinctive genotype can be accurately assembled from a complex metagenome, if they have at least about 20x coverage [55]. Whereas at lower coverage, chimeric sequences are found to be in large quantities within the assembled contigs which explains the high number of hypothetical proteins of the annotated genes in case of metagenomic projects compared to those of genomic ones. The study also suggested a method for estimating error frequency and type of the assembled contigs and genes as well as detecting intrapopulation structure from complex metagenomic datasets.

As previously mentioned, the *De novo* assembly of metagenomic DNA sequences is not only aimed at the recovery of genomes but also considered as the midway to study target genes and metabolic pathways that determine the structure of the microbial community and to identify the taxonomic classification of the assembled contigs that explain the relation between a community members and their surrounding environment. A good example for a functional analysis of a *De novo* assembly is the metagenomic study on a new deep-sea hypersaline lake Thetis located in the Mediterranean sea [56]. Two metagenomic samples from the brine and interface layers of the lake Thetis were sequenced using Roche 454 pyrosequencing technology and a *De novo* assembly was performed. As a result of the analysis of the assembled datasets, three co-existed autotrophic carbon dioxide fixation pathways were discovered in the interface layer which was the major finding of this study. Also in contrary to what was assumed before that autotrophy is not important in hypersaline environments, the genes for the reductive acetyl-CoA and reductive Tricarboxylic acid (TCA) pathways were found in the brine layer proofing that these pathways are functional at the hypersaline condition. Also, it was revealed that acidic amino acid residues are overrepresented in the proteins of brine layer compared to those of interface confirming the typical protein composition characterizing organisms that live in an extreme hypersaline habitat. This study is one of the first metagenomic surveys for a newly discovered lake Thetis which has the saltiest brine water ever reported (348%) [57].

Recently, a single metagenomic sample of Mediterranean deep chlorophyll maximum (DCM) community was subjected to 454 pyrosequencing using both direct sequencing (DS)and fosmid cloning approaches[58]. The sequenced reads from both DNA libraries were assembled independently and the results were compared. Both DS and fosmid sequencing demonstrated the significant abundance of group II Euryarchaeota in this community. However, only DS results indicated the abundance of photosynthetic cyanobacteria *Prochlorococcus marinus* subsp. *pastoris,Synechococcus* sp. and the heterotroph alphaproteobacteria *Canidatus pelagibacter*. Also, it was observed that fosmid library sequencing resulted in a bias against low GC-content organisms which are the most dominant according to DS estimations which suggested a novel method to isolate low abundant organisms from complex communities.

As an attempt to overcome the problem of complexity and limited knowledge of microbial communities, simulated Data studies played an important role in the assessment of the quality and reliability of the resulted datasets from *De novo* metagenomic assembly [59], [60]. As a result of these efforts, it was suggested that high sequencing will raise the coverage of the studied metagenome and accordingly can overcome the functional classification limitations due to metagenomic assembly. Also, although metagenomic assembly increases the possibility of getting chimeric contigs, it improves the annotation of more complete genes and operons. Moreover, it was shown that as the complexity of a metagenomic sample increases, the possibility of reconstructing non-chimeric contigs decreases.

1.3.2.3.2 Annotation

Annotation of metagenomic datasets can be classified into feature prediction and functional annotation[31]. Feature prediction is the process in which genes or genomic elements are identified in any DNA sequence. Although algorithms for the tools of the prediction of coding sequence (CDs) in complete genome sequences are well-developed with accuracy percentage reach 95%[61], few tools were developed specifically for the gene prediction of metagenomic sequences. MetaGene Annotator (MGA)[62] is an examples of metagenomic annotation tools. MGA is a prokaryotic gene finding from environmental genome shotgun sequences using codon frequencies estimated from the GC content of the query sequence as an approach to detect putative coding sequences with 95% sensitivity.

Functional annotation of metagenomic datasets faces a major computational challenge that affect the maximum annotated portion of the sequence to be between 20-50%. This leaves an average of 65% of genes of metagenomic sequences unrevealed by using this annotation method. The genes that cannot be mapped to any of the known reference genes or proteins in databases are called ORFans. Three hypotheses explaining the reasons for having ORFans are software algorithmic errors leading to wrong calls of coding sequences(CDs), absence of comparable biochemical functions for the predicted genes and existence of structural similarity to known protein but accompanied by an absence of sequence similarity to any of the known genes[63].

Protein structural analysis and biochemical characterization methods are recent approaches that can be used to improve annotation of ORFans.

Due to the large size of metagenomic datasets, computational approaches are the most feasible procedures for sequence annotation and complete analysis. MG-RAST server[64] is a fully automated analysis pipeline for metagenomic datasets. Until November 2012, the total number of users on MG-RAST server is about 8000 and the whole number of datasets (private and public) submitted is 64,855 of a total base pairs of 18.46 Tbs and number of sequences 169.4 billion. The public metagenomic projects hosted by the server is 295 project covering 15 different environments and constituting 11,381 metagenomes having 52,590 million sequences of a total number of base pairs 5,326 Gbs. The huge size of data proves the move of scientific community towards centralizing resources and standardizing annotation methods.

1.3.2.3.3 Taxonomic Binning

Binning is defined as the process involving sorting DNA sequences into groups that belong to a single genome or a similar multiple genomes derived from closely related organisms[31]. Algorithms of binning software can be classified into compositional-based, similarity based approaches and hybrid algorithm involving both compositional and similarity approaches. Compositional-based binning uses the conserved nucleotide pattern of genomes, such as GC content and the specific distribution of k-mers, as a measure to compare all sequence fragments into reference genomes of known taxonomic classification. Examples of software using compositional binning are Phylopythia[65], S-GSOM[66] and TACAO[67]. This binning model is not reliable for short reads of NGS platforms since they do not carry enough information for accurate comparisons. On the other hand, homology-based algorithms compare the implicit genetic information in a DNA sequence to those of known genes in reference databases and based on the degree of similarity, a taxonomic group can be assigned to the unknown sequence. Further subdivisions of this algorithm can be found based on the search method applied. CARMA[68] is a similarity-based software relies on Hidden markov model (HMM) search method and MG-RAST[64] software is another similarity-based software but it uses the basic local alignment search technology (BLAST) in the taxonomic classifications of unknown sequences[69]. Despite using BLAST as a search method, MEGAN[70]software improves its results by using lowest common ancestor (LCA) strategy to filter the blast hits according to their bit scores and then selects the best hit species for the assignment of unknown DNA fragment. If redundant BLAST hits are found and MEGAN cannot assign the query sequence to a specific classification, higher taxonomic level (e.g. Phylum) will be used as an assignment for the unknown sequence. Binning of metagenomic reads faces a problem of chimeric bins which occurs when the metagenomic dataset contains two or more genomes that can be assigned to the same high level taxonomic classification [31].

In this study, we present the assembly of datasets comprised of 4.4 million 454 pyrosequencig reads from the metagenomic sample of the lower convective layer of the Red Sea Atlantis II brine environment (ATII-LCL). Also, we evaluated the impact of the change in the assembler computational parameters on the size of the assembled contigs and the pattern of contig extension. In addition, we highlighted the functional and taxonomical classifications of the assembled contigs. Finally, we were able to completely annotate one of the important operons for heavy metal resistance (mercuric reductase) from our assembled dataset as an example of the novel gene discoveries that can be one of valued-outcomes of this research.

2. Materials and methods

2.1 Sample collection

Sampling from Red Sea Atlantis II lower convective layer of the brine pool (ATII-LCL) located at 21º 20.63' N, 38º04.61' E, was carried out during the KAUST/HMR Res Sea expedition onboard R/V Agaeo in April 2010. More than 100 liters of sea water from ATII-LCL were serially filtered through Nitrocellulose/Cellulose acetate filters of sizes 3 µm, 0.8 µm and 0.1 µm respectively. Each filter captures specific range of organisms according to its mesh size. For example, 0.1-filter captures mainly bacteria and archaea which cannot pass through its pores. Filters were stored at -80 ºC in the American University in Cairo (AUC) laboratory until extraction of total DNA from the filters.

2.2 DNA extraction and preparation for sequencing

Metagenomic DNA extraction from ATII-LCL sample was performed at AUC-KAUST genomics lab by dividing the 0.1-filter into four quarters. One quarter was used to extract DNA using a modified protocol based on instructions mentioned in metagenomic DNA Isolation kit for Water from Epicentre (catalog number MGD08420). The membrane was cut into small pieces and placed in a 50 ml sterile conical tube and 5ml of TE buffer (pH 8) was added to the filter pieces. Then, 200 µl of Lysozyme solution (100mg/ml) and 30 µl of RNase A (10mg/ml) were added to the cell suspension. The tube was incubated at 37 ºC in shaking water bath for 1hour. After incubation, 5ml of Meta-Lysis solution (is that a solution from a KIT(2x) and 100 µl of protinase K enzyme (20mg/ml) were mixed by vortexing and incubated at 65 ºC for 2 hours. The cell mixture tube was left to cool in room temperature and placed on ice for 3 to 5 minutes. The supernatant has been transferred into a new falcon tube to which 6ml of protein precipitation reagent was mixed by vortexing for 10 seconds. The tube has been centrifuged for 10 minutes to pellet the debris. Then, the supernatant was transferred to a clean tube and 10 ml of isopropanol was mixed by inverting the tube several times. After centrifugation at 14000x g and 4ºC, the supernatant was discarded and 10 ml of 70% ethanol was added to the DNA pellet. Another round of centrifugation was performed using the previous parameters to pellet DNA once again and the pellet was left for air drying for about 8 minutes and then resuspended in 100 µl of TE buffer. The DNA library was prepared for sequencing using Roche GS-FLX Titanium Rapid Library Preparation Method, kit number 25890110 according to the manufacturer instructions*.*

2.3 454-shotgun pyrosequencing

Roche GS FLX Titanium 454 sequencing platform located in the AUC-KAUST genomics lab was used to sequence the ATII-LCL samples according to 454 pyrosequencing instruction. Three full 454 gaskets were used to generate the ATII-LCL dataset.

2.4 Assembly

Official Roche assembler, Newbler V.2.6, was used to make six assemblies of 454 sequencing data resulted from three independent sequencing runs. A pilot assembly version was performed using all the reads resulted from the first sequencing run (1.3 million reads) to adjust the computation parameters of the assembly program that can be applied for the other assembly series. This was achieved by gradually increasing the stringency of the assembly parameters including minimum identity percentage and minimum overlap length ranging from 90% / 40bp (default values), 90/50, 95/40, 95/50, 98/40 up to 98% / 50bp. To test for the pattern of extension of the assembled contigs due to the change in stringency, three landmark genes were identified on the largest contig (contig 1) using a prokaryotic gene finding program Metagene Annotator (MGA)[71]. These landmark genes were selected such that one is located at the beginning of the contig, the second one is in the middle and the third one is at the end. Distances in base pairs between the landmark genes were measured. The contig1 resulted from each combination of computation parameters was compared to each other using the default parameters of NCBI BLASTN tool. Finally, the separation distance in base pairs among landmark genes of contig1 instances were measured and located.

By using the best assembly parameters assessed from the previous step, six iterative assembly versions were done. The first assembly (Version 0.5) used the reads of only one sequence flowgram format (sff) file of the resulted two from the first 454 sequencing run. The other five assemblies (Versions 1.0, 1.5, 2.0, 2.5 and 3.0) were performed by adding one sff file each time to increase the number of reads incorporated in the assembly gradually until uploading all the six sff files of the three pyrosequencing runs.

2.5 Annotation of the assembled contigs

The gene finding process was applied to the largest assembled metagenomic data set version 3.0 using a prokaryotic gene finding program, Metagene Annotator (MGA)[71] Which is a Linux command line program to extract putative open reading frames (ORFs) from fasta sequences. MGA does not require any option adjustments for its running. Automatic annotation of these ORFs using MG-RAST server version 3.3.0.6 was performed using the default parameters. Also, NCBI Blastx tool was used to align ORFs against NCBI protein database and create a best hit file using E-value of 1e-5.

Artemis[72] is a java application for sequence annotation and visualization developed by Wellcome Trust Sanger Institute, release 13.2.0 has been used for the manual annotation of the genes on contig 287 that carried the operon for mercuric resistance.

2.6 Functional and taxonomic analysis

MG-RAST server version 3.3.0.6 was used for functional and taxonomical analysis of the ORF sequences identified in AT-II LCL assembled dataset. M5nr database of MG-RAST is an integrated collection of multiple databases in a single container that allows performing similarity search for any sequence from many protein databases at the same time. M5nr database has been updated February 22, 2011 including Greengenes;16S rRNA Gene Database, JGI;Joint Genome Institute, KEGG; Kyoto Encyclopedia of Genes and Genomes, NCBI; National Center for Biotechnology Information, RDP; Ribosomal Database Project, SEED; The SEED Project, SILVA; SILVA rRNA Database Project, UniProt; UniProt Knowledgebase, VBI; Virginia Bioinformatics Institute and eggnog; evolutionary genealogy of genes Non-supervised Orthologous Groups. MG-RAST program parameters were adjusted to match the application of assembled sequences by removing preprocessing and dereplication filters that are usually applied to reads.

All ORFs (87,357) from the assembled contigs were compared to NCBI nr database (version date December-2012) using BLASTX tool. Default parameters for BLASTX were applied and the number of best hits retrieved per sequence was set to 100. BLASTX results were uploaded to Megan as an input for the analysis process. Megan software version 4.69.4, built 5 Jul 2012 [73] was used for both the functional and taxonomical analysis of the identified ORFs. The program parameters applied were minScore=50.0, topPercent=10.0, winScore=0.0, minSupport=5, minComplexity=0.0. Megan taxonomy database was downloaded from ftp://ftp.ncbi.nlm.nih.gov/pub/ taxonomy/taxdmp.zip on June 15, 2012. Megan SEED tree was created on May 17, 2010. Megan total reference classifications are KEGG (1,791 classes) SEED (2,607 classes) and Taxonomy (658 classes).

3. Results

Six standard flowgram format files (sff) were produced from the three independent 454 sequencing runs, two files per run, from the single metagenomic DNA library of the Red Sea Atlantis II lower convective layer (ATII-LCL). More than one million reads were extracted from the two sff files of each run, (Table 1). A total of 4,104,994 reads constituting 1.5 Gbp were resulted from the three 454 pyrosequencing runs. The average read length is 560 bp and the average GC percentage of all the reads is 52%.

Table 1: The impact of sequencing depth on the size of the largest twenty four assembled contigs.

Sequencing Run	**Run 1**		**Run 2**		**Run 3**	
Number of reads / sff file	655,292	682,314	579,504	475,690	871,020	841,174
Number of reads / sequencing run (bp)	1,337,606		1,055,194		1,712,194	
Summation of all reads (bp)	1,337,606		2,392,800		4,104,994	
Number of bases / sequencing run (bp)	524,204,317		351,848,418		672,581,733	
Summation of all bases (bp)	524,204,317		876,052,735		1,548,634,468	
Average read length (before QC)	560		576		555	
Average read length (after QC)	310		255		295	
GC%	52 ± 7%		52 ± 7%		52 ± 7%	

3.1 Assessment of the best computation parameters applied in the assembly

The reads of the first 454 sequencing run of ATII-LCL (1.3 Mbp) were assembled using the default computation parameters of Newbler software producing 28,547 contigs with largest contig size of approximately 236 kb. To test for the impact of using more stringent computation parameters on the quality of the assembly results, different combinations of minimum overlap length and minimum identity percentages were applied. The best assembly results were obtained from the setup of 40 bp minimum overlap length and 98% minimum identity percentage where the largest contig size increased to about 350.9 kb and the total number of assembled contigs decreased to 23,843, (Table 2).

Table 2: The effect of modifying the assembly computation parameters on the quality of results obtained.

Computation parameters*	90/40	95/40	95/50	98/40	98/50
Total number of reads	1,337,606	1,337,606	1,337,606	1,337,606	1,337,606
Total number of assembled reads	1,142,385	1,114,351	1,114,655	1,022,137	1,021,609
Singletons	72,261	84,687	84,495	104,048	104,549
Largest contig Size (bp)	236,358	259,994	303,963	350,934	350,934
Average contig size(bp)	1,912	1,831	1,825	1,838	1,837
Total number of contigs	28,547	24,895	25,142	23,843	23,863
N50 contig size	2,622	2,546	2,538	2,563	2,563

* Minimum identity percentage and minimum overlap length are the two computation parameters of Newbler assembler v.2.6 that have an impact on the assembly quality. Default setup is 90% for minimum identity and 40 bp of the minimum overlap length.

The results of the extension pattern evaluation for the assembled contigs due to the increase in stringency of the computation parameters are summarized in Figure 4. The distances in base pairs between landmark genes (100,341 bp and 130,043 bp respectively) remained unchanged regardless of the applied parameters. However, contig 1 has extended almost from only one edge according to the increase in the assembly stringency. This was confirmed by a pairwise alignment of contig1of default parameters with the other high stringent instances of the same contig which showed a 100% similarity of the first version of contig1(shortest contig) with the comparable part in the extended version.

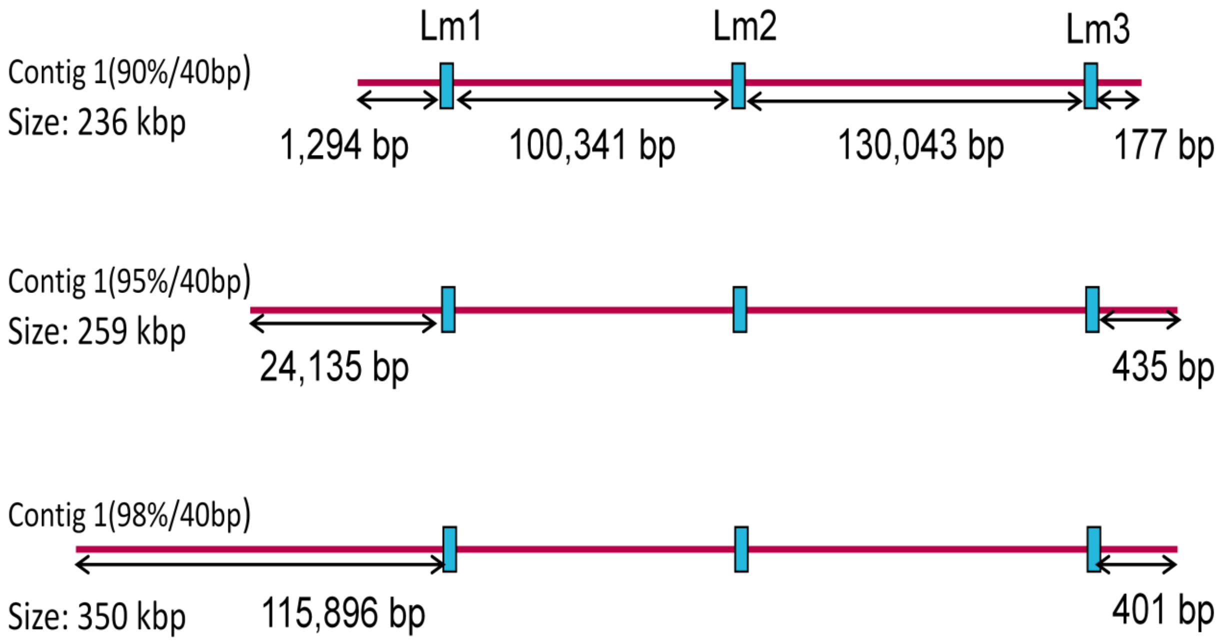


Figure 4: the pattern of extension in contig1 due to the change in assembly computation parameters

Lm1: Landmark gene1, Phosphatidate Cytidylyltransferase (Blast hit data: Positives 86%, e-value 1e-147, Query size 942 bp)

Lm2: Landmark gene2, Transketolase (Blast hit data: Positives 84%, e-value 0.0, Query size 2025 bp)

Lm3: Landmark gene3, FAD dependent Oxidoreductase (Blast hit data: Positives 87%, e-value 0.0 , Query size 1281 bp)

3.2 Analysis of the Assembly results

Reads from the three independent pyrosequencing runs of ATII-LCL were assembled using the optimized computation parameters from the previous step and the resulted datasets were compared in (Table 3).

Table 3: A comparison of the assembly results for the three ATII-LCL independent 454-sequencing runs.

Assembly Version	V1	V2	V3
Total number of reads	1,337,606	2,392,800	4,104,994
Total number of assembled reads	1,022,137	2,214,457	3,844,674
% assembled	88.94%	92.55%	93.66%
Total number of singletons	104,048	127,692	174,011
% singletons	7.7%	5.3%	4.2%
Largest contig size (bp)	350,934	351,272	350,936
Average contig size(bp)	1,838	2,085	2,185
N50 contig size	2,563	3,168	3,525
Total number of contigs	23,843	32,885	40,693

The table shows a slight increase in the percentage of assembled reads from about 89% in version 1.0 to about 93% in version 3.0 which is accompanied by a raise in the average contig length from 1,838 bp to 2,185 bp. However, the largest contig size remained almost constant at size about 350 kb. Also, the total number of assembled contigs has dramatically increased with the addition of more reads in each assembly from 23,843 in assembly version 1.0 to 40,693 in version 3.0. On the other hand, the percentage of singletons has decreased from 7.7% in the assembly of first 454-sequencing run to 4.2% as the reads of the third run were added to the assembly. In addition, N50 value has showed an increase from contig size of 2,563 bps in the assembly version1.0 to contig size 3,525 bps in version3.0. By comparing the distribution of the number of contigs in ten size ranges starting from 10 to 20 kb range to more than 100 kb (Table4 and Figure5), we found that the total number of contigs in all ranges increased from 237 contigs in V.1 assembly to 436 in V.3. However, the percentage of all bases constituting these contigs compared to the total size of the whole dataset only slightly increased from 28.55% in the first assembly version to 29.65% in version three. The plot for the distribution of the number of contigs throughout the size ranges started at 10 kb showed that about 55% of the total number of contigs in each assembly versions fell in the smallest contig size range from 10-20 kb. The remaining 45% were distributed in a descending order among all other size ranges except for the one larger than 100 kb.

Table 4: Number of contigs in different size ranges.

Contig Size(Kb)*	Assembly-V1	Assembly-V2	Assembly-V3
10 – 20	124	171	245
>20	41	51	78
>30	24	21	36
>40	11	10	21
>50	11	14	8
>60	7	9	8
>70	5	7	9
>80	3	4	5
>90	3	2	2
>100	8	17	24
Total number	237	306	436
Total bases in contigs>10Kb	7,281,862	9,793,161	13,640,785
Total bases in all contigs	25,502,497	36,716,981	46,004,852
% bases of contigs >10Kb	28.55%	26.67%	29.65%

By comparing the number of contigs larger than 100 kb in the three assemblies, the results indicated that the number of contigs increased with the addition of more reads from only 8 contigs in assembly version 1.0 to 24 contigs in version 3.0 (Table 4) and (Figure 5).

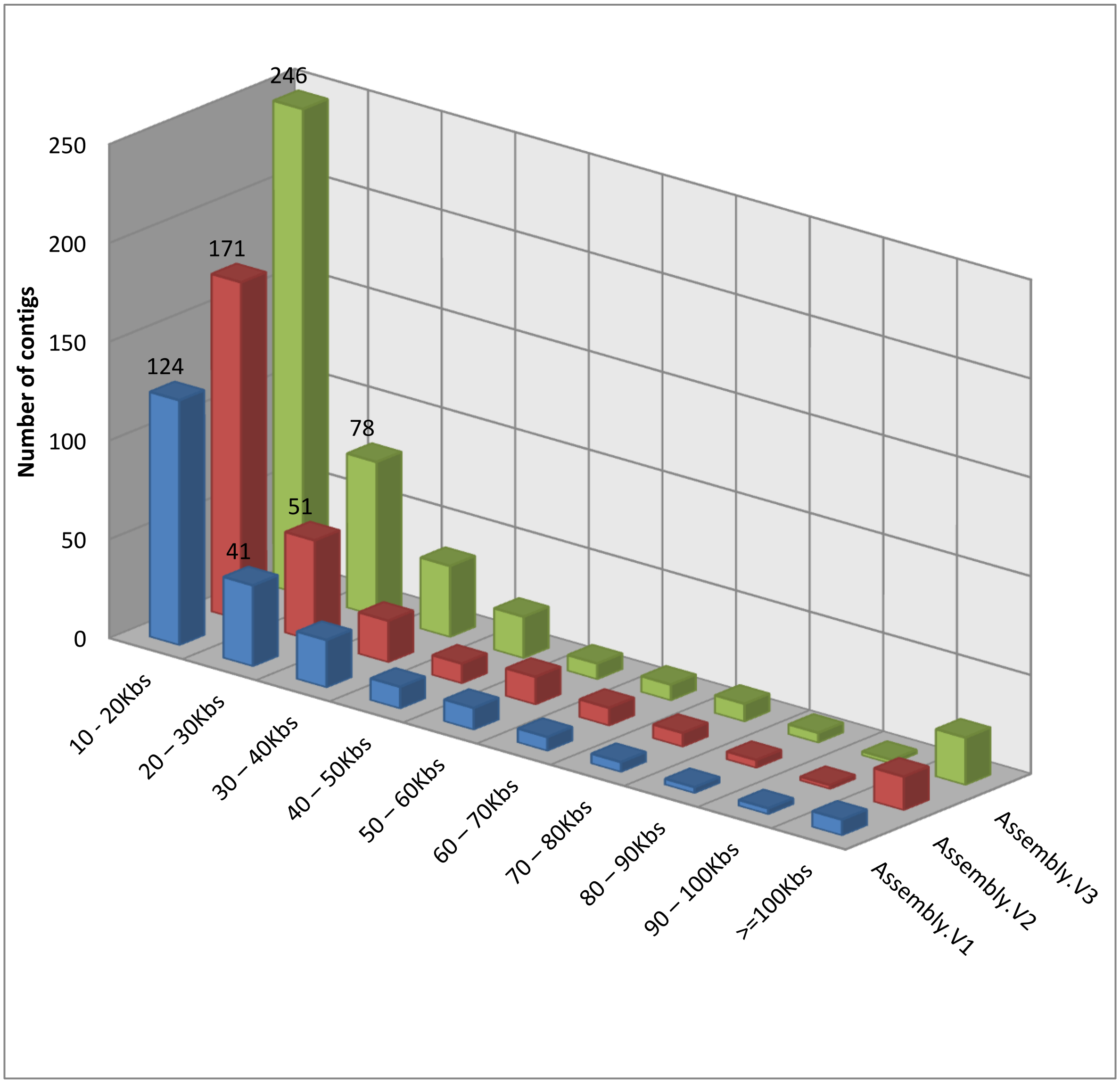


Figure 5: The number of contigs larger than 10 kb found in different size ranges of each assembly

The plot for the effect of sequence depth on the length of the largest 24 assembled contigs (figure 6) indicated a significant increase in the length of all the studied contigs per assembly round, except for contig 1. Interestingly, only contig1 has deviated from the pattern of extension of the assembled contigs by reaching its maximum length of about 350 kb after the addition of just the first two sff files. In other words, a depth of one sequencing run was enough for contig 1 to get its maximum length. Generally speaking, the largest twenty four assembled contigs gained an average length of approximately 30 kb at each sequencing depth starting from an average length of about 110 kb in assembly version 1 to an average length of 170 kb in version 3.

Although the absolute number of reads assembled in the largest twenty four contigs increased from one assembly to another (Appendix A), their relative percentages from the total assembled reads only slightly increased from 9.02% in assembly V 0.5 to 10.78% in V 3.0 with up and down fluctuations of these percentages throughout different assemblies (Figure 7). The relative percentage of reads assembled in contig 1 moved from 0.69% in assembly V0.5 to 1.76% at V1.0 and remained almost constant at such value for the other four assemblies (Table 5). Figure 8 indicates that although contig1 remained almost at size 350 kb from assembly version 1 to version 3 the number of reads incorporated in each assembly continued to increase.

Table 5: The relationship between contig1 length and its assembled reads**.**

Assembly version	V0.5	V1.0	V1.5	V2.0	V2.5	V3.0
Contig1 Size (bp)	112,239	350,934	350,937	351,272	350,934	350,936
Number of Reads/contig1	4,523	23,540	33,595	41,865	57,160	72,122
Total Number of Reads	655,289	1,337,597	1,917,096	2,392,780	3,263,795	4,104,966
% Reads	0.69	1.76	1.75	1.75	1.75	1.76

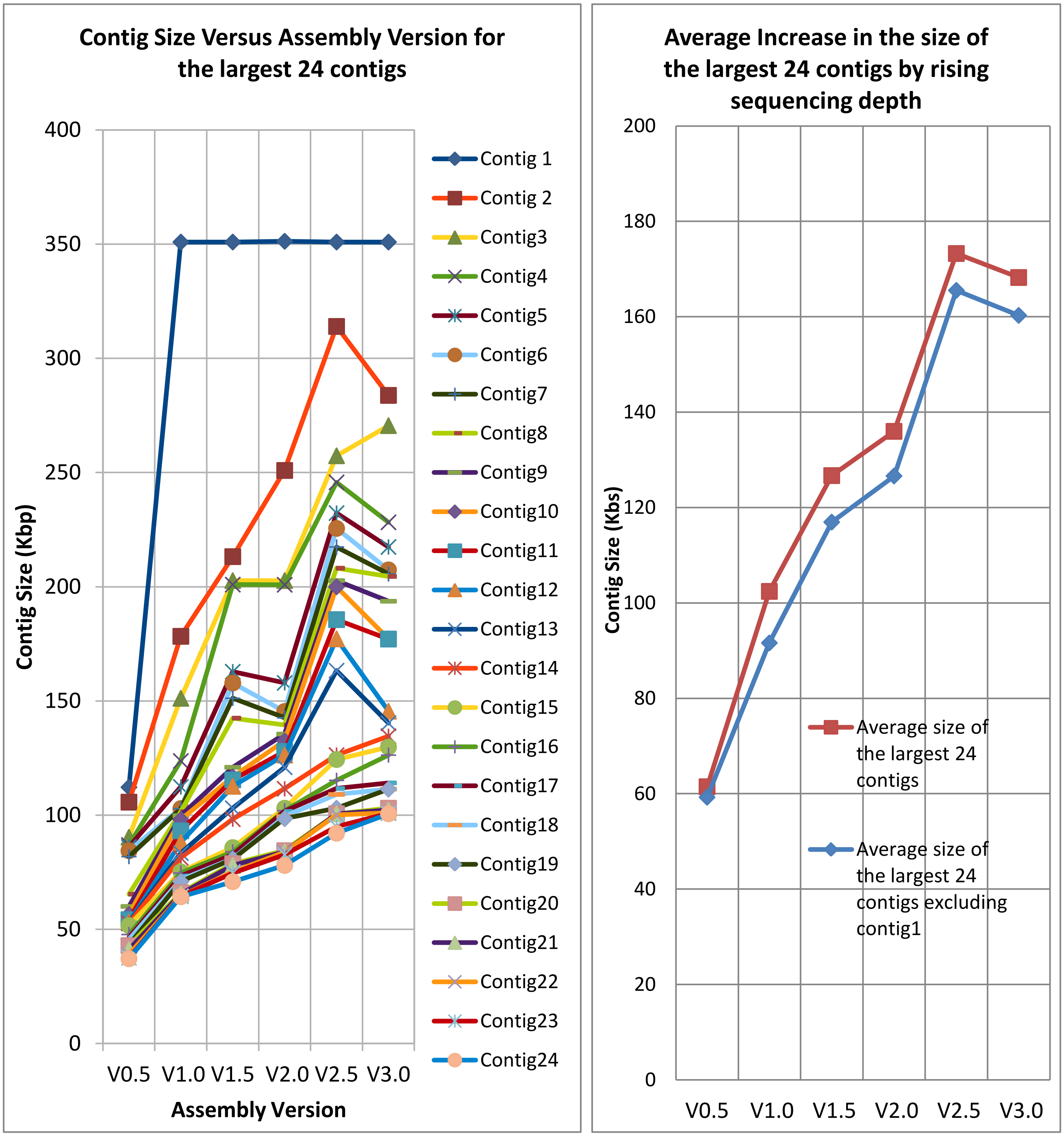


Figure 6: The impact of sequencing depth on the size of the largest twenty four assembled contigs.

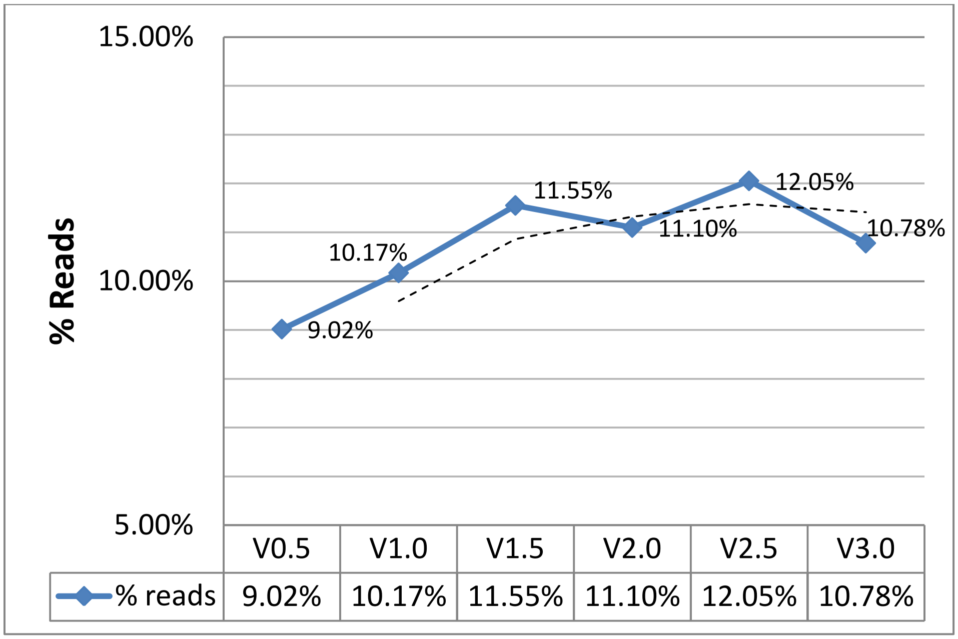


Figure 7: The percentage of reads incorporated in the largest 24 contigs in different assembly versions. Details for the number of reads incorporated are listed in AppendixB.

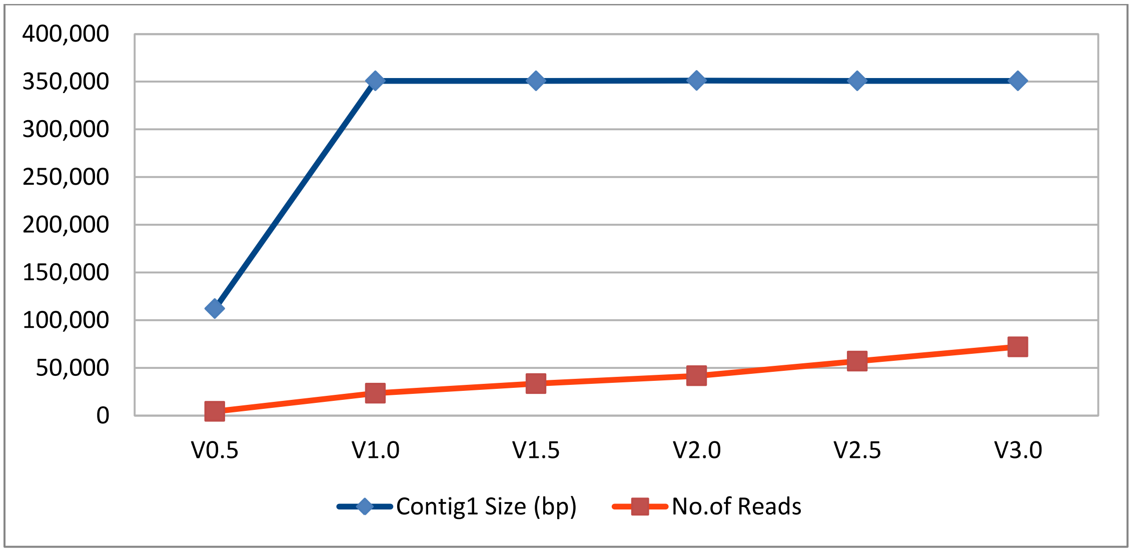


Figure 8: Comparison of Contig1 size(bp) and the Number of reads incorporated in each assembly run.

3.3 Annotation

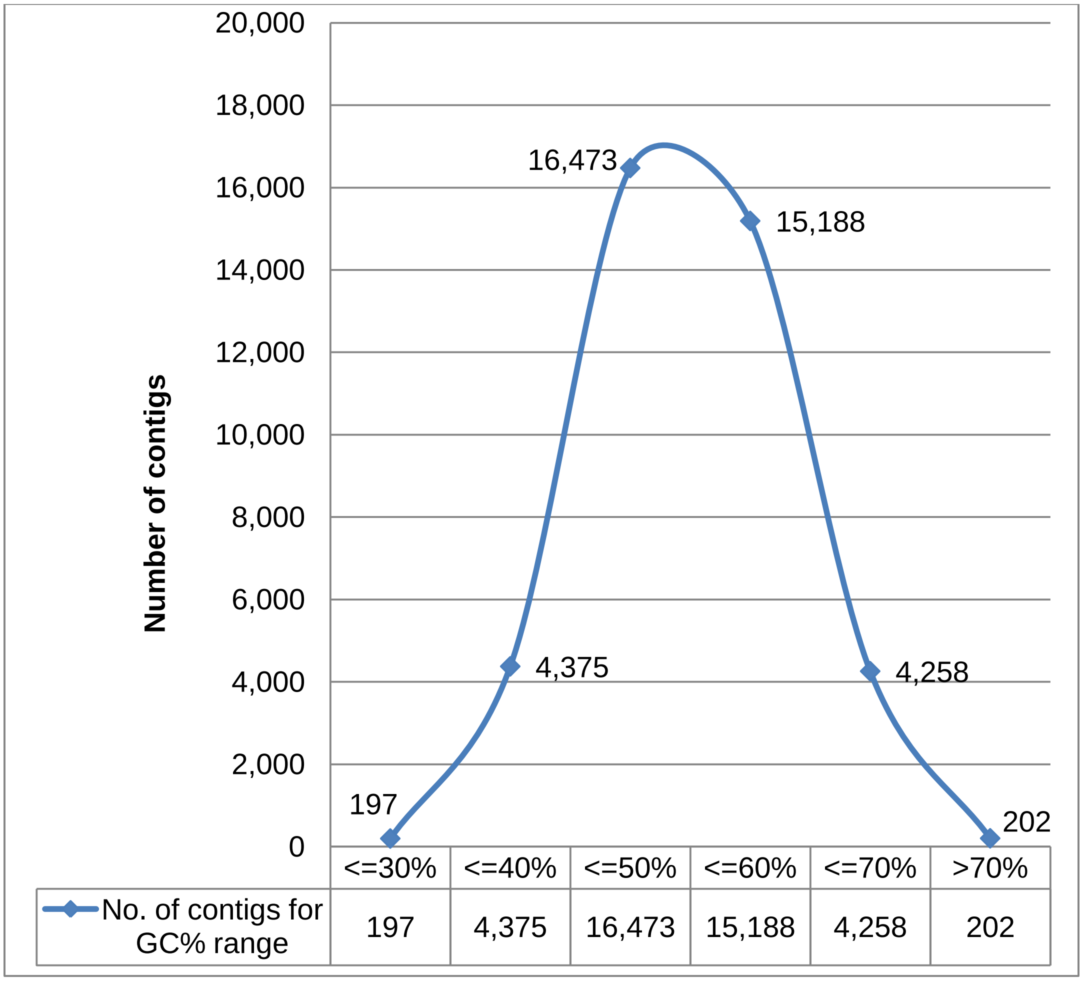
The number of ORFs resulted from the metagenomic assembled dataset v3.0 using MGA program was 87,357 DNA sequences of putative genes. MGA lists a summary for each contig at the top of the ORF table including contig number, its length, number of reads, average GC%, average score for ribosomal binding site (RBS) and the model used for comparison either bacteria (b) or archaea (a) (Table 6). The table created from MGA sequence file lists each ORF start and end positions on the contig, a confidence score for the predicted gene, RBS start and end positions if they are found and another confidence score for the identified RBS. (example in Table 6).

Table 6: Contig1as an example of ORFs identified from the assembled contigs**.**

```
# contig00001 length=350936 numreads=72122
# gc = 0.571352,rbs=0.78852
# self: b
```

GeneID	Start Position	End Position	Gene Score	RBS start	RBS End	RBS Score
gene_1	11	436	30.6199	444	449	2.94686
gene_2	436	1716	142.957	1722	1727	3.90458
gene_3	1726	2661	86.557	2666	2671	3.79324
gene_4	2658	3395	78.7786	3405	3410	5.2284
gene_5	3413	4747	82.1636	0	0	0
gene_6	4744	6222	118.115	6234	6239	0.279796
gene_7	6443	6898	36.1892	6909	6914	8.39031
gene_8	7066	7965	80.8288	7055	7060	6.74128
gene_9	8071	8274	22.7255	8057	8062	7.01345
gene_10	8319	9554	122.104	9562	9567	0.802451
gene_11	9700	11256	156.572	9685	9690	5.82891
gene_12	11273	11872	53.7414	11880	11885	-2.04487
gene_13	11977	13074	150.469	13083	13088	9.20518
gene_14	13071	14333	142.015	14342	14347	7.02112
gene_15	14480	15988	157.721	14470	14475	7.84855
gene_16	15989	17053	144.608	15974	15979	1.46408
gene_17	17114	18166	149.786	17102	17107	7.57462
gene_18	18224	20278	258.648	20285	20290	7.57462
gene_19	20356	21321	111.3	21328	21333	2.52163
gene_20	21490	23172	204.993	23179	23184	-2.02056

The plot of the distribution of different GC% ranges for all 40,693 contigs in the dataset shows a number of 31,266 (77.8%) contigs fall into 40-60 GC% range(Figure 9).



	<=30%	<=40%	<=50%	<=60%	<=70%	>70%
No. of contigs for GC% range	197	4,375	16,473	15,188	4,258	202

Figure 9: Number of contigs in each GC% range.

3.4 Functional classification

The results of the automatic annotation using MG-RAST server showed that 84,050 ORFs (about 96%) produced 58,162 predicted protein coding regions. Only 22,923 (39.4%) of the predicted features were successfully annotated to at least one protein of the M5NR (non-redundant multi-source protein annotation database) of MG-RAST server. The remaining number of features 35,230 (60.6%) had no hits in the protein database and were considered as orfans. Of the 22,923 annotated features, 18,262 (20.9%) were assigned to functional categories (Figure 10). The top ten categories of the plot represent about 60% of the total functional assignments.

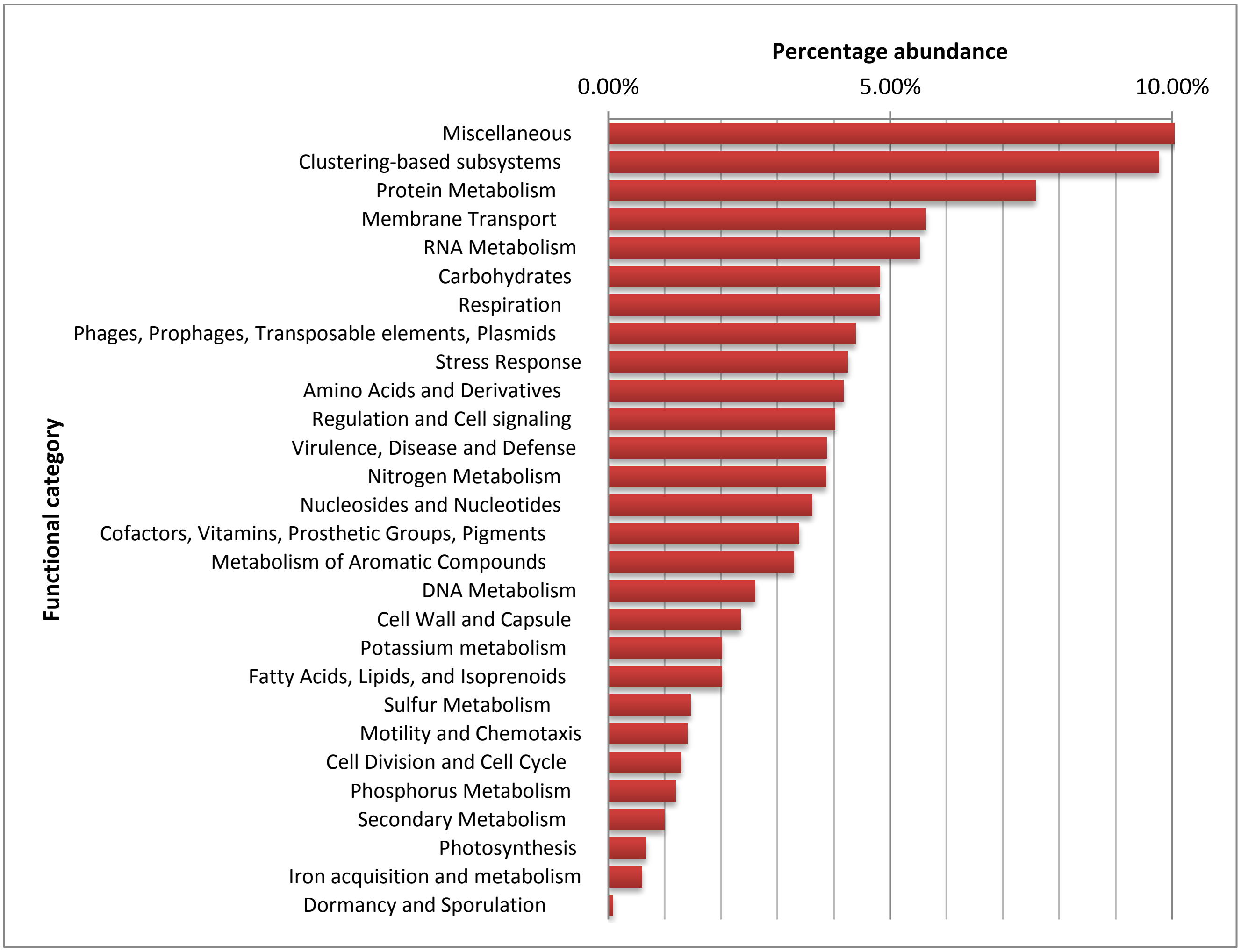


Figure 10: Functional profile for the ORFs of ATII-LCL using Subsystems classification of MG-RAST server. The data was compared to Subsystems using a maximum e-value of 1e-5, a minimum identity of 60 %, and a minimum alignment length of 15 amino acids. See Appendix C for detailed Assignments.

The results of the functional analysis of Megan using Seed subsystems showed that only 11,609 (13%) hits were assigned to functional classifications. The total number of both functional classifications of not assigned and no hits was 75,748 representing about 86% of the whole size of data submitted (87,357). The assigned hits using Megan were significantly less than those of MG-RAST (by 6,653). The functional profile by Megan, shown in Figure 11, lists the most abundant functions assigned to ATII-LCL. The first ten categories of the plot represent about 69% of the assigned functional categories.

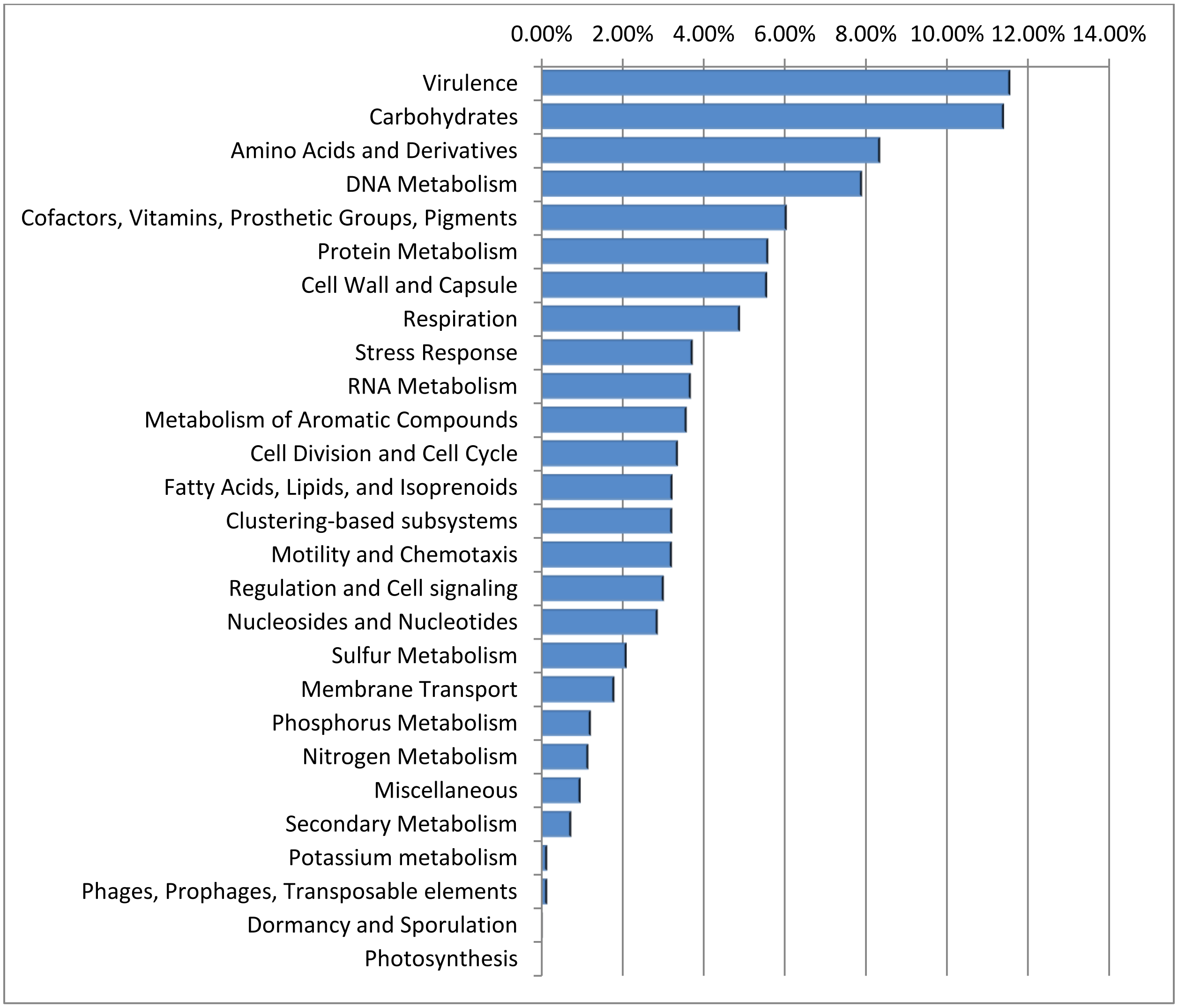


Figure 11: Functional profile of ATII-LCL ORFs using Seed system of Megan software. See Appendix D for the detailed functional assignments using Megan

The results of the functional analysis of the BLASTX hits for the assembled contigs using KEGG (Kyoto Encyclopedia of Genes and Genomes) classification of enzymes and pathways included in Megan indicated that only 9,161sequences (11.4%) were assigned to a KEGG functional category. Figure12 showed that the highest percentage of assignments occurred in the metabolism-related category (60.84%) and in environmental information processing category (18.76%).

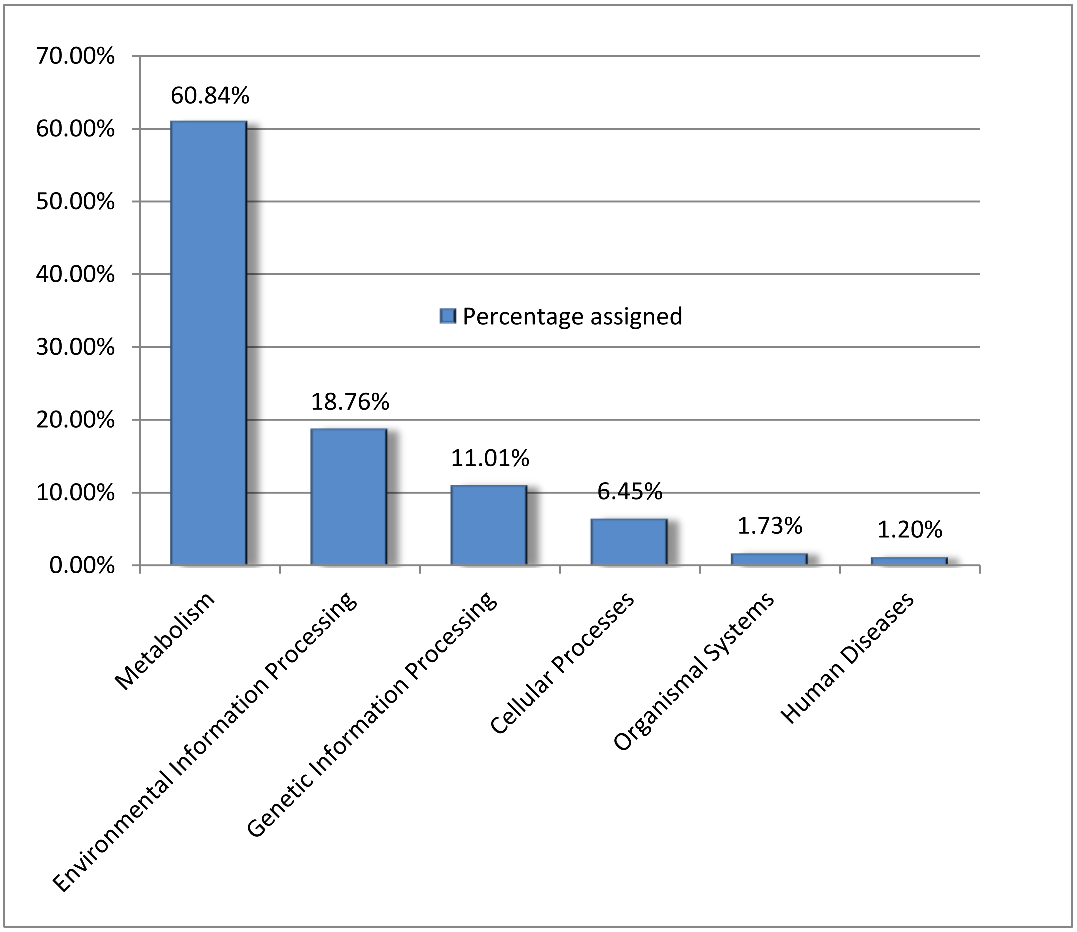


Figure 12: the percentage of KEGG terms identified for ATII-LCL Orfs using Megan

3.5 Taxonomy profile

Megan results showed that only 31,196 (35.7%) ORFs were successfully assigned to different taxonomic nodes from the total number of compared ORFs (87,357). However, the remaining ORFs are distributed between unassigned 24,873 (28.4%) and no hits 31,288 (35.8%) taxonomic categories Figure13. With respect to the phylogenetic tree drawn in Figures13, the distribution of the assigned ORFs to a specific taxonomic category indicated that Proteobacteria and Euryarchaeota are the most dominant phyla in the sample with total hits of 17,925 (57.4%) and 2,236 (6.9%) representing 64.5% of the assigned sequences to cellular organisms. The remaining percentage of the assigned hits (35.5%) was distributed insignificantly among other taxonomic classification (data were not included in Figure 13). The low level tree drawn in Figure14 indicated good recruitments at species level for bacterial strains from Rhizobiales *Phyllobacterium sp.* YR531 with 1,902 hits representing about 42% of the percentage hits assigned to the order (4072). Also, order Burkholderiales showed two representatives at species level *Cupriavidus basilensis* and *Ralstonia pickettii* with 3,013 and 590 assigned hits respectively representing a percentage of 36% of the hits assigned to this order. While in case of Pseudomonadales, the assignment resolution has been stopped at the order level and no hits could be mapped to any lower taxonomic level. The distribution of the hits assigned to different taxa, presented in Figure15, showed high abundance of sequences belonging to the orders Burkholderiales, Rhizobiales and Pseudomonadales with 9,890, 4,072 and 937 hits, respectively.

MG-RAST results showed that 47,747 (54.6%) sequences have been assigned to different taxonomic levels from the total 87,357 uploaded ORFs. Proteobacteria, Firmicutes and Euryarchaeota are the most abundant phyla with percentage assignments of 73.9%, 4.4% and 4.2%, respectively (Figure 16). By comparing results presented in Figures 13 and 14, for the taxonomical classification at the order level using either MG-RAST or Megan, we can clearly observe the similarity in the pattern for the first three most abundant orders in both analysis programs.

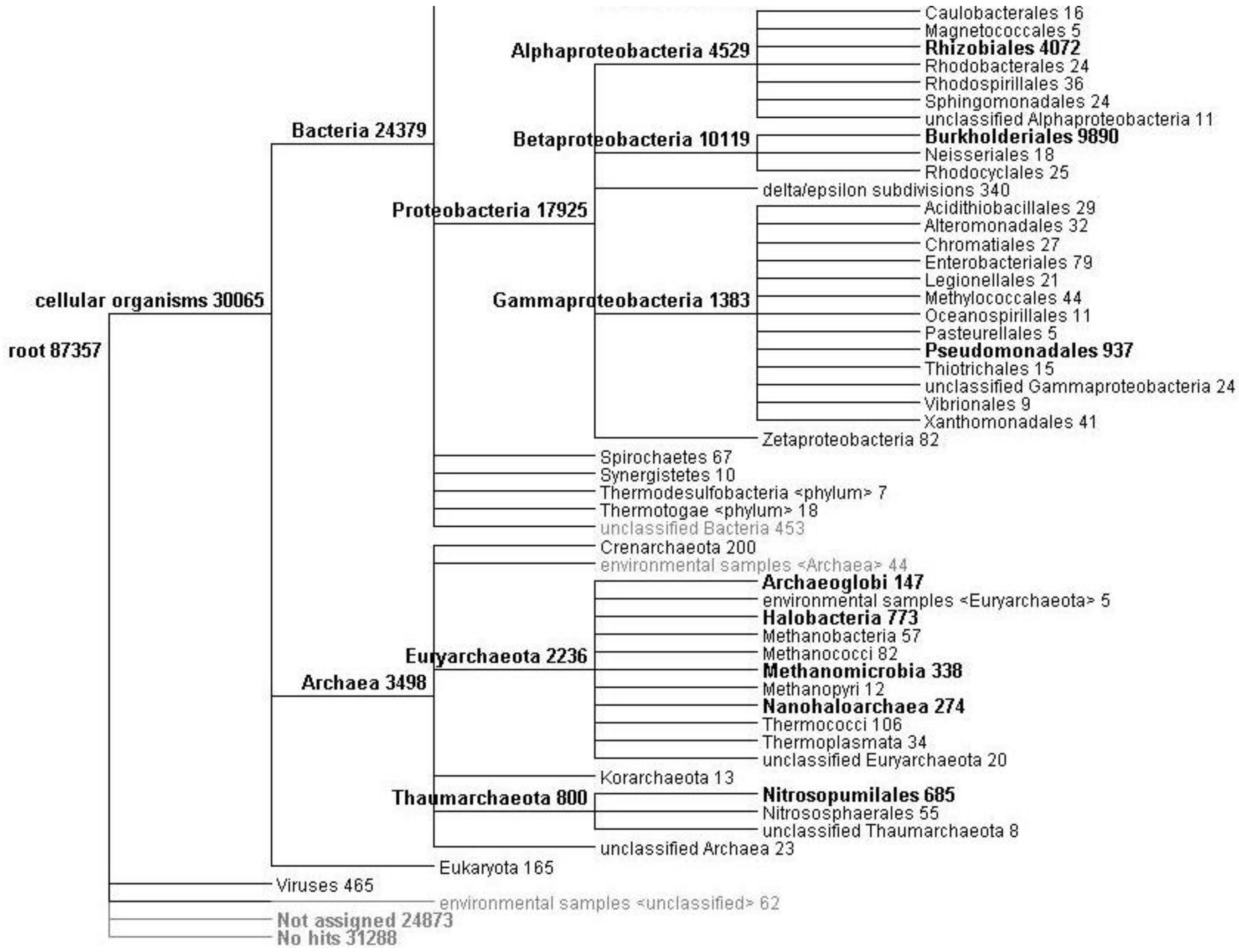


Figure 13: High level view of the Megan taxonomic analysis of the ATII-LCL assembled dataset based on BLASTX comparison of the 87,357 ORF sequences against NCBI-NR (December, 2012) database. Default program parameters are used except for minimum complexity which is set to 0.0 to allow for more hit assignments.

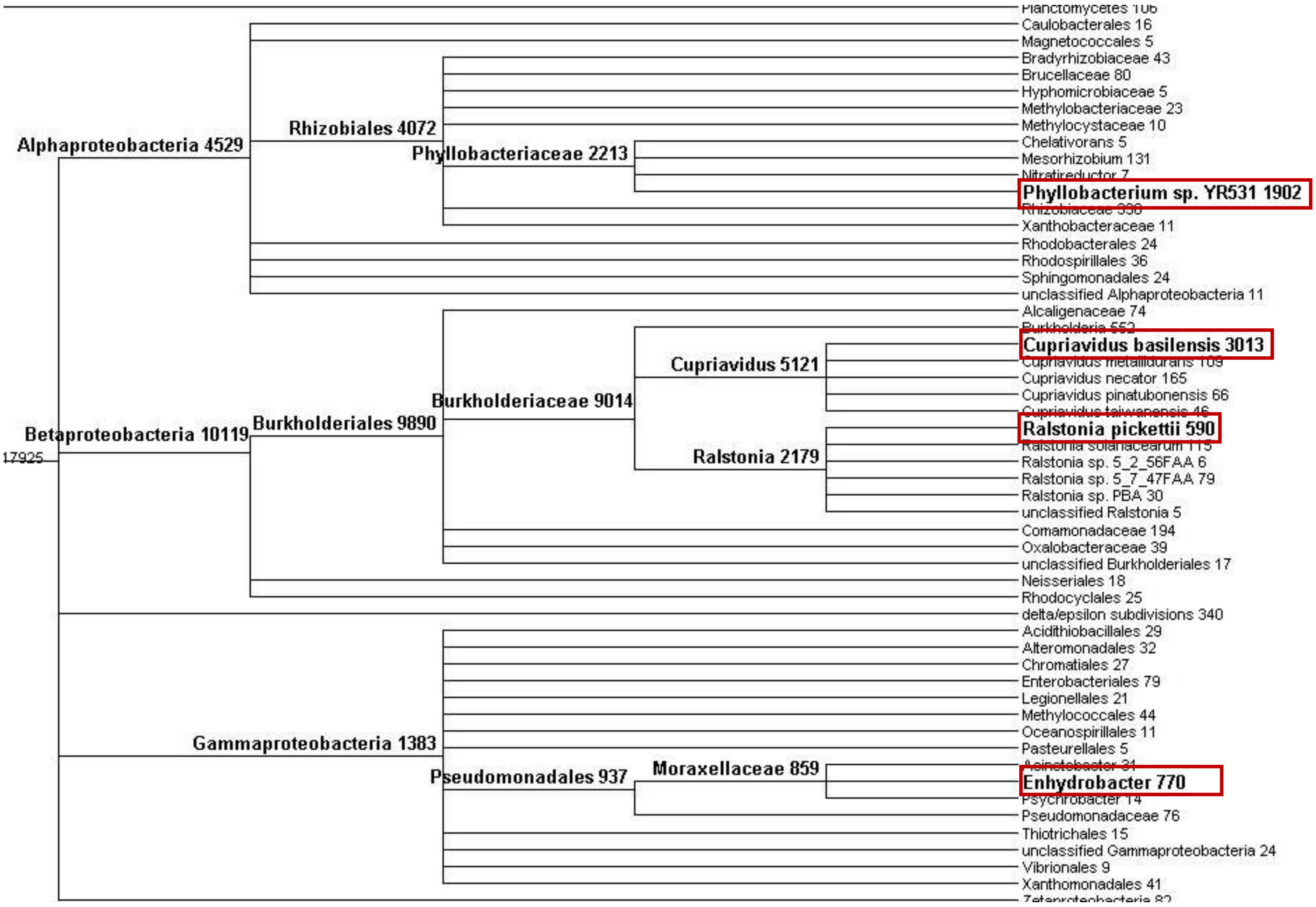


Figure 14: A low level view of the Megan analysis of the ORFs of ATII-LCL assembled dataset.

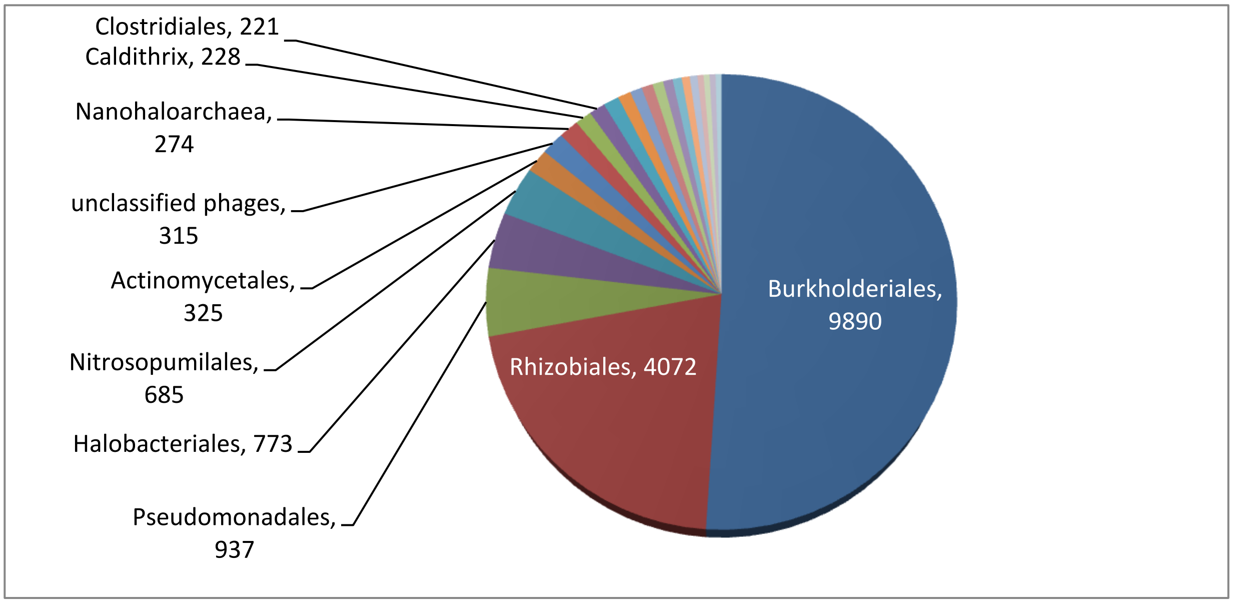


Figure 155: Phylogenetic diversity of the most abundant twenty taxa at the order level of the ORFs identified in the assembled contigs of ATII-LCL dataset based on Megan taxonomical analysis. The detailed table of the number of hits assigned to different taxonomic classifications is shown in Appendix F.

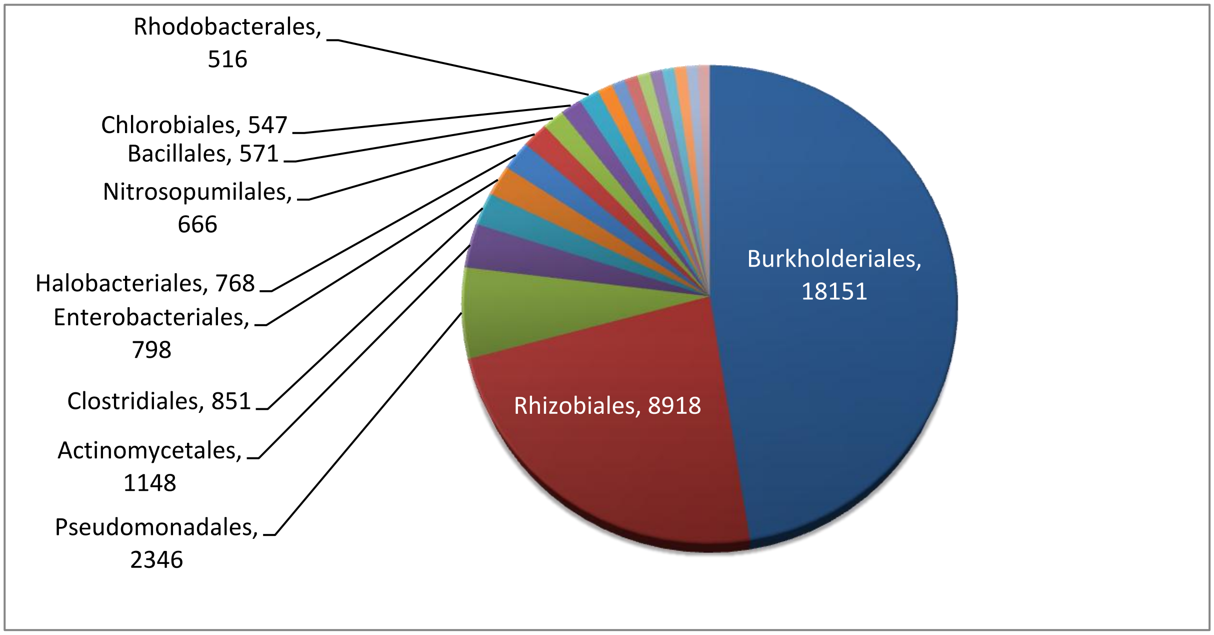
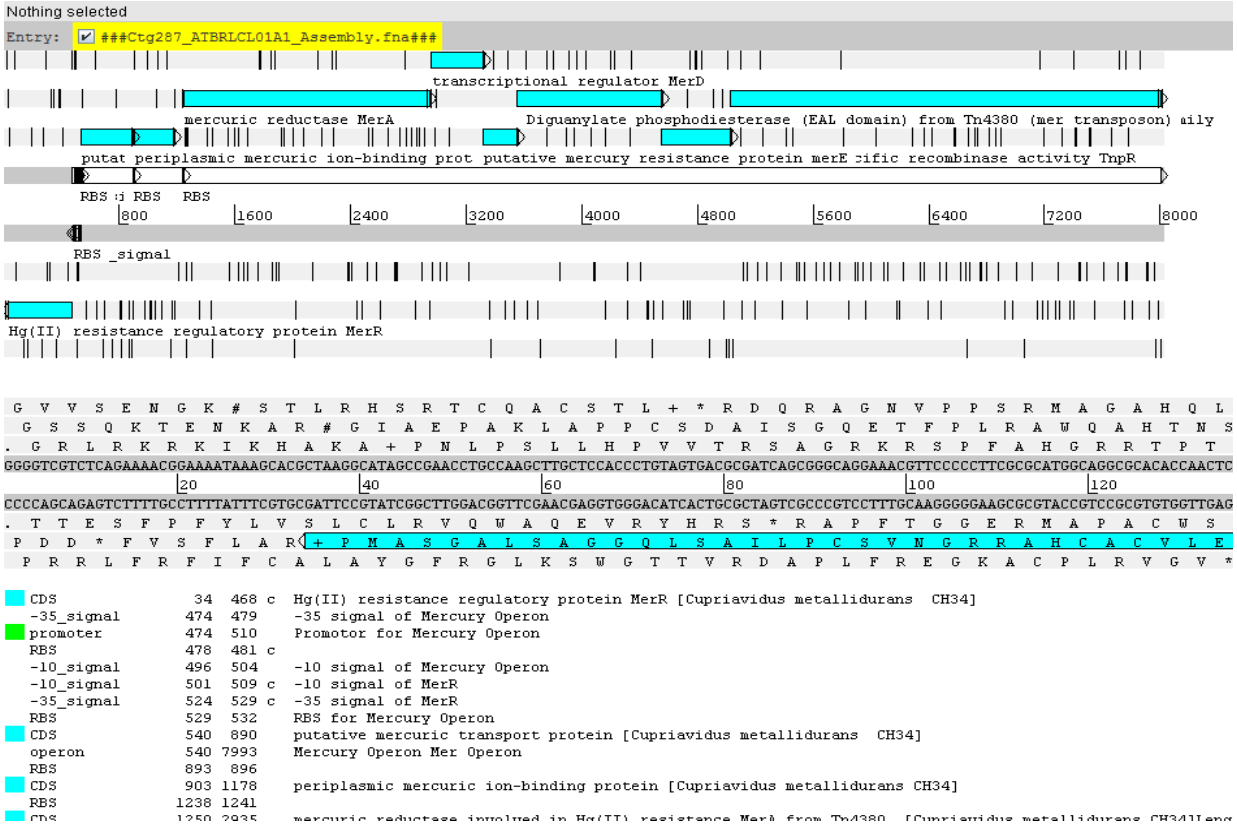


Figure 166: Phylogenetic diversity of the most abundant twenty taxa at the order level of the ORFs identified in the assembled contigs of ATII-LCL based on MG-RAST server taxonomical analysis. The detailed table of the number of hits assigned to different taxonomic classifications for the whole dataset is shown in Appendix E .

3.6 Application of the established dataset in a novel gene discovery, Mercuric reductase operon

BLASTX search was performed for mercury reductase gene (MerA) in the established metagenomic dataset. Unexpectedly, only one instance of MerA gene was retrieved from a small-sized contig (8000 bp). The annotation of this contig (Figure17) showed that it constitutes a complete operon for mercury detoxification process characterizing organisms live in environments with high concentration of heavy metals as in the case of ATII-LCL. The operon consists of 8 successive genes on the forward strand. The promoter area of the operon was located between the nucleotide number 474 and 510 including -10 and -35 sequences located at nucleotide numbers 496 and 474, respectively.


Nothing selected
Entry: ☑ ###Ctg287_ATBRLCL01A1_Assembly.fna###
transcriptional regulator MerD
mercuric reductase MerA
Diguanylate phosphodiesterase (EAL domain) from Tn4380 (mer transposon) aily
putat periplasmic mercuric ion-binding prot putative mercury resistance protein merE :ific recombinase activity TnpR
RBS :j RBS RBS
800
1600
2400
3200
4000
4800
5600
6400
7200
8000
RBS _signal
Hg(II) resistance regulatory protein MerR

```
 G   V   V   S   E   N   G   K   #   S   T   L   R   H   S   R   T   C   Q   A   C   S   T   L   +   *   R   D   Q   R   A   G   N   V   P   P   S   R   M   A   G   A   H   Q   L
   G   S   S   Q   K   T   E   N   K   A   R   #   G   I   A   E   P   A   K   L   A   P   P   C   S   D   A   I   S   G   Q   E   T   F   P   L   R   A   W   Q   A   H   T   N   S
.   G   R   L   R   K   R   K   I   K   H   A   K   A   +   P   N   L   P   S   L   L   H   P   V   V   T   R   S   A   G   R   K   R   S   P   F   A   H   G   R   R   T   P   T
GGGGTCGTCTCAGAAAACGGAAAATAAAGCACGCTAAGGCATAGCCGAACCTGCCAAGCTTGCTCCACCCTGTAGTGACGCGATCAGCGGGCAGGAAACGTTCCCCCTTCGCGCATGGCAGGCGCACACCAACTC
                  |20                 |40                 |60                 |80                 |100                |120
CCCCAGCAGAGTCTTTTGCCTTTTATTTCGTGCGATTCCGTATCGGCTTGGACGGTTCGAACGAGGTGGGACATCACTGCGCTAGTCGCCCGTCCTTTGCAAGGGGGAAGCGCGTACCGTCCGCGTGTGGTTGAG
.   T   T   E   S   F   P   F   Y   L   V   S   L   C   L   R   V   Q   W   A   Q   E   V   R   Y   H   R   S   *   R   A   P   F   T   G   G   E   R   M   A   P   A   C   W   S
 P   D   D   *   F   V   S   F   L   A   R < +   P   M   A   S   G   A   L   S   A   G   G   Q   L   S   A   I   L   P   C   S   V   N   G   R   R   A   H   C   A   C   V   L   E
   P   R   R   L   F   R   F   I   F   C   A   L   A   Y   G   F   R   G   L   K   S   W   G   T   T   V   R   D   A   P   L   F   R   E   G   K   A   C   P   L   R   V   G   V   *
```

```
🟦 CDS               34   468 c  Hg(II) resistance regulatory protein MerR [Cupriavidus metallidurans  CH34]
   -35_signal       474   479    -35 signal of Mercury Operon
🟩 promoter         474   510    Promotor for Mercury Operon
   RBS              478   481 c
   -10_signal       496   504    -10 signal of Mercury Operon
   -10_signal       501   509 c  -10 signal of MerR
   -35_signal       524   529 c  -35 signal of MerR
   RBS              529   532    RBS for Mercury Operon
🟦 CDS              540   890    putative mercuric transport protein [Cupriavidus metallidurans  CH34]
   operon           540  7993    Mercury Operon Mer Operon
   RBS              893   896
🟦 CDS              903  1178    periplasmic mercuric ion-binding protein [Cupriavidus metallidurans CH34]
   RBS             1238  1241
🟦 CDS             1250  2935    mercuric reductase involved in Hg(II) resistance MerA from Tn4380  [Cupriavidus metallidurans CH34]Leng
```

Figure 17: Complete annotation of Mercury resistance operon located on contig 287 of ATII-LCL assembled dataset

MerA gene is the most important component of the mercury detoxification process since it is the gene responsible for converting divalent mercury molecules (HgII) into a volatile reduced form (Hg0) which can be easily eliminated (Figure18). Interestingly, a trnasposase gene (TnpA) was found at position 5027 as part of the mercury resistance operon. Having this mercury transposon and a single copy of the mercury operon in the whole assembled dataset suggests that horizontal gene transfer has a role in the polymorphisms expected of the mercury resistance operons in different organisms inhabiting this environment. Also, it explains the assembly of the closely related MerA genes from different species in one contig despite of using high strigency identity percentage (98%).

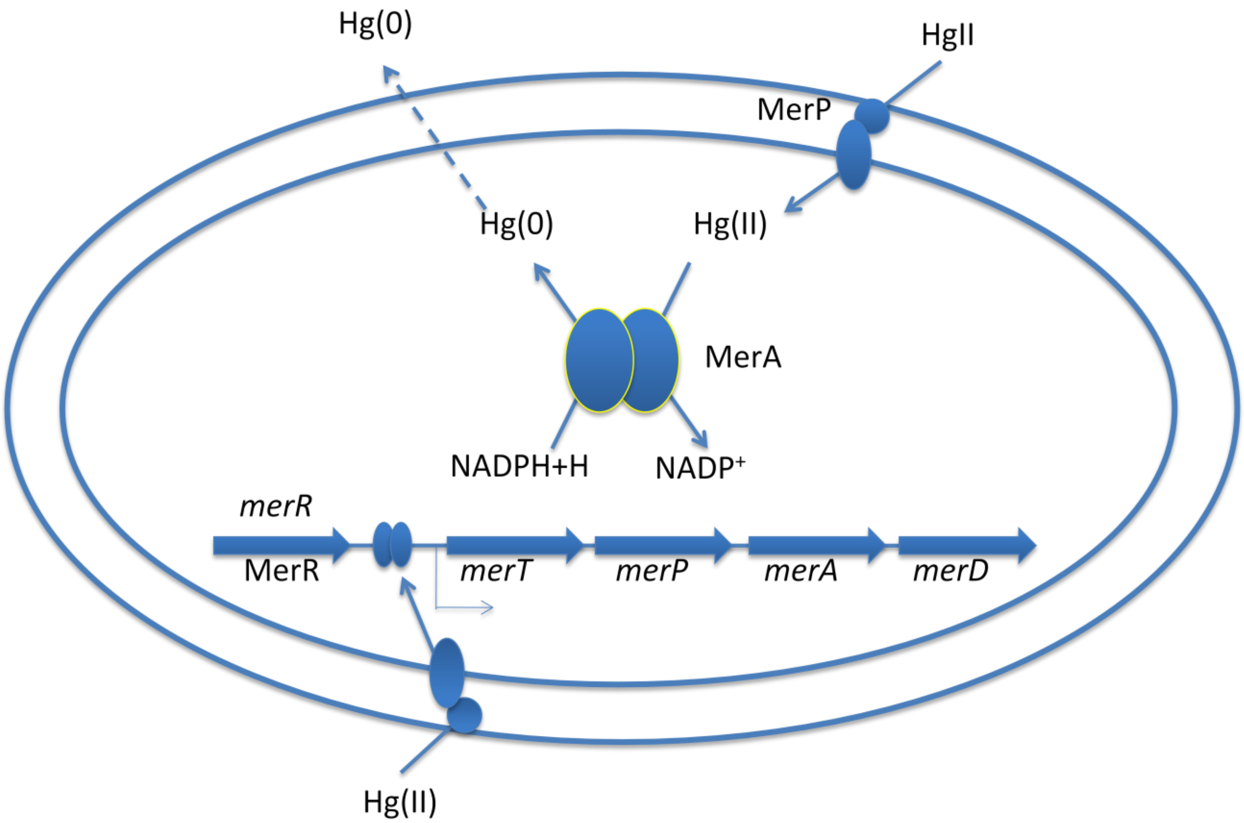


Figure 18: Diagram Showing the role of Mercuric resistance operon in the heavy metal detoxification process in the cell.

4. Discussion:

This is the first study for establishing an assembled metagenomic dataset from the lower convective layer of the Red Sea Atlantis II deep. The constructed dataset can be used as a tool to facilitate finding novel genes and operons for specific functions or even a partial genome assembly from such a unique and unexplored environment.

In an attempt to increase the sequence coverage to overcome the known challenge of undetermined relative abundance and variable species composition in complex metagenomic samples[50], we performed multiple independent pyrosequecing runs from the environmental DNA of the ATII-LCL. The budget limitation was the major constraint for the number of random shotgun sequencing rounds generated. However, we were able to reach a considerable total base pair count of 1.54 billion base pairs which is larger than the data size (1.045 billion base pairs) of a milestone study regarding metagenomic assembly[21].

In this study, the evaluation of the quality of the assembled metagenomic dataset was a challenging step because of the complexity of the community structure and the absence of reference genomes. Accordingly, we started by testing the effect of different computation parameters of the assembly program on the contig length. It was indicated that the more stringent computation parameters used, the better the assembly results with regard to the contig size. This finding agrees with previously described results[50] that at low coverage the assembly settings have a significant effect on the quality of the results [74] .However, the number of singletons increased dramatically with the elevation of stringency, what can be interpreted as the disassembly of reads with lower minimum overlap and minimum identity values from the assembled contigs due to the increase in confidence of overlaps. As a result, a minimum overlap length of 40 bp and a minimum overlap identity percentage of 98% were selected as the best assembly parameters to be applied since they generated the longest assembled contigs. Also, it was important to have an insight on the pattern of contig extension due to the change in the assembly computation parameters. Our results showed that contig1 was extended from the edges, indicated that assembly affected only the extremities of the contig not the middle part. In other words, the reassembled contigs are not completely rearranged when the assembly parameters are modified and the changes occur mostly at the edges of the contigs.

As it was demonstrated from the simulated data studies[59], [60] with regard to the enhancing effect of sequencing depth on metagenomic assembly , we performed three rounds of 454 shotgun sequencing which were used to build six cumulative assembled datasets from AT II-LCL sample. The growth in the average contig size and N50 value throughout different assembly versions confirmed the improvement of assembly quality with increasing the sequence depth. Also, the reduction in the number of singletons which was associated with a raise in the percentage of assembled reads indicated that the sequence coverage improved as more reads were added. Furthermore, the number of the assembled contigs increased with the increase in the read depth which suggested that a full coverage was not reached reached and more sequencing would be required.

Although all the contigs larger than 10 kb maintained an extension pattern as sequencing depth increased, the largest contig (contig 1) reached its maximum unchanged length of 350 kb after adding all the reads of the first sequencing run. The phenomenon can be explained by the accumulation of many repeat sequences from the most abundant genomes of the metagenomic dataset in the largest contig (contig 1) creating a barrier of chimeric sequences that prevented any further growth of the contig. This also was confirmed from our inspection of the reads incorporated in contig1 at each assembly version which showed an increase in the number of reads without an effect on the contig size indicating that these reads are accumulated duplicates. The estimated percentage of the reads in contig1 with reference to the total assembled reads showed an almost the same percentage (1.67%) starting from assembly version 1 to version 3 confirming our hypothesis. Moreover, the previously published work for the assembly of metagenomic simulated data showed that Newbler software has a percentage of forming chimeric contigs ranges from 3.88% to 12.57% of the total assembled reads depending on the degree of complexity of the metagenome[59]. Also, it was demonstrated that chimericity is more concentrated in short contigs of low complex metagenomes and occurs more in larger contigs as the community structure becomes more complex[59].

The GC% of most of the assembled contigs fall in the same range as those of sequenced reads (52% ± 7) which means that assembly does not alter sequence composition and maintains sequence characteristics of the reads incorporated. The large number of identified open reading frames (87,357) from the assembled contigs reflects the structural complexity of the community forming the studied metagenomic sample.

The low percentage of sequences incorporated in the functional analysis either based on MG-RAST 18,262 (20.9%) or, Megan 11,609 (13%) can be interpreted as the lack of enough reference sequences in databases to compare with and suggests a high novelty in the genomes of this community. Also, it can be explained as the insufficient sequence coverage for all the community members that resulted in inaccurate ORF prediction of the assembled contigs which in turn leads to wrong or missing assignments to different functional classifications. The difference between MG-RAST and Megan in the amount of sequences assigned to a functional category is expected to be due to the difference in size of reference databases used by each program where M5NR of MG-RAST is a multisource protein database expected to have more records than those of NCBI-NR database used by Megan.

In case of taxonomical classification of our dataset, the pattern of both programs showed similar abundance level in most cases regardless of the number of assigned hits where Burkholderiales, Rhizobiales and Pseudomonadales are the most abundant orders. This confirms the correctness of the resulted distribution and also, highlights the high complexity of our dataset. It is worth mentioning here that the presence of large volume of the studied sequences in the “not assigned” and “no hits” categories suggests the lack of reference taxa to compare with, in case of “no hits”, but it recommends the need of more sequencing to improve the alignment of the not assigned hits.

5. Conclusion and Future prospects

In conclusion, we succeeded in this study to establish an assembled metagenomic dataset of the Red Sea Atlantis II deep lower convective layer. The work is considered the first step towards the establishment of a large database of assembled datasets from a novel and unexplored environment which will provide a tool for further studies regarding the community structure, function, and mechanisms that adapt microbial community to survive in these exceptional harsh conditions.

We used deep sequencing approach trying to overcome the problem of undetermined coverage and relative abundance of the genomes forming this community. Also, we attempted to standardize computation parameters that can be used to produce high quality assembly. Unfortunately, we couldn't reach enough contig length to assemble potential partial genome of the most abundant organism.

We gave an insight to the functional and taxonomic structure of the microbial community inhabiting this environment which showed a high diversity and structure complexity. Also, large portion (more than 60%) of our data could not be assigned to any protein coding features in any of the protein databases available which points to a high probability of novel genes that have not been studied before. Moreover, we were able to classify the assembled data to the order level only due to high chimericity of the assembled contigs.

In addition, our dataset contributed in a practical application by providing a complete annotation of an interesting operon for the metabolism of heavy metals (mercuric reductase) which have an important biotechnological application in bioremdiation. The protein coded by the mercuric reductase gene was expressed and functionally characterized proving that it has a high activity compared to those of terrestrial origin.

For the future work, I suggest to have an assembly pipeline consisting of high computation resources in addition to a series of different assembly software Metavelvet, MAP, in addition to Newbler. Having different assembly software will help in gaining benefits of each and all the results can be pooled together to have the best assembled dataset. Also, it was demonstrated that paired-end sequencing helps dramatically in improving metagenomic assembly and in genome construction. Moreover, having more samples from the same site ATII-LCL will facilitate the extraction of more DNA and performing deeper sequencing that allow more coverage of this complex community.

References:

1. Wooley, J.C., A. Godzilk, and I. Friedberg, *A Primer on Metagenomics.* PLoS Computational Biology, 2010. **6**(2): p. 1-13.
2. Fiers, W., et al., *Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene.* Nature, 1976. **260**(5551): p. 500-507.
3. Sanger, F., et al., *The nucleotide sequence of bacteriophage phi-X174.* Journal of Molecular Biology, 1978. **125**(2): p. 225-246.
4. Fleischmann RD, A.M., White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al., *Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.* Science, 1995 July 28. **269**(5223): p. 496-512.
5. Amann, R.I., W. Ludwig, and K.-H. Schleifer, *Phylogenetic Identification and In Situ Detection of Individual Microbial Cells without Cultivation.* Microbiological Reviews, 1995. **59**(1): p. 143–169.
6. Pirajno, F., *Metalliferous Sediments and Sedimentary Rock-Hosted Stratiform and / or Stratabound Hydrothermal Mineral Systems*, in *Hydrothermal Processes and Mineral Systems*. 2009, Springer Science+Business Media B.V. p. 727-883.
7. Swift, S.A., A.S. Bower, and R.W. Schmitt, *Vertical, horizontal, and temporal changes in temperature in the Atlantis II and Discovery hot brine pools, Red Sea.* Deep-Sea Research I, 2012. **64**: p. 118-128.
8. ANSCHUTZ, P., et al., *Geochemical dynamics of the Atlantis II Deep (Red Sea): II. Composition of metalliferous sediment pore waters.* Geochimica et Cosmochimica Acta, 2000. **64**(23): p. 3995-4006.
9. Antunes, A., D.K. Ngugi, and U. Stingl, *Microbiology of the Red Sea (and other) deep-sea anoxic brine lakes.* environmental Microbiology reports, 2011. **3**(4): p. 416-433.
10. Antunes, A., et al., *A New Lineage of Halophilic, Wall-Less, Contractile Bacteria from a Brine-Filled Deep of the Red Sea.* American Society for Microbiology, 2008. **190**: p. 3580-3587.
11. Rusch, D.B., et al., *The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific.* PLoS ONE, 2007. **5**(3): p. e77.
12. Handelsmanl, J., et al., *Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products.* Chemistry and biology, 1998. **5**(10): p. R245-R249.
13. Qin, J., et al., *A human gut microbial gene catalogue established by metagenomic sequencing.* Nature, 2010. **464**(4): p. 59-65.
14. Zhu, X.Y., et al., *16S rRNA-Based Analysis of Microbiota from the Cecum of Broiler Chickens.* Applied and Environmental Microbiology, 2002. **68**(1): p. 124-137.
15. Gilbert, J.A. and C.L. Dupont, *Microbial Metagenomics: Beyond the Genome.* The Annual Review of Marine Science, 2011. **3**: p. 347-371.
16. Beja`, O., et al., *Bacterial Rhodopsin: Evidence for a New Type of Phototrophy in the Sea.* Science, 2000. **289**: p. 1902-1906.
17. Beja', O., et al., *Unsuspected diversity among marine aerobic anoxygenic phototrophs.* Nature, 2002. **415**: p. 630-633.
18. Grzymski, J.J., et al., *Comparative Genomics of DNA Fragments from Six Antarctic Marine Planktonic Bacteria.* APPLIED AND ENVIRONMENTAL MICROBIOLOGY, 2006. **72**(2): p. 1532-1541.
19. Nesbø, C.L., et al., *Lateral gene transfer and phylogenetic assignment of environmental fosmid clones.* Environmental Microbiology, 2005. **7**(12): p. 2011–2026.
20. DeLong, E.F., et al., *Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior.* Science, 2006. **311**(5760): p. 496-503

21. Venter, J.C., et al., *Environmental Genome Shotgun Sequencing of the Saragasso Sea.* Science, 2004. **304**(66): p. 66-74.

22. Yutin, N., et al., *Assessing diversity and biogeography of aerobic anoxygenic phototrophic bacteria in surface waters of the Atlantic and Pacific Oceans using the Global Ocean Sampling expedition metagenomes.* Environmental Microbiology, 2007. **9**(6): p. 1464-1475.

23. Karl, D.M., *Hidden in a sea of microbes.* Nature, 2002. **415**: p. 590-591.

24. Wilhelm, L.J., et al., *Natural variation in SAR11 marine bacterioplankton genomes inferred from metagenomic data.* Biology Direct, 2007. **2**(27): p. 1-19.

25. Neufeld, J.D., et al., *Marine methylotrophs revealed by stable-isotope probing, multiple displacement amplification and metagenomics.* Environmental Microbiology, 2008. **10**(6): p. 1526–1535.

26. Palenik, B., et al., *Coastal Synechococcus metagenome reveals major roles for horizontal gene transfer and plasmids in population diversity.* Environmental Microbiology 2009. **11**(2): p. 349–359.

27. Brazelton, W.J. and J.A. Baross, *Abundant transposases encoded by the metagenome of a hydrothermal chimney biofilm.* International Society for Microbial Ecology ISME, 2009. **3**: p. 1420–1424.

28. Huang, Y., et al., *Characterization of a Deep-Sea Sediment Metagenomic Clone that Produces Water-Soluble Melanin in Escherichia coli.* Marine biotecnology 2009. **11**: p. 124–131.

29. Siam, R., et al., *Unique Prokaryotic Consortia in Geochemically Distinct Sediments from Red Sea Atlantis II and Discovery Deep Brine Pools.* PLoS ONE, 2012. **7**(8): p. e42872.

30. Frias-Lopez, J., et al., *Microbial community gene expression in ocean surface waters.* PNAS, 2008. **105** (10): p. 3805–3810.

31. Thomas, T., J. Gilbert, and F. Meyer, *Metagenomics - a guide from sampling to data analysis.* Microbial informatics and Eperimentation, 2012. **2**(3): p. 1-12.

32. Abbai, N.S., et al., *Pyrosequence Analysis of Unamplified and Whole Genome Amplified DNA from Hydrocarbon-Contaminated Groundwater.* Molecular Biotechnology, 2012. **50**: p. 39-48.

33. Scholz, M.B., C.-C. Lo, and P.S. Chain, *Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis.* Current Opinion in Biotechnology, 2012. **23**: p. 9-15.

34. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc. Natl. Acad. Sci. USA, 1977. **74**(12): p. 5463-5467.

35. Sorek, R., et al., *Genome-wide experimental determination of barriers to horizontal gene transfer.* Science, 2007. **318**(5855): p. 1449-1452.

36. Goltsman, D.S.A., et al., *Community Genomic and Proteomic Analyses of Chemoautotrophic Iron-Oxidizing "Leptospirillum rubarum" (Group II) and "Leptospirillum ferrodiazotrophum" (Group III) Bacteria in Acid Mine Drainage Biofilms.* Applied and Environmental Microbiology, 2009. **75**(13): p. 4599–4615.

37. Ansorge, W.J., *Next-generation DNA sequencing techniques.* New Biotechnology, 2009. **25**(4): p. 195-203.

38. Jason R. Miller, Sergey Koren, and G. Sutton, *Assembly algorithms for next-generation sequencing data.* Genomics, 2010. **95**: p. 315–327.

39. Miller, J.R., S. Koren, and G. Sutton, *Assembly algorithms for next-generation sequencing data.* Genomics, 2010. **95**: p. 315-327.

40. Zhenyu Li, et al., *Comparison of the two major classes of assembly algorithms: overlap/layout/consensus and de-bruijn-graph.* Briefings in functional genomics, 2011. **II**(I): p. 25-37.

41. Margulies, M., et al., *Genome Sequencing in Open Microfabricated High Density Picoliter Reactors.* Nature, 2005. **437**(7057): p. 376–380.

42. Myers, E.W., et al., *A Whole-Genome Assembly of Drosophila.* Science, 2000. **287**(5461): p. 2196-2204.

43. Zerbino, D.R. and E. Birney, *Velvet: Algorithms for de novo short read assembly using de Bruijn graphs.* Genome Research, 2008. **18**: p. 821–829.

44. Simpson, J.T., et al., *ABySS: A parallel assembler for short read sequence data.* Genome Research, 2009. **19**: p. 1117–1123.

45. Namiki, T., et al., *MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads.* Nucleic Acids Research, 2012: p. 1-12.

46. Peng, Y., et al., *Meta-IDBA: a de Novo assembler for metagenomic data.* Bioinformatics 2011. **27**: p. i94-i101.

47. Binbin Lai, et al., *A de novo metagenomic assembly program for shotgun DNA reads.* Bioinformatics Advance Access, 2012.

48. Desai, N., et al., *From genomics to metagenomics.* Current Opinion in Biotechnology, 2012. **23**: p. 72-76.

49. Narasingarao, P., et al., *De Novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities.* International Society for Microbial Ecology, 2012. **6**: p. 81-93.

50. Tyson, G.W., et al., *Community structure and metabolism through reconstruction of microbial genomes from the environment.* Nature, 2004. **428**: p. 37-43.

51. Pelletier, E., et al., *"Candidatus Cloacamonas Acidaminovorans": Genome Sequence Reconstruction Provides a First Glimpse of a New Bacterial Division.* Journal of Bacteriology, 2008. **190**(7): p. 2572-2579.

52. Iverson, V., et al., *Untangling Genomes from Metagenomes: Revealing an Uncultured Class of Marine Euryarchaeota.* Science, 2012. **335**: p. 587.

53. Mardis, E.R., *Next-Generation DNA Sequencing Methods.* The Annual Review of Genomics and Human Genetics, 2008. **9**: p. 387–402.

54. Hess, M., et al., *Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen.* Science, 2012. **331**: p. 463.

55. Luo, C., et al., *Individual genome assembly from complex community short-read metagenomic datasets.* International Society for Microbial Ecology, 2012. **6**: p. 898-901.

56. Ferrer, M., et al., *Unveiling microbial life in the new deep-sea hypersaline Lake Thetis. Part II: a metagenomic study.* Environmental Microbiology, 2012. **14**(1): p. 268–281.

57. Cono, V.L., et al., *Unvailing microbial life in new deep-sea hypersaline lake Thetis, Part I: prokaryotes and environmental settings.* Environmental Microbiology, 2011. **14**: p. 268-281.

58. Ghai, R., et al., *Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing.* The International Society for Microbial Ecology, 2010. **4**: p. 1154-1166.

59. Pignatelli, M. and A. Moya, *Evaluating the Fidelity of De Novo Short Read Metagenomic Assembly Using Simulated Data.* PLoS ONE, 2011. **6**(5): p. 1-9.

60. Mende, D.R., et al., *Assessment of Metagenomic Assembly Using Simulated Next Generation Sequencing Data.* PLoS ONE, 2012. **7**(2): p. e31386.

61. Delcher, A.L., et al., *Improved microbial gene identification with GLIMMER.* Nucleic Acids Research, 1999. **27**(23): p. 4636-4641.

62. Noguchi, H., J. Park, and T. Takagi, *MetaGene: prokaryotic gene finding from environmental genome shotgun sequences.* Nucleic Acids Research, 2006. **34**(19): p. 5623–5630.

63. Godzik, A., *Metagenomics and the protein universe.* Current Opinion in Structural Biology, 2011. **21**(3): p. 398–403.

64. Meyer, F., et al., *The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes.* BMC Bioinformatics, 2008. **9**(386).

65. McHardy, A.C., et al., *Accurate phylogenetic classification of variable-length DNA fragments.* Nature Methods, 2007. **4**: p. 63 - 72.

66. Chan, C.-K.K., et al., *Binning sequences using very sparse labels within a metagenome.* BMC Bioinformatics, 2008. **9**(215).

67. Diaz, N.N., et al., *TACOA – Taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach.* BMC Bioinformatics, 2009. **10**(56): p. 1-16.

68. Krause, L., et al., *Phylogenetic classification of short environmental DNA fragments.* Nucleic Acids Research, 2008. **36** (7): p. 2230–2239.

69. Gerlach, W. and J. Stoye, *Taxonomic classification of metagenomic shotgun sequences with CARMA3.* Nucleic Acids Research, 2011. **39**(14): p. e91.

70. Huson, D.H., et al., *MEGAN analysis of metagenomic data.* Genome Research, 2007. **17**(377-386).

71. NOGUCHI, H., 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 2008. **15**: p. 387–396.

72. Rutherford, K., et al., *Artemis: sequence visualization and annotation.* Bioinformatics, 2000. **16**(102000): p. 944-945.

73. Huson, D.H., et al., *Integrative analysis of environmental sequences using MEGAN4.* Genome Research, 2011. **21**: p. 1552-1560.

74. Glockner, H.T.a.F.O. (2012) *Current opportunities and challenges in microbial metagenome analysis--a bioinformatic perspective*. Briefings in bioinformatics **Volume**, 1-15 DOI: 10.1093/bib/bbs039

Appendix A the size change of the largest twenty four assembled contigs in all assembly versions

Contig number	Assembly version					
	V0.5	V1.0	V1.5	V2.0	V2.5	V3.0
Contig 1	112239	350934	350937	351272	350934	350936
Contig 2	105706	178318	213209	250930	313985	283798
Contig3	90356	151167	202718	202719	257364	270569
Contig4	86968	123717	200946	200884	245729	228226
Contig5	86109	112365	162814	157914	232274	217427
Contig6	84467	102909	157876	145417	225556	207406
Contig7	81862	102835	151173	142631	217427	205618
Contig8	65427	101490	142506	139445	208162	204528
Contig9	60054	99797	120971	135535	202718	193681
Contig10	56509	98038	116762	132539	199980	177188
Contig11	54261	93159	115465	127921	185616	177119
Contig12	53127	88064	112650	126308	177188	145513
Contig13	53005	83382	103059	120972	163273	140280
Contig14	52597	81268	98258	111561	126305	134578
Contig15	51715	75201	85828	103072	124387	129834
Contig16	47753	74664	83965	101967	115248	126306
Contig17	46381	73387	82648	101652	111678	114143
Contig18	45573	71708	80949	99462	109070	111417
Contig19	43336	70928	80757	98482	103060	111382
Contig20	42841	66076	78748	84462	100605	103059
Contig21	41178	65529	77977	83971	100602	102216
Contig22	39136	64553	75201	83817	100145	101034
Contig23	37507	64455	74657	82649	94983	100848
Contig24	37154	64288	70875	77983	92101	100606

Appendix B

Number of reads incorporated in the largest 24 contigs in all assemblies

Assembly Version	V0.5	V1.0	V1.5	V2.0	V2.5	V3.0
Contig 1	4523	23540	33595	41865	57160	72122
Contig 2	3243	11629	15788	11481	20959	24963
Contig3	2194	8042	17423	21622	31850	21264
Contig4	6797	9895	17161	21272	17391	18478
Contig5	2186	4550	6515	10109	27814	22272
Contig6	5906	4365	8216	16916	17657	20578
Contig7	4013	2948	11395	9229	17638	14872
Contig8	1689	4486	7502	9323	12474	15559
Contig9	3598	4804	6590	7080	29806	18353
Contig10	3041	6386	12867	9564	11059	11791
Contig11	1452	12332	11942	4773	10885	27748
Contig12	2368	7884	4448	7091	9336	10201
Contig13	2350	3521	6346	8168	24672	10626
Contig14	2102	1751	6540	10286	9729	12074
Contig15	1206	2587	3392	7854	11246	7005
Contig16	1807	2261	5034	4263	20380	12322
Contig17	1280	2043	3136	4399	7404	20479
Contig18	1372	5485	3015	8372	7388	10324
Contig19	1063	3865	3017	10544	10898	8364
Contig20	1501	3847	3545	24095	9170	13774
Contig21	1661	2243	2862	6246	7258	26800
Contig22	1236	1572	3709	3509	3468	10453
Contig23	1542	4347	16189	3938	10727	20520
Contig24	957	1633	11208	3532	7040	11493
Total number of reads /24 contigs	59087	136016	221435	265531	393409	442435
Avgerage number of reads	2,462	5,667	9,226	11,064	16,392	18,435
Total number of all reads	655,289	1,337,597	1,917,096	2,392,780	3,263,795	4,104,966
% reads / 24 contigs	9.02%	10.17%	11.55%	11.10%	12.05%	10.78%

Appendix C Functional abundance of ATII-LCL using MG-RAST Subsystems

Function description	Percentage	Number of hits
Miscellaneous	10.26%	3265
Clustering-based subsystems	9.77%	3109
Protein Metabolism	7.59%	2414
Membrane Transport	5.63%	1792
RNA Metabolism	5.53%	1759
Carbohydrates	4.82%	1533
Respiration	4.81%	1531
Phages, Prophages, Transposable elements, Plasmids	4.39%	1397
Stress Response	4.25%	1353
Amino Acids and Derivatives	4.17%	1328
Regulation and Cell signaling	4.02%	1280
Virulence, Disease and Defense	3.88%	1234
Nitrogen Metabolism	3.87%	1230
Nucleosides and Nucleotides	3.62%	1151
Cofactors, Vitamins, Prosthetic Groups, Pigments	3.38%	1077
Metabolism of Aromatic Compounds	3.29%	1048
DNA Metabolism	2.61%	830
Cell Wall and Capsule	2.35%	747
Potassium metabolism	2.02%	643
Fatty Acids, Lipids, and Isoprenoids	2.02%	642
Sulfur Metabolism	1.46%	464
Motility and Chemotaxis	1.41%	448
Cell Division and Cell Cycle	1.30%	413
Phosphorus Metabolism	1.20%	382
Secondary Metabolism	1.00%	318
Photosynthesis	0.67%	213
Iron acquisition and metabolism	0.60%	190
Dormancy and Sporulation	0.09%	28

Appendix D Functional profile for ATII-LCL ORFs using Seed classification of Megan software

Seed Classification		No. of hits
Not assigned	40.41%	6709
No hits	22.82%	3788
Carbohydrates	4.63%	769
Virulence	4.29%	712
Amino Acids and Derivatives	3.32%	552
Cofactors, Vitamins, Prosthetic Groups, Pigments	2.29%	381
Protein Metabolism	2.20%	365
Cell Wall and Capsule	2.07%	343
Respiration	1.89%	313
DNA Metabolism	1.81%	301
Metabolism of Aromatic Compounds	1.50%	249
Stress Response	1.46%	242
RNA Metabolism	1.33%	221
Regulation and Cell signaling	1.28%	212
Fatty Acids, Lipids, and Isoprenoids	1.28%	212
Motility and Chemotaxis	1.26%	210
Clustering-based subsystems	1.05%	174
Nucleosides and Nucleotides	1.01%	168
Membrane Transport	0.91%	151
Sulfur Metabolism	0.86%	142
Cell Division and Cell Cycle	0.75%	125
Phosphorus Metabolism	0.49%	82
Miscellaneous	0.39%	64
Nitrogen Metabolism	0.36%	60
Secondary Metabolism	0.22%	36
Phages, Prophages, Transposable elements	0.06%	10
Potassium metabolism	0.05%	8
Photosynthesis	0.02%	3
Dormancy and Sporulation	0.01%	1
		16603

Appendix E MG-RAST Taxonomic classification at the order level of the ATII-LCL assembled dataset

Taxonomic Order	Assigned ORFs
Burkholderiales	18151
Rhizobiales	8918
Pseudomonadales	2346
Actinomycetales	1148
Clostridiales	851
Enterobacteriales	798
Halobacteriales	768
Nitrosopumilales	666
Bacillales	571
Chlorobiales	547
Rhodobacterales	516
Desulfuromonadales	395
Methanosarcinales	341
Alteromonadales	340
Thermoanaerobacterales	332
Rhodospirillales	315
Bacteroidales	314
Chroococcales	312
Caudovirales	300
Xanthomonadales	297
Myxococcales	287
Bacteroidetes Order II. Incertae sedis	270
Rickettsiales	247
Nostocales	244

Cont. Appendix E

Flavobacteriales	225
Methanococcales	215
Thiotrichales	203
Desulfovibrionales	196
Caulobacterales	192
Sulfolobales	192
Lactobacillales	185
unclassified (derived from Bacteria)	183
Planctomycetales	177
Oscillatoriales	176
Thermococcales	175
Vibrionales	175
Rhodocyclales	166
Desulfobacterales	157
Diptera	153
Chromatiales	148
Chloroflexales	143
Campylobacterales	140
Cytophagales	140
Neisseriales	127
Syntrophobacterales	127
Archaeoglobales	119
Thermoplasmatales	119
Thermotogales	117
Methanobacteriales	110
Aquificales	104
Total	43438

Appendix F Megan Taxonomic classification at the order level of the ATII-LCL assembled dataset based on BLASTX comparison of 87,357 ORFs against the NCBI-NR database.

Taxonomic order	ORFs assigned
Burkholderiales	9890
Rhizobiales	4072
Pseudomonadales	937
Halobacteriales	773
Nitrosopumilales	685
Actinomycetales	325
unclassified phages	315
Nanohaloarchaea	274
Caldithrix	228
Clostridiales	221
environmental samples <Bacteria>	215
Ignavibacteriales	179
Bacillales	165
Methanosarcinales	156
Archaeoglobales	147
Methanomicrobiales	132
Desulfurococcales	125
Caudovirales	110
Thermococcales	106
Bacteroidales	84
Methanococcales	82
Mariprofundales	82
Chroococcales	81
Enterobacteriales	79
Chlorobiales	77
Planctomycetales	72
candidate division OP1	70
Spirochaetales	67
Flavobacteriales	67
unclassified sequences	62
Bacteroidetes Order II. Incertae sedis	58

Myxococcales	57
Methanobacteriales	57
SAR406 cluster	56
Nitrososphaerales	55
Cytophagales	55
Lactobacillales	51
Desulfobacterales	51
Syntrophobacterales	47
environmental samples <Archaea>	44
Sphingobacteriales	44
Methylococcales	44
Desulfuromonadales	44
Xanthomonadales	41
Thermoanaerobacterales	41
Rhodospirillales	36
Dehalococcoidetes	35
Thermoplasmatales	34
Alteromonadales	32
Thermoproteales	31
Aquificales	30
Acidithiobacillales	29
unclassified Acidobacteria	28
Chromatiales	27
Rhodocyclales	25
unclassified Gammaproteobacteria	24
candidate division NC10	24
Sphingomonadales	24
Rhodobacterales	24
Nostocales	24
unclassified Archaea	23
Nitrospirales	22
Campylobacterales	22
Legionellales	21
unclassified Euryarchaeota	20
Thermales	20
unclassified Deltaproteobacteria	19
Desulfovibrionales	19
Thermotogales	18

Neisseriales	18
Poribacteria	17
Oscillatoriales	17
Acidilobales	17
Solibacterales	16
Chloroflexales	16
Caulobacterales	16
Bdellovibrionales	16
Thiotrichales	15
Selenomonadales	15
Sulfolobales	14
Malpighiales	13
Korarchaeota	13
Halanaerobiales	13
Candidatus Brocadiales	13
Spartobacteria	12
Methanopyrales	12
Herpetosiphonales	12
unclassified Alphaproteobacteria	11
Oceanospirillales	11
Synergistales	10
Methanocellales	10
Caldilineales	10
Vibrionales	9
Thermomicrobiales	9
Aureococcus	9
unclassified Thaumarchaeota	8
Fusobacteriales	8
Enteropneusta	8
unclassified Epsilonproteobacteria	7
candidate division WWE1	7
Verrucomicrobiales	7
Thermodesulfobacteriales	7
Opitutales	7
Haemosporida	7
Eutheria	7
Deferribacterales	7
Kinetoplastida	6

Hypocreales	6
Endopterygota	6
Anaerolineales	6
environmental samples <Green non-sulfur bacteria>	5
environmental samples <Euryarchaeota>	5
Prochlorales	5
Pasteurellales	5
Ophiostomatales	5
Magnetococcales	5
Ktedonobacterales	5
Desulfurellales	5
Deinococcales	5
Acidobacteriales	5
Total	21732