Application of Metagenomics for Discovery of Natural Products and Virophages

by

Jinglie Zhou

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 6th, 2016

Keywords: Metagenomics, Antibiotics, Culture-independent, Virophages
Pathogens, the NGS technology

Approved by

Mark R. Liles, Chair, Associate Professor of Biological Sciences
Paul A. Cobine, Associate Professor of Biological Sciences
Leonardo De La Fuente, Associate Professor of Entomology and Plant Pathology
Yucheng Feng, Professor of Crop, Soil and Environmental Sciences

Abstract

Metagenomics provided a powerful insight into an as-yet-uncultured bacterial community in different environments through construction of clone libraries and shotgun sequencing, therefore presenting potential in drug and novel microbe discovery. However, some limits such as insert size of BAC/Plasmid vector and length of sequencing reads are still big challenges for metagenomic study. In chapter II, we focus on identification of novel genomes of virophages, circular dsDNA viruses that infect giant DNA viruses, from a metagenomic DNA database of yellow stone lake. Three novel virophage genomes from Yellowstone Lake microbial assemblages were found, indicating a potentially high diversity of virophages from Yellowstone Lake in the photic zone or in Lake Floor hydrothermal vents. In chapter III, a metagenomic fosmid library constructed using DNA sample from a petroleum reservoir off the coast of Norway was screened for novel thermal-stable carbohydrate degrading enzymes. Comparison of sequencing results generated from shotgun metagenomic sequencing and fosmid library sequencing was performed as well. In the final chapter, another ~110 kb average insert sizes from soil metagenomic DNA (Cullars Rotation, Auburn, AL) achieved using a broad host range shuttle BAC vector, pSmartBAC-S was applied for discovery and expression of novel type I polyketide synthase (PKS) pathways and other secondary metabolite pathways, which is the largest soil metagenomic library so far

and able to harbor large-size of biosynthetic pathway. The two approaches PCR and macroarray screening targeted in exploitation of BAC clones with type I PKS pathways. An additional approach of *in silico* screening combined with the next generation sequencing (NGS) was aimed to discover diverse secondary metabolite pathways. Identified BAC DNA with PKS pathways were then transformed into a heterologous expression host *E.coli* BTRA for further screening of produced novel polyketide, followed by multiple antibiotic assays.

Acknowledgments

First of all, I would like to express my deepest appreciation to my committee chair, my advisor Prof. Mark R. Liles for his continuous support. All the achievements in the project won't be possible without his professional guidance and persistent support. I will cherish the advices he gave, not only in my research, but also in my life. His patience and erudition always encouraged me to go deeper in my research and finally realize some scientific contributions in my PhD study. I would like to appreciate my committee members, Prof. Yucheng feng, Prof. Leonardo De La Fuente and Prof. Paul A. Cobine for the suggestions provided on my research with their profundity of knowledge in biological science. My sincere gratitude to Prof. Joseph W. Kloepper for being an external reader.

I am grateful to Dr. David Mead and Dr. Scott Monsma who instruct and help me screen the BAC library. Thank Dr. Blaine Pfeifer who provide me the strain E. coli BTRA for expression. I am grateful to Dr. Alexander Wentzel and Dr. Anna Lewin who collaborate with me on discovery of thermo-stable cellulases. Thank Dr. Vu Thuy Trang Pham who worked hard and took the burden from me. Thank my labmate Alinne Pereira who worked together with me in the metagenomic project. Thank Nancy Capps for keeping our lab neat and providing support. I appreciate the help from all other lab members, Dr. Dawei Sun, Dr. Molli Newman, Malachi Williams, Charles Thurlow,

Cody Rasmussen-Ivey and Priscilla Barger. Thank my former group members, Dr. Mohammad Jahangir Hossain, Dr. Shamima Nasrin and Dr. Chao Ran who generously shared with me a lot of experience in the research. I am very thankful to the undergraduate students, Anna Moye and Mohamed El Zeiny who work very hard with me.

I would like to express my sincere love and utmost gratitude to my wife Yuanyuan Zhang for her support and encouragement during my graduate school journey. I greatly appreciate the arrival of my new-born son, Yixin Eason Zhou, who is the greatest gift from God. Thank my parents for their unconditional love and support. I am also very thankful to my parents-in-law who help us take care the baby in the most intense final period of our study.

Table of Contents

List of Tables

List of Figures

Chapter I. Introduction and Literature Review

1.1 Microbial community diversity in natural environments

The development of microbiology in the past 30 years has greatly altered our view of microorganisms, extending our knowledge of their metabolic and phylogenetic diversity. Microorganisms are significant and abundant members of every environment on Earth and they contain vast genetic diversity. Microorganisms almost occupy more than the one-third of Earth's biomass, providing important functions in the biogeochemical cycling of carbon, sulfur, nitrogen, phosphorus [1,2,3,4]. Microorganisms ubiquitously survive in global ocean, lakes, soils, human body and even extreme environments such as hot spring, deep-sea and Earth's Poles [5,6]. Microbial populations in soils vary depending on different physical, chemical and biological conditions, and are estimated to have approximately 10^{10} bacterial cells per gram of soil with > 10,000 genotypes [7,8]. Total population of marine microbes is estimated to be 3.6×10^{29} cells, accounting for > 90% of the total oceanic biomass [9]. These environmental microbes play a vital role on biochemical transformation such as degradation of cellulose and other carbohydrates, ammonification, nitrification, denitrification, nitrogen-fixation in nature, contributing in some cases to nutrient acquisition for multicellular host organisms and other microorganisms. As well, the functions generated by interacting microbial communities have a significant impact on geochemical cycles. For example, the large abundance and diversity of cyanobacteria in nature contribute both to 20–30%

of Earth's photosynthetic productivity and almost half of the nitrogen-fixation in marine systems [10,11]. The human body can be treated as a typical reservoir for microorganisms, e.g., human gut harbor approximately 10^{14} bacteria, ten times more than human body cells and some of them have been proved to be functionally critical for the digestion system [12,13]. Many specific environments will tend to enrich microorganisms with genes influencing the cycles of their own ecosystem, and it can provide a reservoir for scientists to explore functional proteins such as thermo, acid and alkali-stable enzymes.

Natural products generated by diverse microbes have been an invaluable resource for a huge number of medicines, specially antimicrobial and antifungal agents. Comparing to the secondary metabolite synthetized by plants and animals, microbial natural products are easier for screening and industrialization because of the short life cycle and more readily manipulated microbial cells compared to some eukaryotic organisms. Penicillin, the earliest antibiotic developed for clinical use, was found by Alexander Fleming from *Penicillium rubens* in 1928, frequently used until today and opening up the possibility of utilizing microorganisms for the pharmaceutical industry [14]. Following that, a variety of antibiotics such as the cephalosporins, tetracyclines, aminoglycosides and rifamycin were discovered by screening microorganisms collected from soil, water and other environments [15,16,17,18]. In addition to medicines used in antibacterial therapy, great attention also has been paid to the natural products with pharmacological activities synthesized by microorganisms such as curacin A, a metabolite with potent antitumor activity extracted from a marine cyanobacterium [19]. In addition, some microbial natural products are capable of advancing the basic

biotechnology. The commercialized examples capable of promoting biotechnologies include *Taq* polymerase, a thermostable DNA polymerase isolated from *Thermus aquaticus* living in the hot spring by Thomas D. Brock in 1965 [20]. Now Taq polymerase has been widely used in PCR because of its ability to withstand high temperature during chain reaction, greatly revolutionized the way of study in molecular genetics. All these applied natural products greatly promote drug industry and development of biotechnologies, indicating vast commercial value of natural products generated by microbes.

1.2 Contrast between culture-dependent and –independent methods

The study of microorganisms has been greatly dependent on the development of methods, such as the invention of the microscope and now next-generation sequencing. In 1663, Antonie Van Leeuwenhoek used the first home-made microscopy and successfully observed bacteria from his teeth, which he called "animalcules" [21]. His observation and records of microbial morphologies attracted many scientists to observe and study this microscopic world, triggering the establishment of microbiology. During this period of microbiology, botanist Ferdinand Cohn explored many new bacteria and described the life cycle of *Bacillus subtilis* through microscopy [22]. In the following years, microscopy became the most principle approach in microbiology till the next revolutionary shift when pure-culture was applied to study microorganisms. In 1880s German scientist Robert Koch published famous Koch's postulates for isolating

pathogens and developed first innovational media, altering microbiology from only observation to cultivation [23,24]. From then on, a large amount of culture media that aims for growth of different microorganisms have been developed. As a result, a large amount of microorganisms were identified based on pure culture, laying the foundation of modern microbiology [25]. Even now, culture-dependent methods are an important component of modern microbiology because it provides the capacity to study biological functions and a platform for research into microbial physiology in spite of great effort needed to inquire proper media and growth conditions.

However, almost 99% of microorganisms that exist in natural environments are not readily cultured in the laboratory, far out-sizing the extent of known cultured microorganisms and this has limited scientists in the past to discover the diversity of microbes and the natural products they produce using culture-dependent methods [7,25,26]. With the advent of culture-independent methods, this has allowed scientists to study a much greater diversity of microbial life both to understand their diversity and to access their natural products. Reconstruction of phylogenies based on 16S rRNA genes that are highly evolutionarily conserved was the first key breakthrough for studying environmental microbes [27]. It was firstly proposed by Carl Woese and then successfully contributed to defining the phylogenetic taxonomy of the Archaea in 1977 [28]. PCR is a powerful and classical technology for the microbial study, utilizing a heat-stable DNA polymerase to achieve DNA replication in vitro and capable of accessing genetic information of as-yet-uncultured microorganisms [29]. Usually, some universal primers for PCR reaction can be designed to amplify conserved genes (e.g. 16S rRNA gene) or

certain specific gene(s) of interest [27]. After PCR, the sequences will be enriched in PCR production for future identification. Some less precise methods such as Terminal-restriction fragment length polymorphism (T-RFLP), Denaturing Gradient Gel Electrophoresis (DGGE), Ribosomal Internal transcribed Spacer Analysis (RISA) have been applied to analyze the heterogeneity of PCR products [30,31,32]. The development of next-generation sequencing technologies now provides more resolution and through massively parallel sequencing enables a greater extent of knowledge concerning the diversity of microorganisms identified from environmental DNA.

Metagenomics, a generalized culture-independent concept firstly proposed by Jo Handelsman, Jon Clardy and Robert M. Goodman in 1998, provides a postgenomic view to study environmental DNA from a specific habitat without individual isolation and cultivation [25,33]. Metagenomics requires using an appropriate protocol for DNA isolation depending on the environmental sample that is capable of generating DNA with high purity and sufficient yield. With the developments of the next generation sequencing technologies, scientists in metagenomics began to have the option of directly sequencing the extracted DNA from environments instead of the steps of constructing cloning library when only concerning about genetic information of cloned DNA [34]. This made it possible for the next generation sequencing technologies (e.g. Life 454, Illumina Miseq) with low-cost and high-throughput, producing millions of metagenomic sequence reads for each sample [35]. Direct “shotgun metagenomic” DNA sequencing provides an easier method to gain insight into the encoded functions and dynamic change of microbes in specific habitats compared to constructing a clone

library, and this has triggered an explosion of research into collecting genetic information of environmental microorganisms [36]. Large-scale environmental DNA databases can be utilized to discover new microorganisms, functional genes and combined with environmental factors (e.g. temperature, PH, location and nutrients) for further analyzing environmental roles of microbes and microbial interaction with their hosts [25].

1.3 Construction of metagenomic libraries including different vector systems

While shotgun metagenomics has provided a phenomenal wealth of data on the predicted functional diversity of environmental metagenomes, there are some very significant limitations and biases of this approach. The ability to predict the function of a sequenced gene is dependent upon prior knowledge of gene functions, and the GenBank and other sequence databases have a very incomplete record of biological functions for environmental microorganisms. It is typical for a large percentage (30-60%) of the sequence reads to have no significant hit in a GenBank database. The other significant bias for shotgun metagenomics is that complete biosynthetic pathways are difficult, if not impossible, to assemble from short sequence reads. Chimeric contiguous ("contig") sequences can also result from assembly of short sequence reads from a diverse metagenome, resulting in the formation of contigs that do not exist in nature. For these reasons, the use of shotgun metagenomics with current sequencing technology will not be able to provide complete biosynthetic pathways from

environmental microorganisms, and the use of direct metagenomic cloning of large DNA fragments represents the only viable method to accomplish this with culture-independent methods. After DNA extraction, environmental DNA can be sheared and cloned into a suitable vector, and then the vector can be transformed into *E. coli* or other cultured heterologous hosts, generating a metagenomic clone library [25]. The resulting transformants allow screening for the clones with particular functions so that individual clones encoding specific functions or pathways can be separately characterized. Recombinant clones can be engineered for better expression by adjusting the growth conditions to include inducing agents for better expression and the cloned DNA can be sequenced for identification of classification or other functional pathway information [37,38]. Construction of a metagenomic library and isolation of clones containing an intact pathway for the first time may make it possible that the functional pathways derived from as-yet-uncultured microorganisms are able to be expressed and further utilized for their respective bioactivity.

The vector systems in molecular biology are vehicles used to introduce foreign DNA into cells. In molecular cloning, vectors are required to possess at least four components: an origin of replication, promoter, cloning site, genetic markers. The origin of replication is a region of DNA sequence where replication is initiated in a genome. A Promoter is able to initiate transcription of the insert DNA. The cloning site, which may be a multiple cloning site, is the position where insertion of foreign DNA is allowed. These components are tightly associated with the efficiency of replication and expression in the heterologous host. There are four major types of vectors: plasmids,

phage, cosmids, and artificial chromosomes. Among them, cosmids, fosmids and bacterial artificial chromosome (BACs) are three major vectors for construction of metagenomic clone libraries. A cosmid is a hybrid plasmid containing *cos* sites (cohesive ends) derived from λ phage. It is capable of carrying 37 to 52 kb of DNA, while normal plasmids are only able to insert 1–20 kb DNA [39]. The empty clone produced by self-ligation is avoided due to the size limits. Recombinant cosmid DNA can be directly transformed into *E.coli* but also can be packaged into λ phage capsids firstly and then more efficiently transduced into cells such as *E.coli* (*cos* sites are essential) [40]. Fosmid vectors are derived from bacterial F-plasmid and able to insert up to 40kb DNA [41]. The low copy number is an important feature for fosmid vectors, guaranteeing higher structural stability than cosmids [42]. Clones containing insert DNA may express toxic products to the host cell, so it is very necessary to keep low copy number in library. The copy number of fosmid is 1-2 per cell and each host cell (usually *E. coli*) only can hold one fosmid replicon [42]. In some complex genome research such as Human Genome Project, fosmid libraries were used to keep accuracy of insert DNA in the continuous generations [42]. A BAC vector is based on F-fosmid as well, but the average insert size of it can reach ~ 350 kb, much larger than fosmid and cosmid [43,44]. Due to the F factor system, the library constructed by BAC has the same long-term stability as that built by fosmid [43]. The RK2 replicon that has been introduced as a separate origin of replication in some BAC vectors is responsible for the inducible copy number of BAC so that single-copy and high-copy of BAC can be alternated, with the control of the RK2 origin *trfA* gene under the control of an arabinose-inducible

promoter and located on the *E. coli* chromosome [45]. In addition, the larger insert of BAC gives much higher possibilities to cover intact gene pathway from environmental DNA or genome. Combined with innovative screening methods, BAC greatly accelerate discovery of novel pathways and drugs. A shuttle vector fuse the components derived from two different host species, and hence it can propagate in both hosts. By introducing the RK2 origin of replication and the arabinose-inducible *trfA* onto the vector, the Liles laboratory previously demonstrated that a broad-host range BAC vector could be used to shuttle metagenomic DNA between gram-negative heterologous hosts [45]. The BAC vector applied in this research pSMART-BAC-S was developed collaborative with the Lucigen Corporation (Middleton, WI), and supports high throughput conjugation-based transformation into both Gram-negative and Gram-positive hosts, and chromosomal integration or stable episomal maintenance in *E. coli* for heterologous expression.

Preparation of high quality of high molecular weight (HMW) DNA is also a big challenge for many special habitats, especially soil environments. Since HMW DNA is critical if we want to construct a large insert metagenomic library (>100kb) or long-read sequencing, it is very difficult to remove deleterious contaminants that will affect downstream application (e.g PCR, hybridization and Sequencing) while minimize DNA shearing [46,47]. Usually, two strategies, direct lysis and indirect cell methods, are performed to recover environmental DNA from soils. The direct lysis method applies lysis of bacterial cells in soil before DNA extraction and purification. However, this method causes more contact between DNA and contaminants, making it much harder

to obtain high quality DNA [48]. The second strategy, indirect DNA isolation, isolates bacterial cells from soils first, followed by a combination of chemical and enzymatic lysis within an agarose plug. Then DNA isolation is dependent on electrophoresis of DNA from the plug into a new gel. The band of compressed DNA can be excised for further purification. This approach avoids mixture of DNA and contaminants but decrease diversity of recovered DNA comparing to the direct lysis, because many bacteria may not be isolated [48]. In our research, both direct lysis and indirect cell methods have been performed and high quality HMW have been recovered using our protocol of both methods. Recovered DNA has been demonstrated to be pure enough for metagenomic library construction (Cullars Rotation soil, Auburn University).

1.4 Screening of metagenomic libraries by sequence-based methods

16S rRNA gene is an ideal genetic marker for determining bacterial phylogenetic affiliations because it has both highly conserved primer binding sites and hypervariable regions, early metagenomic studies sequenced cloned 16S rRNA gene (enriched by PCR) to profile the phylogenetic diversity of environmental DNA in spite of bias caused by distinct copy number of 16S rRNA gene among different microbes and PCR amplification [49]. As metagenomic methods improved, the vector system used could contain larger-size DNA contain more functional genes. Sequentially, a 16S rRNA gene or other phylogenetic markers can be treated as "phylogenetic anchoring" to determine phylogeny of the interested gene in the rest of the flanking DNA when a larger-size

DNA is cloned. One such example is the discovery of proteorhodopsin, a type of light-harvesting protein, firstly proved to be existed in marine bacteria by DeLong group through identification of both 16S rRNA gene and rhodopsin-like proteins from the same large insert [50].

Another strategy to provide insight into the genomic complexity present within a metagenomic library is random sequencing. When a large number of clones are chosen and end-sequenced, ample information concerned about distribution and abundance of microbial community, gene function and genomic organization is possibly able to be inferred by bioinformatics methods [25]. *De novo* assembly refers to reconstruct the original sequence by linking DNA fragments from random clone by mating overlapping regions [51]. If the amount of sequenced clones is high enough, and the insert size is large enough, then (partial) reconstruction of some genomes in this habitat are possibly achieved. With rapid advance of NGS, more and more labs are capable of sequencing environmental samples of interest with affordable prices, and generating enormous metagenomic sequence databases. Therefore, analysis of these tremendous sequences then becomes a very crucial step, requiring the use of powerful computers for bioinformatics analyses.

There are many kinds of bioinformatics software in the linux system that can be used to analyze metagenomic DNA, involving trimming reads, assembly, annotation and taxonomy, etc. Some bioinformatics analysis platforms such as CLC Genomics Workbench and Geneious also provided all these functions but in much more friendly interface, which is easier to be operated by biologists with few programming

knowledges [52]. Since the low cost of the NGS technologies has enabled massive sequence databases, this becomes a challenge for computational analysis, and therefore supercomputer systems have already become very important equipment for biological research.

Also many website platforms such as GOLD (Genomes OnLine Database), CAMERA (Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis) and the metagenomics RAST server (MG-RAST), the Galaxy project and KBase (The Department of Energy Systems Biology Knowledgebase) provide free access and powerful workflow to analyze metagenomic data, facilitating the research for many scientists without access to high-performance computers [53,54,55,56].

1.5 Screening metagenomic libraries by function

Functional screening a metagenomic library is a method based on identifying a particular phenotype expressed by the cloned gene. The change of phenotype of a corresponding metagenomic clone(s) by incorporation of specific substrate into the growth medium can be applied to recognize the clones producing specific enzymes or metabolites based on a specific phenotypic characteristic [25]. Some clones with expressed carbohydrate-degrading enzymes such as esterases and proteinase can direst present halo around the colony on the indicator medium [57]. Another example here is screening for antibiotic resistance of a clone by growth inhibition assays using top soft agar mixed with a certain pathogen overlays over the clone colonies on the agar medium,

leading to formation of inhibition zone if clone has a corresponding antibiotic activity [58]. In addition, the pathogens can be grown in the broth together with the supernatant extracts from the clone cultures for antibiotic or antifungal assay [59]. Functional screening directly targets to expressed natural products, followed by sequencing insert DNA of isolated clones for bioinformatics analysis.

However, due to the low frequency of metagenomic clones expressing the interested genes and the corresponding phenotype, the high-throughput screening (HTS) is the most efficient approach for active clone discovery [60]. The key testing vessel of HTS is the 96, 384, or 1536-well microplates. Usually a metagenomic clone library is stored in these kinds of microplates as well. Therefore, a simple copy of the stock plate to liquid medium in the same microplate or the agar medium with incorporation of the substrate can achieve an assay by screening for the targeted phenotype among thousands of clones rapidly. A pin replicator is the simplest tool for this transfer, while modern technology also allows researchers combination of robotics, data processing software and sensitive detectors to achieve millions of chemical, genetic or pharmacological tests in a shorter time [61,62]. For example, Mewis *et al* designed a biomining cellulase HTS assay to detect final absorbance of a chromogenic dinitrophenol-cellobioside substrate using qFill3 (Genetix), qPix2 robot colony picker (Genetix), rapidStak automated microplate Stacker (Thermo) and Varioskan Flash Multimode Reader (Thermo) [63]. By adopting these automated device for high-throughput screening, the probability of discovering interested natural products will be greatly improved.

Heterologous expression of cloned genes in a host is another key element for screening metagenomic libraries. A good heterologous host strains requires expression of the cloned genes using the native transcriptional and translational apparatus for appearance of the phenotype in the recombinant clone [64]. *E. coli* and its mutants are the most wildly used heterologous expression hosts because they are well-characterized and easy to culture in the lab condition [65]. For example, an isocyanide-containing antibiotic 1 together with its biosynthetic genes was isolated and characterized from an *E. coli* clone library [66]. In addition to *E. coli*, other bacteria or fungus are necessarily considered to be other choice of expression hosts such as *Pseudomonas*, *Streptomyces* or *Saccharomyces* to promote the possibility of expressing the cloned gene [67,68,69].

1.6 The use of next-generation sequencing methods for metagenomics

The application of sequencing technologies has enabled biologists to glean substantial new genetic information from both cultured and as-yet-uncultured microbes. The first sequencing technology, Sanger sequencing, was developed by Frederick Sanger in 1975 [70]. The reaction components in Sanger sequencing is very like that of PCR reaction, also an *in vitro* DNA replication reaction with a single-stranded DNA template, a DNA primer, a DNA polymerase and normal deoxynucleosidetriphosphates (dNTPs), but DNA sample is divided into four separate reaction, each of which requires additional one dideoxynucleosidetriphosphate (ddNTP), which terminates DNA chain elongation at random positions of the DNA strand [71]. The resulting products containing

DNA fragment with one nucleotide different, and then the sequence can be read from the DNA band by autoradiography or UV light after gel electrophoresis. Recent Sanger sequencing employs dye-terminator or fluorescently labelled dNTPS that facilitate the reading of nucleotide bases.

Different from Sanger sequencing, NGS technologies are massive and parallel, which is able to achieve sequencing large–scale metagenomic DNA with tremendously reduced cost. Therefore, ordinary labs began to have capability to achieve large-scale sequencing, resulting in an exponential increase of sequence data in NCBI and other sequence databases. Life 454 and Illumina are the two common technologies used in metagenomic research. Emulsion PCR is the first step of the 454 sequencing, involving enormous small DNA-capture beads. One adapter-ligated DNA fragment can be attached to one bead and then amplified in a water-in-oil emulsion. Each DNA-enriched bead will be spread into a ~29 μm well on a fiber optic chip, where the sequencing reaction is located, and then a mix of enzymes including DNA polymerase, ATP sulfurylase, and luciferase will be packed into each well. During a sequencing run, the four dNTPs are inputted sequentially, leading to addition of one (or more) nucleotide(s) and releasing a light signal that is recorded by the CCD camera [72]. The sequence can be read relying on the signal strength. Illumina chose a distinct strategy from Life 454. In sequencing length, Life 454 can reach 700bp, longer than Illumina (50-250bp) [73,74]. So Life 454 is used to be dominant in direct sequencing environmental DNA, because longer sequencing length can prevent *do novo* assembling two sequences from different entities [75]. However, with update of Miseq to the length of 2 × 300 bp and 13.2-15 Gb

Miseq gradually replace application of Life 454 in metagenomic sequencing in recent metagenomic studies [74]. On the other hand, Illumina has advantages in lower cost and high-throughput, which is up to 30 billion per run and $0.05 to $0.15 per million bases, while Life 454 just reach 1 million per run and $10 per million bases [76].

In addition, there is some novel sequencing technologies called as "the third generation sequencing". The one already commercialized is PacBio RS RS II system that uses a single molecule real time sequencing (SMRT), of which average read length is more than 10,000 bp with averaged throughput of 500 megabases per SMRT Cell [77,78]. This technology utilizes a chip that contains many zero-mode waveguides (ZMWs), which actually is a circular hole with a nanophotonic confinement structure [79]. In the bottom of a ZMW, a single molecule of single stranded DNA template is immobilized together with a DNA polymerase [79]. During the sequencing process, the light signal generated by the fluorescent dye molecule attached to each nucleotide is visualized by the monitor in real time. Recently, the company launched a new system (Sequel system) announced to be capable of generating seven times as high throughput as and costs about the half price of the old RS II system. Another company utilizing nanopore sequencing is Oxford Nanopore Technologies. It takes advantage of millions of nanopore channels embedded in an array chip. Each nanopore is associated with an electrically resistant polymer membrane and an individual electrode, allowing single DNA to pass through and generating a disruption on the voltage for distinguishing four standard DNA bases A, T, C, G, and also modified bases [80]. Due to detection of bases through electrical signals rather than fluorescence, the device MinION is greatly

miniaturized with an only pocket size and claimed to have more than 5 kb read length with extremely low cost [81]. The technology was accessible for an early-access community in early 2014 and is available in May 2015, and it is expected to cause a revolutionary change in sequencing market. As NGS technology evolves, this will also impact the practice of metagenomics research. Long-read length of both technologies will probably improve *de novo* assembly and help to achieve longer contigs or even complete microbial genomes from metagenomic data. However, the limitation right now is still on the low accuracy rate (80~85%), which may lead to errors during metagenomic assembly [80,82]. Currently there is no sequencing technology that enables sequencing of intact biosynthetic pathways from environmental microorganisms, necessitating a direct cloning method in order to access previously uncharacterized natural product biosynthesis pathways that could encode a great diversity of novel chemical entities.

1.7 Mining microbial natural products and genomes using the metagenomics approach

1.7.1 Carbohydrate-degrading enzymes

Carbohydrate-degrading enzymes such as cellulase, chitinase, proteinase and xylanases are critical for environmental microbes to digest plant and animal residues. Hess *et al* analyzed metagenomic DNA (totally 268 Gb) from microbes attached to

plant fiber incubated in cow rumen, obtaining 27,755 genes encoding carbohydrate degrading enzymes and 90 of them were successfully expressed [83]. Dai *et al* utilized functional screening on a metagenomic clone library to study the fibrolytic microbiome in Yak rumen, resulting in 150 glycoside hydrolase (Glycoside hydrolases, GH) genes for fiber degradation, particularly an endoglucanase of a novel GH5 subfamily occurred frequently [84]. From analysis of 35 contigs larger than 10 kb, 25 of them are from Bacteroides and 4 are from Firmicutes. In addition to cow rumen, composting operation is also a reservoir for Carbohydrate-degrading enzymes. Martins *et al* identified a total of 112 cellulase genes (mostly GH5 family), and 32 genes of which contained cellulose binding domains (Carbohydrate binding module, CBM) through metagenomic shotgun sequencing of two composting samples collected from the São Paulo Zoo Park, in Brazil.

Within all carbohydrate-degrading enzymes, cellulase is a kind of important resource for the biofuel industry due to its ability to decompose cellulose into monosaccharides ("simple sugars"). Cellulose is a polysaccharide with thousands of D-glucose units (cellobiose) linked by β-1,4 glycosidic bonds, which is the most abundant organic polymer on Earth and insoluble in water and organic solvents [85]. Generally cellulases can be divided into three categories [7, 8]: endo-β-1, 4-glucanases (EC3.2.1.4), exo-β-1, 4-glucanases (EC3.2.1.91, CBH I and EC3.2.1.176, CBH II) and β-1, 4-glucosidases（EC3.2.1.21). The main role of endo-β-1, 4-glucanases is to cleave internal bonds at amorphous sites, cutting the long-chain into short chains with exposed new chain ends. Exo-β-1, 4-glucanases further transform short chains generated by

endo-β-1, 4-glucanases into tetrasaccharides or disaccharides (e.g. cellobiose) through cleaving units from the exposed ends of the short chains. Finally, β-1, 4-glucosidases work on exocellulase products and generate monosaccharides. Many fungi, bacteria, actinomycetes are proven with capability of producing cellulases [57]. It involved in up to 68 fungal genera mainly including *Trichoderma, Trichoderma, Aspergillus, Chaetomium, Penicilliu,* etc. Genera of bacteria producing cellulases include *Pseudomonas, Bacillus, Clostridium,* etc.

1.7.2 Secondary metabolites

Microbes harbor abundant and diverse biosynthetic pathways for production of secondary metabolites, which contribute to their growth, development, or reproduction and support their huge potential for drug discovery. These pathways contains a cluster of genes with multiple biosynthetic domains responsible for synthesis of secondary metabolites, including groups of polyketide synthase (PKS), fatty acid synthase (FAS) and non-ribosomal peptide synthase (NRPS), etc. Through metagenomic approaches, it is able to insight diverse pathways of as-yet-uncultured microbes. Kampa *et al* shotgun sequenced the DNA sample collected from both the lichen *Peltigera membranacea* and its symbiotic bacteria, leading to the discovery of a novel polyketide named nosperin (the pederin family) with anticancer activity from the photobiont *Nostoc* sp [86]. In addition to discovery of new metabolites, Wilson *et al* assembled metagenomic reads from a ‘‘metabolically talented’’ marine sponge, and identified an endo-symbiont genus

Entotheonella with pathways annotated to encode nearly all bioactive molecules that have been isolated from its host [87]. However, there are still limitations in study of secondary metabolite biosynthetic pathway using metagenomic approaches. Due to the existence of some repetitious biosynthetic domains with high sequence similarity, it is difficult to assemble complete and correct pathways from metagenomic shotgun reads [88]. This problem will require long read length sequencing technologies such as PacBio RS and Oxford Nanopore, even though improvement on the error rate is still needed for applying these technologies in metagenomics. Another way to study sequences of pathways is to construct a metagenomic clone library using sheared environmental DNA. It provides possibilities to both harbor complete biosynthetic pathways in clones and express interested metabolites in heterologous expression host. Clones with pathways can be distinguished based on the "sequence tag" existed in some conserved domains by PCR or sequencing. In addition, since lots of biosynthetic pathways are large-size (can be more than 100kb), it requires the vector able to support long insert DNA like BAC vectors rather than fosmid vectors (only about 40kb).

Actually, the conserved domains of many biosynthetic pathway classes (e.g. NRPS, PKS, isoprene and shikimic acid) can also be utilized to profile the biosynthetic diversity and capacity of metagenomes. Based on the identity degree of domains comparing to the known database, there are several bioinformatics tools such as eSNaPD, NaPDoS and AntiSMASH 3.0 developed to annotate and classify biosynthetic gene clusters as well as predict structures of their metabolites from metagenomic data [89,90,91]. Although the predicted metabolite structure may not be

totally correct, it still can be used as a reference to match the real corresponding metabolite detected through proteomics [86].

Type I PKS pathwayswhich is focused in this dissertation typically have a multidomain architecture and are capable of synthetizing complex polyketides from simple starter and extender units such as acetyl-CoA and malonyl-CoA through successive rounds of Claisen condensation reactions [92,93]. Usually, type I PKS pathways in bacteria are modular, comprised of multiple set of modules, whereas the majority of fungi have iterative type I PKS with only a single module [94,95]. Each module of type I PKS possesses multiple functional domains as a megasynthase, and a basic module included a ketoacyl synthase (KS), an acyl carrier protein (ACP) and acyltransferase (AT) [96]. Additional domains such as ketoreductase (KR), dehydratase (DH) and enoyl redutase(ER) also aid in generating different functional groups of produced polyketides [96,97]. Modular properties of PKS pathways enable them to synthetize natural compounds with activities relevant to different medical areas such as antibiotic, antifungal and anticancer agents, therefore generating considerable research interests from academic and industrial scientists.

1.7.3 Microbial and viral genomes

Theoretically, a metagenome is able to include the genomic sequences from all the microorganisms present in a given environment, if sequencing depth and DNA sampling is perfect. The high-throughput sequencing and analysis pipelines applied in

metagenomics can be used to elucidate a representative, random fraction of the genome sequences in an environmental DNA sample [98]. The first nearly complete genome of an uncultured microbe is obtained from a metagenomic small-insert plasmid library of a natural acidophilic biofilm [99]. The success of reconstructing microbial genomes was possibly on account of the sampled biofilms with low-complexity and few different species. Many extreme environments also harbor low-complexity microbial community, and hence benefit metagenomics to reconstruct more genome sequences. Narasingarao *et al* obtained recovered two draft uncultured genome sequences from a hypersaline lake in Victoria, Australia through from Sanger sequencing, which represented a totally new branch of uncultured *Halobacteria* [100]. In a recent article, three partial uncultured archaeal genome sequences, comprising a candidate archaeal phylum *Lokiarchaeota*, were assembled from a metagenome of a deep marine sediment[101]. Interestingly, this novel archaeal lineage indicated a monophyletic group with eukaryotes in phylogenomic analyses, encompassing the base of all eukaryotes.

In addition to microbial genomes, reconstruction of viral genomes are another hotspot for metagenomic sequencing due to its relatively smaller genomic size. The sequence diversity of viruses in metagenome is much higher than their host or bacterial genomes, leading to a big challenge here for *de novo* assembly of viral genomes [102]. A more successful strategy to achieve complete viral genomes from mixed samples is iterative assembly, which is depending on a seed sequence (usually a contig) that was initially identified by sequence similarity to a known virus and then applies iterative rounds of contig extension on paired-end or single reads (e.g. softwares PRICE, IVA

and OLC) [103,104]. PRICE was applied to reconstruct the Bas-Congo virus genome, leading to an outbreak of human acute hemorrhagic fever in the Democratic Republic of Congo in 2009, from DNA sampled from a patient's serum [105]. Long contigs of a novel dolphin rhabdovirus were assembled by application of OLC to a metagenome of a cell culture supernatant derived from a dead white-beaked dolphin [106]. The iterative assembly strategy can also be operated manually by numbers of rounds of mapping assemblies so that each round of extension can be validated by naked eyes in case of chimeric contigs. Through this way, Zhou *et al* assembled five circular genomes of virophages that infect the giant DNA viruses from an ataractic lake and fresh water lake [107].

The rapid development of new sequencing technologies and bioinformatics tool is a key to promote assembly of as-yet-uncultured microbial or vial genome sequences from complex shotgun metagenomes. The short read length of the NGS technologies, Roche 454 and Illumina Miseq, is hard to avoid chimeric contigs when assembling random fragments of multiple genomes from different organisms, and therefore limiting its capability to recover complete and correct microbial genomes. In the further, new long-read sequencing technologies such as Pacbio and Oxford Nanopores are expected for a better assembly of metagenomic reads.

Chapter 2. Three novel virophage genomes discovered from Yellowstone Lake metagenomes

2.1 Abstract

Virophages are a unique group of circular double-stranded DNA viruses that are considered parasites of giant DNA viruses, which in turn are known to infect eukaryotic hosts. In this study the genomes of three novel virophages YSLV5, YSLV6 and YSLV7 were identified from Yellowstone Lake through metagenomic analyses. The relative abundance of these three novel virophages and previously identified Yellowstone Lake virophages YSLVs 1-4 were determined in different locations of the lake, revealing that most of the sampled locations in the lake, including both mesophilic and thermophilic habitats, had multiple virophage genotypes. This likely reflects the diverse habitats or diversity of the eukaryotic hosts and their associated giant viruses that serve as putative hosts for these virophages. YSLV5 has a 29,767 bp genome with 32 predicted ORFs, YSLV6 has a 24,837 bp genome with 29 predicted ORFs, and YSLV7 has a 23,193 bp genome with 26 predicted ORFs. Based on multilocus phylogenetic analysis, YSLV6 shows a close evolutionary relationship with YSLVs 1-4, whereas YSLV5 and YSLV7 are distantly related to the others, and YSLV7 represents the fourth novel virophage lineage. In addition, the genome of YSLV5 has a G+C content of 51.1% that is much higher than all other known virophages, indicating a unique host range for YSLV5. These results suggest that virophages are abundant and have diverse genotypes that

likely mirror diverse giant viral and eukaryotic hosts within the Yellowstone Lake ecosystem.

2.2 Introduction

Virophages are circular double-stranded DNA viruses that infect giant viruses and their protist hosts, and were reported to be distributed widely throughout the world, even including an Antarctic lake [108,109,110]. Sputnik was the first described virophage that was found to inhabit a water-cooling tower in Paris, France, infecting a mamavirus in an *Acanthamoeba* species [108]. Three years later, a virophage designated as Mavirus was identified from *Cafeteria roenbergensis*, a marine phagotrophic flagellate from Texas coastal waters [110]. Unlike Sputnik and Mavirus, the genome of Organic Lake virophage (OLV) was identified by *de novo* assembly of a metagenomic shotgun sequencing database from a hypersaline meromictic lake in Antarctica [109]. This was the first example of virophage discovery using culture-independent methods, providing access to novel virophage genomes by exploring metagenomic databases. The fourth reported virophage, almost identical to Sputnik and named Sputnik 2, was associated with contact lens fluid of an individual with keratitis [111,112]. In 2012, we obtained 4 complete virophage genomes (YSLV 1-4) from Yellowstone Lake and one nearly complete genome (ALM) from Ace Lake in Antarctica [113]. In 2014, a virophage named Zamilon was reported to be associated with a *Mimiviridae* host and closely related to Sputnik [114]. Virophages, as parasites of the giant viruses, may play a potential role in

lateral gene transfer, mediating gene exchange between different giant DNA viruses and enlarging their genome size [108,115].

Yellowstone Lake occupies a dominant space in Yellowstone National Park (YNP), USA, and so far the largest number of distinct virophages were discovered from this lake ecosystem [113]. In a major metagenomic survey of this lake, samples were taken at three general different locations, representing the northern region which is rich in lake floor hydrothermal vent activity [116,117,118], the West Thumb region where additional lake floor vents occur but that differ in chemistry relative to the northern lake vents [116,119], and the Southeast Arm region of the lake where as of yet there is no known lake floor geothermal activity. These sampling locations represent: i) different hydrothermal vents; ii) microbial streamers associated with vent openings; iii) mixing zones, where vent waters mix with lake water; and iv) photic zone water column samples, with some taken above sampled vents. Samples were size-fractionated and then extracted DNA subjected to 454 pyrosequencing (approximately 7.5 Gbp), which is available at http://camera.crbs.ucsd.edu/projects.

In this study, to better understand the distribution, abundance and diversity of YSLVs in Yellowstone Lake, the above described Yellowstone Lake metagenomic sequences and targeted unique representatives of virophage major capsid proteins (MCP) were subjected to analysis. Assembly of the distinct virophage genomes present in Yellowstone Lake metagenomes allowed identification of three novel virophages. In addition, a large number of short contigs showing significant similarity with MCPs of known virophages were reconstructed. Our results reveal significantly higher virophage

diversity in Yellowstone Lake than previously recognized, implying the important role played by virophages in this ecosystem as well as the potential possibility to isolate them from this and other freshwater lake ecology.

2.3 Materials and methods

2.3.1 Sampling and DNA extraction

A total of 42 water samples were obtained from different locations of Yellowstone Lake using a remotely operated vehicle in September of 2007 and 2008. Sampling information and methods for biomass collection and DNA extraction have been previously published [120].

2.3.2 Assembly of Yellowstone Lake metagenomic sequences and virophage genomes

The methods for sequence assembly were similar to the ones described in [113] with some modifications. Shotgun metagenomic sequences from Yellowstone Lake samples were assembled *de novo* using Newbler v2.6 (Roche). Contigs derived from assembly of the entire Yellowstone Lake metagenomic sequence database were constructed as a local database for tBLASTx search for MCPs, which are known to be conserved among all known virophages. Contig sequences with significant similarity (E-value < 10^{-3}) to

virophage MCPs were collected as virophage MCP-related sequences. Each contig served as a reference sequence, and then reference assembly was performed using trimmed reads from the Yellowstone Lake metagenomic database with minimum overlap length of 25 bp and minimum overlap identity of 90%. Once an extended sequence with a longer size was obtained, it was used as the next reference sequence for reference assembly of the metagenomic reads. This procedure was repeated until the respective assembled contig sequences stopped extending. All reference assemblies were performed using the bioinformatics platform Geneious Pro (version 7.1.5; Biomatters Ltd).

2.3.3 Genome closure using virophage-specific PCR and sequencing

The existence of the three novel virophages in each water sample was determined by polymerase chain reaction (PCR) with oligonucleotide primers specific to the MCP from each respective virophage genome (Table 1). The PCRs were conducted using 0.2 nM of each respective forward and reverse primer, EconoTaq® PLUS GREEN 2× Master Mix (Lucigen, Middleton, WI) and 10 ng of metagenomic DNA with the following thermal cycling conditions: A touchdown PCR was performed with 10 cycles of 98^{o}C for 20 sec, 75-65^{o}C for 15 sec and 72^{o}C for 5 min, followed by 30 cycles using the same conditions with a 65^{o}C annealing temperature. The PCR amplicons were resolved by agarose gel electrophoresis (SB gel run for 2 hrs at 165 V) and visualized using ethidium bromide staining on an AlphaImager HP gel documentation system

(ProteinSimple, Santa Clara, CA).

Table 1. Primers used in this study

Virophage	Target region	Forward primer (5'-3')	Reverse primer (5'-3')
YSLV5	MCP	ATGAGTGCCGACATTGAGAA	AGCCGAGATAATGTCCTGCT
YSLV5	gap	GGCTCGTGGCAGTCGGGATT	CGGCACCGTCGTCCTTCCAT
YSLV6	MCP	CTAGGCGGTCCTCAACAATC	CAGACATTACACCGCCAGAA
YSLV6	gap	ATGAGTTACGCCTGTGCAATTCTTCCA	ACATTTCATAAACACGCTTTAAGGGCT
YSLV7	MCP	ACGACCAGCGCCGGATTCAA	ACCGTGACGATGGCATTCACT
YSLV7	gap	ATGCGACGCCTATGCAATGGC	GCGGCTGAGAATAATGCAGGGC

In order to close the gaps of the assembled sequences to form a circular genomic DNA for each of the virophage genomes, primers were designed based on the ends of the assembled sequences (Table 1). Genomic DNA extracted from photic zone lake samples that resulted in a PCR amplicon using the virophage-specific MCP primer set was used as the template. A KAPA Hifi^{TM} HotStart ReadyMix (Kapa Biosystems, INc., Wilmington, MA) was used to perform a touchdown PCR as indicated above. As for sequencing, PCR products were purified using a DNA Clean & Concentrator kit (Zymo Research, Irvine, CA) and quantified by Qubit® dsDNA BR Assay Kit (Life Technologies, Grand Island, NY) according to manufacturer protocols and Sanger sequenced (Lucigen Corporation, Middleton, WI). Sequences were trimmed for quality using the CLC Genomics Workbench (CLC Bio, Cambridge, MA) and then used for assembly of virophage genomes using the Geneious Pro bioinformatics package.

2.3.4 Prediction and annotation of virophage encoded genes

Geneious Pro was used to predict virophage open reading frames (ORFs) with a start codon of ATG, minimum size of 150 bp and standard genetic code. Predicted ORFs were compared to the GenBank database by BLASTp and PSI-BLAST programs [121,122]. The translated ORFs were annotated using the InterProScan 5 program and NCBI Conserved Domain Search [123,124]. A local virophage database, containing all predicted ORFs in all known virophages, including YSLVs 1-4 and the three new virophages described in this study, was constructed for further identification and analysis of homologous genes.

2.3.5 Phylogenetic analysis of conserved virophage genetic loci

Alignments of predicted amino acid sequences of three virophage core genes (ATPase, MCP, and Pro) were performed by MAFFT (version 7) and then concatenated [125]. The concatenated alignment was input into RAxML (version 8) for reconstruction of phylogenetic trees using maximum likelihood with 1000 iterations [126].

2.3.6 Nucleotide sequence accession numbers

The genomic sequences of the three Yellowstone Lake virophages (YSLVs) have

been deposited in GenBank under the accession numbers KM502589 (YSLV5), KM502590 (YSLV6) and KM502591 (YSLV7).

2.4 Results

2.4.1 Abundances and distribution of YSLVs in different Yellowstone Lake samples

Based on relative abundance data from the metagenomic datasets, YSLV1 was the dominant virophage, contributing up to approximately 0.17% of the total library reads (n= 526,420), and was roughly up to 4 to 10-fold more abundant than the other virophages in the same metagenomes (Fig. 1). YSLV1 distribution was biased towards mixing and photic zone environments, and potentially interesting, this virophage was either below detection or at relatively very low abundance in all but one of the samples acquired from the West Thumb region of the lake (Fig. 1). In making similar comparisons for the other virophages, there appeared to be no strong evidence of potential lake region provenance, with distributions being of relatively similar abundance across the lake (Fig. 1).

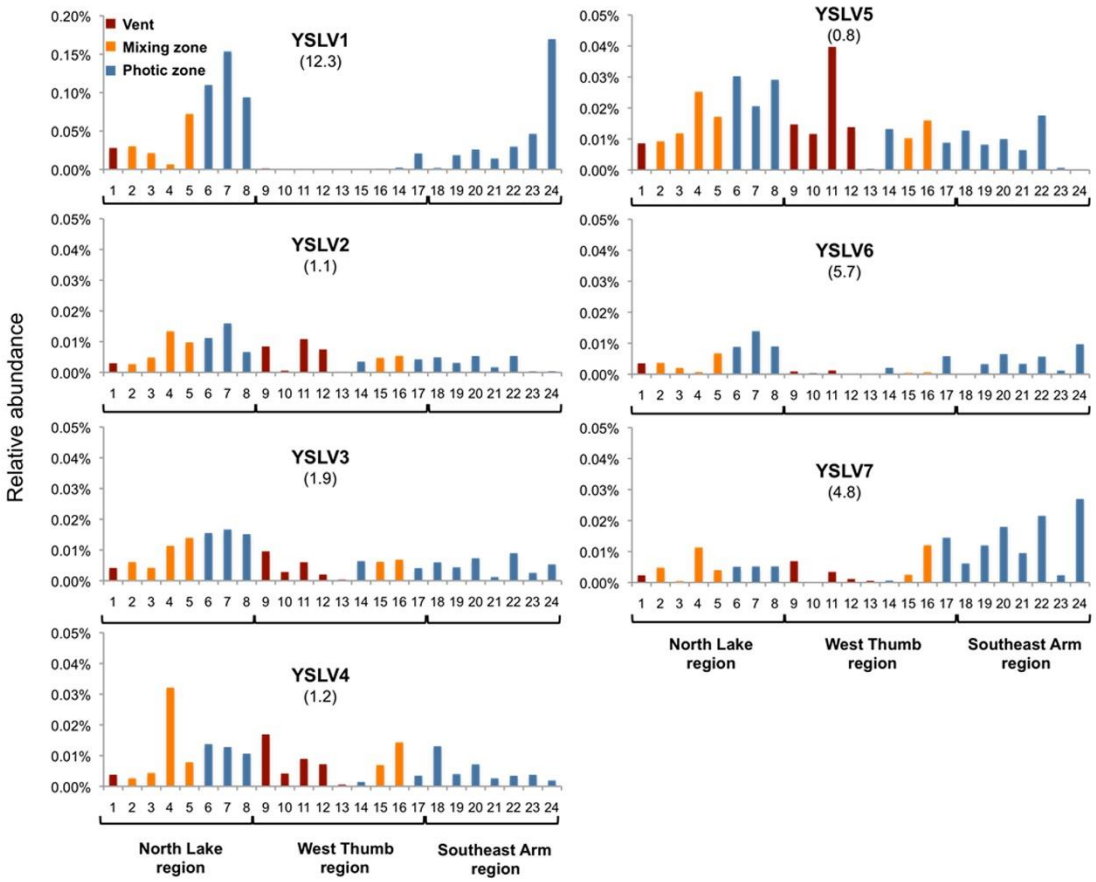


Figure 1. Distribution and abundance patterns of seven dominant virophages in Yellowstone Lake, Yellowstone National Park as determined by relative representation in 24 different metagenomes. The ratio of relative abundance of each virophage in the photic zone water to vent samples (photic : vent) is provided parenthetically below each virophage identifier. Samples 6, 7, and 8 are size-fractioned samples from a single Inflated Plain photic zone water sample, i.e., 0.1-0.8 μm, 0.8-3 μm and 3-20 μm size classes, respectively. Similarly, samples 22, 23, and 24 represent a single size-fractioned photic zone water sample acquired in the Southeast Arm (same respective size classes). Otherwise, all samples are from 0.1-0.8 μm size fractions. Note the difference in Y-axis scale for YSLV1.

Because of the nature of the lake sampling effort, virophage distribution could also

be examined in terms of lake microenvironment and biomass size. To varying degrees, all of the virophage were detected in vent water metagenome samples (40-68 °C), though relative abundances were biased towards the coolest temperatures in the photic zone samples (Fig. 2). There were no virophages associated with streamer samples (libraries not shown), which are extensive macroscopic community assemblages intimately associated with the orifice of lake floor vents, and that were sampled and washed separately from vent waters.

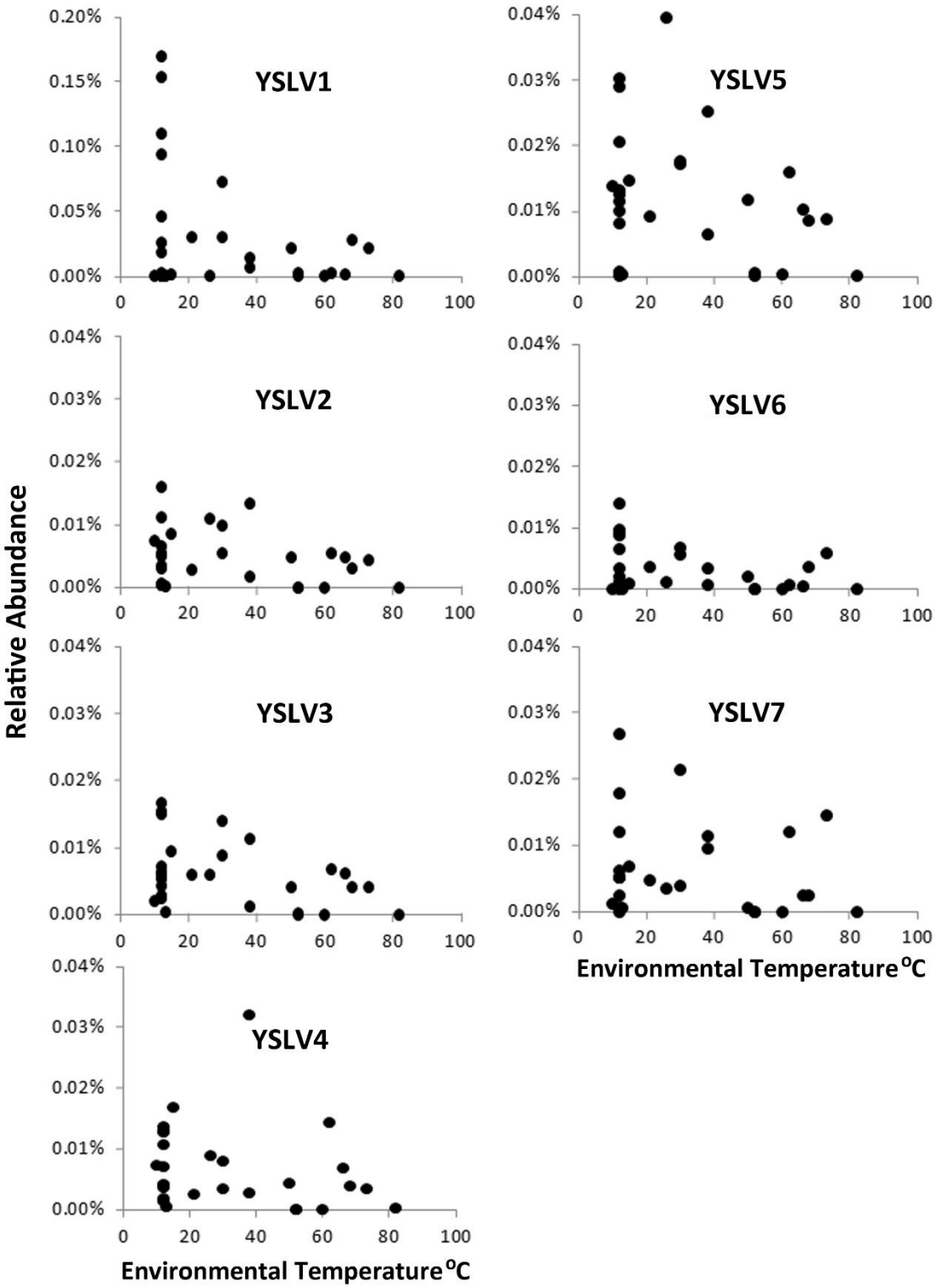


Figure 2**.** Relative abundance distributions of each virophage as a function of environmental sample temperature. Y-axis scales are the same as used for Fig. 1 and note again the scale difference for YSLV1.

Separate metagenome libraries from the Northern and the Southeast Arm lake regions were also developed to represent differing biomass size classes of 0.1 – 0.8, 0.8 – 3.0 and 3.0 – 20.0 μm. Virophage YSLV1 exhibited an abundance pattern in the Southeast Arm (Fig. 1; samples 22, 23, and 24), suggesting it and/or its host were more prevalent in the smallest size fraction category (Fig. 1). However, this pattern did not hold in the Northern lake region (Fig. 1; samples 6, 7, 8). Size distribution of YSLV5 in the Southeast Arm photic samples suggested it predominated in the largest (20.0-3.0 μm) category, but again this also was not demonstrated significantly in the northern lake region samples (Fig. 1).

2.4.2 Complete genomes of three novel virophages

Based on the initial metagenomic assembly, a total of 28 incomplete contigs were identified, revealing a potentially high diversity of virophages that did not have sufficient abundance and genome coverage in order for their respective genomes to be completely assembled. The three largest and nearly complete genomes, ranging from 29-22 kb, were selected for more detailed analysis and genome completion by PCR. These virophage genomes were named as YSLV5, YSLV6, and YSLV7, respectively. The existence of the three novel virophages was first identified by PCR targeting unique MCP gene sequences within different Yellowstone Lake samples, and this enabled identification of specific Yellowstone Lake water photic zone samples that contained

these virophage genomes (data not shown) and thus target for targeted PCRs.

We were able to use the metagenomic DNA extracted from the virophage-containing photic zone samples in order to complete the three novel virophage genomes. The virophage-specific PCR amplicons primed from the ends of the assembled contigs and filled in the predicted gaps. Positive PCR amplicons were first confirmed by agarose gel electrophoresis (data not shown). Sequencing of these amplicons resulted in three DNA sequences that were 1096 bp for YSLV5, 565 bp for YSLV6 and 403 bp for YSLV7. Each of these sequences successfully assembled with their corresponding virophage genomes, forming three complete and circular virophage DNA sequences.

The genome of YSLV5 was 29,767 bp in length, consisting of 32 predicted ORFs. However, YSLV5 had a G+C content of 51.1%, which is much higher than that (ranging from 26.7 to 39.1%) of the other virophages including YSLVs 6-7 determined in this study. This might suggest a unique host range for YSLV5. Eleven predicted ORFs of YSLV5 were homologous to that of known virophages. Among these 11 ORFs, the top hits of all were ORFs of YSLVs except ORF06 and ORF11, which were homologous to V21 (sputnik) or ALM ORF20 (Ace lake mavirus), respectively. YSLV6 had a 24,837 bp genome with 26.8% of G+C content and 29 predicted ORFs. Sixteen ORFs were homologous to that of known virophages, 12 of which had the most significant BLAST hits to other YSLVs based on local BLASTp analysis. YSLV7 had a genome size of 23,193 bp, with a 27.3% G+C content and 26 predicted ORFs. Eleven ORFs were homologous to that of known virophages, 7 of which had top BLAST hits to other YSLV ORFs. Taken together, these results suggest that YSLVs are rather diverse but

more closely related to each other than to virophages identified in other environments instead of Yellowstone Lake.

2.4.3 Conserved virophage genes

Thus far, five conserved core genes have been detected in all known virophages, including a putative DNA helicase (HEL), packaging ATPase (ATPase), cysteine protease (PRO), major capsid protein (MCP) and minor capsid protein (mCP) [113,127]. They were also identified in YSLVs 5-7 through BLASTp and PSI-BLAST analyses. With the exception of HEL, the other four core genes had high aa similarities (42-62%) only with their virophage homolog counterparts, respectively, suggesting an early divergent evolution of virophages. Two hypothetical proteins that were only present in OLV and YSLVs 1-4 were also shared by YSLVs 5-7 and highlighted in red and purple in Fig. 3. In addition, the majority of virophage-homologous genes in YSLVs 5-7 had the highest similarity to ORFs of OLV or YSLVs 1-4. Accordingly, it suggests a closer evolutionary relationship of YSLVs 5-7 with OLV and YSLVs 1-4 than with Sputnik, Zamilon, Mavirus and ALM (Fig. 3).

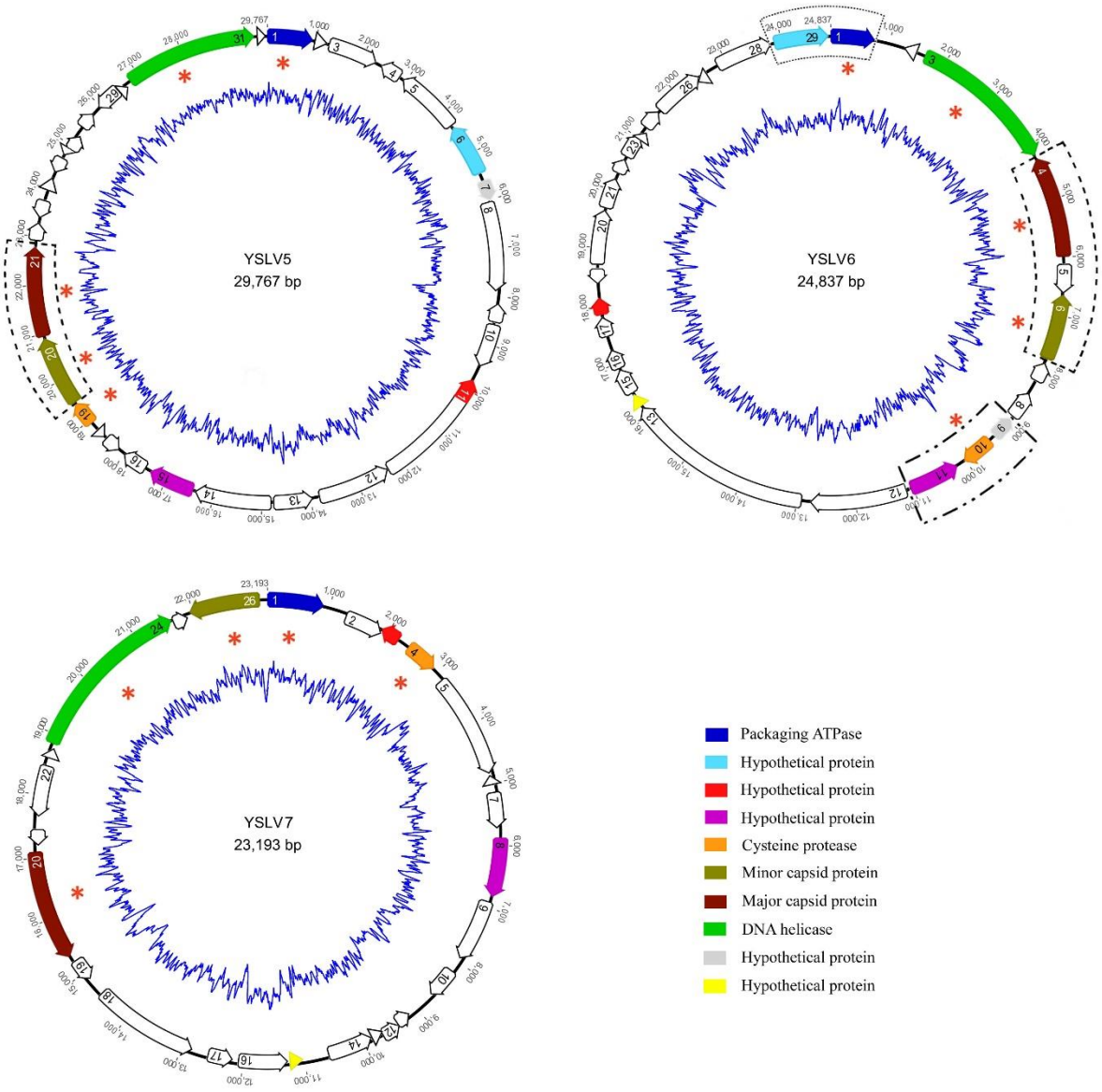


Figure 3**.** Circular maps of the complete genomes of YSLV5, YSLV6 and YSLV7. Homologous genes were labeled by the same color. Gene clusters that have conserved synteny among virophage genomes are highlighted using different kinds of dotted line. Skwiggly blue line in the center presents %GC skew. Red asterisk indicates five conserved genes of virophage including HEL, ATPase, PRO, MCP and mCP.

YSLV6 core genes MCP, mCP, PRO and ATPase revealed the highest sequence similarities (57-67%) to that of YSLV4. YSLV5 ATPase and mCP also showed the

highest similarities to that of YSLV4 with amino acid identities of 38.7 and 26.9%, respectively; MCP and PRO displayed high similarities to that of YSLV3 and YSLV7, with 29.4 and 32.0% aa identities, respectively. As for YSLV7, by contrast, ATPase and PRO were the most similar to that of OLV4 (28.9% aa identity) and YSLV5 ORF19 (32% aa identity); MCP and mCP, however, exhibited the greatest similarity to Zamilon ORF06 (22.1% aa identity) and ORF05 (28.1% aa identity), respectively, rather than to other the YSLVs or OLV.

YSLV5 ORF31, encoding a putative helicase, showed significant BLASTp hits to YSLV3 ORF11, YSLV6 ORF03, Zamilon ORF09, V13 and MV01 (E value < 10^{-5}). It contained a fusion domain of primase to SF3 helicase. The primase-helicase fusion protein is common among viruses [128,129]. The predicted helicase of YSLV6 ORF03 had 30.4 and 24.0% amino acid identity with YSLV3 ORF11 and YSLV5 ORF31, respectively. The primase-helicase fusion domain was also detected in YSLV6 ORF03. By contrast, like YSLV2 ORF10, YSLV7 ORF24 had a superfamily 1/2 helicase domain. These results suggest that virophage helicases underwent multiple recombination events during the evolution of these viruses.

Interestingly, YSLV6 ORFs 05, 16, and 21 contained a GIY-YIG endonuclease domain. This domain was also identified in Mavirus MV06, Sputnik V14, OLV OLV24, YSLV1 ORF09 and YSLV3 ORF12. YSLV6 ORF16 showed significant similarity to MV06 (34.1% aa identity) and YSLV1 ORF09 (35.3% aa identity). YSLV6 ORF05 was homologous to OLV01 (35.4% aa identity, E value 3.01^{-22}) and shared 38.0 and 29.9% amino acid identities with MV06 and YSLV1 ORF09, respectively. Although the GIY-

YIG domain was undetectable in OLV01 using InterProScan (data not shown), OLV01 shared significant sequence similarity with YSLV6 ORF05, suggesting either a potentially distantly related GIY-YIG domain or an alternative functional analog in OLV01 (data not shown).

2.4.4 Differences in gene synteny among newly discovered virophage genomes.

All previously described virophage genomes had a cluster of two adjacent genes encoding MCP and mCP with identical synteny. This conserved gene cluster was also present in YSLV5 and YSLV6 (Fig. 3). Another gene cluster consisting of the ATPase gene (ORF29) and a gene of unknown function (ORF01) was also discovered in YSLV6 (Fig. 3). This gene cluster was previously reported in YSLVs 2-4 and OLV, suggesting a closer evolutionary affiliation of YSLV6 with these virophages than with the others.

Surprisingly, the gene syntenies observed for the MCP and mCP genes and the third one of the HEL gene and an ORF with unknown function in known virophage genomes was absent within the YSLV7 genome. Although the genome of YSLV7 contained both MCP and mCP, they were separated by 4,843 bp (Fig. 3). This observation supports a distant evolutionary affiliation of YSLV7 with other virophage genomes.

2.4.5 Phylogenetic analysis

A phylogenetic analysis was first conducted using the predicted MCP amino acid

sequences for all available MCP gene sequences recovered from the Yellowstone Lake metagenomic datasets. In total, there were 32 full-length or partial virophage-like MCP sequences identified in distinct contigs. Of these 32 MCPs, 7 correspond to complete virophage genomes (four previously published, and three in this study), leaving 25 MCP sequences that hypothetically correspond to as-yet uncharacterized virophages. Of these 25 additional MCPs that are not associated with complete virophage genomes, seven were determined to be complete enough to provide sufficient alignment with other virophage MCP amino acid sequences. Maximum likelihood analysis revealed that the non-genome-associated MCP sequences affiliated closely with the genome-associated MCP sequences (data not shown).

A phylogenetic analysis was also conducted based on the concatenated amino acid sequences of three core genes ATPase, PRO and MCP in order to shed light on the evolutionary relationships of virophages with complete genome sequences. YSLV6 and YSLV4 appear to form a monophyletic group, which is in agreement with the results of gene content analyses (see above). Accordingly, they are the closest relatives, and YSLV6 is a new member of the virophage lineage comprising YSLVs 1-4 and OLV (Fig. 4). YSLV5 was more affiliated with YSLVs, excluding YSLV7, and OLV as well. They seem to have a common ancestor and to diverge from the Sputnik/Zamilon lineage with strong bootstrap support (Fig. 4), which is consistent with previous studies [114]. YSLV7 was distantly related to any other virophages and apparently represented a novel virophage lineage.

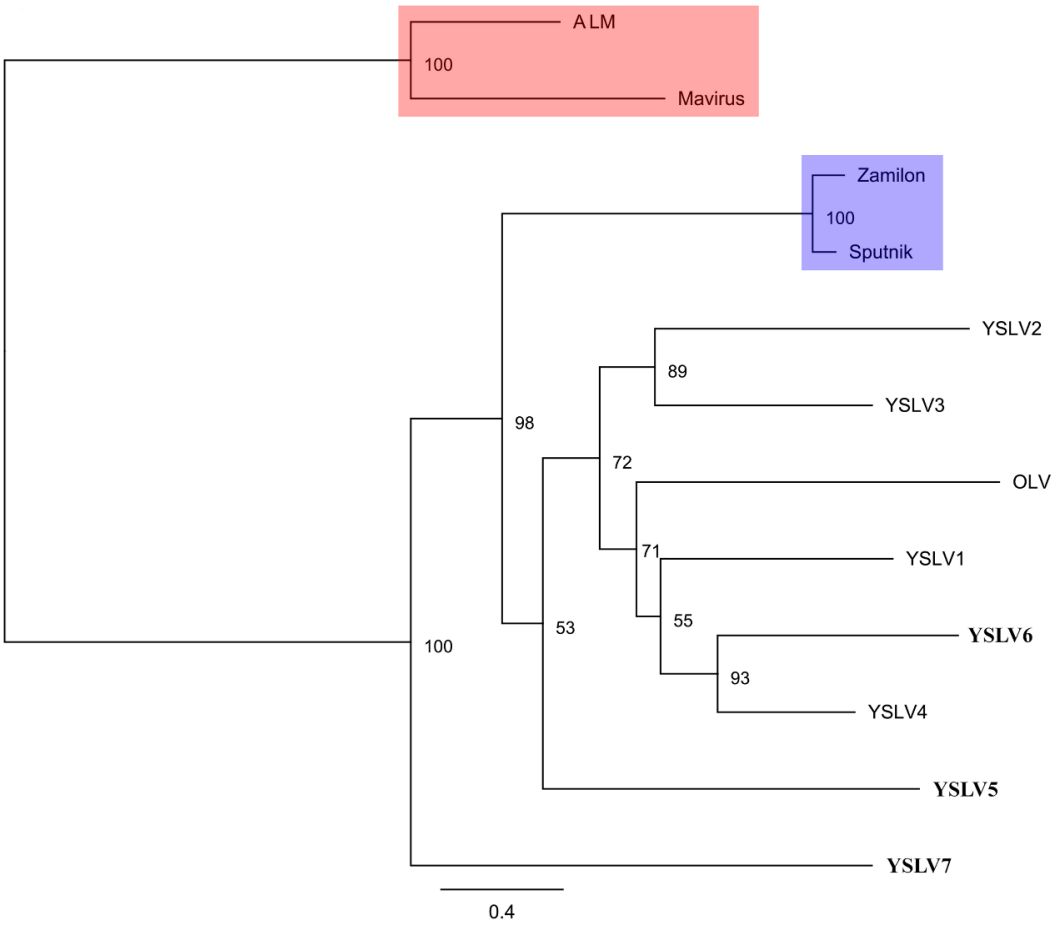


Figure 4**.** Maximum likelihood-based phylogenetic analysis of the seven Yellowstone Lake virophages based on the concatenated alignment of MCP, PRO and ATPase amino acid sequences (1398 aa). Bootstrap values (1000 iterations) are indicated at each node. YSLV5, YSLV6 and YSLV7 obtained in this study are shown in bold. The distinct lineages are labeled on the tree.

2.5 Discussion

Using the strategy of targeting the conserved major capsid protein as a genomic anchor to assemble shotgun metagenomic sequences derived from different sampling locations in Yellowstone Lake, this study revealed the presence of at least 32 distinct

virophage genotypes in the unique Yellowstone Lake ecosystem. Three complete novel virophage genomes (YSLVs 5-7) are presented and are now in addition to four complete virophage genomes (YSLVs 1-4) previously described [113]. This was enabled by *de novo* PCR efforts using virophage-specific PCR primer sets with lake DNA from specific photic zone samples that were found to contain these novel virophages. These PCRs extended and closed each of the respective new virophage genome contigs.

The abundance, distribution and diversity of virophages are somehow associated with water chemistry and temperature (Fig. 2). It appears that non-hydrothermal environments, especially with water temperature < 30°C, are the primary Yellowstone Lake habitats for most of these viruses and presumably linked to the ecology of their eukaryotic and giant viral hosts, which are unable to tolerate and thrive at high water temperature. However, the abundance and distribution of YSLV5 appear unusual by comparison. For example, within a deep vent water sample collected in 2007 (sample 14, 52 m deep, 60-66°C), YSLV1 was below detection, and YSLVs 2-4 and 6-7 were less than 0.01% (Fig. 1). However, this sample recorded the highest abundance for YSLV5 (Fig. 1). Indeed, the general abundance pattern for these virophages suggests potentially real differences with respect to their environment preference or tolerance. YSLV5 occurrence in vent samples was nearly equal to photic zone water (vent: photic ratio, approximately 0.8), whereas average abundance of YSLV1 in photic zone samples was approximately 12 fold greater than in vent water samples. Although it is premature to suggest YSLV5 represents a novel thermophilic virophage, its high G+C content coupled with its divergent phylogenetic position (Fig. 4) and its different lake

distribution pattern (Figs. 1-2) implies the YSLV5 host ecophysiology somehow differs from the other YSLVs as well as all other virophages thus far characterized. In this context, it is important to note that all vent samples likely contain at least some non-vent water due to the inability to attain an absolute seal around vent openings with the rim of the sampling cup device of the remote operating vehicle (previously noted by Clingenpeel et al. 2011). The most prominent example would be sample 9, where the vent water emitted from a rock assemblage, within which lake water could easily circulate (see Suppl. Movie S3 in Clingenpeel et al. 2011). Accordingly, in terms of virophage ecology, we view the occurrence of the YSLVs in vent fluids with considerable caution. Perhaps an alternative and more reasonable view may be to relate relative abundances of these virophage in vent samples as being an indicator of their prevalence in lake bottom waters surrounding these vents, potentially reflecting their tendency to contribute to lake sediment detritus (sinking) or may be important information with respect to ecological differences of their hosts (e.g. spatial relationships in the photic zone).

Among the three novel virophages described in this study, YSLV6 appears to be closely related to OLV and YSLVs 1-4 and represents a new member in this lineage. This is supported by multiple lines of evidence, including the multilocus phylogenetic analysis (Fig. 4), the presence of three conserved gene clusters with similar gene synteny (Fig. 3) and the number of shared conserved genes. The four core genes ATPase, Pro, mCP and MCP present within YSLV6 have high % identities to the corresponding homologous genes of YSLV4, and these two virophages were phylogenetically grouped

together with more than 90% bootstrap support (Fig. 4). This evidence strongly suggests that YSLV6 and YSLV4 are close relatives. In contrast, the phylogenetic analysis indicates that YSLV5 is relatively distinct from the other YSLVs. Additionally, the unusual high % G+C content of YSLV5 also suggests a distinct host range for this virophage compared to other known virophages. Even though the phylogeny of YSLV5 is uncertain, both phylogenetic and gene content analyses indicate it is affiliated with YSLVs rather than with Mavirus or Sputnik lineages. Consequently, like YSLV6, YSLV5 possibly also belongs to the virophage lineage of YSLVs and OLV, albeit with distinct features. Most virophage homologous genes in YSLV7 (7 out of 11), including two conserved core genes of ATPase and PRO, reveal high sequence similarity to OLV and other YSLVs, suggesting the affiliation of YSLV7 with this clade. However, the two capsid protein genes that are usually highly conserved in viruses are most similar to that of Zamilon virophage instead of YSLVs, indicating a complicated evolution of YSLV7. In addition, the conserved gene cluster of MCP and mCP present in all other virophages is absent in YSLV7. Taken together, these results indicate that YSLV7 is the first member of a very distinct lineage from known virophages as shown on the phylogenetic tree (Fig. 4). As more virophages are discovered and characterized, the biological significance of these genomic differences and how this impacts the molecular interactions between virophages and their giant virus and eukaryotic hosts may be better understood.

In conclusion, this study has significantly broadened the perspective of virophage diversity and novelty in nature. The Yellowstone Lake metagenome libraries enabled

the discovery of new phylogenetic lineages that exhibit distinctly different genome structures as well as apparent distribution patterns that presumably are linked to some degree to host ecological preferences. Their discovery contributes to a long history of other novel finds preserved in the World's first national park.

Chapter 3. Novel Archaeal Thermostable Cellulases from an Oil Reservoir Metagenome

3.1 Abstract

In a previous study microbial assemblages had been sampled from an offshore deep sub-surface petroleum reservoir 2.5 km below the ocean floor off the coast of Norway under conditions of high temperature and pressure. In this study we used both shotgun sequencing and fosmid library construction to survey the functional and phylogenetic diversity of this extreme habitat. In addition, the metagenomic fosmid library containing 11,520 clones was screened using function- and sequence-based methods to identify recombinant clones encoding and/or expressing carbohydrate-degrading enzymes. Annotation of metagenomic shotgun and library sequences confirmed that Euryarchaeota taxa were dominant in this habitat. ORFs encoding carbohydrate-degrading enzymes were predicted through a batch BLAST against the CAZy database, and many fosmid clones expressing carbohydrate-degrading activities were discovered by functional screening of the library within an *E. coli* heterologous host. Each complete ORF predicted to encode a cellulase identified from sequence-based or function-based screening was subcloned into an expression vector. Each of the resulting five subclones were found to have significant activity using a fluorescent cellulose substrate, and three of these expressed cellulases were observed to be highly thermostable up to at least 80^{o}C. Based on phylogenetic analyses, the thermostable

cellulases were derived from thermophilic Archaea and are distinct from known cellulases. The cellulase F1C contains two distinct cellulase domains, perhaps resulting from the fusion of two archaeal cellulases, which is a novel protein structure that may result in enhanced cellulase activity and thermo-stability.

3.2 Introduction

Extremophiles have attracted great interest for their ability to thrive in conditions that sometimes include multiple environmental extremes and for their specific enzymes that have adapted to these extreme conditions and therefore may have industrial applications [5]. *Taq* DNA polymerase is the classic example of an enzyme from a thermophile, which was derived from *Thermus aquaticus* isolated in culture from hot springs at Yellowstone National Park [20]. The traditional approach used to identify and exploit these enzymes is culture-dependent, which greatly restricts the diversity of thermophilic natural products that are discovered. The relatively small percentage of microorganisms that can be readily cultured under laboratory conditions has led to the development of novel culture methods and the adoption of culture-independent methods [130,131,132].

However, the application of novel cultivation strategies is slow and will not enable cultivation of the extant diversity of microbial genomes from extreme environments, necessitating new approaches to exploit natural products from as yet uncultured microbes for industrial application [133]. The use of culture-independent metagenomic

methods permits access to microbial genomes and their biologically active molecules through isolation of DNA from environmental microbes followed by direct sequencing or cloning DNA to generate a metagenomic library [25,134]. The library can then be screened by both sequence-based and function-based methods for natural product discovery [135,136]. Each of these approaches have their own sets of biases, including the potential inability to heterologously express a cloned gene due to differences in the transcriptional and/or translational apparatus, and biases associated with annotation of shotgun sequences. Even with these well-recognized methodological biases, a metagenomics approach has been previously shown to be effective in discovery of enzymes with novel activities [135,137]. In fact, the very first published example of a functional metagenomic approach for enzyme discovery involved the cloning of cellulases from "zoolibraries" [138]. Recently, metagenomic approaches have been used to discover many novel carbohydrate-active enzymes (CAZymes) from soil [139,140], cow rumens [141], sediments [142], biochemical reactors [143] and aquatic environments [144,145].Petroleum reservoirs contain complex hydrocarbons trapped in porous rock formations and contain varying amounts of formation water that is high in salt content. In a previous study, samples were collected from a reservoir 2.5 km below the sea floor of the Norwegian Sea with an *in situ* temperature of 85°C and pressure of 250 bars [146,147]. The oil-producing well used for sampling never experienced any injection of microorganisms from the surface or other habitats, and during sampling contamination and cell lysis due to rapid shifts in temperature and pressure were avoided via use of a pressurized chamber that slowly transitioned samples to a moderate temperature and

pressure [146,147]. The sampled habitat is under multiple extremes with high pressure, temperature and salt, and contains abundant carbon sources that in principle could support microbial growth [148]. Microorganisms have adapted to such extreme conditions through the evolution of thermostable enzymes [5].

This study focused on discovery of thermo-stable CAZymes, especially cellulases, from the oil reservoir metagenome using both sequence-based and function-based screening. Cellulases are glycoside hydrolases (GH) that are able to hydrolyze 1,4-beta-D-glycosidic linkages in cellulose, hemicellulose, lichenin, and cereal beta-D-glucans. In the CAZy database, there are a total of 131 GH families, 17 of which include enzymes with cellulase activities [149]. Cellulases have potential application in biofuel production (e.g. cellulosic ethanol) by decomposition of plant-derived cellulose into monosaccharaides before fermentation [150]. However, biofuel production processes such as simultaneous saccharification and fermentation (SSF) or separate saccharification and fermentation (SHF) require high temperature and will inhibit the activity of currently available cellulolytic enzymes [151]; therefore, identification of novel thermostable cellulases could be a vital step in improving cellulosic ethanol production. The availability of specific carbon substrates (e.g., cellulose) within these oil reservoir samples was unknown, but given the formation of these petroleum reservoirs from residues of phytoplankton and zooplankton, we hypothesized that carbohydrate-degrading enzymes would be encoded by the metagenomes in the habitat and would have evolved enhanced thermal stability[152].

3.3 Materials and Methods

3.3.1 Oil reservoir sampling, DNA isolation and handling

Oil reservoir samples were collected and metagenomic DNA isolated in a previous study [147]. Isolated DNA was used in direct 454 pyrosequencing for metagenomic analysis of the phylogenetic and functional diversity of this oil reservoir microbial assemblage [147]. For preparation of a fosmid library, DNA was amplified using Phi29 polymerase (WGA; Qiagen REPL midi kit) in individual reactions of 50 µl and pooled after amplification. Two individual rounds of amplification were conducted using DNA from each sample (water and oil phases). The amplified DNA was isolated and purified using Qiagen QIAamp DNA mini kit using the manufacturers protocol and used in fosmid library construction using the Epicentre pCC1FOS system, as described in Aakvik *et al* [153]. The 11,520 clones of the resulting fosmid library were arrayed in 30x 384-well microtiter plates.

3.3.2 Extraction of fosmid DNA from the fosmid library

Each plate of the 384-well formatted *E. coli* metagenomic library was used to inoculate a deep-well 384-well plate containing 170 µl per well of LB broth containing 12.5 µg/ml chloramphenicol and 0.01% (w/v) arabinose for plasmid copy-number induction. After 24 hours of growth at 37ºC while shaking at 200 rpm, the clones of every two deep 384-well plates were pooled into a single 250 ml centrifuge bottle, and

the pooled fosmid DNAs were extracted using a QIAGEN Large-Construct Kit, resulting in 15 separate samples (i.e., two 384-well plates per pool). These fosmid DNAs were then incubated with Plasmid-Safe™ ATP-Dependent DNase (Epicentre) to reduce chromosomal contamination.

3.3.3 Sequencing of the fosmid library

Each pooled fosmid DNA prep was used as a template in a Nextera DNA Sample Prep Kit reaction (Illumina, San Diego, CA), and a unique set of bar-codes was used for each pooled library plate. Then the fragmented DNA was purified using a DNA Clean and Concentrator Kit (Zymo Research, Orange, CA) and further used for an amplification reaction according to standard Illumina protocols. The amplified library was then purified using a Size-Select IT kit (Omega Biotek, Norcross, GA) to isolate the desired DNA size fraction (~500 bp on average). The purified, bar-coded DNA fragments were quantified using a Qubit fluorimeter followed by pooling at equimolar concentrations and denaturing using 0.1 N NaOH. Finally, the pooled fosmid DNAs were used for an Illumina HiSeq sequencing run with a 2 x 100bp paired end sequencing kit (Illumina, San Diego, CA).

3.3.4 Bioinformatic analysis of the fosmid library and shotgun sequences

Two different strategies were used for sequence-based screening. In the first

strategy, only HiSeq reads from the fosmid library were used for assembly and further sequence-based screening. The raw sequences generated by each HiSeq run were imported into the CLC Genomics Workbench (Qiagen, Cambridge, MA), and trimmed at a stringency of 0.01 (equivalent to Q score of >40). Trimmed sequences were assembled *de novo* using the CLC Genomics Workbench to generate a set of contigs per each fosmid pool. ORF prediction was then performed using "ORF finder by six-reading-frame" on Camera Portal 2.0 [53]. The predicted ORFs were used for a batch BLASTp against the CAZy database using the tool dbCAN for identification of carbohydrate-degrading enzymes as well as lipases/esterases [149,154]. In addition, all raw sequence reads recovered from the fosmid library were also exported to MG-RAST to profile microbial diversity and abundance based on phylogeny and function [155]. In order to compare microbial diversity present within the fosmid library (plate 3, 4, 9, 10, 13, 14, 17 and 18) to that of shotgun sequences from the same sample (Well II), trimmed sequence reads from direct 454 pyrosequencing [147] were also uploaded to MG-RAST for analysis. Organism abundances were predicted using "Best hit classification" (Max. e-Value Cutoff: 1e-5, Annotation Sources: M5NR). Functional abundances were predicted using "hierarchical classification" (Max. e-Value Cutoff: 1e-5, Annotation Sources: Subsystems).

In the second strategy, we combined fosmid library HiSeq reads from plates 3, 4, 9, 10, 13, 14, 17 and 18 with shotgun sequences generated from 454 pyrosequencing from the same sample (Well II) in order to achieve longer contigs. The raw 454 sequence reads from shotgun sequencing were imported into the CLC Genomics

Workbench and trimmed at a stringency of 0.01 (equivalent to Q score of >40). Trimmed reads were randomly subsampled into 10 separate files at 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100% of total reads. Each subsample was *de novo* assembled separately, followed by comparison of assembly statistics. The best assembly was obtained from the use of 90% of the shotgun reads, based on average length, N80, N50, N20 and maximum contig sizes, and these contigs were then used for assembly with fosmid library reads. HiSeq reads from the selected fosmid library plates were also subsampled at 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100% of total reads, and separately *de novo* assembled with contigs resulting from the 90% sub-sampling of shotgun reads using SPAdes 3.5.0 [156]. The best assembly result was generated using 20% of the sub-sampled fosmid library reads together with contigs from the shotgun reads, evaluated as above. The contigs resulting from each of the above assembly strategies were used for ORF prediction using Prodigal [157]. The predicted ORFs were used for a batch BLASTp against the CAZy database using the tool dbCAN for identification of CAZymes as well as lipases/esterases [149,154].

3.3.5 Functional screening of the fosmid library for carbohydrate-degrading activity

Assays for five different hydrolase enzymatic assays were conducted with five substrates to functionally screen the library. In each assay, the *E. coli* fosmid clones were grown overnight at 37ºC in 96-well plates with each well containing 200 µl of LB

broth including 12.5μg/ml of chloramphenicol, while shaking at 200 rpm. After overnight growth the *E. coli* cultures were inoculated onto the respective agar medium that included 0.01% arabinose to induce plasmid copy-number using a pin replicator. Cellulase and xylanase activities were screened using LB agar containing 0.1% CMC and 0.1% xylan (beechwood), respectively [158,159]. The amylase assay medium was comprised of 1% tryptone, 0.25% yeast extract, 0.5% K_2HPO_4, 0.3% starch (soluble) and 1.5% agar [160]. The protease assay utilized 2% skim milk, 0.5% yeast extract, 0.08% sodium citrate dehydrate and 1.5% agar [161]. LB agar with 1% tributyrin was used to detect the activity of esterases/lipases [162]. After 37°C incubation overnight, all agar plates except starch agar plates were incubated at 60°C overnight again and further fumigated with chloroform for 1 hour to lyse *E. coli* cells. Halos of clones expressing proteases or esterases/lipases could be directly observed. For the three other enzymatic assays, colonies were first removed using 95% ethanol and dH_2O. Then CMC and xylan agar plates were stained using 1% Congo red for 15 min and de-stained using 3M NaCl. Clones with cellulase activity could be identified from the yellow halos around the clone. For the cell lysis step, starch agar plates were fumigated for 1 hour with chloroform at room temp. Then an iodine solution (0.3% iodine and 0.6% potassium iodine) was used to stain starch agar plates. After 15 min staining, clones showing obvious halos were identified as amylase-positive clones. The positive clones were re-streaked from original wells onto agar plates with their respective substrates, and tested for validation. Only clones that were validated as positive upon re-testing were selected for further analyses.

3.3.6 Sequencing fosmid clones that express cellulase activity

Fosmid clones with reproducible cellulase activity were selected for next-generation sequencing. Fosmid clones were inoculated into 500 ml LB broth with 12.5μg/ml chloramphenicol and 0.01% arabinose for plasmid copy-number induction. After incubation overnight at 37^{o}C, each fosmid clone DNA was separately extracted using the Large-Construct DNA isolation kit (Qiagen). A Nextera DNA Sample Prep Kit (Illumina, San Diego, CA) was employed for preparation of bar-coded fosmid DNA clone sub-libraries, with each clone separately bar-coded, purified and quantified as described above. The pooled fosmid clone DNAs were then sequenced using an Illumina MiSeq with a 2 x 300 bp paired-end sequencing kit (Illumina, San Diego, CA). After sequencing, the clone sequences were trimmed, assembled *de novo* and ORFs were predicted using the CLC Genomics Workbench. Cellulase ORFs of each clone were annotated by a BLASTp search.

3.3.7 Subcloning of cellulase genes

Predicted cellulase encoding ORFs from six clones expressing cellulase activity along with complete or nearly complete cellulase gene ORFs identified from pooled library sequencing were selected for subcloning. Each respective ORF was PCR amplified and subcloned into the Expresso Rhamnose SUMO subcloning system

(Lucigen, Middleton, WI) and electroporated into *E. coli* 10G. Subclones able to express cellulase were selected after growing on CMC agar and staining (1% Congo red, 15 min).

3.3.8 Thermal stability test of subclones with cellulase activity

Two methods were used to evaluate the thermal stability of subclones expressing cellulase activity. A broth culture of each clone (0.2% rhamnose and 30 μg/ml Km, 37°C overnight) was collected and heated at a series of temperatures for different incubation times. In the first method, supernatants of cell lysates (using chloroform) of clones were heated at 37°C, 60°C, 70°C and 80°C for 1h, 2h, 3h or 6h, then spotted onto CMC agar plate, and those with yellow halos were recorded. The second method utilized 4-Methylumbelliferyl-β-D-cellobioside (MUC), a fluorescent cellulase substrate, to quantify cellulase thermal stability [163]. MUC has been widely used to assay exo-cellulase activity, and has also reported to be capable of detecting endo-cellulase and beta-glucosidase activities [164,165,166,167,168]. Equal volumes of 100 μM MUC was added into heated supernatants of cell lysates from subclones (37°C, 60°C or 80°C) for 6 hours in a 96-well plate followed by incubation at 37°C overnight. Next day, the fluorescence of each well was monitored using an excitation at 375 nm and emission of 445 nm with a BioTek Cytation 3 plate reader (Thermo Fisher Scientific Inc).

3.4 Results

3.4.1 Functional and phylogenetic classification of shotgun and fosmid metagenomic sequences

Metagenomic sequences from both pooled fosmid and direct shotgun sequencing of an oil reservoir [147] were uploaded into MG-RAST. A series of metagenomic analysis tools (e.g. “Best hit classification” and “hierarchical classification”) in MG-RAST were applied to compare the functional and phylogenetic composition of the two sequence databases. In both shotgun sequences and fosmid library sequences there was a very high abundance of genes derived from the domain Archaea (E value < 10^{-5}), followed by hits to the domain Bacteria, and with very few hits against viruses or Eukaryotes. At the phylum level, *Euryarchaeota* and *Proteobacteria* were found to be the most abundant in both fosmid and shotgun databases (Fig. 5A). However, approximately 20% of sequence reads in both databases were allocated to the category “unassigned”, indicating that the oil reservoir harboured a large number of unknown microbial taxa (Fig. 5A). In contrast, two bacterial phyla had apparent differences in relative abundance in the two databases, with taxa affiliated with the phylum *Proteobacteria* found more frequently in the shotgun database, and taxa affiliated with the *Bacteroidetes* found more frequently in the fosmid library (Fig. 5A).

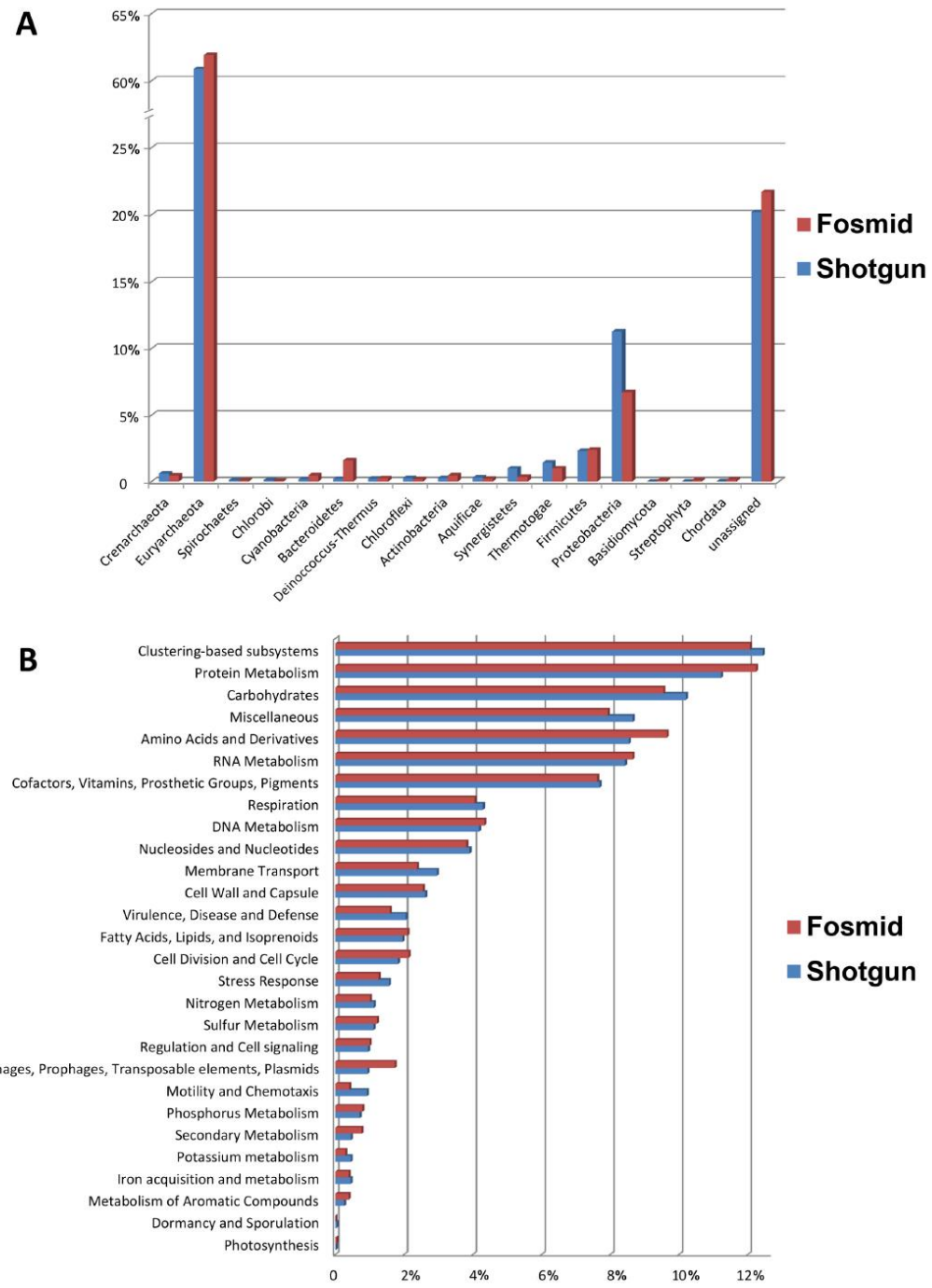


Figure 5. Relative abundance of shotgun sequences and fosmid metagenomic library sequences at the phylum level (Panel A) and based on functional classification as compared to the SEED database [169] (Panel B).

In addition to a phylogenetic analysis, a functional classification indicated that the

fosmid and shotgun sequences had a similar distribution of functional category relative abundances (Fig. 5B). The category "Carbohydrate", including many carbohydrate-degrading enzymes, was the third most abundant (Fig. 5B), and indicated that many carbohydrate-degrading enzymes were encoded within the oil reservoir metagenome in both databases.

3.4.2 Identification of carbohydrate-degrading enzymes by sequence-based screening

The Illumina HiSeq sequencing of the fosmid library generated 40.1 Gbp of sequence reads; after trimming, we obtained 37.0 Gbp of quality sequence with an average read length of 92bp. These HiSeq reads were assembled *de novo* yielding 697,947 contigs, with an average coverage for these contigs larger than 1kb ranging from 3.4x to 112.5x. The ORFs predicted from these contigs were queried against the CAZy database using a local BLASTp search, leading to the discovery of 29,764 ORFs with significant BLAST hits (E-value < 10^{-5}). These ORFs were derived from 28,913 contigs and included six CAZy families including auxiliary activities (AA), carbohydrate-binding modules (CBM), carbohydrate esterases (CE), glycoside hydrolases (GH), glycosyltransferases (GT) and polysaccharide lyases (PL). Based on the results of a local BLASTp against the CAZy database, we obtained 101 significant hits for cellulases, 21 hits for xylanases, 174 hits for amylases, 39 hits for proteinases/peptidases and 102 hits for esterases/lipases (Table 2). All cellulase,

xylanase and amylase hits were described as members of the GH group.

Table 2. Number of positive hydrolase hits from the oil reservoir metagenomic library identified from either functional screening using specific substrates, or by sequence-based screening using BLAST searches against a local CAZy database.

CAZyme	Functional screening	Sequence-based screening (Only fosmid library reads)
Cellulase	6	101
Xylanase	2	21
Amylase	85	174
Protease	33	39
Esterase/Lipase	9	102

In order to achieve greater contig lengths, HiSeq reads were sub-sampled from shotgun sequencing of Well II as well as the corresponding plates from the fosmid library. Assembly of 90% of the shotgun reads yielded 4,938 contigs with an average length of 2752 bp (N50: 9,913 bp, N20: 56820 bp and maximum contig size: 480,455 bp). The resulting contigs were further assembled with 20% of the fosmid library reads, generating 655 contigs with average length of 6273 bp (N80: 22,852 bp, N50: 63,318 bp, N20: 139,660 bp and maximum contig size: 279,442 bp). Both of these contig sets were used for ORF prediction, followed by a local BLASTp search against the CAZy database. This resulted in a total of 1,432 ORFs with significant BLASTp hits (E value $< 10^{-3}$). Within these significant BLASTp hits, there were 13 significant hits for cellulases, 14 hits for xylanases, 34 hits for amylases, 35 hits for proteinases/peptidases

and 29 hits for esterases/lipases (Table 2). The numbers of hits for each enzyme class were much less than that identified using either the fosmid library HiSeq reads or the shotgun reads. This was likely due to the pooling together of library and shotgun reads to generate longer contigs, resulting in less repetitive hits.

3.4.3 Identification of carbohydrate-degrading enzymes by function-based screening

For each of the targeted CAZymes we discovered a greater number of enzymes via sequence-based compared to function-based screening (Table 2). For the comparison between function- and sequence-based screening results, only the CAZyme hits identified using assembly of the entire fosmid library were used so that these would be directly comparable to the function-based screening. For cellulases, a total of six validated clones were identified from functional screening (0.052% hit frequency), whereas 101 putative cellulase-encoding ORFs were identified from sequence-based screening (0.88% hit frequency). There were two validated clones that expressed xylanase activity (0.017% hit frequency), whereas 21 putative xylanase-encoding ORFs were identified by sequence-based screening (0.18% hit frequency). There were 85 clones that expressed amylase activity (0.74% hit frequency), with 174 putative amylase-encoding ORFs found by sequence-based screening (1.51% hit frequency). There were 33 protease-expressing clones (0.29% hit frequency), with 39 putative protease-encoding ORFs identified by sequence-based screening (0.34% hit frequency).

Lastly, nine clones were discovered that expressed an esterase or lipase activity (0.078% hit frequency), with 102 putative lipase- or esterase-encoding ORFs identified from sequence-based screening (0.89% hit frequency). For each of the enzyme classes there was an increase in the number of identified genes via sequence-based screening relative to function-based screening, ranging from 1.2-fold higher in the case of proteases to 16.8-fold higher for cellulases, suggesting that many clones identified from sequence-based screening were not expressed or active in an *E. coli* heterologous host.

Among the different CAZy classes, we selected cellulases for further characterization due to their potential industrial application in the generation of cellulosic ethanol. All six clones that expressed a cellulase were tested for their thermal stability. Three clones P16O17 (indicated hereafter as "F1"), P17M3 (indicated hereafter as "F4") and P4C10 (indicated hereafter as "F6") gave obvious halos on CMC agar assays after the clone supernatants had been incubated at elevated temperatures, whereas P17H9 (indicated hereafter as "F2"), P8E14 (indicated hereafter as "F3") and P17C9 (indicated hereafter as "F5") did not show any evidence of cellulase activity at higher temperature. Among the temperature-resistant clones, clone F1 had cellulytic activity at all temperatures (37°C, 60°C, 70°C and 80°C) and at 1, 2, and 3 hours of incubation. In contrast, the clone F6 supernatant lacked activity when heated at 80°C (1, 2 and 3 hours), and clone F4 supernatant did not have observable activity after 80°C incubation for 3 hours. The cellulase activity expressed by F1 was observed to be the most thermostable and the most efficient at cellulose degradation based on this CMC substrate assay. Clone F1 also had cellulase activity in the quantitative assay using the

MUC substrate (Fig. 6). Interestingly, although clone F5 did not have apparent activity against the CMC substrate, it did demonstrate strong activity when tested using the MUC substrate at 37°C or 60°C.

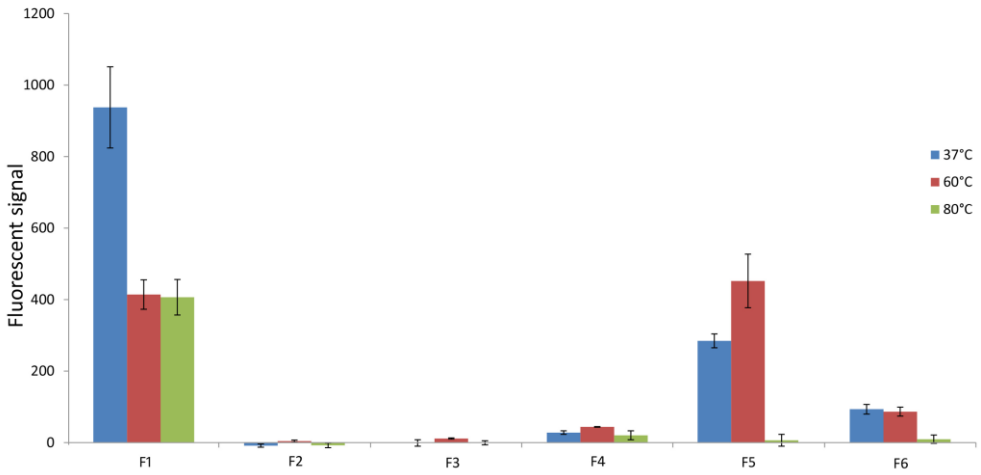


Figure 6. Quantitative MUC assay for six fosmid clones that were identified using a functional assay with the CMB substrate. The activity is reported as units of fluorescent signal intensity. Supernatants of the six clones were incubated at 37°C (blue), 60°C (red) or 80°C (green) to test the thermal stability of each cellulase.

The thermostability of each expressed cellulase was evaluated in the quantitative MUC assay. The supernatant of each respective clone was heated at 37°C, 60°C or 80°C for 3 or 6 hours, and then incubated with the MUC substrate. Clones F1, F4, F5 and F6 showed a strong fluorescent signal in the MUC assay (Fig. 6). The activity of F1 using the MUC substrate was still the highest among all clones and activity remained high, but this signal was reduced when the temperature was increased to 60°C or 80°C (Fig. 6). Interestingly, the broth culture of clone F5 heated at 60°C showed increased activity

in the MUC assay relative to the activity at 37°C or 80°C (Fig. 6), suggesting that the cellulase expressed by clone F5 has a temperature optimum around 60°C. The cellulose-degrading activities of clones F4 and F6 were relatively weak as determined in the MUC assay, but since these data were not normalized per mg protein this may reflect lower expression of the cellulase rather than low enzymatic activity (Fig. 6). The MUC activity of clone F6 gradually reduced when temperature increased, whereas that of clone F4 remained stable at all three temperatures (Fig. 6). The cellulase activities from these clones identified by functional screening were observed to be distinct in terms of their thermostability, and the differences in activities observed probably reflected changes in their protein structure and activity at different temperatures.

3.4.4 Sequence analysis of cellulase ORFs identified from both sequence-based and function-based screening

The six cellulase positive fosmid clones identified by functional screening were used to prepare sub-libraries with Nextera bar-codes and pooled together for sequencing using an Illumina MiSeq. The respective fosmid clones were separately analyzed and a set of contigs were obtained for each clone, from which cellulase-encoding ORFs were detected. Interestingly, clone F1 has three predicted glycoside hydrolase (GH) domains that represent two different GH classes (Fig. 7). A domain at the N-terminus, F1_1, exhibits 86.8% amino acid identity to the endocellulase of the archaeon *Pyrococcus horikoshii* (residues 67 to 385), and affiliates with the GH5 class. In contrast, the two

domains at the C-terminus affiliate with the GH12 class, with domain F1_2 exhibiting 59% amino acid identity to one glycoside hydrolase of the archaeon *Ignisphaera aggregans* DSM 17230 (residues 537 to 749), and domain F1_3 exhibiting 84.1% amino acid identity to the endo-1,4-beta-glucanase of the archaeon *Pyrococcus furiosus* DSM 3638 (residues 856 to 1,081) [124]. . In addition to the GH5 and GH12 domains cellulase F1 is predicted to have a carbohydrate-binding module 2 (CBM2, residues 1,227 to 1,315) that may assist in binding cellulose substrates (Fig. 7). The predicted 3D structure of F1 was estimated using the Swiss-Model server using an endo-1,4-beta-glucanase (3axx.1.A) and Endoglucanase A (3vgi.1.A) as templates to model the GH5 and GH12 domains, respectively (Fig. 8) [170]. Since the F1 domain F1_2 (in yellow, Fig 8) only has 35.8% amino acid identity to 3vgi.1.A, the weak homology of domain F1_2 to known cellulase domains may preclude an accurate *in silico* model. The cellulase genes from clones F2 and F3 were predicted to have a GH12 domain that is identical to the C-terminus of F1 (residues 866 to 1,322) including the F1_3 domain (Fig. 7). But for clones F2 and F3, they lacked the GH5 domain present in F1 (F1_1), and this could explain their apparent lack of thermostability. In the case of clone F4, there were two predicted overlapping cellulase ORFs (F4_1 and F4_2) that had 78.8% amino acid identity to the endo-1,4-beta-glucanase of the archaeon *Pyrococcus furiosus* and 79.6% amino acid identity to the endocellulase of archaeon *Pyrococcus horikoshii*, respectively. Interestingly, F4_1 had one cellulase domain of GH5 identical to that of F1_1 and F4_2 had two cellulase domains of GH12 (named F4_2_1 and F4_2_2) that were also identical to GH12-related domains of F1, but the direction of F4_2 was

reversed compared to the region of F1 covering both F1_2 and F1_3 domains (Fig. 7). The overlapping region of F4_1 and F4_2 was actually the end of their C-terminus, which was 72aa length and distinct with that of the F1 GH12-related domain (only 9.8% amino acid identity, data not shown) (Fig. 7). Clone F5 contains a predicted cellulase that is identical to that of the cellulase predicted from clone F6, which has 99.7% amino acid identity to the endoglucanase of the bacterium *Thermosipho africanus*.

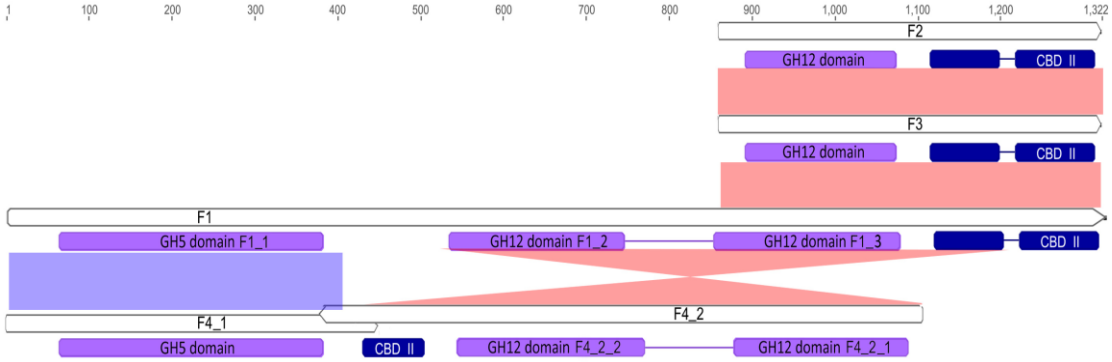


Figure 7. Domain annotation of the cellulases F1, F2, F3, F4_1 and F4_2, which was predicted by interproscan [124]. Cellulase domains were labeled in purple and CBM domains were labeled in blue. Identical regions between different cellulase sequences were presented in light blue and red shallow.

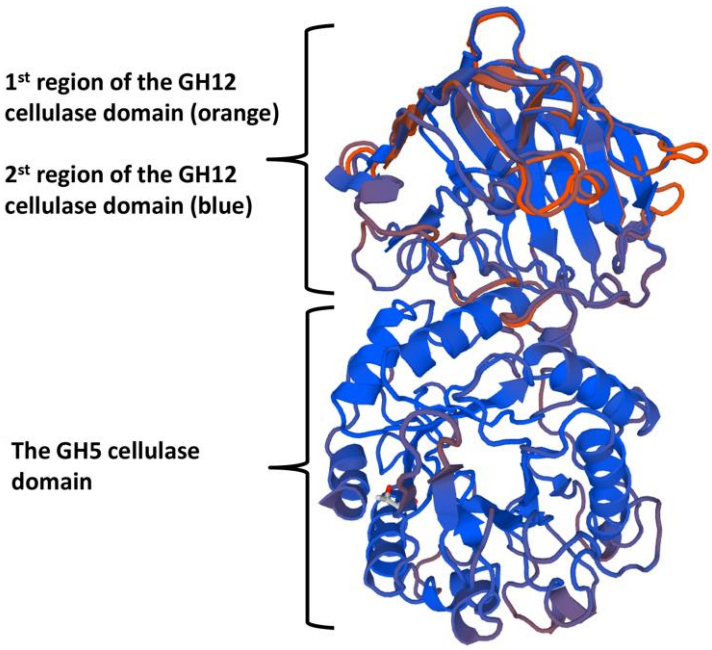


Figure 8. The 3D model of the cellulase F1 was predicted using the Swiss-Model server. The GH5-related cellulase domain F1_1, residues 44 to 412, was modeled using 3axx.1.A (87.3% amino acid identity) as the template. The GH12-related cellulase domains were modeled using 3vgi.1.A . The first GH12-affiliated domain (F1_2; residues 539 to 747, 35.8% amino acid identity) is depicted in orange and the second GH12-affiliated domain (F1_3; residues 830 to 1,089, 82.7% amino acid identity) is depicted in blue.

Despite a large number of cellulase ORFs discovered from the fosmid library (n=101), there were only 7 complete or nearly complete ORFs identified. This is likely due to the use of shorter Illumina HiSeq read lengths and the large number of fosmids in each pool, reducing the coverage per clone. Five of the ORFs identified from this search, from the pooled fosmid clones in plates 1 and 2 (P1+P2_contig 4468.4, designated as “S1”), plates 5 and 6 (P5+P6_contig 43387.3, designated as “S2”) plates

9 and 10 (P9+P10_contig 1829.4, designated as “S3”), plates 12 and 24 (P12+P24, _contig 94750.15, designated as “S4”) and from plates 15 and 16 (P15+P16_contig 25805.3, designated as “S5”) had an identical DNA sequence with varied length and 70.6% amino acid identity to the endoglucanase of *Pyrococcus abyssi* GE5 (NP_126623). Interestingly, the predicted cellulase ORF from plates 19 and 20 (P19+P20_contig 79977.13, designated as “S6”) had 85.3% amino acid identity to the endo-1,4-beta-glucanase found from a tomato plant (*Solanum lycopersicum*). Since it is highly unlikely that a relative of a tomato plant is present in the oil reservoir and the coverage of the cellulase-encoding ORF is low, this perhaps represents a case of lateral transfer or contamination. All of these predicted ORFs except S6 were successfully amplified from pooled fosmid DNA. Only one sequence from a predicted cellulase ORF in plates 3 and 4 (P3+P4_contig 223.1, designated as “S7”) corresponded to the same sequence identified from a fosmid clone (F6) expressing a cellulase activity, and in this case we used the fosmid DNA as template for the PCR.

The hybrid assembly of sequence reads derived from both the fosmid library and shotgun reads resulted in a larger number of complete cellulase-encoding ORFs (n=13), compared to only using fosmid library reads (n=7), and these were contained within larger contigs and that included mostly intact ORFs. Putative cellulases discovered from the hybrid assembly include all cellulase ORFs discovered above except for S6. In addition to the cellulase ORFs described above, three additional cellulases were identified from the hybrid assembly that are predicted to be beta-glucosidases. The first of the beta-glucosidases, designated as “S8” had 77.5% amino acid identity to the beta-

galactosidase of *Pyrococcus furiosus* (WP_011011185). The second beta-glucosidase, designated as "S9", had 73.1% amino acid identity to the beta-glucosidase of *Thermococcus kodakarensis* (WP_048053751); whereas the third beta-glucosidase, designated as "S10", had 96.1% amino acid identity to this enzyme from *T. kodakarensis*.

Since contigs derived from the hybrid assembly were much longer, all of the cellulase ORFs described above were mapped to these long contigs. This led to the discovery that two long contigs contain multiple cellulase ORFs. Contig_A (480,455bp, the largest contig obtained from any of the assemblies) contained two cellulase ORFs, S8 and S10, as well as a complete rRNA operon (Fig.9). The 16S rRNA gene of Contig_A has a top BLASTn hit to *Thermococcus celer* (98.5% nucleotide identity, M21529). Contig_B (279,442bp) had three predicted cellulase ORFs, specifically S4, F1 and S9 (Fig. 9).

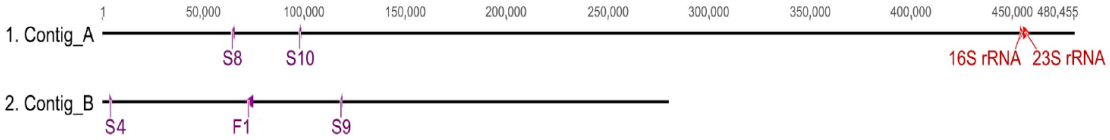


Figure 9. Sequence annotation for contig_A and contig_B, indicating the predicted cellulase ORFs in purple. The rRNA operon, including 16S rRNA and 23rRNA genes on contig_A were annotated in red, was predicted using RNAmmer v.1.2.

3.4.5 Thermal stability of subcloned cellulases

Cellulase genes identified from both sequence-based and function-based screening were subcloned into the inducible expression Expresso-Rhamnose subcloning system. The resulting subclones were streaked on CMC agar to assay for cellulase activity. Two of these subclones, S3C and S5C (the letter "C" denotes that these are subclones), showed apparent cellulytic activity (data not shown). In contrast, using the MUC substrate, activity was detected from five subclones (Fig. 10). The first four subclones were derived from sequence-based screening (S1C, S3C, S8C and S5C) and the last one was from function-based screening (F1C). The sequences of S1, S3 and S5 are identical but have different lengths. S5 is 63 bp and 117 bp shorter than S1 and S3, respectively. Despite its shorter length compared to these other two subclones, S5C was observed to express a higher cellulytic activity than S3C and S5C and did not have noticeable reduction in cellulase activity after being heated at 80°C (Fig. 10). In contrast, the cellulase activity of the subclones S1C and S3C was lower and decreased after heating at 80°C (Fig. 10). Interestingly, the subclone F1C showed a significantly higher cellulase activity after heating at 60°C and 80°C compared to its activity at 37°C (Fig. 10). The subclone S8C had the highest activity against the MUC substrate among all five subclones and also had the highest activity after heating at 60°C compared to the activity observed at 37°C or 80°C, indicating adaptation of this cellulase to a temperature close to 60°C.

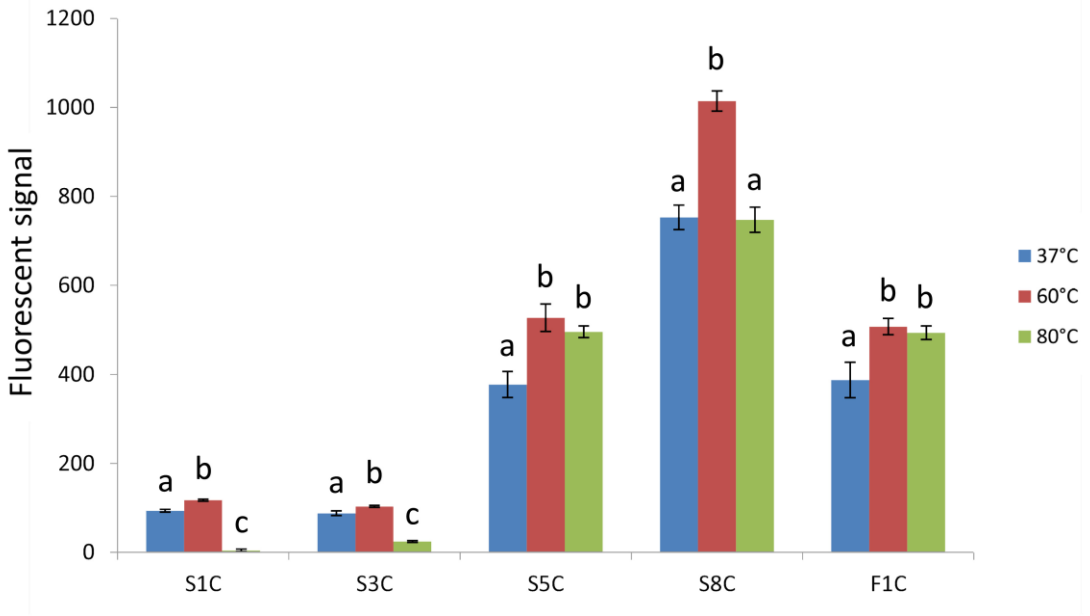


Figure 10. Quantitative MUC assay for supernatants of cell lysates from four subclones, in units of fluorescent signal intensity. The supernatants from each of the four subclones were incubated at 37°C (blue), 60°C (red) or 80°C (green) for 6 hours to test the thermal stability of each respective cellulase. Values for a subclone with different superscripts (a, b, ab) were significantly different ($P < 0.05$) by one way ANOVA followed by Turkey multiple comparison.

The thermal stability and activity of the cellulase enzymes were analyzed in crude cell extracts as well as isolated protein (Ni-NTA affinity chromatography). Activity per mg protein in extracts containing cellulases from S1C, S5C and F1C all showed significantly higher activity compared to the negative control. Activity of the S3C cellulase was lower compared to the other cellulases, but still higher than the observed background (Fig. 11), indicating heat stability as well as cellulase activity in all enzymes characterized. An activity assay was also performed using isolated protein, where the first elution fraction (eluted with 100 mM imidazol) both untreated as well

as heat incubated (65°C, 20 min) were used (Fig. 12). Isolated cellulase from S3C did not show any significant activity in the assay, suggesting poor yield in isolation in combination with a lower level of activity. Cellulases from S1C and S5C both showed activity as isolated proteins; however, a small decrease in activity was found after heat incubation of the isolated proteins (Fig. 12). For the cellulase from F1C the measured activity in the isolated protein sample was found to be remarkably higher compared to the other two candidates. In addition, the cellulase candidate from F1C is apparently very heat stable and the observed activity per mg protein increased notably after heat incubation of the isolated protein, indicating it is a very active as well as thermostable cellulase.

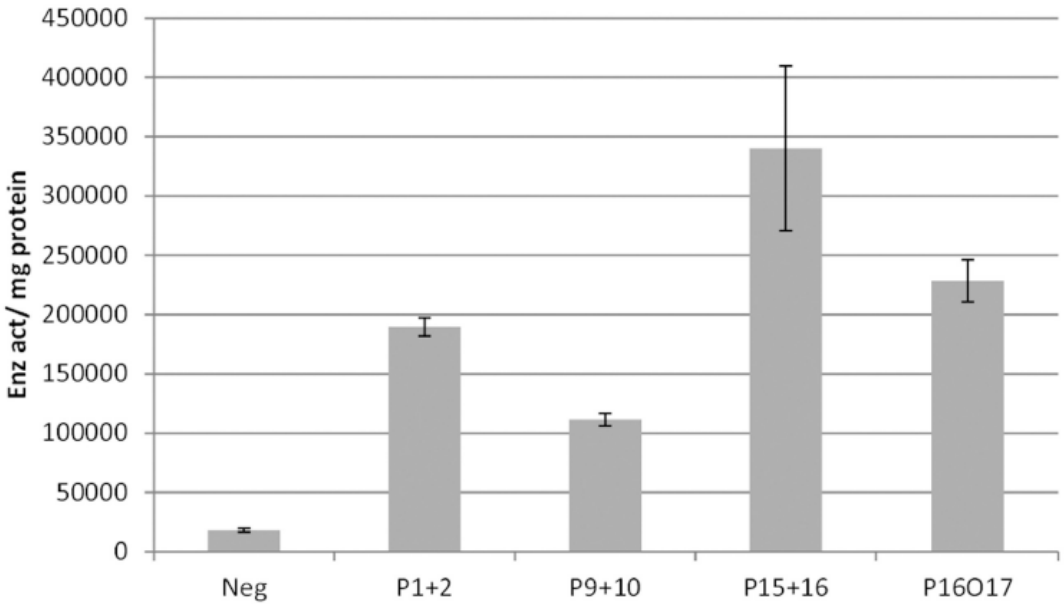


Figure 11. Activity assay using crude cell extracts, with activity/mg protein plotted for extracts originating from *E. coli* expressing the four cellulase variants, in addition to the *E. coli* negative control.

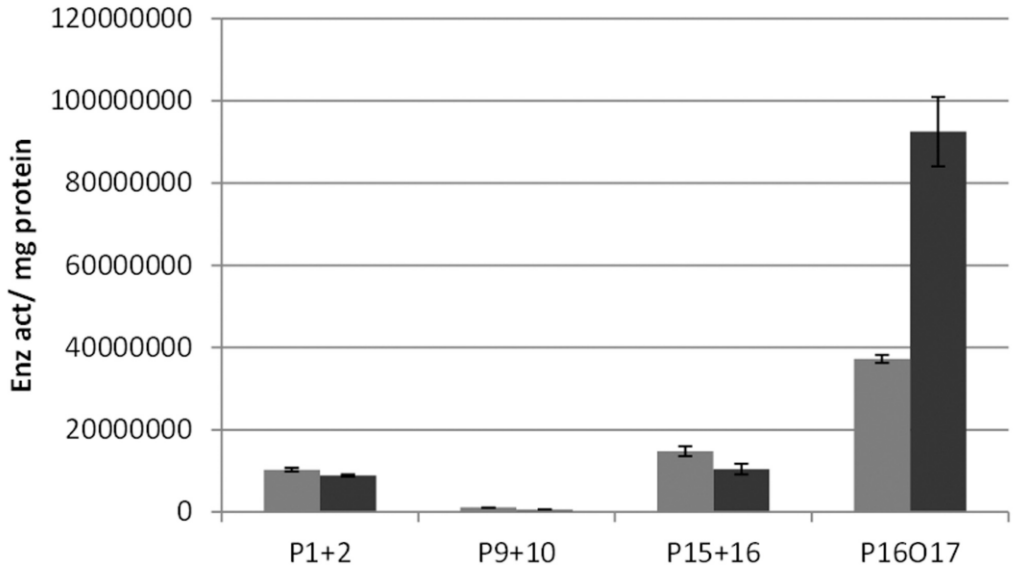


Figure 12. Activity assay of the Ni-NTA isolated protein, with activity/mg protein plotted for untreated (lighter bars) as well as heat treated (darker bars; 65^{o}C, 20 min) protein samples.

3.4.6 Cellulase phylogenetic analysis

All cellulase ORFs were collected together with a database of cellulase gene sequences and a phylogenetic tree was constructed using PHYML to shed light on the evolutionary relationships of these cellulases. In the tree, the cellulase from S6 was distantly related to other identified cellulases but clustered together with two known eukaryotic cellulases, presenting high similarity to that of *Nicotiana tabacum* and *Solanum tuberosum* that are affiliated with GH9 (Fig. 13). The cellulase gene sequence that was repeatedly identified from five different fosmid pooled plates (S1, S2, S3, S4, and S5) was affiliated with GH5 and formed a monophyletic group with an archaeal

cellulase from *Pyrococcus abyssi* (Fig. 13). Cellulases with an identical amino acid sequence from clones F5 and F6 as well as from S7 were affiliated with GH5, which included bacterial cellulases from *Thermosipho africanus* and *Fervidobacterium nodosum* (Fig. 13). The N-terminal domain from clone F1 (F1_1) and cellulase of the ORF F4_1 (clone F4) were identical and both affiliated with the GH5 class and were in a clade together with an archaeal cellulase identified from *Pyrococcus horikoshii* (Fig.13). In contrast, the C-terminal domains from clone F1 (F1_2 and F1_3) and two domains from the ORF F4_2 (F4_2_1 and F4_2_2) were all clustered with archaeal cellulases of the GH12 class. Two cellulase domains F1_2 and F4_2_1 were identical and affiliated with an archaeal cellulase identified from *Ignisphaera aggregans* DSM 17230. The domain F1_3 had an identical amino acid sequence with the domain F1_2_2 as well as cellulases from clones F2 and F3, and the closest relative of them is an archaeal cellulase from *Pyrococcus furiosus* (Fig. 13). In addition, three addition cellulases S8, S9 and S10 that are putative beta-glucosidases formed an independent clade with known archaeal cellulases from *Pyrococcus furiosus*, *Thermococcus sibiricus* MM 739 and *Thermococcus kodakarensis* KOD1 repectively that affiliate with GH1. The phylogenetic analysis supports the monophyly of these bacterial and archaeal cellulases, and indicates that the thermostable cellulases identified in this study from Archaea represent novel clades, whereas the bacterial-derived cellulases were closely related to previously identified cellulases.

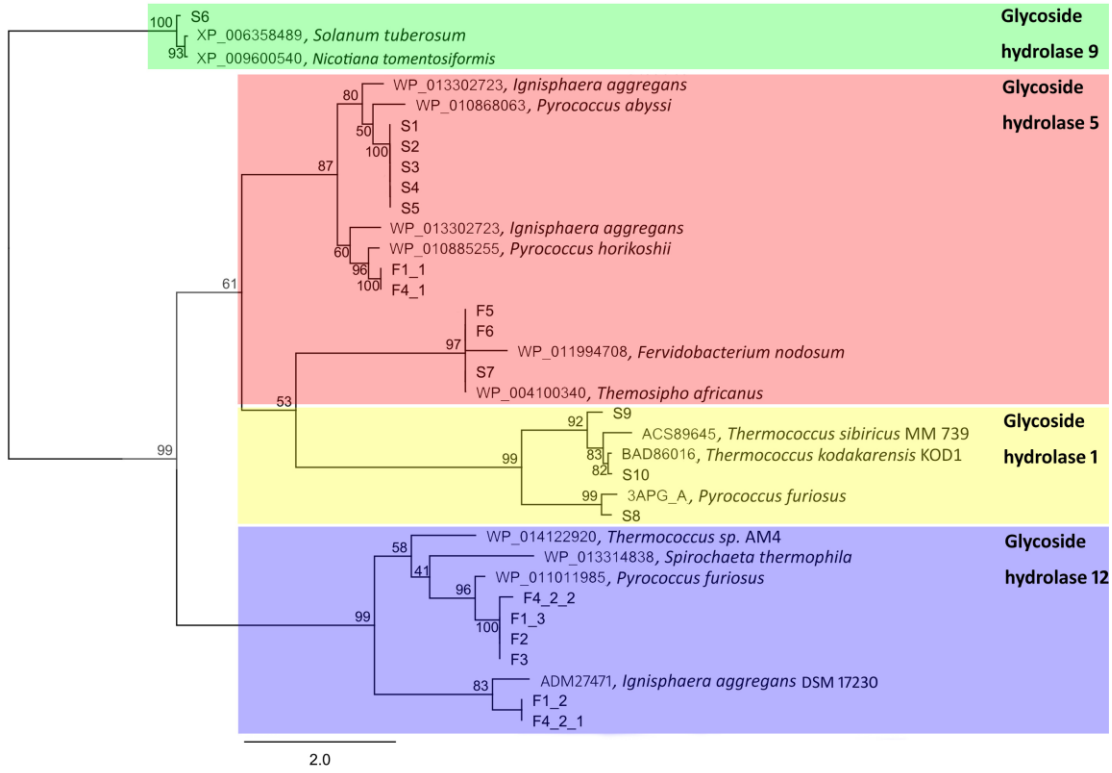


Figure 13. A maximum likelihood phylogenetic analysis using amino acid sequences of cellulases identified in this study (in bold) and previously described cellulases derived from members of the domains Eukaryota, Archaea and Bacteria. 1000 iterations were conducted for bootstrap support, and bootstrap values are indicated at each node. Cellulases affiliated with GH9 are highlighted in green, GH12-affiliated sequences are highlighted in blue, GH5-affiliated sequences are highlighted in red and GH1-affiliated sequences are highlighted in yellow.

3.5 Discussion

We observed that the oil reservoir sample was dominated by members of the domain Archaea, phylum *Euryarchaeota*, with sequences recovered either from

shotgun sequencing or from a fosmid metagenomic library indicating more than 60% of all significant hits to these archaeal taxa (Fig. 5A). For other phyla, only three were present in shotgun or library sequence databases at greater than 0.1% relative abundance, with a range of 6.7% to 11.2% *Proteobacteria*, 2.3% to 2.4% *Firmicutes* and 0.1% to 0.14% *Thermotogae*, respectively. At a genus level, *Thermococcus* and *Pyrococcus* were the most abundant genera with about 22% and 4% of the significant hits, respectively (data not shown). The results obtained from shotgun sequences and from the metagenomic library were highly comparable in terms of phylogenetic and functional composition (Fig. 5) and exhibited an overall low diversity as expected due to the high pressure (250 bars) and temperature (85^{o}C) in this environment. The alpha-diversity of the shotgun sequences database as determined based on Shannon's Diversity Index is 42.08 species, which as expected is relatively low compared to non-extreme environments. We did observe that taxa affiliated with the bacterial phyla *Proteobacteria* and *Bacteroidetes* had different abundances between these two sequence databases (Fig. 5A), which could reflect a bias in the amplification and/or cloning of genomic DNA from these bacteria. Given the observations that the Archaea are dominant in this environment, that most of the enzymes obtained in this study are derived from taxa affiliated with the *Euryarchaeota*, and the extreme nature of these habitats, we conclude that our sampling of these oil reservoir microbial assemblages has been inclusive of much of the extant phylogenetic and functional diversity. The large number of unassigned sequences from both shotgun and library sources indicates that even though this is an extreme habitat with limited phylogenetic breadth that there

is a considerable amount of previously unknown metagenomic diversity in the sampled environment.

Inferences of the functional capacity of these oil reservoir microorganisms gleaned from MG-RAST output indicated that carbohydrate-degrading enzymes are frequently encoded within the archaeal and bacterial genomes (Fig. 5B). However, crude oil consists primarily of hydrocarbons of various molecular weights, and one would predict that only small amounts of carbohydrates such as cellulose, starch and xylan, if any, exist in deep sub-surface oil reservoirs. While the concentrations of these carbohydrates were not determined from oil samples, the assumption is that these carbohydrates are present in limited amounts and are probably from remnant biomass. Alternative functions of polysaccharide hydrolases in organisms from oil reservoir samples may be in the metabolism of storage polysaccharides or extracellular polysaccharides (EPS) formed by many organisms, including hyperthermophilic archaea [171]. In addition, we observed that some of the CAZymes discovered from sequence- or function-based screening were redundant, indicating that the methods we used in this study had sufficiently exhausted much of the enzymatic diversity present in these samples and that there is a limited overall diversity of CAZymes in this hyperthermal habitat. Surprisingly, a cellulase gene S6 was discovered to have homology with a cellulase from *Solanum lycopersicum*, the garden tomato. The contig from which S6 is derived is very short with very low coverage for its assembly, and was only assembled from reads derived from the fosmid library. It may indicate a potential gene transfer event or

a sample contamination during DNA extraction. The lack of additional sequences linked to this predicted cellulase gene precludes a more precise conclusion.

Five subclones were generated that were observed to have significant cellulase activity in the quantitative MUC assay. Subclones F1C, S5C and S8C showed good thermal stability in the MUC assay at both 60°C and 80°C, and were therefore the best prospects for thermostable cellulases for future applications. Interestingly, these three cellulase genes were all present on large contigs identified from a hybrid assembly of fosmid library and shotgun reads. The S8 ORF was located on contig_A, which was 480,455 bp and the longest among all assembled contigs (Fig. 9). Fortunately, this contig included an intact rRNA operon, with the 16S rRNA gene from contig_A having 98.5% identity with the 16S rRNA gene from *Thermococcus celer*. Among the predicted ORFs from contig_A, 85% of these ORFs (446 out of 523) had a top BLASTp hit to a gene product from the genus *Thermococcus.* This is very robust evidence that contig_A and the cellulose-encoding genes present on this contig were derived from a *Thermococcus* species. Contig_B (279,442 bp) contained sequences of the S4 and F1 ORFs (

(Fig. 9). The S5 cellulase is identical to but 42aa shorter than the S4 cellulase, and it is possibly just an incomplete ORF (belongs to part of S4) assembled from a short contig. Interestingly, shorter one S5 present a significant cellulase activity against MUC in the subclone but the complete ORF S4 didn't. No 16S rRNA gene was identified on contig_B, but 88% of the top BLASTp hits (274 out of 312) from predicted ORFs on contig_B also had significant similarities to the genus *Thermococcus*. Furthermore, an

analysis of codon usage from clone F1 sequences also supports an origin from *Thermococcus* spp. rather than from *Pyrococcus* spp. (data not shown). Taken together, these results suggest that both of the highly active and thermostable cellulases identified from this study were derived from thermophilic Archaea within the genus *Thermococcus* that was predicted to be the dominant genus in the environment by MG-RAST, and that expression of these archaeal cellulases was possible (at least in some cases) from native archaeal promoters expressed in an *E. coli* heterologous host.

A phylogenetic analysis of the cellulases obtained from this study supports the affiliation of these cellulases with cellulases previously obtained from all three domains of life. Our data indicates that the majority of the cloned cellulases, and other CAZymes, are affiliated with Archaea and Bacteria, particularly with taxa affiliated with the phylum *Euryarchaeota*. Interestingly, all of the bacteria-derived cellulases are affiliated with the GH5 category, with many archaeal-derived cellulases also in this clade (highlighted in red shadow, Fig. 13). GH5 is one of the largest of the CAZy GH families and contains cellulases widely derived from bacteria, archaea and even eukaryotes. In contrast, other archaeal cellulases were classified within the category of GH1, GH5 and GH12. Archaeal cellulases affiliated with GH12 were all identified from function-based screening (highlighted in blue shadow, Fig. 13) and were subsequently discovered independently from the hybrid assembly of fosmid library and shotgun sequencing reads. Conversely, the three archaeal cellulases affiliated with GH1 are putative beta-glycosidases, which were only identified from sequence-based homology searches (highlighted in blue shadow, Fig. 13). Therefore, there were examples of bias in the

discovery of cellulases depending on whether a function-based or sequence-based screening method was used. Presumably the cellulases not identified via sequence analysis initially were due to a low sequencing depth of the fosmid library reads for particular clones in the pooled library format. The inability to identify some cellulases from function-based screening was anticipated due to lack of expression of the enzymes from their native promoters in an *E. coli* host, an inability to be translated due to differing codon usage, or an inability to be secreted and/or active under the conditions used for functional screening. The observation that we obtained distinct cellulase types using function- and sequence-based screening methods highlights the potential biases associated with metagenome mining methods and supports the use of multiple approaches to identify novel natural products from environmental metagenomes.

The cellulase F1 was the longest ORF among all of the identified cellulases and its subclone had the highest cellulase activity using MUC as a substrate. Interestingly, F1 is predicted to have three distinct cellulase domains that affiliate with glycoside hydrolases within the GH5 (one domain, F1_1) and GH12 (two domains, F1_2 and F1_3) classes. Homologies of these three domains against known cellulases are low, especially F1_2 with only 59% amino acid identity to its top BLASTp hit. From the phylogenetic tree, it is also easy to observe that the three domains have distinct lineages (Fig. 13). Although both of F1_2 and F1_3 belong to a same class GH12, they are affiliated to two different independent clades (Fig. 13). Furthermore, it is likely that the archaeal cellulase F1 was evolved from fusion of two cellulases with distinct families, potentially resulting in its strong cellulase activity and thermal stability. However, no

known cellulase containing these three domains has been identified previously to our knowledge. Also, CBM2 modules mainly exist in bacterial enzymes, and only six of the modules (out of 1953 described to date) were reported to be of archaeal origin in the CAZy database [149]. These data suggest that the fusion of these two cellulytic domains generated a novel protein structure with enhanced thermal stability. In addition, clone F4 was found to contain two cellulases with a total of three domains including a GH5 and GH12 domains. The three F4 cellulase domains were almost identical to the corresponding domains from the F1 cellulase, with the exception of partial amino acid sequence of the CBM domain and an inter-domain amino acid sequence (residues of F1 from 449 to 519) that was lacking in the F4 cellulase. Also ORFs of F4_1 and F4_2 overlapped each other in 72aa and the orientation of F4_2 was reversed compared to the GH12 domain of F1 (Fig. 7). The close relationship of the F1 and F4 cellulases, that also have a distinct organization, suggests a common ancestry of these multi-domain cellulases that have since diverged in their structure. Future studies will investigate the three-dimensional protein conformation for the F1 and F4 cellulases and structure-function relationships important for cellulase activity against multiple substrates and stability under environmental extremes.

In conclusion, this study revealed that an oil reservoir microbial assemblage harbored novel metagenomic diversity and could be mined for thermostable cellulases and other CAZymes using both function- and sequence-based methods. The results of this study have provided novel thermostable archaeal cellulases that are stable up to at least 80^{o}C. These thermostable enzymes could be used in the degradation of

lignocellulosic biomass for biofuels applications and will be subject to more detailed studies in the near future to evaluate their potential applications.

3.6 Further work

Right now the last work that has not been finished in this chapter is Purification of active cellulases from subclones using SDS-PAGE. Purified cellulases can then be processed for structure analysis using LC/MC and a more accurate quantitative assay, which will be an incontrovertible evidence for their existence.

For production of cellulase enzymes the four *E. coli* 10G strains harboring the subcloned genes of interest, along with a negative control *E. coli* 10G, will be cultivated in 1000 mL batches. A 5 mL LB culture containing 0.5% glucose and 30 µg/mL kanamycin (except for the negative control) will be used as inoculum for 1000 ml LB-kanamycin media, and the cultures will be incubated at 37°C until the OD600 was 0.3-0.5 (3 hours). Cultures will then induced using 0.2% rhamnose (final concentration) and cultivated for another 9.5 hours. Crude cell extracts will be prepared by sonication in 5-10 mL buffer (50 mM KPO_4, pH5.5) for 7 minutes (50% duty cycle and output control 4) followed by centrifugation at 20,000 x *g* for 30 minutes at 4°C. Isolated extracts will be used in heat stability analysis (extract incubation at 70°C for 3 hours), activity assay (as described above) and used for isolation of the enzymes by Ni-NTA affinity chromatography. For enzyme purification, 450 µl of sterile filtered (0.2 µm) cell extracts will be incubated with 1 ml Ni –NTA agarose (equilibrated with native

binding buffer; 50 mM NaH_2PO_4 buffer (pH8.0) with 0.5 M NaCl, 10 mM imidazol and 1 mM DTT) for 60 minutes at RT in a Rotamixer. Agarose beads will be washed in native wash buffer (50 mM NaH_2PO_4 buffer (pH8.0) with 0.5 M NaCl, 20 mM imidazol and 1 mM DTT), resuspended in 2 ml of the same buffer and applied in a plastic column. The beads will be washed three times with 5 ml wash buffer and the bound proteins thereafter eluted using elution buffer (50 mM NaH_2PO_4 buffer (pH8.0) with 0.5 M NaCl, and 1 mM DTT) with increasing concentrations of imidazol (100, 150, 200, 250 and 500 mM) in 1 ml fractions. Isolated proteins will be subjected to heat incubation (65°C for 20 minutes), and used in a cellulase activity assay.

3.7 Appendix

3.7.1 All cellulase ORFs

>F1

MYRQKALAVFVLFVVLAGVAGSIPAGYAATNTSTYTTPTGIYYEVRGDTIYMI
NVATGEETPIHLFGVNWFGFETPNYVVHGLWSRNWEDMLLQIKSLGFNAIRLP
FCTQSVKPGTMPTGIDYAKNPDLQGLDSVQIMEKIIKKAGDLGIFVLLDYHRI
GCNFIEPLWYTDSFSEQDYINTWVEVAQRFGKYWNVIGADLKNEPHSSSPAPA
AYTDGSGATWGMGNNATDWNLAAERIGKAILEVAPHWLIFVEGTQFTTPEID
GSYKWGHNAWWGGNLMGVRKYPVNLPRNKLVYSPHVYGPDVYDQPYFDP
AEGFPDNLPDIWYHHFGYVKLDLGYPVVIGEFGGKYGHGGDPRDVTWQNKII
DWMIQNKFCDFFYWSWNPNSGDTGGILQDDWTTIWEDKYNNLKRLMDSCS
GNATAPSVPTTTTTTSTPPTTTTTTTSTPTTTTQTPTTTTPTTTTTTTTTPSNNVP

FETVNVLPTSSQYEGTSVEVVCDGTQCASSVWGAPNLWGVVKIGNATMDPN

VWGWEDVYKTAPQDIGTGSTKMEIRNGVLKVTNLWNINMHPKYNTMAYPE

VIYGAKPWGNQPINAPNFVLPIKVSQLPRILVDTKYTLEKSFPGNNFAFEAWLF

KDANNMRAPGQGDYEIMVQLYIEGGYPAGYDKGPVLTVDVPIIVDGRLLNQT

FELYDVIADAGWRFFTFKPTKNYNGSEVVFDYTKFIEIVDNYLGGGSLTNHYL

MSLEFGTEIYTNGCTSFPCTVDVRWTLDKYRFILAPGTMATEEAMRVLVGEV

QPPASTTTSQTTTSTTTPTPTTTTTTQTSTTTTTTSPPTTTAPAQDVIKLRYPDDG

QWPEAPIDRDGDGNPEFYIEINPWNILSAEGYAEMTYNLSSGVLHYVQALDSI

TLKNGGAWVHGYPEIFYGNKPWNNNSATDGEVPLPGKVSNLSNFYLTVSYKL

LPKNGLPINLAIESWLTREPWRNSGINSDEQELMIWLYYDGLQPAGSKVKEIV

VPIVVNGTPVNATFEVWKANIGWEYVAFRIKTPIKEGTVTIPYGAFISAAANVT

SLANYPELYLEDVEVGTEYGTPSTTSAHLEWWFYNVSLEYRPGEPLLSQPPAE

GSAPSEGGQTPSEGATTGTLDVKLVNSWGTGAQYEVSVNLDTSSTWKLLIKIK

DGKISDIWGASIVGTQGDYVVVQPSSPTASATVGFVTSGNAPLVEEAVLLSGD

KVLATWTAPTASASDLNVTIKIDSEWDSGFVVKIYVTNNGNAPVSSWQIKLR

MTSLISSIWGGTYTASGDVVTIVPTGNNTVINPGDTVEIGFVASKQGAYVYPEL

IGVEIL

>F2

MSAEGYAEMTYNLSSGVLHYVQALDSITLKNGGAWVHGYPEIFYGNKPWNN

NSATDGEVPLPGKVSNLSNFYLTVSYKLLPKNGLPINLAIESWLTREPWRNSGI

NSDEQELMIWLYYDGLQPAGSKVKEIVVPIVVNGTPVNATFEVWKANIGWEY

VAFRIKTPIKEGTVTIPYGAFISAAANVTSLANYPELYLEDVEVGTEYGTPSTTS

AHLEWWFYNVSLEYRPGEPLLSQPPAEGSAPSEGGQTPSEGATTGTLDVKLVN

SWGTGAQYEVSVNLDTSSTWKLLIKIKDGKISDIWGASIVGTQGDYVVVQPS

SPTASATVGFVTSGNAPLVEEAVLLSGDKVLATWTAPTASASDLNVTIKIDSEW

DSGFVVKIYVTNNGNAPVSSWQIKLRMTSLISSIWGGTYTASGDVVTIVPTGN

NTVINPGDTVEIGFVASKQGAYVYPELIGVEIL

>F3

MSAEGYAEMTYNLSSGVLHYVQALDSITLKNGGAWVHGYPEIFYGNKPWNN

NSATDGEVPLPGKVSNLSNFYLTVSYKLLPKNGLPINLAIESWLTREPWRNSGI

NSDEQELMIWLYYDGLQPAGSKVKEIVVPIVVNGTPVNATFEVWKANIGWEY

VAFRIKTPIKEGTVTIPYGAFISAAANVTSLANYPELYLEDVEVGTEYGTPSTTS

AHLEWWFYNVSLEYRPGEPLLSQPPAEGSAPSEGGQTPSEGATTGTLDVKLVN

SWGTGAQYEVSVNLDTSSTWKLLIKIKDGKISDIWGASIVGTQGDYVVVQPS

SPTASATVGFVTSGNAPLVEEAVLLSGDKVLATWTAPTASASDLNVTIKIDSEW

DSGFVVKIYVTNNGNAPVSSWQIKLRMTSLISSIWGGTYTASGDVVTIVPTGN

NTVINPGDTVEIGFVASKQGAYVYPELIGVEIL

>F4_1

MYRQKALAVFVLFVVLAGVAGSIPAGYAATNTSTYTTPTGIYYEVRGDTIYMI

NVATGEETPIHLFGVNWFGFETPNYVVHGLWSRNWEDMLLQIKSLGFNAIRLP

FCTQSVKPGTMPTGIDYAKNPDLQGLDSVQIMEKIIKKAGDLGIFVLLDYHRI

GCNFIEPLWYTDSFSEQDYINTWVEVAQRFGKYWNVIGADLKNEPHSSSPAPA

AYTDGSGATWGMGNNATDWNLAAERIGKAILEVAPHWLIFVEGTQFTTPEID

GSYKWGHNAWWGGNLMGVRKYPVNLPRNKLVYSPHVYGPDVYDQPYFDP

AEGFPDNLPDIWYHHFGYVKLDLGYPVVIGEFGGKYGHGGDPRDVTWQNKII

DWMIQNKFCDFFYWSWNPNSGDTGGILQDDWTTIWEDKYNNLKRLMDSCS

GNATAPSVPTTTTTTSTPPTTTTTTTSTPTTTT

>F4_2

MDPNVWGWEDVYKTAPQDIGTGSTKMEIRNGVLKVTNLWNINMHPKYNTM

AYPEVIYGAKPWGNQPINAPNFVLPIKVSQLPRILVDTKYTLEKSFPGNNFAFE

AWLFKDANNMRAPGQGDYEIMVQLYIEGGYPAGYDKGPVLTVDVPIIVDGRL

LNQTFELYDVIADAGWRFFTFKPTKNYNGSEVVFDYTKFIEIVDNYLGGGSLT

NHYLMSLEFGTEIYTNGCTSFPCTVDVRWTLDKYRFILAPGTMATEEAMRVL

VGEVQPPASTTTSQTTTSTTTPTPTTTTTTQTSTTTTTTSPPTTTAPAQDVIKLRY

PDDGQWPEAPIDRDGDGNPEFYIEINPWNILSAEGYAEMTYNLSSGVLHYVQ

ALDSITLKNGGAWVHGYPEIFYGNKPWNNNSATDGEVPLPGKVSNLSNFYLT

VSYKLLPKNGLPINLAIESWLTREPWRNSGINSDEQELMIWLYYDGLQPAGSK

VKEIVVPIVVNGTPVNATFEVWKANIGWEYVAFRIKTPIKEGTVTIPYGAFISA

AANVTSLANYPELYLEDVEVGTEYGTPSTTSAHLEWWFYNVSLEYRPGEPLL

SQPPAEGSAPSEGGQTPSEGATTGTLDVKLVNSWGTGAQYEVSVNLDTSSTW

KLLIKIKDGKISDIWGASIVGTQGDYVVVVGVDVVVVVVVGGVLVVVVVVG

TDGAVAFPEQLSMSLFRLLYLSSHIVVQSSCRIPPVSPLLGFQLQ

>F5

MPTSYNNTSTKTGNGVNCCTTFNRKNTWKRPKDTKKCCSSYKINRENVCKE

GGTMRNFFKVFTLVLVVISVMLFGENKKLTAFDYNKMIGIGINMGNALEAPFE

GAWGVVIKDEYFEIIKEKGFDSVRIPIRWSAHILDKPPYTIEKDFLERVKHVVD

KALENDLIVIINCHHFEELYENPEKYGEVLLEIWKQVSSFFKDYSDKLYFEIYN

EPAKNLTPEKWNDLYPKVLKEIRKTNPSRIVIVDVPHWGNYNYINQLKLVNDP

YLIVSFHYYEPFNFTHQGAEWINPRLPVGVKWSAKSYEIEQIKSHFEYVDSFS

KKYNVPIFLGEFGAYSKADMDSRIKWTKAVSQIAREFGFSICYWEFCSGFGLY

NKITNTWNEGLLNAVFGK

>F6

MPTSYNNTSTKTGNGVNCCTTFNRKNTWKRPKDTKKCCSSYKINRENVCKE

GGTMRNFFKVFTLVLVVISVMLFGENKKLTAFDYNKMIGIGINMGNALEAPFE

GAWGVVIKDEYFEIIKEKGFDSVRIPIRWSAHILDKPPYTIEKDFLERVKHVVD

KALENDLIVIINCHHFEELYENPEKYGEVLLEIWKQVSSFFKDYSDKLYFEIYN

EPAKNLTPEKWNDLYPKVLKEIRKTNPSRIVIVDVPHWGNYNYINQLKLVNDP

YLIVSFHYYEPFNFTHQGAEWINPRLPVGVKWSAKSYEIEQIKSHFEYVDSFS

KKYNVPIFLGEFGAYSKADMDSRIKWTKAVSQIAREFGFSICYWEFCSGFGLY

NKITNTWNEGLLNAVFGK

>S1

MATTAWGAGDRPPSEFWEVRVMRALKKFFPIFIGLLFLLSPVSAVEYRAENGK

IYADGNEIHLYGVSWFGFELKDHVVFGLTQRNWKEILQDVKRLGFNAVRLPF

CSESIKPGTKPNLNKINYELNPDLKNLTSLEIMEKIIAYANELGIYVLLDYHRIG

CAYIEPLWYTDEYPEEQYIADWVFLAERFGRYPNVIGADIKNEPHDEASWGT

GDETDFRLFAERVGKAILEKAPHWLIFVEGVQYTHLSEIDSKNPYPCFWGENL

MGVREYPVRLPEGKVVYSPHVYGPSVYEMPYFSDPSFPDNLLEIWELHFGYL

KDLNYTLVIGEWGGNYEGKDKVWQDKFSEWLVEKGIHDFFYWCLNPESGDT

KGVFLDDWKTVNWEKMRVIYRVIKASNPEFEEPLYIILKANTTSRVLDKGERI

KLYWYTSGEVVDSNFADLSEGEIEIELNQSTTFYIAARKGGEVKNESIRFSVIEP

NTPSGEETETPTVPETTPKSGEHSSTSWLFLALLLLAAVAVLAKLRR

>S2

MEKIIAYANELGIYVLLDYHRIGCAYIEPLWYTDEYPEEQYIADWVFLAERFGR

YPNVIGADIKNEPHDEASWGTGDETDFRLFAERVGKAILEKAPHWLIFVEGVQ

YTHLSEIDSKNPYPCFWGENLMGVREYPVRLPEGKVVYSPHVYGPSVYEMPY

FSDPSFPDNLLEIWELHFGYLKDLNYTLVIGEWGGNYEGKDKVWQDKFSEW

LVEKGIHDFFYWCLNPESGDTKGVFLDDWKTVNWEKMRVIYRVIKASNPEFE

EPLYIILKANTTSRVLDKGERIKLYWYTSGEVVDSNFADLSEGEIEIELNQSTTF

YIAARKGGEVKNESIRFSVIEPNTPSGEETETPTVPETTPKSGEHSSTSWLFLAL

LLLAAVAVLAKLRR

>S3

MLPGRAGMTSRLPPQPLAMATTAWGAGDRPPSEFWEVRVMRALKKFFPIFIG

LLFLLSPVSAVEYRAENGKIYADGNEIHLYGVSWFGFELKDHVVFGLTQRNW

KEILQDVKRLGFNAVRLPFCSESIKPGTKPNLNKINYELNPDLKNLTSLEIMEKI

IAYANELGIYVLLDYHRIGCAYIEPLWYTDEYPEEQYIADWVFLAERFGRYPN

VIGADIKNEPHDEASWGTGDETDFRLFAERVGKAILEKAPHWLIFVEGVQYTH

LSEIDSKNPYPCFWGENLMGVREYPVRLPEGKVVYSPHVYGPSVYEMPYFSD

PSFPDNLLEIWELHFGYLKDLNYTLVIGEWGGNYEGKDKVWQDKFSEWLVE

KGIHDFFYWCLNPESGDTKGVFLDDWKTVNWEKMRVIYRAIKASNPEFEEPL

YIILKANTTSRVLDKGERIKLYWYTSGEVVDSNFADLSEGEIEIELNQSTTFYIA

ARKGGEVKNESIRFSVIEPNTPSGEETETPTVPETTPKSGEHSSTSWLFLALLLL

AAVAVLAKLRR

>S4

MLPGRAGMTSRLPPQPLAMATTAWGAGDRPPSEFWEVRVMRALKKFFPIFIG

LLFLLSPVSAVEYRAENGKIYADGNEIHLYGVSWFGFELKDHVVFGLTQRNW

KEILQDVKRLGFNAVRLPFCSESIKPGTKPNLNKINYELNPDLKNLTSLEIMEKI

IAYANELGIYVLLDYHRIGCAYIEPLWYTDEYPEEQYIADWVFLAERFGRYPN

VIGADIKNEPHDEASWGTGDETDFRLFAERVGKAILEKAPHWLIFVEGVQYTH

LSEIDSKNPYPCFWGENLMGVREYPVRLPEGKVVYSPHVYGPSVYEMPYFSD

PSFPDNLLEIWELHFGYLKDLNYTLVIGEWGGNYEGKDKVWQDKFSEWLVE

KGIHDFFYWCLNPESGDTKGVFLDDWKTVNWEKMRVIYRAIKASNPEFEEPL

YIILKANTTSRVLDKGERIKLYWYTSGEVVDSNFADLSEGEIEIELNQSTTFYIA

ARKGGEVKNESIRFSVIEPNTPSGEETETPTVPETTPKSGEHSSTSWLFLALLLL

AAVAVLAKLRR

>S5

MRALKKFFPIFIGLLFLLSPVSAVEYRAENGKIYADGNEIHLYGVSWFGFELKD

HVVFGLTQRNWKEILQDVKRLGFNAVRLPFCSESIKPGTKPNLNKINYELNPD

LKNLTSLEIMEKIIAYANELGIYVLLDYHRIGCAYIEPLWYTDEYPEEQYIADW

VFLAERFGRYPNVIGADIKNEPHDEASWGTGDETDFRLFAERVGKAILEKAPH

WLIFVEGVQYTHLSEIDSKNPYPCFWGENLMGVREYPVRLPEGKVVYSPHVY

GPSVYEMPYFSDPSFPDNLLEIWELHFGYLKDLNYTLVIGEWGGNYEGKDKV

WQDKFSEWLVEKGIHDFFYWCLNPESGDTKGVFLDDWKTVNWEKMRVIYR

VIKASNPEFEEPLYIILKANTTSRVLDKGERIKLYWYTSGEVVDSNFADLSEGEI

EIELNQSTTFYIAARKGGEVKNESIRFSVIEPNTPSREETETPTVPETTPKSGEHS

STSWLFLALLLLAAVAVLAKLRR

>S6

MFFNAQRSGKLPKHNNVSWRGNSCMRDGKSDDATIFKDLVGGYYDAGDAIK

FNFPQSFALTMLSWSVIEYSAKYEAAGELNHVKDIIKWGTDYLLKTFNSSADT

IDRVVAQVGSGDTADGSTTPNDHYCWMRPEDIDYDRPVFECHSCSDLAAEMA

AALASASIVFKDNKAYSQKLVHGARTLFKFSRDQRGRYSAGGTEAAIFYNSTN

YYDEFVWGATWLYYATGNSSYLQLATTPGIAKHAGAFWGGKYYGVMSWDS

KLPGAQVLLSRLRLFLSPGYPYEEILSTFHNQTSIVMCSYLPLFTSFNFTKGGLI

QLNYGAPQPLQYVANAAFLAALFSDYLAAADAPGWYCGPNFYSVDVLRNFA

ESQIDYILGKNPRKMSYVVGFGNHYPKHVHHRGASIPKNKVRYNCKGGWK

WRDSKKPNPNILVGAMVAGPDVHDGFHDVRTNYNYTEPTLAGNAGLV

>S7

MPTSYNNTSTKTGNGVNCCTTFNRKNTWKRPKDTKKCCSSYKINRENVCKE

GGTMRNFFKVFTLVLVVISVMLFGENKKLTAFDYNKMIGIGINMGNALEAPFE

GAWGVVIKDEYFEIIKEKGFDSVRIPIRWSAHILDKPPYTIEKDFLERVKHVVD

KALENDLIVIINCHHFEELYENPEKYGEVLLEIWKQVSSFFKDYSDKLYFEIYN

EPAKNLTPEKWNDLYPKVLKEIRKTNPSRIVIVDVPHWGNYNYINQLKLVNDP

YLIVSFHYYEPFNFTHQGAEWINPRLPVGVKWSAKSYEIEQIKSHFEYVDSFS

KKYNVPIFLGEFGAYSKADMDSRIKWTKAVSQIAREFGFSICYWEFCSGFGLY

NKITNTWNEGLLNAVFGK

>S8

MKFPSNFLFGYSWSGFQFEMGLPGSEVESDWWAWVHDKENIFSGLVSGDLPE

NGPAYWHLYKQDHDIAESLGMDAIRGGIEWARIFPKPTFDVKVDVERDENGNI

ISIDVPESAIEELEKLANMDALNHYREIYSDWKERGKTFILNLYHWPLPLWLH

DPIAVRKLGPDRAPSGWLDERSVVEFTKFAAFIAYHLDDLVDMWSTMNEPNV

VYEQGYTRPQSGFPPGYLSHEAAEKAKLNLMQAHARAYDAIKEHSDKPVGVI

YAYKWIDAEDEAAEESVLELRRRDYDFVDGLYSGKSLTAGEREDFKGRVDW

VGVNYYSRLLFGKAGDSVRLLEGYGFVSPRGGYAKSGRPASDFGWEIYPEGL

EKLLVELSGRYELPLFITENGMADAVDRYRPYYLVSHLAAIRRAMEKGADVR

GYLHWSLTDNYEWAQGFRMRFGLVMVDFETKKRYLRPSALVFREIVTRKEIP

EELEHLADVNAITAR

>S9

MEIKFPDKFLFGTSTAAHQVEGDNKWNDWWYYEQIGKLPYKSGKACNHWE

LYREDIELMAELGYNAYRFSIEWSRLFPEEGKFNEDAFNRYREIIELLLEKGVT

PNVTLHHFTSPRWFMEKGGFLKEENLKYWEEYVDKAAELLKGVKLVATFNE

PMEMIIEGYLTGNWPPFLRNAEKAFTAEANILKAHRIAYEVLSGEFKVGVVKS

SPLIRPISPEFREVAGEVDNLQNWYFLNAIFSGELVTPFGTVRTGESDADFIGVN

YYTLHLIGDVSPVEGLYRYEFGGYGRTQMGWKIYPEGIYEVLKRASGYGRPL

YVTENGIATLNDSERVDFIARHLHQVWRAIEDGVDVRGYFYWSLMDNYEWD

KGFEPRFGLIEMDFETFERRPRKSAYFYGEIAREKKVGLAPPRKIRKVQSSASP

SSSRRSSNLGSFTFSMASFASSSSISPSL

>S10

MFPEKFLFGTSTAAHQVEGDNKWNDWWYYEQIGKLPYKSGKACNHWELYR

EDIELMAELGYNAYRFSIEWSRLFPEEGKFNEDAFNRYREIIELLLEKGVTPNV

TLHHFTSPRWFMEKGGFLKEENLKYWEEYVDKAAELLKGVKLVATFNEPLVY

VTMGYLTAYWPPFIKSPFKSFRVAANLLRAHAIAYELLHGKFKVGIVKHIRVM

LPERKGDEKAAQKADNLFNWYFLDAIWSGKYRGAFKTYSVPESDADFVGVN

YYTASTVRRSLNPLKMFFEARDAEIGERRTQMGWSVYPEGIYLALKRASEYG

RPLYVTENGIATLDDEWRKEFIIQHLRQVLRAIEDGIDVRGYFYWSLMDNYE

WREGFEPRFGLIEVDFETFERRPRGSAHLYGEIARKRKLPGEEA

3.7.2 All Amylase ORFs

>NODE_1_length_435385_cov_101.67_ID_1_61

MMSLVFHGNLQYAEIPKSEIPKVIEKAYLPVIGRLLKEEIPFGLNITGYTLEILPR

EVIELVREGIASGLIEIIGTSYTHAILPLLPLDRVEAQVRKDRELKEELLEVSPKG

FWLPELAYDPIIPAILKDNGYGYVFADGEAMLLSDHLNSAVKPIKPLYPHLIKA

QRGEGNVFLNYLLGLRELKKAVEKAFPGKVTLEAVKNIDAVPVWVAINTAVM

LGIGRFPLMSPKKAANWLRGKDDVLLYGTDIEFIGYRDLAGRRMTIDGLLEV

VKALNVELSLPSELKHSGRKLYLRTSSWAPDKSLRIWTEDEGNARLNMLTPFV

GEPLAFLAENSDARGWEPLPERRLDAFRAIYNDWRGSK*

>NODE_1_length_435385_cov_101.67_ID_1_416

MPGIPIEVRTHTALHVVKGAVVKVLGEEAKWTASTYVKGNRGVLIVKFSRKP

TAEEVAEIERLANEKVEEDVPIEVYKLPREEAEKRFGEDMYDLFPIPPDVRTLR

VVVIEGWNVNACKEEHAGTTGEVGEIRIRKVRFRPNKELLEVSFEVL*

>NODE_13_length_76116_cov_6.14248_ID_25_69

MQVNIEQYLSQALEMARRGEGRTRPNPAVGAVIVKDGVVIGRGYHPQAGQPH

AEIFALREAGDRARGADMYVTLEPCSHHGRTGPCTDAIIKAGLARVFVGTPDP

NPQVAGAGIRKLRAAGIDVRCGILENECRRIIAPFAKHILEKLPFVILKSAMTLD

GKTATTSGHSQWVSNAASRLEVHRLRDRVDGIMVGIGTVLRDDPQLTTRLPE

GGRDPERIVVDSHLRIPVDAAILHLDSTAATLIATTASAEPDKIDAVRGTGAQV

LILPEKGGRVDLHALMTVLGERGLQSILLEGGAELNGALWRAGLVDRVMMF

VAPKIVGGEGRGVFNGPGAATMTEAAILRDVRVRRFGEDTMIEGEVEKCSPA*

>NODE_26_length_55303_cov_102.387_ID_51_47

MTEKLFYEDPYLREAKARIVEIKPLGKGKVAVLLDRTVFYPEGGGQPSDRGVI

EGEGFRLLVDKVKGKEEIWHEGRLKGRFPEEGEEVSLILDWEWRYENMRQH

TGQHILSAVIKELYDANTTGFQIFEDHNKIEIDYPGEVSWDMVLEIERKANEVV

WRDLPVEVQVYSKLPDELRKELRKELSEKVTPPIRIVSIPGVDVIPCGGTHVRS

TGEVGIVKVTNFYRKSTKLWRIEFACGNRALKYLDGILEDYWLSLERMPNKN

RPLVERVEELRNEISRLEMEKKDLRMELWDWKARALLGSSKKIGGLNVVVHR

ESWSLKDAQAFVVYLVDKNPNTVALFAGDNYVIFAKNKEFEGLNMKELLKE

VLFAVGGGGGGSENLARGGGFRAAPEEVLERAFRALEKKVVKL*

>NODE_34_length_49436_cov_6.06376_ID_67_3

MTASLDRLSLSVPGAEPYFSLSFQMSAKARRQLALPPAPVDASRVYRLRQVAD

RLGQLHADSACTAGRLEMLAVFSGVLRWLAYRYLQSRNCEVGTESIMIGRQT

VQLPGLRRCQEAFADLYPPDMVFSGMKPQVFLYGPFGEKGRTQTLIELLVLSI

QTANPAATAFRQLYDDVELSRRIAYRATLQALDGSLRQGGISGFLGMPLLELL

QAPVKASPDSLEGQLHYVRKHWQGLLPDALLLAIDAGLALRSEETRMRGGG

PGPQAVPDYSALPLTEEERFSSDKDWMSDLVLLAKSVHVWLDQLSRQYGRFI

KRLDEIPDEELDVLARRGFSGLWLIGIWKRSPASRQIKRLCGNAEAEASAYALY

DYAVAEEFGGQDGVNDLDRRCRQRGIRLACDVVPNHTGIDSRWMREHPDWF

VQASYPPYPGYRFTGPDLSTDGGMSLYLEDGYWDHSDAAVVFKRVDRHSGT

VHYIYHGNDGTHLPWNDTAQLNFLLPQVREAMIRTIVEVARTFRVIRFDAAMT

LAKKHYQRLWFPLPGGGAGVPSRSEYSMSSEEFERLFPQEFWREVVDRVTAE

VPDTLLLAEAFWLMESYFVRTLGMHRVYNSAFMHMMKHEDNAKFRQLLKK

TLAFNPEILKRFVNFMNNPDEATAVEQFGKGDKYFGVAVLLATLPGLPMFGHG

QVEGFREKYGMEYRRAYLQEDIDRGFVEHHEKMIFPLLHRRFLFSGARYFELY

DFISNGQVVEDVLAYSNRWGEQRALVVYHNRPLQTGGWVRKTCPRGEEGAV

QDEKPLSRALDLQVGENIFYRFRDHCSDVHYLRSSHELTEQGLFVTLGPYQAH

VFLDFAAVTDTDGCWRQLWQKLGGQPVADLDREYRRLFHAGVLQALNQVL

KAGADFFRLGDGGCARVVRIYGEFLQILQRDLHLGGDIERLTDCLQQQLQTV

KGLEEEIAGDIREILWPWLVLLSLERLSHGGAEPVPVADWLDRYPFEEVLMGR

WPEDVAVENMQLLRLLLRHRTCGCLEREHWLKRLFDDDDVRAFLLVHRYQD

CNWFSRERFERLVNGLMTTAVLEKQDSARQWFERCRHKIVLILARAEQTGYR

LDKFLTLV*

>NODE_68_length_31877_cov_6.39008_ID_135_25

MILPSFHLCLPTWLADMLPPPERTFTTPESRMSLVIEMARYNILRGSGGPFGAA

VFDLDNGQLIAPGVNLVTSANCSVAHAEMVALMLAQKRVNNFSFAAAGLNA

ELTTSTVPCAMCLGAIPWAGLKQLACGARESDARAVGFDEGDKPPRWPQLLE

KRGVRVLRDICRDKAVGVLQEYRRLGGEIYNGKG*

>NODE_77_length_27986_cov_11.8809_ID_153_7

MLVSFNFEVHQPHRLKKSVDIGCKNLWERYIDTNLNKDIFNKVANKCYIPANR

TILDLIDEYDIKVSYSITGVFLEQAMEFNDEVLDLFKDLVKTGNVELIGETYHH

SLSSFFEDHGEFREDIKLHQKMIKELFGYKTEVFRNTELIYNNKIAETIKDLGF

KGIFTEGTERILGWKSPNYVYNALCGLKVLLRNYKLSDDIGFRFSCQDWEEY

PLTADKYATWLSKTPGDCINLYMDYETFGEHQWKDTGIFEFLRHLPQELQKY

DHIEYATPSEILERCIPRGNIDVHEFSTISWADTERDVSAWIGNRMQQLSFSKLK

EIREHLGRYMSMKDKENPEKSKQISKNSPIETNSERNLSIHEPKGFVETLEYKI

YKNLQTSDNLYYMCTKGFNDMNVHSYFSHFESPFDAYAAYLDVMYDFKNHL

MIEPILRKYLKYKNHK*

>NODE_105_length_21092_cov_6.02199_ID_209_17

MMRRRPQAMPLLVLLLMFGLLAGCGRRGPVRPVRQPLPAAPEQLVLRQQGT

QMLLSWSMPQRNQDGTELTDLAGFKVMRMDYDPTEDCPDCRDTSILLRQIEL

EYLRDVQSVDGRFYLADPDLEEGRGYQYRIIPYNRWGQDGTPVSGREVFSIIP

PAPENVQAETTDGVLTLTWRAPQDMGSDMQLLGYNVYRRRPGRPFAVAPLN

RQPLSATLFEDHSFKSGNTYLYAVRAVVLLHHRGVESRLSKAVVATPWTSSRA

PF*

>NODE_147_length_13579_cov_9.74591_ID_293_12

MEIKHDYRVKLFDEMGFMRKKCKECGQYFWTLDPDRETCGDSPCDKYSFIG

NPITKKKYTYNEMVKEYIKFFEENGHTPIKRSPVIARRWRDDILLTIASIAVFQP

WVTKGIVEPVANPLVIAQPCIRLNDIDNVGRTGRHMTCFTMAAHHAFNKEDD

FKYWTDKTVELCFNFMKRLGIDEKSITFIESWWEGGGNAGPCYEVITHGVEL

ATLVFMQYEKIGDSYKEIPLKIVDTGYGIERFVWASQGTPTAYDAVFGNIVKKL

KENAGIDKIFESQDSSSIASAQDVERILAESATLAGLMDIENVGDLRVLRKKVA

EKIGMDVNELDKILSPLEYIYAIADHTRCLSFMFGDGIVPSNVREGYLARLVLR

KTLRYMDKVGISIPLKEIICMQLEDLKDLYPELMEMKDYIMDVVDAEEKKYIQ

TINRGRGIVERMVKSKTEISLDDLIELYDSNGLPPEVVQDIVEEINKKGKKEIK

VTVPDNFYTIVAERHEEEGVEEPKAKKQELPDVDAPETELLFYKNPKQKEFEG

KVLRTVGDYVILDKTIFYPEGGGQKYDIGYLNDIKVLEVQKKNGIVYHKVPE

VSRFKEGDIVKGTIDWINREKLMRNHTATHIINAAAQKVLGKHVWQTGSNVD

TEKGRLDITHYERITREQLKEIEKIANDIVLKGINVKSSFMSRNEAEQKYGFRI

YQGGVVPGNTLRIIEIEGTDVEACGGTHCENTSEVGFIKILKTERIQDGVERLE

YTSGTNSVEEVQKIEDFLIESADILGVPTTQLPKTVKRFFEEWKEQKKTIEELQ

KKIGEFKKYELANKFEKVGEYDVLVELVTGTQKELMSIADNITGENSIVVLLN

ENNYILCKRGKNVNLSMKELIRTIAKGGGKDDLAQGKYSDDIESIKAKVIGAL

QS*

>NODE_156_length_12705_cov_6.61345_ID_311_12

MWATEWANPSLSAIYMTKNSAGSAKKALTKKLKSAISTPRKHAPLAQLDRAS

DYGSEGREFESSAARHFTNIDPVDLDRKQDRGFMQEALVEASAAARLGEVPV

GAVVVKDGEIIGRGHNLRETSNDPTTHAEMIAIRQAAETLGSWRLIGCTLYVT

LEPCVMCMGAIILARIPRLVFGCRDPRVGAVGSVFDFSRDERFNHRVAVTEGVL

DQECSDLLSGFFRRLRAAKKQRRQQAVEDPPDDSS*

>NODE_259_length_5034_cov_5.48298_ID_517_4

MNRESEEMIKSWMTFALDLARQGGAEGEIPVGAVVVQSGRLVGQGFNRREG

TLDPTAHAEIVAIRQAANELGRWRLDDCDLYVTLEPCPMCAGALVMARIGRV

FYGAKDPKWGACGTLYDIPRDARWNHSCKIRGGILAEDCAKMLREFFRARR

QPEEQGGEMAERSKAGAWRASERRRSVGSNPAFSAIGLGAFSSVGQSVRLITG

RSGVRVPEGPP*

>NODE_413_length_3053_cov_3.95796_ID_825_3

MDKFLEAAIQEARQGMDAGGIPIGSVLVIDNQIVGRGHNQRVQKGSAVLHAE

MDCLENAGRIKASDYRRATLYSTLSPCDMCSGAVLLYGIPKVVIGENRTFCGP

EDYLRSRGVEVEVVDNPECYQLMETFIKKNLNCGTKILAFKNSHNKLINRTEN

ASFLREECDGSQNFALPPDIMKSWLTWMA*

>NODE_720_length_2231_cov_2.03279_ID_1439_4

MKTSTRILKVVESEGKRALVALEKNIFHPSGGGQPGDTGTLKSDVFEADVTDC

RYKDGIPLLVLSVKKGRPEQGMPVEAEVNLERNHLLSRMHTGEHILSRILEDS

HRGLQVYKVAIGEEESTISISYDGNVDWDILFDAEKKALEIIRADLPVNVEVVE

REKAETMEGLKINWERVGAPEIRVVSIPDFDIIACSGTHTDST

>NODE_1205_length_1673_cov_2.06468_ID_2409_2

MSPLLNEYKDEEGRIRHIWSTFSKDQVDLNYANYKVLLAVLDALLYYVKKGA

TLIRLDAIAFVWKEIGTSCVHLPQTHEVIQLLREAIHEVAPEVIIVTETNVPHDE

NISYFGSGEDEAQMVYNFALPPLIAHAIISENAHFLTSWAKTLSLPNDKVCFFN

FTASHDGIGMRPARGILPQSGLDLLQTTCIDHGGEVSYRKMPDG

>NODE_1807_length_1350_cov_2.07931_ID_3613_1

MRERNIMMNENAQWYKDAIIYQVHVKAFFDSNNDGIGDFEGLIRKLDYLKN

LGVNTLWLLPFYPSPLRDDGYDIADYKNIHPEYGSLRDFRRFLRRAHDMGFR

VLTELVINHTSDQHPWFQRARKAKPGSVWRDFYVWSDDPNRFSEARIIFQDFE

TSNWTYDPVAGTYYWHRFYSHQPDLNFDNPRVRRAVLKILDHWMNMGVDG

FRLDAVPYLYERDGTNCENLP

>NODE_1807_length_1350_cov_2.07931_ID_3613_2

DAGLSCRIAETELVEYFQPNFWPNTPDILPEFLQIGGRSAFVIRLVLAATLSSCY

GIYGPPFELLVSDALPGHEEYLNSEKYEIRDWNWDQPRNLKDLVFRVNAIRHE

NKALQSTRNLKFLETDNENILFFLKESTEEDNLLLIGVSFDPFTNQSCHVTLPL

EMLDIERNQPYLLHDLLGDEKFVWQGETNMVGFDPAVLPAKIFRVYRRMHRE

EDFDYFM*

>NODE_1930_length_1308_cov_1.23116_ID_3859_2

SKAIMDFPLEKARYLMNLALKEAGLGAKKGEVPIGAVLWDLKHHQILAKAH

NQSIALKDPSAHAEILALRQAGRKNKQLSAIK*

>NODE_2198_length_1221_cov_2.31536_ID_4395_1

MISVCFYFQVHQPMRLDKNYSFFQMGRSHHYRDEAANRDIMRKVADKCYLP

ANRMMLDLIELHKGKFRISYAITGVAMEQFQEFCPEVLDSFRALADTGCVEFI

GETHYHSLAFLFSREEFRRQVKMHSRILQEFFGAKPVTFRNTELIYNNDLALEI

EKMGYKAILAEGADQVLGWRSPNFVYQPAGCSKLKALLKNYRLSDDVAFRF

SDRGWSEWPVTVEKYADWVHKVAGSGEIINLFMDYETIGEHQWADTGIFEFF

RKLPEAVLSRGDFAFETPGEAAAHLDPMAQLD

>NODE_2245_length_1207_cov_1.59074_ID_4489_1

DRIALLVARYENLSSKVDPSLWEQNVTILISYGHMIQEKGTVPLQSLLQFLDSQ

LQEYIQGIHILPFFPYSSDDGFSVIDYRKVNPELGDWQDIERIARKYILMVDLV

LNHVSKCSSWFQDYIGGILPALDFFIEVDPKTDLSQVARPRTTPLLTPVDTVLG

TKWVWTTFSEDQIDLNYKNPDVLIEMLDILLFYILKGARIIRLDAIAYLWKEIG

TSCLNLPQTHEVVKLFRDIIDYLAPGVLLLTETNLPQKDNYSYFGEQDEAHMV

YQFSLPPLLLYSFHHQDSQLLTQWLQNISPPENCTFFNLQLLTMVLALGLWKG

LFLKKNWKI*

>NODE_2612_length_1106_cov_1.38407_ID_5223_2

MVKIIETVLMKISAITMGFEGETDDPKINEIRKRQVKNFITILMVSHGTPMILM

GDEIYRTQHGNNNAYCQDNEKTWLDWTLKEKHQDIFRFFKKMIEFRKKQSRI

KEKTFLYR*

>NODE_2761_length_1074_cov_2.9736_ID_5521_1

MLLTTPGIPFIFYGDELGMKGVYDPYFTESVIEPFPWYSSLSGDGQTLWKSVG

FNHAFTGVSVEEQSQREDSLLNTVKRWIKFRKENEWMTNSWIVDLRTSEFVV

GYTVTNGEKSLRIYHNISGHEEEFEGIRLKPFESKVL*

>NODE_2761_length_1074_cov_2.9736_ID_5521_2

EFWLNMGIDGFRFDAAKHIYDYDLQQKKFSYNHEKNIQFWKKVMDKARSIK

SDVFAVTEVWDAPEIVAEYAKTIGCSFNFYFTEALRESINHGNTYKIWDCFSRT

LTDDRGLYIPSNFSSNHDMTRLASSLQSEDQRKVFFCNAPYNARNSIYFLWR*

>NODE_3522_length_926_cov_1.61827_ID_7043_2

MGPKGKAEVVNTLKVGKSLILHQVKVQEGNLQEKETVLLQVDEGDRIATAR

NHTATHLLHAALRKVLGEHVKQAGSLVEPKRLRFDFTHIKPVSEEELKKIEAE

VNRVILSGVEVETEVMDYEEAVQQGAMALFGEKYEDKVRVVKVPGFSAELC

GGTHLKNTSQAGLFVIASEEGVAAGVRRIEALTGFEAYKYVQEVRDTLKEVQ

SSLEVGARQVVDKVKSLVQENKQLVRENNDSPKNLLPGREQTC

>NODE_3590_length_916_cov_2.20913_ID_7179_1

CAPALIFLSEAIVHPDEVGRYIRKDECQLSYNPQLMALLWNTLATRDVGLLHR

AMDRWFEIPSDCAWVNYVRSHDDIGWVFSDDDAQALNINPYDHRRFLNDFF

TGKFEGSFARGLPFQDNPITGDMRISGTTASLAGLELALEHDDQHEVNLAIQRI

LLLYGVILTIGGIPLLYLGDELGVLNDYGYERDPQLAGDSRWAHRIKTDWDKA

SSRHKRETVEGEIFLGLLRLIQIRQQNLAFTRADTEIISTGNSHVFGYFRRHENQ

SVLVLANFSEEEQTISARRLRALGLRKTVRIYLQGTSLL

>NODE_3615_length_912_cov_2.24331_ID_7229_1

MEYLQNLGVSAVELMPVHQFVHDARLVEKGLRNYWGYNTIGYFAPHNEYAV

YGQTGQQVQEFKQMVKTLHEADIEVILDVVYNHTAEGNHLGPTLSFRGIDNA

AYYRLNAEDPRYYVDYTGTGNSLNMRHPHVLQLIMDSLRYWVTEMRVDGFR

FDLASTLARELHDVDKLSAFFDLVQQDPVVSQVKLIAEPWDVGEGGYQVGN

FPPLWSEWNGKYRDCMRDFWRGKDQTLGEFAYRFTGSSDLYENTGRRPFASV

NFVTAHDGFTLQDLVSYNEKHNEANGEGNADGTDDNRSWNCGAEGP

>NODE_3747_length_895_cov_4.79948_ID_7493_1

MTSRKKDESWISEQSKLTLARLMPRLEDRFAAEIEPASWRSYRDRLKQNFPKL

FSALMSIYGTRYDFFYHLEQIIITATEYWANRSDSLKALDASRAADPSWFQSQR

MVGAMCYVDLFADDLETLRAKIPYLSEMGITFLHLMPPFQSPDGDNDGGYAI

SSYRDIDPKLGTMEQMADLATELRHYGISLVLDFVFNHTSDEHKWAQHALQG

EEEYQNYYRMFPDRTEPDLYEKHIRDVFPDEHPGCFTYRNRIRKMGLDNIP*

>NODE_4201_length_835_cov_0.951977_ID_8401_1

MWQKGVDILLRVLPQILEMGYQVVLQGTGDPRIEQMCREAAADARGQVAVN

LGYDEAFAHSIIAGSDLLAVPSRFEPCGLTQLYALRYGTLPFVRKTGGLADSVI

DARENETGTGFVFQGENPDEVVQILTEARHMFDQPAVWQSIQQRAMKQDYS

WDTAAKQYVKVFENVHP*

>NODE_4833_length_768_cov_3.47114_ID_9665_1

FWFRDACLMLNVLLAFGLTDQAKRILDTFPARQKNNGYFQSQEGEWDSNGQ

VLWIMDRYQRITGEKPVTAWMNAVEKGAEWIVRKRIRKKDAGLHAGLFPPGF

SAEHLGLNDYYYWDDFWGVAGLRSASRLLELTGKIEQARRFETEARNFEEAL

FRSIDAIPESRSQGAIPASPYRRIDAGAVGSMVADYPLQITPPANERITKTAELM

LQRFFRKGAFFQDIIHSGINAYLTLDIAETLLRGKDPRYRDLLHN

>NODE_4903_length_759_cov_1.41456_ID_9805_1

MDVWPGKPYPLGATYDGMGTNFSIFSEAAEKVELCLFDDEGKETRVELPEMT

AFCWHGYLHGIGPGQRYGYRVYGEWAPDRGLRCNGAKLCSIPMQKPLKEVL

RGIPPSSGIHWENPTNS

>NODE_2_length_194384_cov_26.6735_ID_3_133

MMSLVFHGNLQYAEIPKSEIPKVIEKAYLPVIGRLLKEEIPFGLNITGYTLEILPR

EVIELVREGIASGLIEIIGTSYTHAILPLLPLDRVEAQVRKDRELKEELLEVSPKG

FWLPELAYDPIIPAILKDNGYGYVFADGEAMLLSDHLNSAVKPIKPLYPHLIKA

QRGEGNVFLNYLLGLRELKKAVEKAFPGKVTLEAVKNIDAVPVWVAINTAVM

LGIGRFPLMSPKKAANWLRGKDDVLLYGTDIEFIGYRDLAGRRMTIDGLLEV

VKALNVELSLPSELKHSGRKLYLRTSSWAPDKSLRIWTEDEGNARLNMLTPFV

GEPLAFLAENSDARGWEPLPERRLDAFRAIYNDWRGSK*

>NODE_3_length_148265_cov_21.0578_ID_5_72

MPGIPIEVRTHTALHVVKGAVVKVLGEEAKWTASTYVKGNRGVLIVKFSRKP

TAEEVAEIERLANEKVEEDVPIEVYKLPREEAEKRFGEDMYDLFPIPPDVRTLR

VVVIEGWNVNACKEEHAGTTGEVGEIRIRKVRFRPNKELLEVSFEVL*

>NODE_15_length_69714_cov_14.0118_ID_29_34

MTEKLFYEDPYLREAKARIVEIKPLGKGKVAVLLDRTVFYPEGGGQPSDRGVI

EGEGFRLLVDKVKGKEEIWHEGRLKGRFPEEGEEVSLILDWEWRYENMRQH

TGQHILSAVIKELYDANTTGFQIFEDHNKIEIDYPGEVSWDMVLEIERKANEVV

WRDLPVEVQVYSKLPDELRKELRKELSEKVTPPIRIVSIPGVDVIPCGGTHVRS

TGEVGIVKVTNFYRKSTKLWRIEFACGNRALKYLDGILEDYWLSLERMPNKN

RPLVERVEELRNEISRLEMEKKDLRMELWDWKARALLGSSKKIGGLNVVVHR

ESWSLKDAQAFVVYLVDKNPNTVALFAGDNYVIFAKNKEFEGLNMKELLKE

VLFAVGGGGGGSENLARGGGFRAAPEEVLERAFRALEKKVVKL*

>NODE_37_length_32450_cov_0.630468_ID_73_10

MLVSFNFEVHQPHRLKKSVDIGCKNLWERYIDTNLNKDIFNKVANKCYIPANR

TILDLIDEYDIKVSYSITGVFLEQAMEFNDEVLDLFKDLVKTGNVELIGETYHH

SLSSFFEDHGEFREDIKLHQKMIKELFGYKTEVFRNTELIYNNKIAETIKDLGF

KGIFTEGTERILGWKSPNYVYNALCGLKVLLRNYKLSDDIGFRFSCQDWEEY

PLTADKYATWLSKTPGDCINLYMDYETFGEHQWKDTGIFEFLRHLPQELQKY

DHIEYATPSEILERCIPRGNIDVHEFSTISWADTERDVSAWIGNRMQQLSFSKLK

EIREHLGRYMSMKDKENPEKSKQISKNSPIETNSERNLSIHEPKGFVETLEYKI

YKNLQTSDNLYYMCTKGFNDMNVHSYFSHFESPFDAYAAYLDVMYDFKNHL

MIEPILRKYLKYKNHK*

>NODE_60_length_16859_cov_0.515711_ID_119_16

MEIKHDYRVKLFDEMGFMRKKCKECGQYFWTLDPDRETCGDSPCDKYSFIG

NPITKKKYTYNEMVKEYIKFFEENGHTPIKRSPVIARRWRDDILLTIASIAVFQP

WVTKGIVEPVANPLVIAQPCIRLNDIDNVGRTGRHMTCFTMAAHHAFNKEDD

FKYWTDKTVELCFNFMKRLGIDEKSITFIESWWEGGGNAGPCYEVITHGVEL

ATLVFMQYEKIGDSYKEIPLKIVDTGYGIERFVWASQGTPTAYDAVFGNIVKKL

KENAGIDKIFESQDSSSIASAQDVERILAESATLAGLMDIENVGDLRVLRKKVA

EKIGMDVNELDKILSPLEYIYAIADHTRCLSFMFGDGIVPSNVREGYLARLVLR

KTLRYMDKVGISIPLKEIICMQLEDLKDLYPELMEMKDYIMDVVDAEEKKYIQ

TINRGRGIVERMVKSKTEISLDDLIELYDSNGLPPEVVQDIVEEINKKGKKEIK

VTVPDNFYTIVAERHEEEGVEEPKAKKQELPDVDAPETELLFYKNPKQKEFEG

KVLRTVGDYVILDKTIFYPEGGGQKYDIGYLNDIKVLEVQKKNGIVYHKVPE

VSRFKEGDIVKGTIDWINREKLMRNHTATHIINAAAQKVLGKHVWQTGSNVD

TEKGRLDITHYERITREQLKEIEKIANDIVLKGINVKSSFMSRNEAEQKYGFRI

YQGGVVPGNTLRIIEIEGTDVEACGGTHCENTSEVGFIKILKTERIQDGVERLE

YTSGTNSVEEVQKIEDFLIESADILGVPTTQLPKTVKRFFEEWKEQKKTIEELQ

KKIGEFKKYELANKFEKVGEYDVLVELVTGTQKELMSIADNITGENSIVVLLN

ENNYILCKRGKNVNLSMKELIRTIAKGGGKDDLAQGKYSDDIESIKAKVIGAL

QS*

3.7.3 All lipase/esterase ORFs

>NODE_1_length_435385_cov_101.67_ID_1_238

MFVGHYKEVPEKDTGFEGVTIRWLVSPKLGAKNFAMRYFVMKKGSEIPIHQH

DWEHEIFIVKGEGVITDGKEEYPVKAGNFLYVPPNEPHGYKATGETFEFLCIIP

AKKEAIPEDEWA*

>NODE_9_length_97431_cov_87.7446_ID_17_8

MGENPFGNPPTNLLPIEVPPHVLMLRGIGWDSNIYLVRDGEEALIIDTGTGINW

HVYAGMWERGGYLEGVRRVIIFNTHEHFDHVGGNNVVKRWLERHGIEVYFA

AHKITANVIERGDEGVILSYFYGRRFEPHQVDFKLEDGDKLRVGSLELLVIHTP

GHTAGSSCLYLDDGKHRVMFTGDTVFKGTVGRTDLPTGDGWALRESLERLA

GFDVDFGFPGHGGYIDDWEGNLKEVLRWLA*

>NODE_14_length_75743_cov_96.3567_ID_27_19

MWERNRVIVLGHRGYMSDYPENTLLSFRKAVEAGVDGIELDVWLTKDGRLV

VMHDETIDRTSNMKGRQKDMTLEELKKADVGQGERIPTLEEVFEAIPRNALV

NVELKDREAAREVAEIVAENNPERVMISSFDIDALREYRKYDDETTMGLLIDR

EEVVPLIPKLKDELNLWSVNVPMEAIPLIGLEKTLQALHWIRSLGLKVVLWTE

NDVLFYKDDNLAKLKGLFEVVIANDVVRMIDYLRKLGLR*

>NODE_14_length_75743_cov_96.3567_ID_27_26

MIWQIALLLILVFLAFVAFVGYKMVKPPRLVEDWTPKDFGFEYEDIEFTTEDG

VKLSGWWVENGSDKTVIPLHGYTASRWYSLYMKPTVEFLLKEGYNVLVFDF

RAHGESEGKYTTVGDKEILDVKAAVKWLKGTHSEKARKIGLIGFSMGAMVTI

RSLAEIEDVCCGVADSPPMDLDKTGARGLKYFANLPEWLYVFVKPFTKLFSG

GKEISPMEYADRVKKPLLLIAGEKDPLVKVEEVREFYERNRKINPNVELWVTD

APHVRTLKFHPEEWKARVREFFNRAFNNP*

>NODE_15_length_74913_cov_43.2825_ID_29_60

MAKGLTEKDLGKFKLLGNIDALGRKLVFQVTEISVEKDDYFSRLYLYDGRRV

KPFTSGKKDGNPRFSPNGKLVAFTSKRDKESKEAELYVIPTDGGEARLLAKFK

YGIKNLRFTEDGKGIAVVTPVDVEKKPKDDVHIIKELPFWFNGTGWVYGKRS

VVYLVDVESGKKKRLTPKNLDVGQIRFHNGKLYFTAQEDRERKPMVSDLYVL

EGRKAKRLTPGKWSISDFIPLDDGTFILKANTRERGIPTNTHIYHYNPETGEMR

KLTRDLDRSAYNSLNSDVRGAQRAELVFKNGWVYYVATDGPRANLFRVNLD

GKIERVIGGDRSVESFAIGDYIAFTAQDAVTPLELYLLRDGKEKKVTDFNGWIR

DYSLSKPEHFTVKASDGVEIDAWVMKPTNFEPGKKYPAVLEIHGGPKTAYGY

AFMHEFHVLTAKGFVVIFSNPRGSDGYGEEFADIRGHYGERDYQDIMEVVDE

ALRRFDFIDPERIGVTGGSYGGFMTNWIVGHTNRFKAAVTQRSISNWTSFFGT

TDIGYFFAPDQIGGDPWSNTEGYWEKSPLKYAPNVETPLLIIHSMEDYRCWLP

EALQFFTSLKYLGKTVELALFPGENHDLSRGGKPKHRVRRLELIVGWMEKWL

KG*

>NODE_17_length_70503_cov_89.6033_ID_33_70

MEIYKSKFGTPERGWVILVHGLGEHSGRYGKLISMLNDAGFAVYTFDWPGHG

KSPGKRGHTSVEEAMEIIDSIIDEIGEKPFLFGHSLGGLTVIRYAETRPERIRGVV

ASSPALAKSPKTPGFMVVLAKVLGRIVPGLTLSNGIDPNLLSRNPDAVKRYIED

PLVHDRTSGKLGMSIFTNMEKAHEDAGRIKVPILLLVGTGDLITPPEGSRKLFE

ELKVEDKEIKEFEGAYHEVFEDPEWGEEFHRAIVEWFVEHSERA*

>NODE_31_length_53957_cov_6.22129_ID_61_11

MIDIWQATMVTGQLGVNCYLLGCPRTRQAIVIDPGGDGSRILALLEEQDFVLK

TVVNTHGHFDHIGANRTLIQKTGAELLLHEADLPLLRGAADHAASFGCQPIEP

SPEPTRLLKGGDRVEVGTIGLDVLHVPGHSPGSVCLKSGEDLFVGDVLFAGSI

GRTDLPGGDHLLLLKGLHSRLMTLEDSVRVYPGHGPETSIGRERKSNPF*

>NODE_33_length_50065_cov_6.70844_ID_65_7

MVINQLESLSCNKSFGGWNKRFRHYSPTLNCDMAFAIYLPPQAEQQRVPVLY

WLSGLSCTEENFVQKAGAQRIAAELGIALVAPDTSPRGECVPDDPQGDYDFGL

GAGFYLNAVQEPWARHYRMYDYVVKELPALIEELFPVTGERSISGHSMGGHG

AIVCALRNPERYRSVSAFAPIANPMNCPWGQKAFSGYLGKERKLWLDYDSSV

LIGEAKDQLPLLVDQGGKDQFLSEQLKPEALRQAAEENGYPLVYRQQDGYDH

SYYFIASFIEDHLRFHADFLNGRSVEQWLAEQE*

>NODE_60_length_34884_cov_6.20968_ID_119_2

MRRTCNIPLSTMILAVMLFFILTTGCRDRTPRLRALPNHAVILAFGDSLTAGNG

ADHEASYPARLQQITGWRTINAGVPGEISAEGVERLPGLLQRYSPDLVVLCHG

GNDLLRHIAGDTTAAHLATMIEMIRENGAQVILLGVPRPGILPRPATFYKKVAE

QYGVPLESETLTEILRDDSLKSDLIHPNAAGYDRLAKAVAAILKRNGA*

>NODE_81_length_26871_cov_7.57804_ID_161_13

MRFSALFCLFSVCLFSGIAQSATLPTAFDLRNIDGRSYIGPVRNQEQCGSCWSF

GTLAAAETTWNRTRGLYDDQTIDFSEAFLTWSLSPLYDGLHGCNGGNLELQQ

NTALIEYGVPLESDFPYTMTDPGEDLHWDATRYSFLDWYRIPPGDIETTRRVL

YHIGAVTAGVLVEDDFYDYTGGIYTNGTTAITSKIPYNTAVNHLIALVGWNDD

PGDGGLGYWILRNSWSDRWGLDGYMNIRYTTAGVTLHSSYMTLEPWDGASI

ALENNNDLTATPWSAGGTLNAHGVDLWGGAASSVANRGAILAEAYSADELAT

ARGVYLWGGPEGAVSNAGTIVGLAGSENQQASAYAVCLQGGLVNNAGLLGAI

AESYADQALAFGIWAANGGSAAEITNSGEIIAWAHESAMNAAYGIWADSRSLI

KVTNTGSIEAYADDYAIGVLLTGGPALLQNSGTIRGSYASVYALQNTLMVLGT

GSDLFGRVFLKGDEDTLVLTGNGTEDTAFYDVETLLMAGNDWSLSGDSTFDT

IEIALGRLGMDGNLTGETSILEDGILGGNGSLTGMVTNTGTVAPGHSVGHLTIA

GDFIQTSGGTLEIEIGDGIGDLLTVSGTADLAGSLLVLPDGYASAGSYTFLEAGT

IAGAFDNLVSAAVFSVTLNDDTASTLSLDLARNSYLSLAAPHNRGLADTLDDL

RPTADSDFGALLDRLDLALSRQALNDGLAALTPRIHGLASTVLIGDAQERLAG

LRRHLQQIDPAMLLEGNPSGKISAWFDILGQYNRYGSDGGYFGARENLYGLLL

GVERTTAKGLTLGVAASVSECRLEARDSDDDSEIETRQGYLYAAWRDPRRVG

GLHLNAAMGGGLSQLDSERVIPFAGRKTRSEHDGTLLGATIGGGYAVAAGGW

IFDPTVGLSFVHLREESFCESGADSADLKIAARDNDSLQSLLGLRLRRPIQLTAF

SLEPELRLEWRHEFNRKTESLRARLAGGGTFATPGRYLAGDGVRLGASVKTIL

GDSVSGMLDYDCDLQSHGATGHALRLQLAVAF*

>NODE_85_length_26589_cov_6.44785_ID_169_13

MKSRKITITILVDNQTHEGLHAEHGLAMWIEADGRRILFDTGQGAALASNAQ

ALGVDLKDTDMLVLSHGHYDHTGGIPELLRQSGKVDVYCHSGVVQPRYAIRE

GRPKPIQMPTKAMAALDRLPSQRLHWIRQPTWLTEDIGITGFIPRGSNYEDTG

GPFYLDPEGRYADPINDDLALWIRQDDELVVCVGCSHAGLVNTLNYIRDLNH

GQRIRTVIGGFHLLDAGSERLERTISALTELEPETVIPCHCTGDNAVAMLRQTFG

KAVLPGAAGMIFQF*

>NODE_88_length_25483_cov_6.18761_ID_175_24

MKIGTLEIIQIPAGEMNNFSYLLFCPATRRGLAVDPSLEPQRLLDAIDAHGVDLT

WLVNTHGHRDHVAGNDLILQATGASLAAHPLAVPKIDRPLTEGDALMVGDTE

VKVLHTPGHTPADITLNPPGVLLTGDTLFVTKVGRADMTGSDPVALYDSLRRL

AQFPGETLVFPGHDYGPKAYSSIEYERRNNPYLRCPDLESFLALRMG*

>NODE_105_length_21092_cov_6.02199_ID_209_7

MKGIVIRTAEQEEYPHPNHDRFFLRDVVTAATNPALSLHRGRIEAGGEILPHTH

EGQTETFYILGGEALCTVNGEEHTFGPGCCVVAPPGVRHSLKNIGDEPVDLLA

IFTPPLK*

>NODE_116_length_18753_cov_6.42371_ID_231_5

MKFNKFKVIAVCAGMMLSFTYGVVTVQYKIFPFEQLRAIKQIASPSPTYSDYF

YHKKSFFEQHGGRNYDVVFIGDSITDGAEWEDLFPSLKIANRGIGGDRTDGVL

KRLDSIYSTSAGRAFIMIGINDFNSGMSVDEVFENYRSIVNKLSEHGMKIYIQS

TILAGKRRENLNIKIVELNKRLELFATQNDSIIYIDLNAGLAQDSLLNPMYSRD

DVHLNGKGYAVWKGIISPYVQ*

>NODE_193_length_8683_cov_2.56837_ID_385_12

MPSSLNLWVVGSVGTNIELNGDDLGLHGGIGLTWERGEWRVGGGLFGDSRD

LDTSYHGNQDIKAVGPGAFVAYSPEGTGLEFRVTGLWQTVDLDLKRGYANGA

GYATSDGSTNANVLGLSGRVQWTRGVTDQLALTPFAEYTWQTTHIDGYSESG

GPFPASFNSRDETSNSLRTGLRADAALFADADTWAWIAWDHRFEDKSSGLGG

TALGLGSFTYAGSRVDQDWADVGVGASWDISERLSANTALGFALGCDDETMS

DVTATVGFSYQLW*

>NODE_402_length_3127_cov_2.42867_ID_803_2

MTIEAPMSLMRKTLAGILSLLLLTGCAHSTVVTDMASDARWHRLDTTNPLPA

VGWVRGRADVIHIYIEGDGVAYSTPTKPSPDPTPITPTALLLAQQDGAPAAAYL

GRPCQYVSGEACNNGCWTSGRFSEAVLRTMNELVDAAKLEAGAGRVVLIGFS

GGGAVAALLAERRPDVAELVTVCGNLDPAEWTSMHGVTPLHGSLNPADRAA

ALSDLPQTHLLGGADTNVTRRVTDAFVSRLTPGAPVTVRVVPGLGHGGADW

AEAWPALLSGLRVSD*

>NODE_743_length_2180_cov_1.67414_ID_1485_3

MFFKLGEMKGTRRENLQNGPGGAMYYSCIAPGEKPEGSRLKMVARIELDPGA

AVGEHRHSGDEEVYIALSGEGVFTDDGVRHDVGPGDVMITLDGHRHSLENTG

TGPLVFVAVIAE*

>NODE_1237_length_1650_cov_1.2088_ID_2473_2

MQEIYKLEKAEKQVTFFNERLVNERKISKPEKFTFTNKDGIELEGWIIKPVDFE

EGKKYPAILDIHGGPKTVYGEVFFHEMQIWASEGYVVMFTNPRGSDGRGNKF

ADIRGKYGTVDYEDLMSFVDEALKRYPFIDKERLGVTGGSYGGFMTNWIIGH

TDKFKAAVSQRSISNWISKFATTDIGYYFVADQQSATPWDNFEKLWWHSPMK

YADKVKTPTLFIHSDEDYRCWIAEAIQMFTSFKYFGVESRLVILKGENHDLSRS

GKPKNRITRLREITNWFNKYLKE*

>NODE_1864_length_1327_cov_2.27333_ID_3727_2

MPDLGLPLDGEFYPTQKMAPAKSLVVFLHGYGANGLDLINIASYLDKLLPDT

AFYAPDGPEECDLTPMGRQWFSLASVDPHQMRTDPRSLPLAMESLHHGVCQ

AAPILNDFLDAVLDRHNLTADKLALFCFSQGTMMGLHVALTRPDTVAGVFGFS

GLFTGGNKRLPEKPAGDMPPVWLVHGAADDVVPPQAMRMSQDALAEYGIK

AKTHMRPTLGHAIDDDGLFIARDGLLEVLR*

>NODE_1875_length_1322_cov_1.159_ID_3749_2

MEFIEAHGLDLEWVLLTHGHADHICGLEKLRPLARSGVAVHRLDAPMLASPE

QNLSAFMQDRCRSKPPERELEDGDMIHAGGLVIAVIHTPGHTPGSVCFHVKED

GEEALLSGDTLFAQSVGRTDLPGGDQGKLLRSLEKIASFPDGMAVYPGHGPET

TIGAERRKNPFWPGEQQ*

>NODE_3033_length_1015_cov_1.55405_ID_6065_1

GDSLTEGLGVNKEDAFPKLVETMIQNELQKDITVINGGVSGSTTSDGLARLKW

YMKKKPYLVFLALGANDGLRGLNLQQSEQNLEEIIKYAQKHNAKVLLAGMLI

PPNYGVEYSQQFKKMYQQLKNKYNLGSMPFLLDGVAGKKELNQRDGIHPNE

AGHQHIAKKFLNF*

>NODE_4_length_144450_cov_36.9995_ID_7_9

MAKGLTEKDLGKFKLLGNIDALGRKLVFQVTEISVEKDDYFSRLYLYDGRRV

KPFTSGKKDGNPRFSPNGKLVAFTSKRDKESKEAELYVIPTDGGEARLLAKFK

YGIKNLRFTEDGKGIAVVTPVDVEKKPKDDVHIIKELPFWFNGTGWVYGKRS

VVYLVDVESGKKKRLTPKNLDVGQIRFHNGKLYFTAQEDRERKPMVSDLYVL

EGRKAKRLTPGKWSISDFIPLDDGTFILKANTRERGIPTNTHIYHYNPETGEMR

KLTRDLDRSAYNSLNSDVRGAQRAELVFKNGWVYYVATDGPRANLFRVNLD

GKIERVIGGDRSVESFAIGDYIAFTAQDAVTPLELYLLRDGKEKKVTDFNGWIR

DYSLSKPEHFTVKASDGVEIDAWVMKPTNFEPGKKYPAVLEIHGGPKTAYGY

AFMHEFHVLTAKGFVVIFSNPRGSDGYGEEFADIRGHYGERDYQDIMEVVDE

ALRRFDFIDPERIGVTGGSYGGFMTNWIVGHTNRFKAAVTQRSISNWTSFFGT

TDIGYFFAPDQIGGDPWSNTEGYWEKSPLKYAPNVETPLLIIHSMEDYRCWLP

EALQFFTSLKYLGKTVELALFPGENHDLSRGGKPKHRVRRLELIVGWMEKWL

KG*

>NODE_8_length_109724_cov_30.3658_ID_15_83

MFVGHYKEVPEKDTGFEGVTIRWLVSPKLGAKNFAMRYFVMKKGSEIPIHQH

DWEHEIFIVKGEGVITDGKEEYPVKAGNFLYVPPNEPHGYKATGETFEFLCIIP

AKKEAIPEDEWA*

>NODE_11_length_97240_cov_21.356_ID_21_8

MGENPFGNPPTNLLPIEVPPHVLMLRGIGWDSNIYLVRDGEEALIIDTGTGINW

HVYAGMWERGGYLEGVRRVIIFNTHEHFDHVGGNNVVKRWLERHGIEVYFA

AHKITANVIERGDEGVILSYFYGRRFEPHQVDFKLEDGDKLRVGSLELLVIHTP

GHTAGSSCLYLDDGKHRVMFTGDTVFKGTVGRTDLPTGDGWALRESLERLA

GFDVDFGFPGHGGYIDDWEGNLKEVLRWLA*

>NODE_12_length_86170_cov_30.6128_ID_23_20

MIWQIALLLILVFLAFVAFVGYKMVKPPRLVEDWTPKDFGFEYEDIEFTTEDG

VKLSGWWVENGSDKTVIPLHGYTASRWYSLYMKPTVEFLLKEGYNVLVFDF

RAHGESEGKYTTVGDKEILDVKAAVKWLKGTHSEKARKIGLIGFSMGAMVTI

RSLAEIEDVCCGVADSPPMDLDKTGARGLKYFANLPEWLYVFVKPFTKLFSG

GKEISPMEYADRVKKPLLLIAGEKDPLVKVEEVREFYERNRKINPNVELWVTD

APHVRTLKFHPEEWKARVREFFNRAFNNP*

>NODE_17_length_63320_cov_4.74196e-05_ID_33_38

MHAMQESVMKSRKITITILVDNQTHEGLHAEHGLAMWIEADGRRILFDTGQG

AALASNAQALGVDLKDTDMLVLSHGHYDHTGGIPELLRQSGKVDVYCHSGV

VQPRYAIREGRPKPIQMPTKAMAALDRLPSQRLHWIRQPTWLTEDIGITGFIPR

GSNYEDTGGPFYLDPEGRYADPINDDLALWIRQDDELVVCVGCSHAGLVNTL

NYIRDLNHGQRIRTVIGGFHLLDAGSERLERTISALTELEPETVIPCHCTGDNAV

AMLRQTFGKAVLPGAAGMIFQF*

>NODE_44_length_27172_cov_26.2319_ID_87_23

MEIYKSKFGTPERGWVILVHGLGEHSGRYGKLISMLNDAGFAVYTFDWPGHG

KSPGKRGHTSVEEAMEIIDSIIDEIGEKPFLFGHSLGGLTVIRYAETRPERIRGVV

ASSPALAKSPKTPGFMVVLAKVLGRIVPGLTLSNGIDPNLLSRNPDAVKRYIED

PLVHDRTSGKLGMSIFTNMEKAHEDAGRIKVPILLLVGTGDLITPPEGSRKLFE

ELKVEDKEIKEFEGAYHEVFEDPEWGEEFHRAIVEWFVEHSERA*

>NODE_57_length_17552_cov_9.55604_ID_113_10

MSAVFQSRLIKLVVLGGLSFGVFLGIWWADRFRETPPDQYHIKGKFGPQFGWP

IIPIHVVLLPDGRLLSYGTDERGRQGAQFQYDVWDPKRGVGPDSHLTLPNKV

GTDLFCSGQLIVPADDSVLLVGGDRTVNGVRNWSSPDINFFDGKTNVLRSAG

RTMERPRWYPTVTTMPNGEIVVTAGRLDPEHYAPLPEIYNPKTGWRTLPGAE

DVAAFGVHNWDYPRQFVQPNGKVFVMSVEGHAYEFDTSGEGSVRPLNVSIF

RGHPYLPSVMYLPGKILTVRWLGLTYDLDINGKEPVVKSAAWSGLARFNGSM

TVMADGTVLLNGGSMLNNASKWYLAPNYESKIWHPDTGKWTDAAIAKRMR

LYHSISLLMQDGSILTGGGGATGPETNLNGEIYYPPYLFKKDGSGERAVQPVLQ

EAPDFLDWGQSFKIKADTPNISKIHLIKTGSVTHTNNFEERFIPLNFKALGEGRF

SVEAPANANIAPPGYYHLFVLNSDGVPSYSKLIKFGG*

>NODE_77_length_8684_cov_0.00312898_ID_153_1

MGQVGPAVTGMGQLSMSRLGGVAGGQGMHFAVSNPGPASGAPSAGTGSETG

LSSGDEMPGRLTLWVVGSVGTNIELNGDDLGLHGGIGLTWERGEWRVGGGLF

GDSRDLDTSYHGNQDIKAFGPGAFVAYSPEGTGLEFRVTGLWQTVDLDLKRG

YANGAGYATSDGSTNANVLGLSGRVQWTRGVTDQLALTPFAEYTWQTTHID

GYSESGGPFPASFDSRDETSNSLRTGLRADAALFADADTWAWIAWDHRFEDK

SSGLGGTALGLGSFTYAGSRVDQDWADVGVGASWDISERLSANTALGFALGC

DDETMSDVTATVGFSYQLW*

3.7.4 All protease ORFs

>NODE_3_length_227264_cov_100.645_ID_5_149

MNRKILGLLIAVVMLLSVVPAGLSIGAVSASPAAPSITTGGSSSQETPKFIPGDV

VEKQIERILRNKHGGKVRLIIAPEKDRAMEVYEALKNLGRIDPISRPEYQFIVV

EMPAGNLEKLKEIPGILRVWEDRTVKLLEPVEPEEALKTLKENSPAKPDMFMS

VFEINAVNVWNNYSILGDDVVVAVLDTGIDVSHPFLQTTLDGRPKIIDIYDASD

EGIAQLYYATNTTLNGTIVVNMEVPVYWGAYYPYYGHDKITTYYNMTSYFV

GNITGDEYYLGLLPERYFDLNNFFGTPNDPYNLGLFGDLSDVYPVLIVENNGT

YTAYIDYNLNNNFTDDQPIGLFTETGDYFQTPDTLVSIALAKVHIGDMDNPDN

YPYIVPYGDGIGYAMFMWDPHGHGTHVSGTVAGVGQPDDPMFRGVYGVAP

NAQLIEVKVLPGEVGFGRTSWIINGMFYAALKGADVISMSLGGGGEINDGIEN

PENFYANMITDLFGVVFAIAAGNEGPSTNTVHSPGTSDLVITVGNYVGNEREA

YWYRVDLGIMSGPSMSSSRGPRDDGMLDPDVMAPGTDIFSSLPMWYTVLYE

NPYRYYGIWSGTSMATPHVSGAVALMISYAKAQGLRYDPFMIKRALELSAKPT

NQTLIDQGFGLIQVDKAIEKLEELSHEPTTYIYGGTTFTSFKNPIEVPQIPISPAYI

EFNSYFYFMFGLPYLYRGVYIRNEYPGSVPLYFYPMDYIPGFGLWYTESEKTY

TISTNVDWIIPSTNAVVAGNNTIGEFSINIDYSRLHGSNTYVGLVYIDDPDTSYV

DGFIPVIVDYPMNPNGETHVKLSDTEKPGEARHYFVYVPRGTKELRVTLRVPA

DENGVPMGRTKLVIARPLGGVEYDGVPDYYYVGANPSGYLYNYTWVVENPV

EGTWEITAYSSTFTKYYSGYSESHYEIEVELASVSISPQLIKKDVGSPSLVDVSA

VVTNNYGTFNASAVGYGVGRLDEAYAFVSEVNQSDWDVIGLVYVDPTTYFIR

FGITQPEDPNADLDLYVYYFPTYDDYLNFENYTLYDEQIGPTSDEVFEQFMPA

PGYYLIMVYGYDTVGYNPIHYLFYYQILGDNGDVTIDSTPFTFSPGSKTINAQ

VNLSDEGTYLGVLGLVDADTGDSMVYAPMIFQVGQPEMLIVGYGEATLGQES

TLHIKVLNKTSMEPVDAPATVTVNGRPYYTTTGEVEATFIPEELGQISFNVKAS

SPVFKDAETTITVNVKEPLSSPVEKINDAKFFIAGSGRITKTVISERRFNDGYRG

AAYLITADGPSGEVGYITIAAPVDSEVYQVSGEPHLLDYTVVKGKNAVYIILK

VQYASPVTVGVTVKVKPQRMGVAFNVLNFLYYQWYQNKVKEFEELYQKAL

EAGVDESVLQEALQYNQTAAEHYQKALDIAGGNILLRINDFKLFGHLRNAYF

TELQAVEILKEALGE*

>NODE_4_length_193355_cov_97.1752_ID_7_2

MRKVLGLLVAFLMLGFVVASVAALPSPDTKPYTQPKNYGLLTPGLFRKAQRM

DWEQEVSTIIMFDTPRNQRIALKILKALGAEVKYQYEVIPAIAVKMKVRDLLVI

AGFLDATSSGRSKVQIPGIQFIQEDYKVKVAVETEGLDESAAQVMATNMWNL

GYDGSGITIGIIDTGIDASHPDLQGKVIGWVDYVNGRSSPYDDNGHGTHVASI

AAGTGAASNGKYKGMAPGAKLVGIKVLGADGSGSISDIIAGVDWAVKNKDK

YGIKVINLSLGSSQSSDGTDSLSQAVNNAWDAGLVVCVAAGNSGPNKYTVGS

PAAASKVITVGAVDKYDVITDFSSRGPTADNRLKPEVVAPGNWIIAARASGTS

MGQPINDYYTAAPGTSMATPHVAGIAALLLQAHPSWTPDKVKRALIETADIVK

PDEIADIAYGAGRVNAYKAAYYDNYAKLTFTGYVANKGSQTHQFTISGAGFVT

ATLYWDNSGSDIDLYLYDPNGNQVDYSYTAYYGFEKVGYYNPAAGTWTIKV

VSYSGSANYQVNVVSDGTLGQPGGGEPAPEPTPEPTVDEKTFTGTVHRYYDR

SDTFTMTVNSGATKITGDLTFDTGYHDLDLYLYDPNKNLVDRSESSNSYEHVE

YTNPAPGTWYFLVYAYYTYGYASYQLDVKVYYG*

>NODE_5_length_166447_cov_96.5284_ID_9_114

MRKITALLLSTVLLAALFAAPTSAGEDKVRVIVSVDSAKFNPHEVLGIGGHIVY

QFKLIPAVVVDVPANAVGKLKKLPGVEKVEFDHQATLLKKPPWAGGGGGSTQ

PAQTIPWGIERVKAPSVWSITDGSVGVIQVAVLDTGVDYDHPDLAANIAWCVS

TLRGRVSTKLRDCKDQNGHGTHVIGTIAALNNDIGVVGVAPGVQIYSIRVLDA

RGSGSYSDIAIGIEQAILGPDGVADRDGDGIIAGDPDDDAAEVISMSLGGPADD

SYLHDMIIQAYNAGIVIVAASGNEGASSPSYPAAYPEVIAVGATDSSDQVPYWS

NRQPEVSAPGVDILSTYPDDTYNTLSGTSMATPHVSGVVALIQAAYYQKYGTI

LPVGTFDDMSKNTVRGILHVTADDLGPGGWDADYGYGVVRADLAVQVALG*

>NODE_9_length_97431_cov_87.7446_ID_17_8

MGENPFGNPPTNLLPIEVPPHVLMLRGIGWDSNIYLVRDGEEALIIDTGTGINW

HVYAGMWERGGYLEGVRRVIIFNTHEHFDHVGGNNVVKRWLERHGIEVYFA

AHKITANVIERGDEGVILSYFYGRRFEPHQVDFKLEDGDKLRVGSLELLVIHTP

GHTAGSSCLYLDDGKHRVMFTGDTVFKGTVGRTDLPTGDGWALRESLERLA

GFDVDFGFPGHGGYIDDWEGNLKEVLRWLA*

>NODE_29_length_54411_cov_103.044_ID_57_5

MHKAAAIFVAMLVLFSPLVVAVGQNTTQDKDVVRVVVGVRGNAGKSFMGVA

GGKGGEYSISKEELITYKAQLRGYGKLKKEIPELGVIVLEVPRKALQKIKRLPF

VTYVEEDVKYHALGEVKWDVQYIYAPNVWNNYYKTYGYAAYGYHPSIQVA

VLDTGVDYTHPDLQGAFSWCVRVLNNGGSYYKGTDLRYCWDDNGHGTHVA

GTIGASLNGQGIAGVAPYVQMYVVKVLDSQGSGYLSDIAQGIIDATKGPDGIP

GTADDADVISMSLGGSGSTTLYNAVRYAYSYGVVLVAASGNEAASYPSYPAAY

SEVIAVGAIDSNYRIASFSNRRPDVVAPGVNVYSTLPGGTYGTMSGTSMACPH

VSGVVALMQALRLAAGKPKLTPYQVRNLLIGTAIDLGSRGYDVYYGYGLVDA

EYAVYYALNS*

>NODE_30_length_54336_cov_6.01531_ID_59_47

MISVLTTAHNVYEGGVGLTMFRVNEDVKFAGVQVNDYFFPAYRQDNGVYAC

LFAWPHNVERSAFTPRVVVEDLAGNPRTGGFYYHSIPRPPRHDRINVSQRFLD

NKMPDFQHYYPETTDPLRLFLHVNRELRAKNVARLHEFATKTADHPLWKGAF

LRLPNAAPRANFNDNRDYIFNGRKVDNQTHLGLDLASLAASPVPASNSGTVI

MAEDFGIYGQCIIIDHGLGLQSLYSHLSSIDVAVGDQVAKGQIIGRTGATGMAG

GDHLHFGIVLSGLQVNPREWLDGHWIKDNFTDKWKRAMSQP*

>NODE_32_length_51873_cov_10.9874_ID_63_30

MIYRLNGIGFDSNTYFITGKKNILVDPGTPRTFKFLKEEIEKLADRIDYIINTHC

HYDHCGSDYLFQEHFGAPVLINSLEIEHLKKGDNVTVASLFGDELIPPKEIIPVN

EVTEELNKLGIDVIETPGHTKGGITLAYEDNLITGDTLFAYGVGRYDLPTGNLA

ELRGSVERLERLAFEKRVNNILPGHGETGNLGAFANARLFL*

>NODE_46_length_42132_cov_5.92468_ID_91_32

MPAEKYSIIVIPEGHRRTRRFQVKRSWARFLAAALVVAVLTVGGLVYHCCRIHL

DRAELQRLRAERRDYRQNLRLMAGQLQALQKEMVVLANNDTKVRLMANL

AEPADKVPVGIGGPLDTDPAVELSGVQRQIDEIRHSIDLRRESLEELQGALNDQ

RSLFAAKPSIKPVKGWITSGFGLRNSPFGKGRKMHHGLDIAARTGTPVLAPAD

GVVKRVRTASDYGKMVVIDHGYGYQTLYGHNSKVFVKVGQRVKRGDRISAI

GNTGRSTGPHLHYEVRLNGVPVNPRKFF*

>NODE_52_length_37561_cov_11.2155_ID_103_28

MDNLKKLMVLLMVVLMPTAFGKSVILVSDNFADGLAAEVVQSIFNDTEIVRA

PWGEYNESIVNQIMEISPENIIIIGGPVAVPVEYEEALNESNVSIERLYGPTRYET

NKRILERFKEQLKNRKIVIVYGEDGEITVENNVTVVLSNGTNITIDEDELEELN

TSEVVIVENPMLNKNLIMNRFKNRGFNVSTKAMPQEVLKKIVEKRLDRLESK

VQRLKGLPISAQEETSIAELEQNLEEIQTLIDNGEYQEAYKLEIITERMILNKIRV

KAEAHRGKGKVIKNEIKNKIKNKVKEKLKKNQK*

>NODE_71_length_30196_cov_6.09245_ID_141_8

MSRTRLFMLVLPLVLVACSTLDERGTIAQLRNQQIEIEEEKIVGGLEKAMEGYQ

QFLRETPDTPLAPEAIRRLADLKVENEYGLITDRSAPADGAQALLSAPETARPA

EIAQTVSASAPGESEAEFERRTTTATAAAATKTATAQAADDLERTGALEAIALY

RNLLDEYPLYDRNDQVLYQMSRAYEELGRVDEAMEVMTRLVRTCPKSRYFD

EVQFRRAEYFFTRRKYLDAEDAYKSIVDIGVASSFYQLALYKLGWTFYKQELY

YNALDNFIALLDYQVKTGYDFEQTEDEPERKRVEDTFRVVSLSFSYLGGAEA

MDEFFNEEGRRLYEDRVYGNLGEYYFGKRRYADATATYSAFVSRHPFHKMAP

HFDMRVIEINMAGGFPTLVIDAKKAFATRYGLKSEYWRHFEPADYPEALGLLK

TNLHDLANHYHALYQSPEQAEAKPANFAEALHWYREYLASFPQDEKSPAINM

QLADLLLENRSFAEAAVEYEKIAYDYPRHEKSSMAGYAAVFARREHLVTVSD

DRQEPVRREVVASSLKFAETFPEHEKAAIVLGAAADDLYAMQDYTPALAAAQ

KLIETFPAAEQDIVRAAWLVVGHASYELQHYSDAEQAYGKVLTMLPASDESR

GELVDNLAAAIYKQGEEANATEDYRGAADHFLRVGRMAPTSRIRPAAEFDAA

TALIQLKDWGAAATVLAGFRNSFPGHALQPEVTKKMAFVYREDGKAAQAAA

EYERIETESDDDAIRQEALQVAAELYVEADKQAQALKVYRRYVANFPDPVAL

NLETRNKIADILKANGDRSGYLEQLRQIVALDAAGEGRTERTRYLAGNAALVL

AENTYEMFRAIKLAKPLEESMPKKKRMMKESIKRFNGLIEYESGELTAAATFY

LAEIYAHFSKALMASERPELTFDYYTIKPGDNLIRIARRSKCDISRLLNANNLK

RSSVIVAGKKLKIPRGLYPEELEEYEWALEEQAYPFEERAISVHERNLQLMARG

IYNEWIERSLRKLAEVVPARYARSEEPSPVLTSLDGYGFEIYRPLPVVAGPADGP

STDTARPGAEAPAVDAAQVVEEVRAADEEAMTAGKGQAVKEGPSTVEEPQA

GEEASSTAEVQSSAADGVTAADTAIQTTRSEVEVDVERN*

>NODE_86_length_26087_cov_5.43355_ID_171_15

MGDLSKNFSRSEFACKGTNCCGHSAPVHPELISALQALRDQLNLPLSITSGFRC

NRHNESVGGAARSFHTLGMAADVACPDGMTAEDLAQAAEIIPAFQQGGIGIY

PSWVHLDVRTTGKARWRND*

>NODE_100_length_21523_cov_5.96116_ID_199_19

MNRRQFLKFGLATAGSILLPWPAFAGLSDNKNRCLSLYNTHTGEHLRSLVYW

ETGSYLDDSLKRINHLLRDHRTGEVKAIDPNLLDLLSALHHRMPADSPFEIISG

YRSPATNRKLRKHSKGVAKSSLHMVGRAIDIRLPGCSLANLRKAAASMKGGG

VGYYPQSDFIHVDTGRVRYW*

>NODE_112_length_19392_cov_5.24843_ID_223_6

MRVAAGTYGQPSFAQKVCQDLSMTLLRRHHPAWLLVCSILLVIILGWPARADD

IDRQRRNLREIQGRIERLSQQLASSKQREGAVKDELGQLEKELADLEKNSRRL

EKRLRKVKEQIKEKERSVDLLKRDIADREQYVKRRLGALYRRGEMRLFKVLF

EKESPSRVAENYFYLSRLVREDRKLLNGYRDDWRALQSTLAELEDLRSGQQK

RLDVIDAGQKTLKKGRKTRQAVLRSLKDSRETLSKQITELKEKARRLRGLLKT

LESEKTGEYSGPSAGFKRQKGTLPWPGDGALMVRFGTNFNEQAGSRIESQGIE

LAQNPGTPMHAVAKGKVIFAKPFRGFGNLLILDHGDGYYTLYAQARQLLKKV

GDIVDAGERIGLSGFGGSDTIYFEIRQRGKPLDPLRWLKPRR*

>NODE_115_length_18945_cov_6.42656_ID_229_8

MGDLSKNFNRSEFACKGTNCCGHSAAVHPDLVDALQTLRDRIGKPLSITSGFR

CNRHNKAVGGAEQSFHTLGMAADVSCPAGVSPGELAVIAEEIPLFREGGIGVY

ASWVHLDVRQSGKARWRS*

>NODE_132_length_16776_cov_6.46411_ID_263_6

MKPKHPGLKYGLLLGGVALACFLFGLWIGGPSGNRQQAIETTVSEPETPKVPV

VDRKIIKGTIQPGDTVTSLLGDFFSPQQILSLSQSSKRVFPLTRLCAGQPYELCL

LDGHFESFIYDIDREEQLVVRQDDDGFAVSKIPIEYTVKTDMVTGVIDSSLFEA

VVASGEKEVLAISLADIFAWDVDFIRDIQVGDSFEALVEKRFREGAPAGYGRIL

AARFTNQGHIYNAYLFKDGDQPAAYYDENGRSVRKAFLKAPLSFSRISSGFN

MRRRHPITKRISPHPAIDYAAPTGTPIKTVADGTVIFAAYKRYNGNCVKIRHPG

GWMTMYNHMSRFGRGIKRGVKVRQGQLIGYVGTTGRSTGPHLDFRMYKNG

VVVNPLKVKSPPSAPVSSKHLAEFKQTVNMLAARFENQAPVQTARLDSATKP

SPSENQALAN*

>NODE_375_length_3282_cov_2.09097_ID_749_4

MKQYQDKYFKRAKKENYAARSVYKLKEMDKRFGIVKRGQTVGLTGMTGRA

TGPHLHFSLSVLGELVDPAPLFTTTADNMLQ*

>NODE_658_length_2323_cov_1.16485_ID_1315_3

MDTLSILQAQVQGQKDRIASVTSDLAASHKQMKTLFVRKAAMQKETARESS

QAARRMQKLAEQASSLKDLIAKLEAERKRQKEEARMRVVQEAQKQMAEER

AHRQKIARLKREKKIAEARAEERKEKARLAKVKAKQEAEAQKRAALAAIKPP

PRSFNKARGRLPMPAVGKVSLQYGELTQAGVHAKGISIQTRSSAQVITPFDGT

VLFSGPFRGYGQLLIIEFGGGYHILLAGMSRIDATTGQSLLAGEPVGIMTNISSP

ELYVELRHNGQPINPLPWLTAGKTSGNKRKG*

>NODE_1192_length_1682_cov_1.56913_ID_2383_2

MRKFLVLTVCCLLLLPILARAQAQDEVLSESLQKEHQKADENEQKVRALTQK

AGQISTRLADIEDDVKLLKRRIRDQETGLADIRKDERQAEQDHFALEKEKERI

ALELSGLMRTLWPVHLQNVRSRFEGVEDWAMFDRRFNWLADIYSATSRKLD

EARANAEKIALNLENQRQLAEEAEKQLAQVNQSKDRLLRNQYALRRNLKKIN

KQKENAEEELTEILATIEDLKYQLQSQKTKRFALYKRALPWPVRGQVVAGFNL
KAKPPARGLAIGAPDGSTVQSIFWGKVVHNDTLRGFGHVVIIYHGYNYYSLY
AYLSDTFVRNGQEVEKNEPLGTVGYFPKLDGTGLYFELRFHQKPINPESWLTA
LR*

>NODE_1620_length_1435_cov_2.84327_ID_3239_2
SWRGIEVPLDFRKEENGYLALVMLGSQADEEPITDSAMEVRLNIRGKTKVLK
KQVTLKAVTYPVQHLSVPKKMVTPPEEVLDRIRRESAMVEKALAVSKPGRLW
EIPFVPPLEGAVTSPYGTTRTMNGVKKGIHTGVDLRAAVGTPVVAPAGGIVVL
ADDLYFAGKCVYIDHGNGVFSVMMHLSEIAVRVGDTVRPGDLVGKTGRSGR
VTGPHLHFGVRVHGKWVNPLPLVAADQ*

>NODE_1821_length_1342_cov_1.27737_ID_3641_3
MRKTVPKEGYLHSKEEDMYLSLFDIELSKIMAKSGGIGLSELIFEQLKERVQH
LNSPAEVVDKEAQNIDEQVDRLARKIVEQKQSPKGGIVAFPQSLAPDKRPGEN
FVWPVKGEISSSFGWRIDPFTGQKAWHSGIDIAVPKNTEVKACWSGKVIFAGE
KAGYGKSVILKHKNGLISLYAHNSKLLVEEGQFVRFGQKIALSGSSGRSTGPH
LHFELRKGELALDPLAWQNKEDNSLT

>NODE_1864_length_1327_cov_2.27333_ID_3727_2
MPDLGLPLDGEFYPTQKMAPAKSLVVFLHGYGANGLDLINIASYLDKLLPDT
AFYAPDGPEECDLTPMGRQWFSLASVDPHQMRTDPRSLPLAMESLHHGVCQ
AAPILNDFLDAVLDRHNLTADKLALFCFSQGTMMGLHVALTRPDTVAGVFGFS
GLFTGGNKRLPEKPAGDMPPVWLVHGAADDVVPPQAMRMSQDALAEYGIK
AKTHMRPTLGHAIDDDGLFIARDGLLEVLR*

>NODE_1875_length_1322_cov_1.159_ID_3749_2

MEFIEAHGLDLEWVLLTHGHADHICGLEKLRPLARSGVAVHRLDAPMLASPE

QNLSAFMQDRCRSKPPERELEDGDMIHAGGLVIAVIHTPGHTPGSVCFHVKED

GEEALLSGDTLFAQSVGRTDLPGGDQGKLLRSLEKIASFPDGMAVYPGHGPET

TIGAERRKNPFWPGEQQ*

>NODE_2506_length_1136_cov_2.85134_ID_5011_1

MRINDIRVAPHFMLREFECRCCNCVRLHPELVKKLEAIRQYLDMPIIVTSGYRC

EKHNKEVGGVQRSLHLVGQAADVAIPASGQKNFIAIAKK

>NODE_3692_length_902_cov_3.17161_ID_7383_2

ASLTTRFAHLNRIAIKSGQEVTRGELIGYVGNTGRSTGPHLHYEVRLNGVPVN

PKRYILN*

>NODE_1_length_279442_cov_30.9628_ID_1_159

MNRKILGLLIAVVMLLSVVPAGLSIGAVSASPAAPSITTGGSSSQETPKFIPGDV

VEKQIERILRNKHGGKVRLIIAPEKDRAMEVYEALKNLGRIDPISRPEYQFIVV

EMPAGNLEKLKEIPGILRVWEDRTVKLLEPVEPEEALKTLKENSPAKPDMFMS

VFEINAVNVWNNYSILGDDVVVAVLDTGIDVSHPFLQTTLDGRPKIIDIYDASD

EGIAQLYYATNTTLNGTIVVNMEVPVYWGAYYPYYGHDKITTYYNMTSYFV

GNITGDEYYLGLLPERYFDLNNFFGTPNDPYNLGLFGDLSDVYPVLIVENNGT

YTAYIDYNLNNNFTDDQPIGLFTETGDYFQTPDTLVSIALAKVHIGDMDNPDN

YPYIVPYGDGIGYAMFMWDPHGHGTHVSGTVAGVGQPDDPMFRGVYGVAP

NAQLIEVKVLPGEVGFGRTSWIINGMFYAALKGADVISMSLGGGGEINDGIEN

PENFYANMITDLFGVVFAIAAGNEGPSTNTVHSPGTSDLVITVGNYVGNEREA

YWYRVDLGIMSGPSMSSSRGPRDDGMLDPDVMAPGTDIFSSLPMWYTVLYE

NPYRYYGIWSGTSMATPHVSGAVALMISYAKAQGLRYDPFMIKRALELSAKPT

NQTLIDQGFGLIQVDKAIEKLEELSHEPTTYIYGGTTFTSFKNPIEVPQIPISPAYI

EFNSYFYFMFGLPYLYRGVYIRNEYPGSVPLYFYPMDYIPGFGLWYTESEKTY

TISTNVDWIIPSTNAVVAGNNTIGEFSINIDYSRLHGSNTYVGLVYIDDPDTSYV

DGFIPVIVDYPMNPNGETHVKLSDTEKPGEARHYFVYVPRGTKELRVTLRVPA

DENGVPMGRTKLVIARPLGGVEYDGVPDYYYVGANPSGYLYNYTWVVENPV

EGTWEITAYSSTFTKYYSGYSESHYEIEVELASVSISPQLIKKDVGSPSLVDVSA

VVTNNYGTFNASAVGYGVGRLDEAYAFVSEVNQSDWDVIGLVYVDPTTYFIR

FGITQPEDPNADLDLYVYYFPTYDDYLNFENYTLYDEQIGPTSDEVFEQFMPA

PGYYLIMVYGYDTVGYNPIHYLFYYQILGDNGDVTIDSTPFTFSPGSKTINAQ

VNLSDEGTYLGVLGLVDADTGDSMVYAPMIFQVGQPEMLIVGYGEATLGQES

TLHIKVLNKTSMEPVDAPATVTVNGRPYYTTTGEVEATFIPEELGQISFNVKAS

SPVFKDAETTITVNVKEPLSSPVEKINDAKFFIAGSGRITKTVISERRFNDGYRG

AAYLITADGPSGEVGYITIAAPVDSEVYQVSGEPHLLDYTVVKGKNAVYIILK

VQYASPVTVGVTVKVKPQRMGVAFNVLNFLYYQWYQNKVKEFEELYQKAL

EAGVDESVLQEALQYNQTAAEHYQKALDIAGGNILLRINDFKLFGHLRNAYF

TELQAVEILKEALGE*

>NODE_4_length_144450_cov_36.9995_ID_7_96

MHKAAAIFVAMLVLFSPLVVAVGQNTTQDKDVVRVVVGVRGNAGKSFMGVA

GGKGGEYSISKEELITYKAQLRGYGKLKKEIPELGVIVLEVPRKALQKIKRLPF

VTYVEEDVKYHALGEVKWDVQYIYAPNVWNNYYKTYGYAAYGYHPSIQVA

VLDTGVDYTHPDLQGAFSWCVRVLNNGGSYYKGTDLRYCWDDNGHGTHVA

GTIGASLNGQGIAGVAPYVQMYVVKVLDSQGSGYLSDIAQGIIDATKGPDGIP

GTADDADVISMSLGGSGSTTLYNAVRYAYSYGVVLVAASGNEAASYPSYPAAY

SEVIAVGAIDSNYRIASFSNRRPDVVAPGVNVYSTLPGGTYGTMSGTSMACPH

VSGVVALMQALRLAAGKPKLTPYQVRNLLIGTAIDLGSRGYDVYYGYGLVDA

EYAVYYALNS*

>NODE_6_length_120078_cov_19.8474_ID_11_17

MRKITALLLSTVLLAALFAAPTSAGEDKVRVIVSVDSAKFNPHEVLGIGGHIVY

QFKLIPAVVVDVPANAVGKLKKLPGVEKVEFDHQATLLKKPPWAGGGGGSTQ

PAQTIPWGIERVKAPSVWSITDGSVGVIQVAVLDTGVDYDHPDLAANIAWCVS

TLRGRVSTKLRDCKDQNGHGTHVIGTIAALNNDIGVVGVAPGVQIYSIRVLDA

RGSGSYSDIAIGIEQAILGPDGVADRDGDGIIAGDPDDDAAEVISMSLGGPADD

SYLHDMIIQAYNAGIVIVAASGNEGASSPSYPAAYPEVIAVGATDSSDQVPYWS

NRQPEVSAPGVDILSTYPDDTYNTLSGTSMATPHVSGVVALIQAAYYQKYGTI

LPVGTFDDMSKNTVRGILHVTADDLGPGGWDADYGYGVVRADLAVQVALG*

>NODE_11_length_97240_cov_21.356_ID_21_8

MGENPFGNPPTNLLPIEVPPHVLMLRGIGWDSNIYLVRDGEEALIIDTGTGINW

HVYAGMWERGGYLEGVRRVIIFNTHEHFDHVGGNNVVKRWLERHGIEVYFA

AHKITANVIERGDEGVILSYFYGRRFEPHQVDFKLEDGDKLRVGSLELLVIHTP

GHTAGSSCLYLDDGKHRVMFTGDTVFKGTVGRTDLPTGDGWALRESLERLA

GFDVDFGFPGHGGYIDDWEGNLKEVLRWLA*

>NODE_14_length_72106_cov_0.000138791_ID_27_10

MISVLTTAHNVYEGGVGLTMFRVNEDVKFAGVQVNDYFFPAYRQDNGVYAC

LFAWPHNVERSAFTPRVVVEDLAGNPRTGGFYYHSIPRPPRHDRINVSQRFLD

NKMPDFQHYYPETTDPLRLFLHVNRELRAKNVARLHEFATKTADHPLWKGAF

LRLPNAAPRANFNDNRDYIFNGRKVDNQTHLGLDLASLAASPVPASNSGTVI

MAEDFGIYGQCIIIDHGLGLQSLYSHLSSIDVAVGDQVAKGQIIGRTGATGMAG

GDHLHFGIVLSGLQVNPREWLDGHWIKDNFTDKWKRAMSQP*

>NODE_25_length_42044_cov_0.501822_ID_49_32

MDNLKKLMVLLMVVLMPTAFGKSVILVSDNFADGLAAEVVQSIFNDTEIVRA

PWGEYNESIVNQIMEISPENIIIIGGPVAVPVEYEEALNESNVSIERLYGPTRYET

NKRILERFKEQLKNRKIVIVYGEDGEITVENNVTVVLSNGTNITIDEDELEELN

TSEVVIVENPMLNKNLIMNRFKNRGFNVSTKAMPQEVLKKIVEKRLDRLESK

VQRLKGLPISAQEETSIAELEQNLEEIQTLIDNGEYQEAYKLEIITERMILNKIRV

KAEAHRGKGKVIKNEIKNKIKNKVKEKLKKNQK*

>NODE_33_length_34124_cov_0.140039_ID_65_16

MIYRLNGIGFDSNTYFITGKKNILVDPGTPRTFKFLKEEIEKLADRIDYIINTHC

HYDHCGSDYLFQEHFGAPVLINSLEIEHLKKGDNVTVASLFGDELIPPKEIIPVN

EVTEELNKLGIDVIETPGHTKGGITLAYEDNLITGDTLFAYGVGRYDLPTGNLA

ELRGSVERLERLAFEKRVNNILPGHGETGNLGAFANARLFL*

>NODE_61_length_16356_cov_8.12189_ID_121_20

MRKVLGLLVAFLMLGFVVASVAALPSPDTKPYTQPKNYGLLTPGLFRKAQRM

DWEQEVSTIIMFDTPRNQRIALKILKALGAEVKYQYEVIPAIAVKMKVRDLLVI

AGFLDATSSGRSKVQIPGIQFIQEDYKVKVAVETEGLDESAAQVMATNMWNL

GYDGSGITIGIIDTGIDASHPDLQGKVIGWVDYVNGRSSPYDDNGHGTHVASI

AAGTGAASNGKYKGMAPGAKLVGIKVLGADGSGSISDIIAGVDWAVKNKDK

YGIKVINLSLGSSQSSDGTDSLSQAVNNAWDAGLVVCVAAGNSGPNKYTVGS

PAAASKVITVGAVDKYDVITDFSSRGPTADNRLKPEVVAPGNWIIAARASGTS

MGQPINDYYTAAPGTSMATPHVAGIAALLLQAHPSWTPDKVKRALIETADIVK

PDEIADIAYGAGRVNAYKAAYYDNYAKLTFTGYVANKGSQTHQFTISGAGFVT

ATLYWDNSGSDIDLYLYDPNGNQVDYSYTAYYGFEKVGYYNPAAGTWTIKV

VSYSGSANYQVNVVSDGTLGQPGGGEPAPEPTPEPTVDEKTFTGTVHRYYDR

SDTFTMTVNSGATKITGDLTFDTGYHDLDLYLYDPNKNLVDRSESSNSYEHVE

YTNPAPGTWYFLVYAYYTYGYASYQLDVKVYYG*

>NODE_118_length_3398_cov_0.000299133_ID_235_3

MFCEAKLPGQGLLEGAGQVRNVRVQDLAVGQGGAQGGRLRDIADPGAVLSG

EQGLEPFAVGRGYRNHQPAGGFGIEQAGQVELAAGGHLVQIHFRAEPSAEQLE

QIRRDIGASAENRMRGVYVFRSRSMDAMKMMGYFRKKWNPIYVEPHYLYLT

NEAPADVRPIVPNDRLYSDYQWNLPSIETEIGWNLSKGSDGVKVAVLDTGVQL

DHPDLEGKLAEGYNVVTSGQ

>NODE_303_length_437_cov_0.992147_ID_605_1

GKGVAGVNWGGYGIKIMPIRVLGADGSGTLDAVAQGVRYAADHGAKIINMSL

GGGGNSQILRDAIQYADNKGVTIVCAAGNENGPVSYPAKYPETIAVAAVRYDL

QRAPYSNYGPEVDVAAPGGDTSVDQNGDGYADGVLSTAWT

>NODE_386_length_268_cov_2.723_ID_771_1

PSNGDTYMFLQGTSMAAPHVAGVAALLYASGKTQPEEIRAALKNTAKDLGPT

GEDDYYGAGLIDAYAAINYNGGSTNPPNPEPQPSGE

3.7.5 All Xylanase ORFs

>NODE_12_length_82694_cov_97.5998_ID_23_85

MRVSLTFDVEQDCPPYLTTTRGVEGGLPRLLDLMAEKKVRATFFFTAEMARR

FPQLVRRVLDEGHELGSHGYNHERLDRLPKDDAAKVIEKSLNVLREFGEVVS

FRAPNLQLPETYYDILERHGVLVDSSKATYKGYRLGIRFFGEVLEVPASTTSSV

LRLPWKLQAVIHSRLKEPRIYFAHPWEFVPMQREKIRWDCRFNTGEMALELL

GKLIDHYKSMGAEFLTMREYYDLYGNLKRE*

>NODE_49_length_41468_cov_6.1669_ID_97_12

MRIAGILLLLLMCVARPVRGDEFISLCFHEVRADIGRGDDLSMSTDRFVALLT

WLRQHDYRPVGIDDLLRAREGSKPLPEKAVLLSFDDGYRSFYNQVYPLLKAY

RYPVLLAVVGSWLDAPPGATVDYGGKRVPREKFLSWEQLREMTESGLVEIAS

HSYNGHRGIDANPQGNRQPALTARAYAAATRTYEDDIAYAERINADLQANADL

IEQKLGIRPRVMVWPFGKYSLPAIEAARQAGMPVTLGLGDGPGDTDHLTAVK

RLLIEGRLDLPGLSWRIRHLMANDPQRVVQVDLDYVYDPDPQQVERNLGRLL

DRIKAMQITTVYLQAFADPDGDGVADALYFPNAWLPVRADLFNRVAWQLKTR

AGVKVYAWLPVAWLRCSAAGLVGSKL*

>NODE_179_length_10741_cov_5.34257_ID_357_8

MRILIITPGQPRTTGNWVSANRQRKGLQSLGHTVRIVETTETSGNLERITEDFV

PDVVNLLHAFRSGTFWLASRYAKTIPMVVTLTGTDINHGVDNPAQKPIIQQIM

TEAGAIITIRATTKETLSRDFPEHSSKLHHIAPAVDFGQRPFDLRRKCHIPRHVV

LFLHPAGIRPVKGNLELLEMFEKVVREQPCRLLFCGPLLDKEYGQRFLNAVQK

RTWADYAGQIPADAMPAAMRAADVILNNSISEGFANALQEAACLGIPILARDIP

GNISAFEPGHGGLLYDSPEHFIRQAVLLAKHPEMRRRLSRPTSPAHTLMEEAR

QMERIYHSLISPL*

>NODE_187_length_9821_cov_6.14298_ID_373_9

MTIPPPPPRNEPNLFWLLAISLALHVAVFLVFSGVFFNGRHIERRPVYYVDLSK

MPVLNPQAGRPDGGPAPKAKKTAKPKAAVPAPKPAAPPAKTKKPKTVPAAPK

PTKPKPTPSKTATKAAKPKPQPAQSPASTGQNYQSVQEKLAAMREKQQRQQE

LAALKNKIAALSGDAGTDSGSHGSGAPLGMPDGKGDEIGVDQQTWLRAFYK

ENWSLSKYQVTRLDLEATVSITYNSEGHLINYRFVKSSGDSTFDDSLKRAILK

DRILPFKPGRTLQLDVVFNLKDLMD*

>NODE_810_length_2073_cov_1.49846_ID_1619_1

MLDAGLKPEADNFSRFVLTDEGVAFFFGPYQVAPYAAGEQIVTIPYGNLGGFL

APDIAAGVGSE*

>NODE_1772_length_1368_cov_1.36906_ID_3543_3

RPKFFRPPGGRWNDQVLRIVSDEGLLSVLWTVNGYDVPPRPPEELAELILRRT

RPGAIILLHDGGGATIEALPIIIEKLLFEDYLFVTLDQMFPVRVTPIPLVAPTDSK

R*

>NODE_2372_length_1173_cov_1.63862_ID_4743_1

QGRDPTLFAYPYGEANADTLALAKSHYAAAFGQHSGTMYAGDNAYYFPRYA

LNEHYGDAKRFRLVINTKPLRVFDIVPKDPTLRHNPPSFGFTLANEDTSLESLK

CYSSASGPVQTENLGNRIEVRLSHPFAGARGRINCTTQDHKGQWQWFGQLYY

IPKELRKTSAAGN*

>NODE_2676_length_1090_cov_1.80582_ID_5351_1

MLSVVGAFVFGVELLKTEAGRNAPLPLEEINKRYGNLTPKFWGQDVPGVVTR

FDPSGLQVALTLDACGGKGGFGYDKEVVDFLVENNIKATLFLNARWIRDNPD

EAKTLAENPLFLIANHGFRHKSCSVNGKTAYGIKGTSSVEEVWEEVEKGARA

VEALTGKRPRFYRSGTNYYDEIAVKIVYDLGMKPVGYSILGDGGATFSSERVM

EAVIKARPGDIIICHMNHPEGETAEGLKEAVPILLERGYTFIRLDEAIN*

>NODE_4507_length_800_cov_1.12481_ID_9013_1

HLFFGTYLAYSRPCWMRWLQKLALVAPFNASLRNIFGKLLLRPKMWNKSVNI

QSIMVIQPCDLLENGRQDMCDGCPDSIVHDGKMVWSCRVDELEKFGAFIQCA

PRGCCGGKPAPAPAPEAAAEPAAKPEPAQEAKPEPVPEPAPEPKPEPAPEPTPEP

KPEAVPAPRSETAPAAREMHQPAVKIDVPPAVPETKAVAVEPQARTEEQPAAAA

APEAAPEVQAETVAKAPEAQPTPEPETKPATEAETADKTAEPQPEPAPEPAPK

>NODE_13_length_82572_cov_45.6275_ID_25_86

MRVSLTFDVEQDCPPYLTTTRGVEGGLPRLLDLMAEKKVRATFFFTAEMARR

FPQLVRRVLDEGHELGSHGYNHERLDRLPKDDAAKVIEKSLNVLREFGEVVS

FRAPNLQLPETYYDILERHGVLVDSSKATYKGYRLGIRFFGEVLEVPASTTSSV

LRLPWKLQAVIHSRLKEPRIYFAHPWEFVPMQREKIRWDCRFNTGEMALELL

GKLIDHYKSMGAEFLTMREYYDLYGNLKRE*

>NODE_17_length_63320_cov_4.74196e-05_ID_33_52

MRIAGILLLLLMCVARPVRGDEFISLCFHEVRADIGRGDDLSMSTDRFVALLT

WLRQHDYRPVGIDDLLRAREGSKPLPEKAVLLSFDDGYRSFYNQVYPLLKAY

RYPVLLAVVGSWLDAPPGATVDYGGKRVPREKFLSWEQLREMTESGLVEIAS

HSYNGHRGIDANPQGNRQPALTARAYAAATRTYEDDIAYAERINADLQANADL

IEQKLGIRPRVMVWPFGKYSLPAIEAARQAGMPVTLGLGDGPGDTDHLTAVK

RLLIEGRLDLPGLSWRIRHLMANDPQRVVQVDLDYVYDPDPQQVERNLGRLL

DRIKAMQITTVYLQAFADPDGDGVADALYFPNAWLPVRADLFNRVAWQLKTR

AGVKVYAWLPVAWLRCSAAGLVGSKL*

>NODE_47_length_25731_cov_95.7681_ID_93_8

MKRLRVGVDFHEWDGIFQGSRNHVLGIYRHAINQAPDIDFFFFLESTESLREA

HEEFRRSNVQLVRMPRRNGLIRLGLQLPWLRFRLGIDVLHAQYRLPFIKTGRS

VCTIHDILFETHPEFFPSGFVKEARLTYRLAVRQADLIFTVSEFSKQEITRIYQVP

LTKVQVTYNGVDRAKFFPGDEGRERVQGLGLVPGQYILIVGRLEPRKNHLALI

NAWAQLGASAPPLVIVGQEDPNFPDVREAIDAMADTHKVIRFKQMGDDVLPD

VMRHAAVFVYPAFAEGFGMPVAEAMACGVPVITSNSTSLREVAADGAVLFEP

GDQQGLYMALKATLAMPVLARQALIACALKRVAHFDWNQSAAVLLAGLRG

VSSAVTSERVVHAKPHKID*

>NODE_81_length_7441_cov_8.51435_ID_161_2

MYVVKIPWWLRLLYPSLIWEMPVTAEKKIYLTFDDGPHHEATPFVLDQLKNY

DAKATFFCIGKNVRKHPEIYQRIIAEGHTIGNHTNNHLNGWKVKDAAYIANIQ

EAGNVIASDLFRPPYGRIKRAVIRRLQSKSGSPGVRESENGLRSMVNGLSSPVS

GLRSPVLNLPSKIVMWTVLAGDFDIALSKEKCLHNVVKHARNGSIVVFHDST

KAWERMSYALPKVLAYFTEQGYAFEKL*

>NODE_198_length_1146_cov_58.0752_ID_395_2

LTTTRGVEGGLPRLLDLMAEKKVRATFFFTAEMARRFPQLVRRVLDEGHELGS

HGYNHERLDRLPKDDAAKVIEKSLNVLREFGEVVSFRAPNLQLPETYYDILER

HGVLVDSSKATYKGYRLGIRFFGEVLEVPASTTSSVLRLPWKLQAVIHSRLKEP

RIYFAHPWEFVPMQREKIRWDCRFNTGEMALELLGKLIDHYKSMGAEFLTMR

EYYDLYGNLKRE*

Chapter 4. Recovery and expression of intact secondary metabolite biosynthetic pathways from a large-insert soil metagenomic library

4.1 Abstract

Soil microorganisms express diverse bioactive natural products; however, the majority of soil microbes are recalcitrant to cultivation. We used a metagenomic approach to bypass cultivation and directly capture DNA from diverse microbial genomes from an agricultural soil. The metagenomic library was constructed in a broad host-range bacterial artificial chromosome (BAC) vector and contained 19,200 clones with an average insert size of 110kb. Identification of secondary metabolite biosynthesis clusters was conducted using multiple methods, including PCR, DNA hybridization and next-generation sequencing. In the latter case a pooling strategy was used to multiplex clones for sequencing that enabled identification of individual pathway-containing clones. Contigs were assembled for each pool and screened for secondary metabolite gene clusters using antiSMASH3.0, resulting in identification of 358 clones that contain a polyketide synthase (PKS) and/or a non-ribosomal peptide synthetase (NRPS) pathway, among 1,910 total pathways identified. The cloned pathways are very divergent from known pathways, with the %G+C content varying from 41 to 76% and the nearest BLAST hit of keto-synthase domains ranging from 32 to 83% amino acid identity. Biosynthetic clusters identified via PCR were a subset of the clones identified via sequencing, which were both numerically more abundant and

represent novel pathways that have no currently characterized product. Introduction of the PKS pathway-containing clones into *E. coli* strain BTRA engineered for polyketide expression resulted in the identification of multiple clones that heterologously expressed the cloned PKS pathway, and in some cases produce a compound(s) with anti-bacterial activity. These results indicate that highly novel biosynthetic clusters can be cloned intact from complex metagenomes and heterologously expressed to produce secondary metabolites, thereby expanding our available resources for natural product discovery.

4.2 Introduction

There is a tremendous degree of bacterial diversity in soils and natural environments described at a phylogenetic level. A single gram of dry soil contains more than one billion microbial cells that have diverse functions important in ecological processes and can play significant roles in interactions with other microbes and with plant or animal hosts [135,172]. These microorganisms also harbor a great diversity of secondary metabolite biosynthetic pathways, which contribute to generation of microbial-derived natural products with distinct functions. The vast majority of clinically relevant secondary metabolites (SMs) are derived from soil microorganisms, but only a relatively small percentage of the extant diversity of these biosynthetic clusters has already been exploited for their clinical utility due to the inability to culture the majority of soil microorganisms in the laboratory. Despite recent advances in

methods to culture previously uncharacterized environmental microorganisms [130], and identify their bioactive natural products, there remains a significant gap between the existing phylogenetic and biochemical diversity present within environmental microbiomes and our knowledge derived from cultured isolates. Continuing efforts to expand our culture-dependent access to microbial natural products are highly useful, and will be complemented by the identification and expression of biosynthetic pathways using culture-independent methods. Therefore, the microbial metagenome in natural environments, and in particular soils, represents an as-yet underexploited reservoir for discovering novel microbial natural products.

Next-generation sequencing (NGS) technologies can be used to more thoroughly sample the encoded metagenomes, but even deep sequencing of complex microbiomes with current NGS methods fails to assemble complete biosynthetic clusters [173]. In order to circumvent this limitation, construction of metagenomic clone libraries allows for the cloning of larger contiguous genomic fragments, some of which will contain full-length biosynthetic pathways, from the entire microbiome. The use of cosmid or fosmid libraries, while more facile in their construction and having been used successfully to clone and express many biosynthetic clusters [174,175,176], are inherently limited in their insert size (~40kb) and preclude the cloning of larger full-length pathways (estimated 80% or more missing chance of 30 kb or larger) in a single clone. This research effort adopted unbiased BAC cloning as a means of cloning the greatest diversity of intact biosynthetic clusters.

Polyketides are a group of secondary metabolites that have been widely used in

medicine and have application as antibiotics, antifungals and other bioactivity compounds. To date, a large number of polyketides have been successfully commercialized such as erythromycin, rapamycin, spiramycin and tylosin [92], driving enormous interest in the discovery and clinical development of novel polyketides. Polyketide synthases (PKSs) are responsible for synthesizing polyketides through assembling acyl-coenzyme A and building blocks of amino acids [92]. They are derived from both eukaryotic and prokaryotic systems, and can be further divided into Types I, II, III and IV [177]. The structure of the Type I PKSs are modular, assembly line enzymes that resemble fatty acid synthases and contain multiple domains [178]. The enzymatic domains such as β-ketoacyl synthase (KS), acyltransferase and acyl carrier protein (ACP) usually form a single module, and a complete Type I PKS typically contains several modules with distinct and non-iterative functions. Each module can independently catalyze one cycle of chain extension [179]. With different module organization, distinct polyketide structures with diverse biological functions are capable of being synthetized. Therefore, a complete biosynthetic pathway usually requires a large genetic region greater than 40 kb to encode multifunctional modules. While individual modules of Type I PKS pathways have been previously cloned, shotgun sequenced or PCR amplified from metagenomic sources [180,181,182,183], no Type I PKS pathway has been previously reported to have been isolated intact from a metagenomic library.

There are multiple methods that can be used to identify novel biosynthetic pathways from metagenomes, and each method will have its own inherent biases.

Function-based screening of metagenomic libraries can yield novel gene products that would have been unlikely to have been recognized solely by sequence analysis [184,185,186,187,188]. However, many genes derived from metagenomic sources will not be heterologously transcribed, translated and/or their protein product(s) may not be post-translationally modified as in their source organism [189,190]. Sequence-based methods avoid the constraints of heterologous expression, yet have their own set of biases. For example, degenerate PCR primers have been used successfully to amplify diverse KS domains from environmental metagenomes [191,192] yet each PCR primer set has its own degree of specificity and will only amplify a subset of the targeted domains present in a metagenome. Likewise, hybridization-based approaches are dependent upon the sequences used as a probe; therefore, while SM pathways obtained by hybridization may originate from previously uncharacterized soil bacteria [180], this approach is also not inclusive of metagenomic SM pathway diversity.

There are multiple bioinformatics approaches that have been developed in order to identify biosynthetic clusters from microbial genomes, such as antiSMASH3.0 [193] and IMG-ABC [194], that permit identification of SM clusters using homology-based approaches and use the cluster architecture to improve the statistical probability of identifying SM cluster hits. While these bioinformatic approaches will also be biased in favor of previously described genetic loci due to their dependence upon recognition of sequence motifs present in GenBank or other databases, it is more likely that highly divergent genes or SM pathways may be identified using homology searches across the entire genetic coding region.

In this study we constructed a soil metagenomic library that contained insert sizes sufficiently large to contain full-length SM biosynthetic clusters using a broad-host BAC vector. Furthermore, we mined this library for the presence of SM clusters using multiple sequence-based approaches, including an NGS multiplex pooling strategy wherein each clone was represented within 3 distinct pools (column, row and plate) so that clones containing complete or nearly complete SM clusters could be precisely located within the library. The SM clusters obtained from this study represent completely novel clades of PKS, NRPS and other SM biosynthetic clusters, and a subset of these clusters could be expressed in a heterologous host to produce bioactive compounds.

4.3 Methods

4.3.1 Soil collection and DNA isolation

Bulk soils were sampled from 10-30 cm below the soil surface, in a plot that had not received fertilizer additions for at least 100 years in the Cullars Rotation (Auburn, AL), and was at the time planted with a soybean crop. The soil samples were transported to the laboratory and sub-samples were immediately frozen at -80°C, while the majority of the soil samples were maintained at 4^{o}C until processed within one week for DNA isolation. A 10 g soil sample was processed for HMW DNA isolation and purification as previously described [46,47], and then randomly sheared, resulting in fragmented DNA

with average size of > 100 kb.

4.3.2 Large-insert BAC library construction

The sheared soil genomic DNA (>100 kb) was end-repaired with the DNA terminator kit (Lucigen, Middleton WI) in a total volume of 500 µl with 10 µl of enzymes and then heat killed at 70°C, for 15 min. The end-repaired DNA was ligated with BstXI adaptors (10µl of 100 µM each) in a 700 µl ligation reaction containing 10 µl ligase (2U/µl, Epicenter), followed by gel-fractionation of large DNA fragments ranging from 100 to 200 kb purified by pulse-field gel electrophoresis. Purified large DNA fragments (about 100 µl, 1~3ng/µl) were ligated into the cloning-ready BstXI shuttle vector pSMART-BAC-S (16°C for ~18 hours). The ligated DNA mixture was electroporated into competent *E. coli* cells (BAC-Optimized *E. coli* 10G Replicator Cells, Lucigen). Small-scale ligations and transformations (1 µl DNA per 20 µl cells) were used to judge the cloning efficiency. The insert sizes of 50 BAC clones were determined to find conditions that contained an average size of 100 kb or larger. Once the suitability of the trial ligation reactions was confirmed, large-scale ligations and transformations were conducted to achieve 19,200 clones for the unbiased BAC library (50 X 384-well plates arrayed).

4.3.3 Screening libraries via hybridization

Based on the KS domain sequences amplified from the pooled library DNA using the universal primer set, 4UU and 5LL , we designed 7 overgo oligo pairs (PKSc1~4, 9, 11, 20), labeled the overgo oligoes using Klenow DNA polymerase with ^{32}P to generate labeled probes. We also labeled target DNAs including PKS clones and pooled PKS PCR products by PCR with ^{32}P to generate labeled probes. The first 48 plates (18,432 clones) of this soil shuttle BAC library was gridded onto 22x22cm positively charged nylon filters for hybridization screening purposes. Each filter contains 18,432 independent clones that have been spotted in duplicate in a 4x4 clone array. Filters were hybridized with 32P-labeled probes using standard conditions. After washing to remove unbound probe, filters are wrapped in plastic film and exposed to x-ray film (Kodak Xar, Amersham Hyperfilm) for 2 to 24 hrs. The positive clones were identified and further confirmed by PCR with specific PKS primer sets.

4.3.4 PCR amplification and sequencing of 16S rRNA genes from the library and soil

16S rRNA genes were PCR amplified using pooled library DNA template with the "universal Bacteria"-specific primer set 27F and 1492R [27], using the following conditions: 94°C for 2 min, followed by 30 cycles of 94°C for 30 seconds, 55°C for 15 seconds and 72°C for 1 minute, after which a final elongation step at 72°C for 5 minutes was performed. PCR products were visualized through gel electrophoresis, and the 16S rRNA gene amplicons were purified using the EZNA Cycle Pure kit (Omega Biotek,

Norcross, GA) and cloned in *E. coli* using the TOPO-TA cloning kit (Invitrogen, Carlsbad, CA). Transformants were picked into a total number of eight 96-well plates and Sanger sequencing reactions were conducted with primer 27F (Lucigen Corp., Middleton, WI). Sequences were trimmed using the CLC Genomics Workbench (CLC bio, Cambridge, MA) and a BLASTn was conducted against the GenBank nr/nt database. All 16S rRNA gene sequences with a top affiliation with *E. coli* were presumed to have been derived from host genomic DNA that contaminated the BAC DNA isolation and were eliminated from the analysis. A total of 318 non-*E. coli* 16S rRNA genes were recovered from the cloned amplicons and were used for phylogenetic inference.

To compare the ribotype diversity derived from the library with the original soil sample, we also conducted a 16S rRNA gene survey of the same soil sample (stored at -80^{o}C) that was used for library construction. Metagenomic DNA was isolated from 0.25 g of the soil sample using an EZNA Soil DNA kit (Omega Biotek, Nocross, GA) and the gDNA was used as a template for bar-coded 16S rRNA gene sequencing targeting the V4 variable region with PCR primers 515F and 806R [195] and were used in a single-step 30 cycle PCR using the HotStarTaq Plus Master Mix Kit (Qiagen, USA) under the following conditions: 94°C for 3 minutes, followed by 28 cycles (5 cycle used on PCR products) of 94°C for 30 seconds, 53°C for 40 seconds and 72°C for 1 minute, after which a final elongation step at 72°C for 5 minutes was performed. Sequencing was performed at Molecular Research (Shallowater, TX) on an Ion Torrent PGM following the manufacturer's guidelines. Sequence data were processed by removing

barcode and primer sequences, then sequences were removed that were ambiguous, less than 150 bp, or had homopolymer runs exceeding 6 bp. Sequences were denoised, OTUs generated and chimeras removed. Operational taxonomic units (OTUs) were defined by clustering at 3% divergence (97% similarity). Final OTUs were taxonomically classified using BLASTn against a curated database derived from GreenGenes, RDPII and NCBI [196].

4.3.5 Screening libraries via PCR

All 19,200 BAC clones were screened for the presence of putative polyketide synthase genes using degenerate PCR primers 5LL and 4UU (Table 3). Each 25 μl PCR reaction contained 10 pmol of 5LL and 4UU, 12.5 μl CloneID 1X colony PCR master mix with Taq DNA polymerase (Lucigen Corp.), and 1 μ1 overnight growth of supernatant containing *E. coli* BAC DNA. Amplification was performed for 30 rounds of thermal cycling at an annealing temperature of 60°C for 30 s, extension at 72°C for 1 min, and denaturation at 94°C for 30 s. Reactions were considered positive if an ~750 bp amplicon was visualized upon agarose gel electrophoresis.

Table 3**.** List of oligonucleotide sequences used in this study

Oligonucleotide name	Oligonucleotide sequence
PKSc1OligoA	CGAGCATGCCGTGTCGATCGCCA
PKSc1OligoB	TCAAAGGTCCGTGCATGGCGATC

PKSc2OligoA	GGGCCTCTATATATGTCACGTCT
PKSc2OligoB	CGAAGTCCATGCTCAAGACGTGA
PKSc3OligoA	CTCCAGTTGTGCATTCAGCAGCG
PKSc3OligoB	GGTTAATTCGGGAAGCGCTGCTG
PKSc4OligoA	TCGCGCAACGCGTCGGAAAGCCT
PKSc4OligoB	ATCCTCGTGCTCAAGAGGCTTTC
PKSc9OligoA	CGTATAGCCGACTTTCGCCGACC
PKSc9OligoB	CCATCAACAACGACGGGTCGGCG
PKSc11OligoA	ACATAGCCGATTTCATTGGGGTT
PKSc11OligoB	GCTCGCAACGGCATCAACCCCAA
PKSc20OligoA	TGGCCATCCGGAGAAAGGATCAT
pKSc20OligoB	CTTTTCCAGGAAGGGATGATCCT
5LL	GGRTCNCCIARYTGIGTICCIGTICCRTGIGC
4UU	MGIGARGCIYTICARATGGAYCCICARCARMG

4.3.6 Metagenomic library pooling and sequencing

For all pooling approaches, individual clones were grown in triplicate in 96-deep well plates using 1 mL LB + arabinose to amplify BAC copy number. Pools were made by combining the liquid cultures as appropriate, pelleting the cells and purifying BAC DNA as previously described [197]. For plates 41-50, the initial pooling strategy merged all 384 clones from each original library plate into a single plate pool (10 plate pools);

row clones from the 10 original library plates into single row pools (16 row pools A-P, each pool containing 240 clones); and column clones from the 10 original library plates into single column pools (24 column pools, each pool containing 160 clones. For the remainder of the library (plate no. 1-40), the 384 well plates were replicated in batches of 10 plates into 96 well quadrants. For each batch, 40 plate pools were made from each 96-clone quadrant; 8 row pools A-H were made, one from each 480-clone row (40 quadrant plates x 12 wells/row); and 12 column pools were made, one from each 320-clone column (40 quadrant plates x 8 wells/column).

Fragment libraries for sequencing on an Illumina instrument were constructed with (100 ng) purified BAC DNA from each pool using the multichannel protocol and reagents from (Lucigen, Middleton WI). Unique indexes were used for each library pool within each batch of 10 library plates. Libraries were multiplexed and sequenced on Illumina HiSeq 2500 with v3 chemistry at 2x 150bp.

4.3.7 Assembly *de novo* of metagenomic contigs

The raw HiSeq reads per each pool of columns, rows or plates were imported into the CLC Genomics Workbench (Qiagen, Cambridge, MA) for trimming at a stringency of 0.01 (equivalent to Q score of >40). The trimmed sequences from each pool were exported and assembled *de novo* separately using the SPAdes genome assembler using default settings [156], leading to a set of contigs per each pool.

4.3.8 AntiSMASH screening and deconvolution of contig hits to clones

The contigs from each plate pool were selected first for prediction of SM pathways due to a higher average sequencing coverage for the pooled clones in each plate, compared to rows or columns. The contig sequences were used to predict the presence of SM pathways using a local version of antiSMASH 3.0.4 with prodigal (meta) for gene prediction running on the Alabama Supercomputer with the LINUX operating system to afford high-throughput detection. The contigs from each pooled plate that had predicted SM clusters were then used to conduct a local BLASTn search using contigs obtained from column and row pools, respectively. Those contigs from rows and columns with ~100% identity and sufficient length to indicate a shared origin on the same BAC clone were used to locate the exact library coordinates (plate, row and column) for each SM pathway-containing clone. Finally, PCR using clone-specific primer sets (data not shown) was used to validate these identified clones (primers were designed from plate contigs using Primer 3.0).

4.3.9 Re-sequencing identified clones and annotation

Each clone identified to contain a SM pathway was individually grown and re-sequenced by standard paired-end fragment sequencing. Trimming and assembly was conducted with CLC Genomics Workbench 8.5 followed by manual inspection and reassembly, and antiSMASH 3.0.4 was applied for annotation of fully assembled clone

sequences. In addition, prediction of the ORFs contained on each of the resulting contigs was employed using GeneMark.hmm [198]. For each predicted ORF, the nucleotide sequences were compared by BLASTn and BLASTx against the GenBank nr/nt database.

4.3.10 Phylogenetic analysis

Alignment of 16S rRNA genes and KS domain sequences were generated using MUSCLE. The phylogenetic trees were reconstructed using PhyML 3.0 [199] with MEGA6 [200]. For the analysis of 16S rRNA diversity within the soil metagenomic library, a maximum parsimony analysis was conducted, with 1000 iterations conducted for bootstrap support, and results were expressed in an unrooted tree with each bacterial phylum color-coded.

4.3.11 Electroporation into *E. coli* BTRA

Each pathway-containing BAC clone identified from the pooling strategy was picked from the 384-well plate, and then BAC DNA was extracted from each of the clones. The isolated BAC DNA from each clone was transformed into *E. coli* BTRA through electroporation (Table 4). *E. coli* BTRA is a derivative of strain BL21 that contains an insertion of the *B. subtilis sfp* gene, has an inactivated *Prp operon*, an integration of the *trfA* gene, a deletion of the *recA* gene, and is optimized for post-

translational modification of PKSs/NRPSs and accumulation of precursors.

Table 4**.** Bacterial strains and plasmids used in this study

Bacterial strains	Plasmids
BAC-Optimized *E. coli* Replicator Cells *E. coli* BTRA	pSMART-BAC-S

4.3.12 Induction of pathway-containing clones for expression and extraction of potential secondary metabolites

Prior to inducing the recombinant clones for expression, the *E. coli* BTRA cultures that contained the BAC clones with SM pathways were grown in 20 ml LB containing 12.5 µg/ml Cm for 24 hours, along with an *E. coli* BTRA culture containing the empty BAC vector. An expression induction mix was added to each broth culture to achieve a final concentration of 40 mM sodium propionate, 40 mM sodium butyrate, 0.02% arabinose and 200 µM IPTG for expression of potential metabolites, followed by incubation for an additional 24 hours. After a total of 48 hr incubation, 20 ml broths of transformed BAC clones were extracted using 4 ml ethyl acetate and vortexed for 10 min. After centrifugation for 10 min at 10,000 x *g*, the ethyl acetate layer was removed and completely dried using a nitrogen evaporator. The dried culture extracts were then re-suspended overnight in 200 µl DMSO and diluted by adding 200 µl H2O, leading to a total of 400µl culture extract. This was stored at 4^{o}C until a bioassay was conducted or it was biochemically characterized.

4.3.13 Antibacterial bioassays

To test for antibacterial activity, 10 µl of a culture extract was mixed with 190 µl of a 1:500 dilution of the bacterial tester strain broth culture (*Pseudomonas aeruginosa* strain PA01, *Acinetobacter baumannii*, *Klebsiella pneumoniae*, MRSA #30, *Enterobacter cloacae*) into replicate wells (n=3) of a 96-well plate. After incubation at 37^{o}C overnight with shaking at 250 rpm, the optical density at 600 nm was quantified for each well using a BioTek Synergy HT Microplate Reader (BioTek Instruments Inc., Winooski, VT). The mean percent inhibition of growth of the respective bacterial tester strain for each clone was determined relative to the empty vector negative control. BAC clones that expressed significant inhibition (>20% relative to the negative control) were re-tested multiple times in order to identify BAC clones that consistently expressed an anti-infective activity, using the same method.

4.4 Results

We constructed a metagenomic library in the vector pSMART-BAC-S (Fig. 14) using high molecular weight DNA isolated from the Auburn University's Cullars Rotation agricultural soil (data not shown). A total of 19,200 independent clones were picked as isolated colonies into 384-well formatted plates to format the library. To determine the average insert size (flanked by NotI sites in the vector), 215 randomly chosen clones were expanded for DNA isolation and Not I digestion followed by pulsed

field gel electrophoresis. A range of insert sizes from ~12 kb to >200 kb were found, with an average insert size of 110 kb. Approximately 10 percent of inserts (23/215) contained zero NotI sites, suggestive of a low average GC content, while 15% (33/215) had many NotI fragments <12kb in size, consistent with a high average GC content.

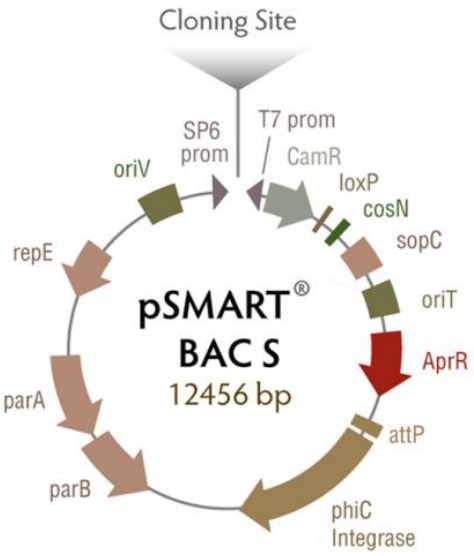


Figure 14**.** Map of the pSmart-BAC-S shuttle BAC vector for broad host range expression in gram-positive and gram-negative bacterial strains. pSMART BAC S contains multiple features to aid in functional screening of inserts. The cloning site is flanked by SP6 and IPTG-inducible T7RNA polymerase promoters, enabling transcription of both strands of the insert. A loxP sequence is included for stable integration into a host bacterial genome, and Gram +/- expression hosts are enabled by inclusion of oriT, oriV and repE. Stable low copy maintenance is aided by parA/parB.

The diversity of bacterial genomes represented within the cloned metagenomic library DNA was assessed by surveying the diversity of 16S rRNA genes and KS

domains from library-derived amplicons. A pooled library BAC DNA template was used for PCR amplification of 16S rRNA genes and KS domain sequences, with each clone grown separately in 384-well format prior to pooling to minimize differences in clone DNA relative abundance within the pooled library template DNA. A total of 318 non-*E. coli* 16S rRNA amplicons were cloned and Sanger sequenced, and the results of BLASTn comparisons and a maximum parsimony analysis of these 16S rRNA genes reveal ribotype affiliations with 9 bacterial phyla including Acidobacteria, Actinobacteria, Bacteroidetes, *Chloroflexi*, *Gemmatimonadetes*, *Firmicutes*, *Planctomycetes*, *Proteobacteria*, and *Verrucomicrobia*, with the phyla *Acidobacteria*, *Actinobacteria*, *Bacteroidetes*, *Gemmatimonadetes*, and *Proteobacteria* being the most dominant taxa accounting for 90.9 % of the total observed ribotypes within the soil metagenomic library.

In addition to surveying the 16S rRNA genes derived from the soil metagenomic library, an analysis was also conducted on the same soil sample used for library construction that had been preserved at -80^{o}C. Metagenomic DNA isolated from this soil sample was used as a template for bar-coded next-gen sequencing of the V4 region of 16S rRNA genes, and the relative abundance of OTUs at 97% or greater identity were affiliated with known 16S rRNA gene sequences in GreenGenes, RDPII and NCBI. The percent relative abundance of the soil ribotypes associated with bacterial phyla was in most cases similar to the ribotype diversity observed in the soil metagenomic library, with the bacterial phyla *Acidobacteria*, *Actinobacteria*, *Bacteroidetes*,

Gemmatimonadetes, and *Proteobacteria* accounting for 83.8% of the observed bacterial ribotype diversity. Among these dominant bacterial taxa there were some notable shifts in percentage relative abundance, particularly the higher percentage relative abundance of *Actinobacteria* taxa within the original soil sample (16.7%) compared to the metagenomic library (7.9%), perhaps reflecting the less harsh lysis conditions used for library construction compared to the bead-beating lysis used for sampling the original soil sample; conversely, the relative abundance of *Gemmatimonadetes* taxa was much higher within the library (12.9%) compared to the original soil sample (1.4%), which may reflect a more facile lysis and cloning of gDNA from members of this phylum.

We initially screened the library for KS genes by macro-array hybridization with a pooled collection of radiolabeled probes corresponding to conserved regions of the KS condensation domain (Table 5). Twelve out of 19,200 arrayed clones hybridized with the probe, and 11 clones were isolated and fully sequenced. All 11 clones contained a KS domain that was 70% identical (or higher) to one of the KS-specific probe sequences, for a 0.06% hit rate. Indeed, the insert DNA of P11P20 is about 138kb much higher than the average.

We next used KS domain-specific degenerate primers for PCR screening of the library. Initial screening of a pooled library DNA template using multiple degenerate primer sets [180] resulted in only the primer set 5LL/4UU (Table 5) giving amplicons of the correct size and sequence (data not shown). We therefore used the 5LL/4UU primer set to PCR screen individual DNA isolates from all 19,200 clones, resulting in ~3000

positives (15.6% hit rate, data not shown); however, >70% of these hits were false positives. After re-screening the PCR-positive clones, 925 of the PCR-positive clones were selected, pooled and sequenced. Contigs from the assembled pools were screened by antiSMASH 2.0 resulting in identification of 110 PKS or PKS/NRPS hybrid gene cluster-containing clones (0.57% hit rate). Although this was a ~10-fold improvement in hit rate over macroarray hybridization screening, the degenerate primers would still be expected to be biased in their ability to amplify divergent KS domain sequences that are not represented within existing databases.

With the goal of performing a complete *in silico* screening of the entire 19,200 member library, we first screened 1/5th of the library (plates 41-50 of the 50 plate library) using the first pooling strategy. 3,840 clones were pooled by row (A-P, 240 clones per pool), column (1-24, 160 clones per pool) and plate (41-50, 384 clones per pool) and indexed pool libraries were constructed and sequenced in a single HiSeq lane. Contigs from the assembled pools were screened by antiSMASH 3.0 resulting in the identification of 74 PKS or PKS/NRPS hybrid gene cluster-containing clones (1.92% hit rate in this section of the BAC library). Of the 74 clones, only 11 had previously been discovered by degenerate PCR screening, indicating that the PCR screening method had missed ~85% of the PKS/NRPS containing clones.

To complete the NGS screening of the metagenomic library, we used a second pooling strategy (decomposition of each 384 well plate into 4 x 96 well quadrants) for the remaining 40 of the 384-well plates, with the clones being processed in four sets of 10 plates (3,840 clones per set). Libraries were generated from row, column and plate

pools and multiplexed on a single HiSeq lane. Contigs from the assembled pools were screened by antiSMASH 3.04 resulting in tatally 1910 SM pathways, and 867 of which were PKS/NRPS clusters. After deconvolution of PKS/NRPS contigs to clones and PCR validation, 284 clones were identified, representing 1.5% of the metagenomic clones in this library. 43 clones of them are also included in 94 PKS/NRPS clones discovered by PCR-screening, indicating a 46% coverage of the NGS screening over PKS/NRPS clones from PCR-screening. All these validated PKS/NRPS clones were then individually processed for resequencing of the whole BAC DNA. Right now we are working on re-sequencing and bioinformatics analysis.

In order to provide a complete insert sequence for each of the pathway-containing BAC clones, a pure culture of the respective clones were recovered from the primary library plates and used for DNA isolation. The identity of the targeted SM pathway-containing BAC clone was confirmed by PCR using a clone insert sequence-specific PCR primer set, and only confirmed clones were used for NGS. Bar-coded fragment libraries of 300-600 bp were prepared from each clone and multiplexed for sequencing using an Illumina MiSeq. Clone sequences were assembled *de novo* and inserts were identified by searching for pSMART-BAC-S vector flanking sequences. Insert-only sequences were then annotated by another pass through antiSMASH 3.04 and gene cluster identifications were confirmed. Right now the first 56 BAC DNA identified through the NGS screening were re-sequenced and recovered, resulting in full-length insert sequences with an average length of (95 kb, ranging from 30kb to 143kb), an average GC content of (58.1%, ranging from a low of 40.4% to a high of 73.3%), and

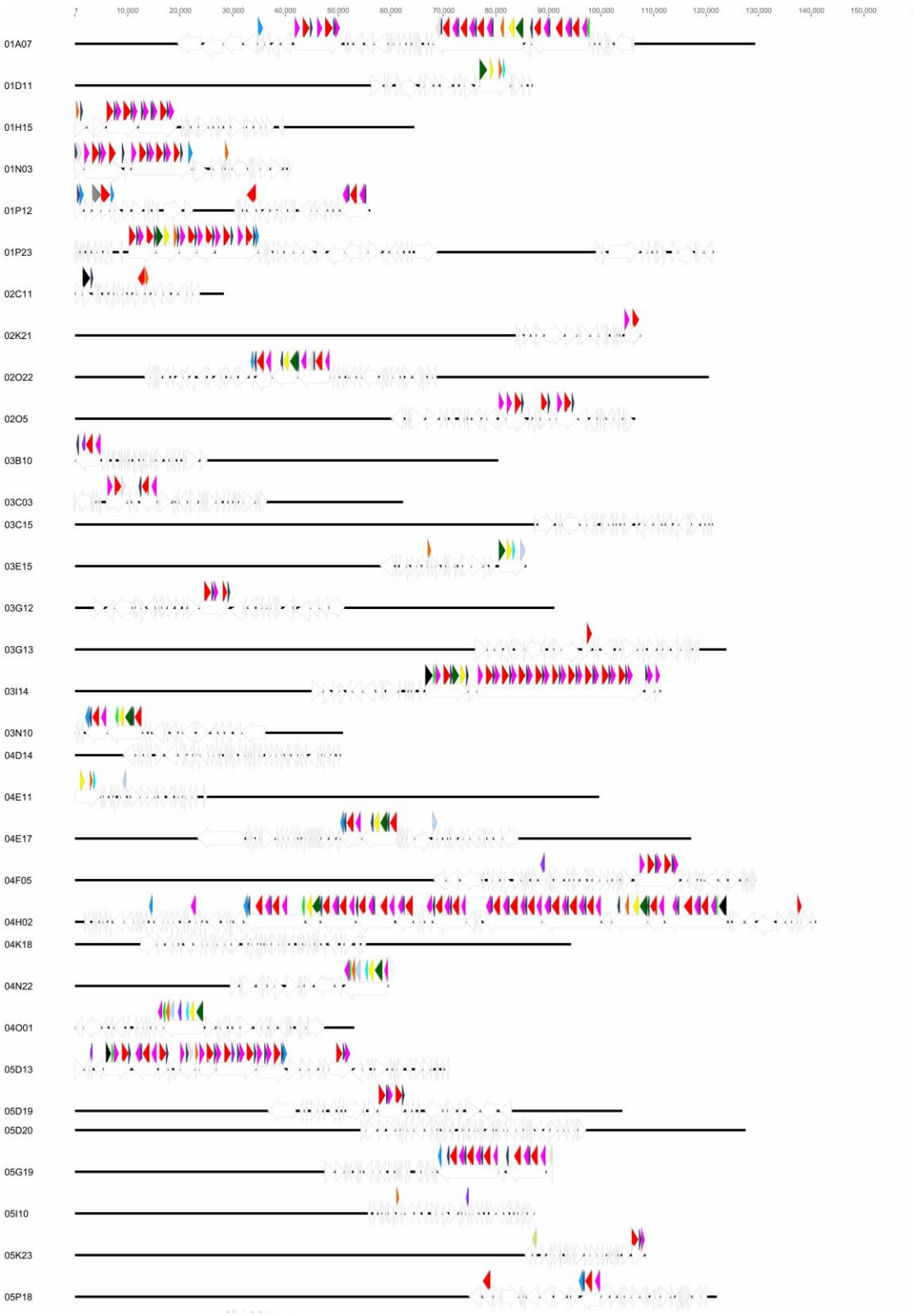
an average of 1.1 gene clusters per insert.

To summarize, the full-length insert of 140 PKS/NRPS containing clones have been recovered, and subsequent analyses were performed either with the collection of contigs used to identify the clones (where indicated), or with the full-length clone insert sequences where available (described below). From 140 PKS/NRPS containing clones, 149 PKS/NRPS clusters were identified by antiSMASH 3.0, 27 of which were predicted to be type I PKS, 48 were hybridized clusters with both type I PKS and NRPS 1 were type II PKS, 14 were type III PKS, 40 were NRPS (Table 5). Architectures of all clusters recovered were presented using antiSMASH plugin on the anaylsis platform Geneious (Fig. 15).

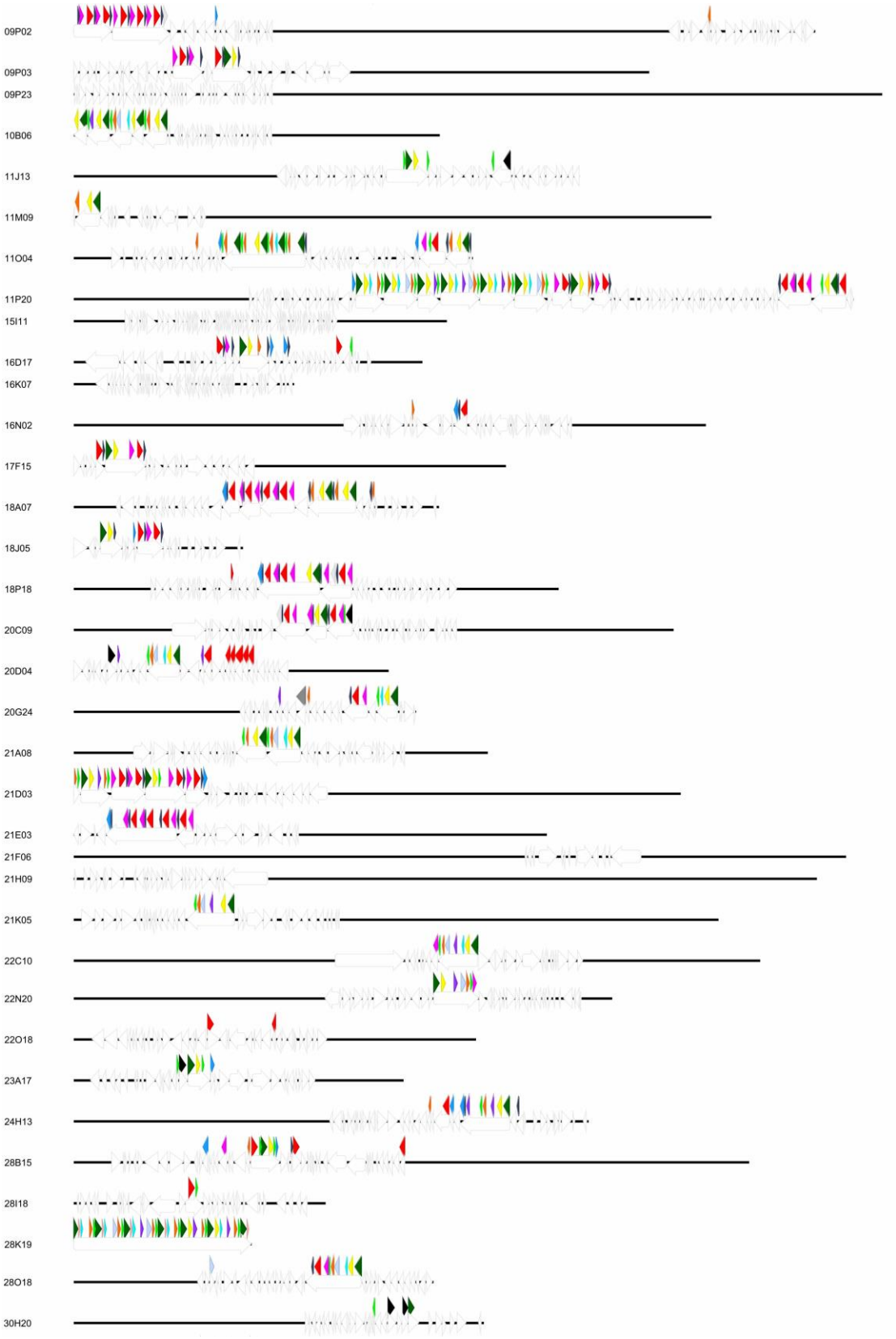
Table 5**.** antiSMASH-identified biosynthetic clusters from the soil metagenomic library as derived from PCR-screening or NGS-screening.

Cluster Type	Macroarray-Screening	PCR-Screening	NGS-Screening
Type I PKS	11	19	78
Type I PKS-NRPS		34	55
Type II PKS			10
Type III PKS		5	117
Type IV PKS			1
NRPS		14	603
Bacteriocin-NRPS			3
Arylpolyene			75
Bacteriocin			210
Homo-serine-lactone			23
Indole			13
Ladderane			17
Lantipeptide		1	49
Lassopeptide		1	46

Microviridin			3
Phosphonate			10
Resorcinol			19
Siderophore			4
Terpene			320
Other KS		1	17
other		10	237
Total	11	86	1910


1
10,000
20,000
30,000
40,000
50,000
60,000
70,000
80,000
90,000
100,000
110,000
120,000
130,000
140,000
150,000
01A07
01D11
01H15
01N03
01P12
01P23
02C11
02K21
02O22
02O5
03B10
03C03
03C15
03E15
03G12
03G13
03I14
03N10
04D14
04E11
04E17
04F05
04H02
04K18
04N22
04O01
05D13
05D19
05D20
05G19
05I10
05K23
05P18


06A05
06D23
06G05
06G11
07D12
07E15
07H03
07H07
07H11
07H14
07J16
07L16
07M04
07O04
07O09
07P02
08A19
08C17
08I08
08J07
08J17
08L11
08M24
08O22
09D5
09G10
09I20_09L06
09J07
09K06_09L06
09L06
09L07
09M07
09M22


09P02
09P03
09P23
10B06
11J13
11M09
11O04
11P20
15I11
16D17
16K07
16N02
17F15
18A07
18J05
18P18
20C09
20D04
20G24
21A08
21D03
21E03
21F06
21H09
21K05
22C10
22N20
22O18
23A17
24H13
28B15
28I18
28K19
28O18
30H20

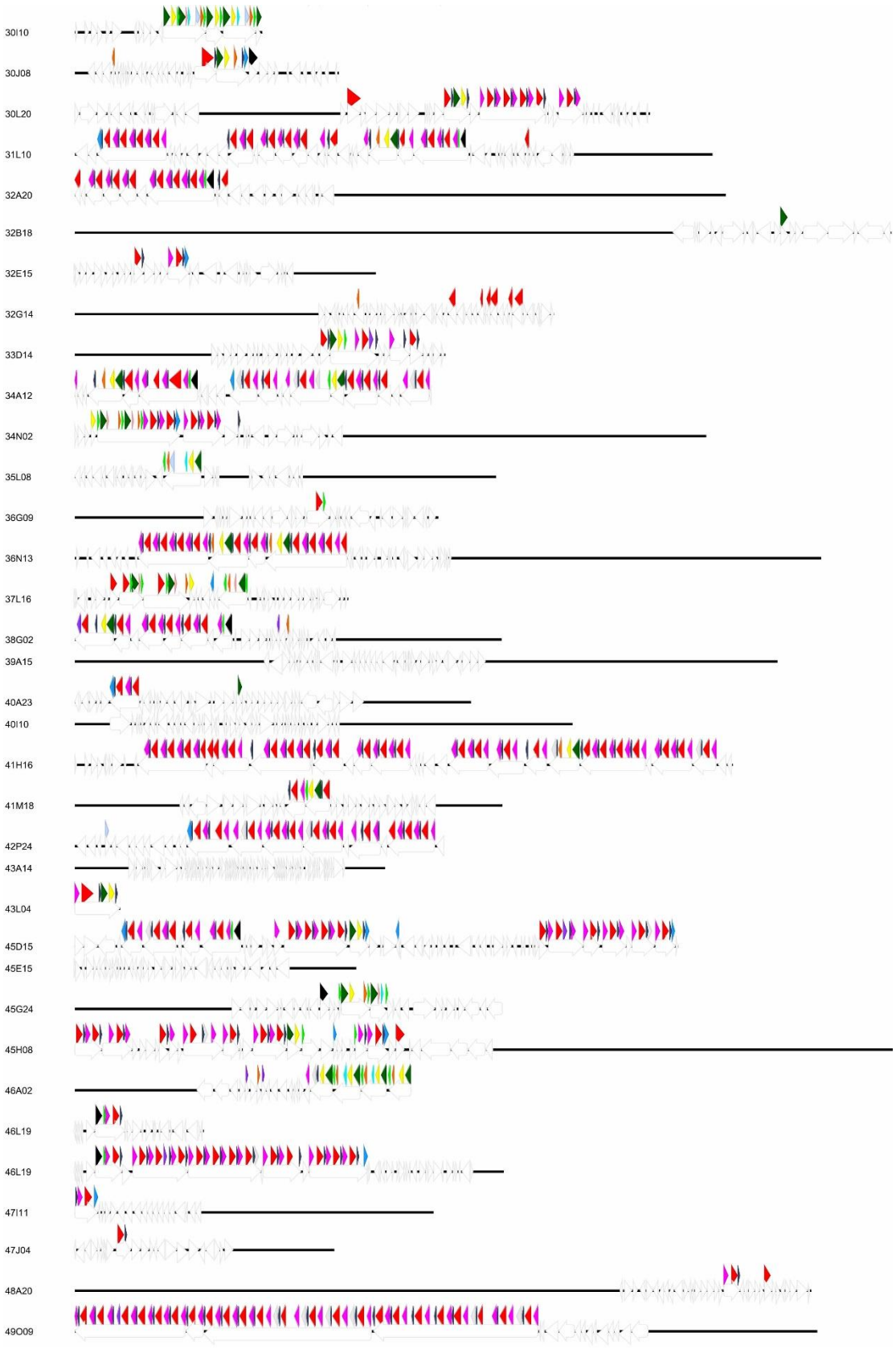


Figure 15**.** Annotation of selected PKS pathways contained within BAC clones.

From the dendogram of identified KS (Fig. 16), the use of NGS to provide high coverage of sequencing for pooled clones provided a higher diversity and number of PKS and hybrid NRPS/PKS pathways than PCR screening, indicating that this method is more sensitive and able to recover more divergent pathways. Theoretically if BAC Sudoku sequencing recovered all genetic information from the library, all PCR-acquired clones should also be identified again by BAC Sudoku using antiSMASH. However, part of PCR-acquired clones rather than all of them were also presented by BAC Sudoku (Fig. 16), implying sequencing depth we applied is still not high enough.

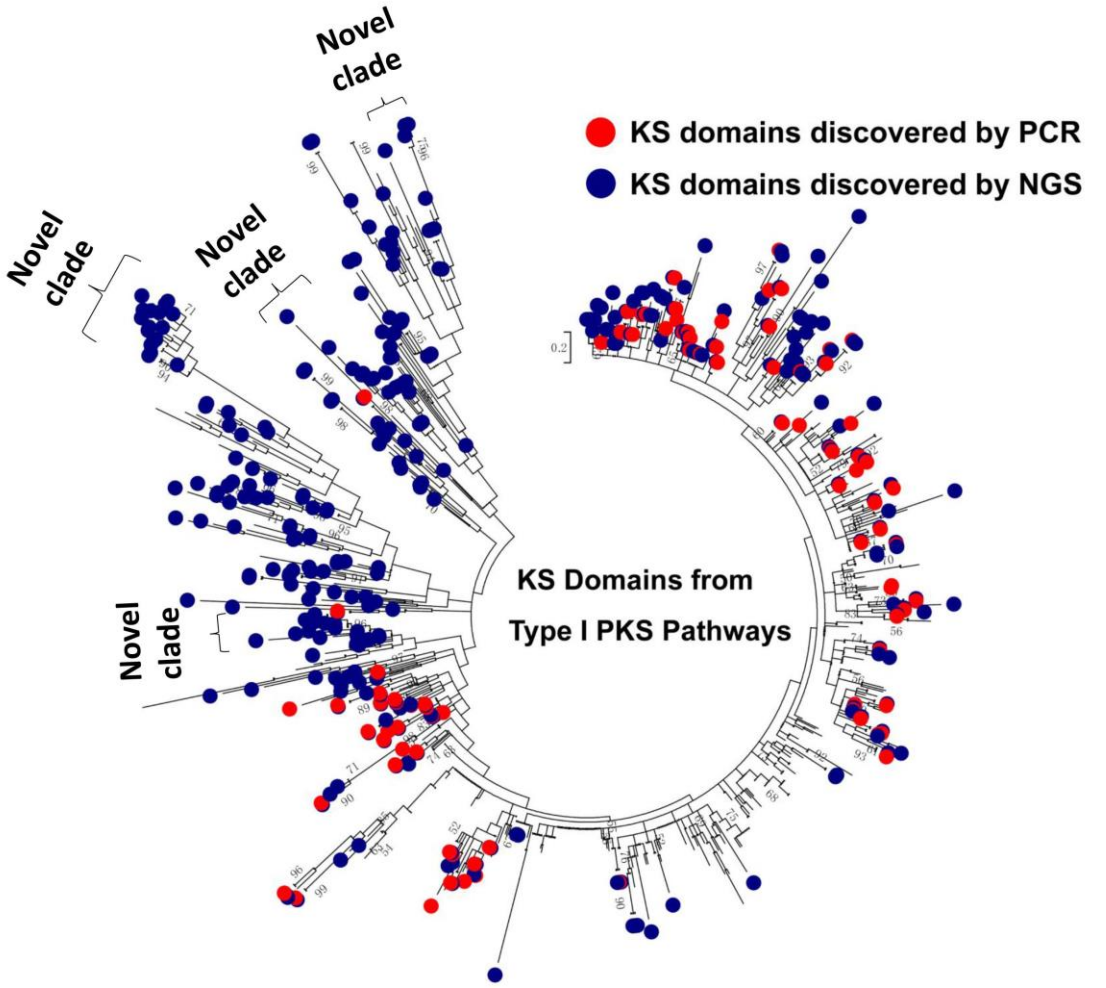


Figure 16**.** KS domain dendogram recovered from soil metagenomic library BAC

clones with a complete insert sequence (predicted polyketide products can also be added with some pathways and label clones identified from PCR- and NGS-screening).

110 PCR screening-identified and 56 the NGS screening identified Clones containing PKS/NRPS clusters have been transformed into *E.coli* BTRA, followed by inoculation in LB with 12.5 μg/ml Cm and induction compounds for expression. Then extracts from broth of clones were added into diluted pathogens, *Pseudomonas aeruginosa* strain PA01, *Acinetobacter baumannii*, *Klebsiella pneumoniae*, MRSA #30, *Enterobacter cloacae* for antibiotic assay. Several candidates of clones with slight inhibition against certain pathogens (>20% inhibition comparing to the negative control) were then collected for large-scale culture (1L) and extraction for potential polyketides. The increased concentration (100X) of two clones P20G24 and P48L9 showed significant inhibition against MRSA #30 (Fig. 17). OD values were measured for seven times during 27 hours incubation. Extracts of P48L9 almost totally inhibited the growth of MRSA and 20G24 also gave a great suppression, indicating a high possibility of producing potent antibiotics. AntiSMASH 3.0 annotated a nrps-t1pks cluster in the insert DNA sequence of 20G24and predicted its 23% of genes showing similarity to the biosynthetic gene cluster encoding an antifungal agent named jawsamycin.

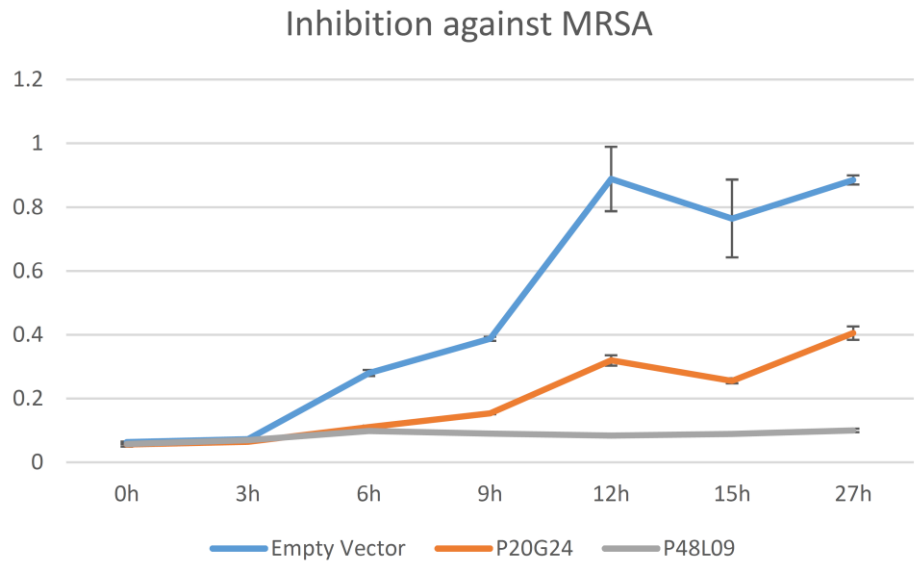


Figure 17**.** Growth of MRSA #30 with 100X extracts of potential compounds from the PKS/NRPS pathway-containing clones P20G24 and P48L9 in *E. coli* BTRA

4.5 Discussion

In this report, we described a strategy of homology-guided metagenomic and the NGS screening to identify and express complete PKS/NRPS and other pathways from soil metagenomes. The large insert size of the soil metagenomic library was critical to the ability to successfully identify complete or nearly complete PKS/NRPS pathways. This is the first report in the literature of the recovery of intact PKS pathways of this size, the majority of which are very divergent from previously described PKS pathways. This opens the potential for exploiting these novel pathways for the discovery of novel chemical entities.

In order to identify BAC clones producing polyketide products from the whole

library, different screening strategies can be performed. The strategy PCR screening using 4UU and 5LL is able to detect both expressed and unexpressed PKS/NRPS pathways in BAC clones, even uncomplete pathways. But problem of PCR screening is labor-intense and generates many false-positive. Application of macroarray hybridization with KS domain probes in this study was more time-saving but less sensitive than PCR screening, leading to only eleven positive clones. Sequencing screening utilizing the next generation sequencing technologies (NGS) was a more accurate identification strategy, capable of recovering PKS sequences and supporting prediction of derived polyketides, which usually required bar-codes for each clone of the whole library that is time-consuming, expensive and unaffoadable for most of laboratories. A wise promotion combined with NGS was reported recently by Jeremy *et al* to reduce number of bar-codes [201]. In this strategy, BAC DNA of each column or row will be bar-coded and sequenced individually, resulting in two coordinate axes that can be used to locate target clones with PKS pathways. Here we reported a more advanced three-dimension pooling strategy that BAC DNA of each plate were bar-coded uniquely as well so that each clone can be located in three coordinate axes. The pipeline we reported applied PCR screening, macroarray hybridization and the NGS screening to narrow number of clones for PKS/NRPS identification firstly, and then validated by colony PCR and re-sequenced to analyze biosynthetic pathways, followed by functional screening for polyketide products. Comparing three approaches with each other, the NGS screening gave 385 PKS/NRPS clones that was higher than 110 identified by PCR screening and 11 by macroarray hybridization. 46% clones from PCR

screening is also included in the result of the NGS screening. In the KS domain dendogram, it showed a much higher diversity presented by the NGS screening than PCR screening, indicating the NGS is much more sensitive (Fig. 16). Due to bias of the primer set 4UU/5LL, the diversity of identified pathways from PCR screening were limited and focused on type I PKS and NRPS with much fewer type I, II and III PKS pathway comparing to the NGS. In addition, if you looked at all metagenomic contigs, not only PKS/NRPS pathways but also bacteriocin, Terpene, Lantipeptide, Lassopeptide, etc. were discovered through the NGS screening (Table 5). All these contigs with interesting pathways will be used for deconvolution to exact clones as well in the future, leading to exploitation of other kinds of SM pathways and their natural products. Therefore, the NGS screening is very powerful and can be a new gold standard for drug discovery, providing a more efficient tool to enable research into novel anti-infective, anti-proliferation, and anti-inflammatory compounds with large synthetic pathways.

The full-length insert DNA of totally 166 clones have been recovered, ranging from 28kb to 200kb. 40% of pathways from these clones were predicted by antiSMASH 3.0 to have low similarity (5%-50%) with biosynthetic pathways of known secondary metabolites such as nostopeptolide, puwainaphycins and nostopeptolide, but the rest of pathways didn't have any significant hits, indicating that all these clones contain novel and distinct from the known. KS domain dendogram also represented some novel clades that were not affiliated any known KS domains (Fig. 16). Therefore, screening approaches used here greatly expanded our knowledge on the diversity of PKS/NRPS

pathways. The extract of the P20G24 broth had a large inhibition against MRSA, and the full length of its insert DNA is 66kb with a predicted 34kb nrps-t1pks cluster containing 25 ORFs and four modules (Fig. 15). Two modules encoded a KS-AT-DH-ACP domain and a stand-alone ketoreductase (KR) respectively are likely responsible for constructing the polyketide backbone (Fig. 15). The cluster had a 23% similarity to that of jawsamycin, a known polycyclopropanated polyketide-nucleoside hybrid. This compound is a potent antifungal agent produced by *streptoverticillium fervens* HP-89 and probably share some medical and expression mechanisms with polyketide from P20G24 that may help us promote our heterologous expression [202]. The cluster in the clone 8C17 (68kb and 33 ORFs) were predicted to a nrps cluster with 40% similarity to that of puwainaphycin A, a known cardioactive cyclic peptide from the blue-green alga Anabaena BQ-16-1. The produced peptide here may share some similar structures or functions. For potential medical application, extracts from the clone 8C17 would be directly processed by LC/MC for characterization.

Cloning and identifying intact secondary metabolite pathways is of course insufficient to provide a resource for drug discovery. Here the use of the engineered *E. coli* strain BTRA by courtesy of Dr. Blaine Pfeifer were selected for heterologous expression, because it was optimized for PKS expression (e.g. erythromycin) and provides a mechanism by which these cloned pathways may be expressed and mined for their novel polyketide products. It is also possible, using the broad host range capabilities of the pSmartBAC-S vector, to introduce and express these libraries in other heterologous hosts such as novel cultured soil *Acidobacteria* [203] to increase the

possibilities of PKS pathway expression in future research.

Functional selection approach is a direct screening strategy aiming to collect positive clones with change of phenotype changed by produced novel secondary metabolites. Right now, 122 BAC clones have been transformed in *E. coli* strain BTRA for polyketide expression, and then processed for antimicrobial assay. The softagar overlay against different pathogens was the first method, which was reported by *et al* in 2009 [204], and leaded to no positive BAC clones with formation of clear inhibition zones around. Limitation of this approach is that many PKS pathways was not able to be expressed in the heterologous host. For the rest of them might express a small amount of polyketide and were hard to present an enough inhibition to form a clear "halo" since origin of these pathways may be phylogenetically distant from the heterologous host. The second method utilized the extract from each transformant to suppress the growth of pathogens in liquid medium, which can be detected using the plate reader in OD value. Therefore, this assay is sensitive enough to catch even slight inhibition caused by produced polyketides. These clones with low antibiotic activities were further cultured in a larger scale to obtain a higher concentration of extracts, followed by incubated together with pathogens for validation. It resulted in two clones P20G24 and P48L9 with significant inhibitions against MRSA, which might be caused by produced polyketides (Fig. 17).

4.6 Further works

Among totally 506 clones identified in this study, there are still full-length insert DNA of 340 clones that have not been obtained but will be done in several months. After all BAC DNA are recovered, we might remove some false-positive clones and do more bioinformatics analysis like we mention above. Some vignettes of “interesting” clones such as high PKS homology hits, the largest cluster found, most dissimilar cluster found and those with related known clusters from MiBig (eg stigmatellin-like, crocacin-like, puwainaphycin-like clusters) will be selected for analysis of their architecture, phylogeny and origin in detail. A higher diversity of PKS/NRPS is also expected from the rest of clones.

The rest of clones will also be transformed into *E.coli* BTRA and screened using both softagar overlay and liquid assay. The clones with significant inhibition against pathogens as well as P20G24 and P48L9 will be further characterized by LC/MS analysis, Bugni PCA, or fluorescent extracts image plus teaser of anti-microbial/antifungal activity. Expression and biochemical work will be combined with sequences analysis for discussion of the constraints on heterologous expression, and the potential bioactive compounds expressed and their application(s). Full-length BAC DNA of some “interesting” clones containing PKS/NRPS pathways can be applied for discussion of their origins that might provide information for us to select other bacteria for better heterologous expression.

Summary

Our study here revealed a possibility for metagenomics application in exploitation of novel natural products and viral genomes. Many microbes in the environments were reluctant for cultivation in lab conditions, making us difficult to understand their communities, interaction and metabolic. However, both sequence-based and clone library-based approaches applied here are able to explore this unknown world and help scientists utilized this great resource reservoir.

Chapter 2. Virophages are a unique group of circular dsDNA viruses that infect the giant DNA viruses, and their genomic sizes range from 18~25kb. This size is the first reason we chose it as target to explore in the lake metagenome. If genome size is too large, it will be hard to avoid chimeric contigs. The circular genomes of them also provide a signal for us to know whether assembly of the complete genome is achieved. In this chapter, iterative assembly together with *de novo* assembly is applied to reconstruct the whole genome. Therefore, the contig can be extended and checked after each round of extension until the repeat sequence (>200bp) is found in both ends. Interaction of virophages with their viral and eukaryotic hosts is very interesting relationship and might change our basic understanding of biology. The giant viruses are also called nucleocytoplasmic large DNA viruses (NCLDVs). They have a huge genomic size, a maximum up to 2.5Mbp, and a large amount of genes previously thought only to be encoded by cellular organisms such as DNA polymerases, topoisomerases and endonucleases. Because of their genomic size comparable to that

of several bacteria and many characteristics placing it at the boundary between cellular and non-cellular organisms, the giant viruses were suggested to be remnants of a "fourth domain" of life. The existence of virophages utilizing their enzymes for reproduction further supports this hypothesis. Virophages also influence the evolution of their cell hosts. Mavirus and ALV were reported to have a similar structure and significant homologies of some conserved genes with a transposon of the Maverick/Polinton family, leading to a presumption that eukaryotic dsDNA transposons are derived from virophgaes due to integration of viral genomes into the ancestors of contemporary eukaryotes. This particular triangular relationship might be clarified more clearly once more virophage genomes were released.

In the previous publications, totally ten virophages have been identified from various aqueous environments, and six genomes of them were obtained through screening different public metagenomic sequence databases including habitats of fresh lakes and Antarctic environment, implying a wide distribution of virophages on the earth. In chapter 2, we reported additional three genomes YSLV5-7 discovered from a metagenome of Yellowstone Lake. Together with four virophages identified before, Yellowstone Lake now has contributed the most genotypes from both mesophilic and thermophilic habitats. Through analysis of virophage in different locations of the lake, YSLV5 is found to be pretty abundant in a water sample with 60-66°C, indicating its thermo-stability in high temperature. In addition, the Yellowstone Lake metagenome is a large database (totally 7.5 Gbp) generated by Roche 454, which has the advantage of long read length for a better assembly than short-read sequencing like Hiseq. Many

public Hiseq metagenomes were also tried for assembly of virophage genomes but the result was not satisfactory. The reason could be the generation of chimeric contigs due to linking short reads with high similarity from distinct entities.

YSLV5-7 genomes all contain five conserved core genes have been detected in all known virophages. YSLV6 also has three additional gene clusters with conserved synteny that are composed of adjacent genes. Combined with the phylogenetic tree and BLASTp results of core genes, YSLV6 indicated close evolutionary relationship with YSLV1-4 and OLV, representing a new member in this lineage. Other two virophage lineages are Mavirus-ALV and Sputnik-Zamilon respective. The phylogeny of YSLV5 was uncertain, locating between Mavirus-ALV and Sputnik-Zamilon lineages in the tree. However, the most interesting feature of YSLV5 is a 51.1% G+C content that is the highest among all known virophages. Since YSLV5 is a thermophilic virophage, it is probably able to stabilize the DNA structure in the higher temperature. YSLV7 formed a distinct lineage from three other. No any gene cluster with conserved synteny can be found in the genome of YSLV7, even the cluster of MCP and mCP that have been identified from all known virophages. Homologs of its core genes were obscure, and the top BLASTp hit of each gene were corresponding to different lineages.

Overall, our research demonstrated the feasibility to recontruct virophage genomes using metagenomic sequencing. Based on average genomic size of virophages, long-read sequencing technologies such as PacBio and Oxford nanopore were expected to promote assembly of their complete genomes from a metagenome even if coverage is relatively low. So far there is no long-read metagenome released, however, it is believed

that this kind of data will greatly help us to understand molecular interactions between virophages and their giant virus and eukaryotic hosts.

Chapter 3. Extreme environments are the edges of this world with unique ecosystems that is different from normal habitats with full of oxygen, range of temperatures from 5°C to 40°C and protection of our atmosphere from most damaging radiation. The organisms in these environments are called “extremophiles”, most of which are microbes. Because of harsh living conditions, abundance and diversity of extremophiles are usually much lower than the normal, which facilitate metagenomic sequencing to achieve more complete microbial communities. In order to survive, microbes in these habitats evolved to have various mechanisms to protect their protein and DNA, which may provide valuable application in our industries. Therefore, great efforts have been made by microbiologists to explore natural products from these microbial extremophiles through metagenomics. In this study, microbial assemblages were sampled from an offshore deep sub-surface petroleum reservoir 2.5 km below the ocean floor off the coast of Norway under conditions of high temperature and pressure, and extracted DNA were used to construct a metagenomic fosmid library for exploitation of carbohydrate-degrading enzymes. The reservoir is 250 bars and 85°C, so thermal stability was expected from identified enzymes.

Both function and sequence-based approaches were applied to screen the fosmid library. The functional screening utilized medium incorporated with corresponding substrates to visualize activities of clones. The 96-pin replicator was employed to inoculate the library onto the medium agar in order to achieve high throughput. The

identified clones were then collected for sequencing to study their insert DNA. For sequence-based screening, fosmid DNA extracted from the whole library were all sequenced using Hiseq (one unique barcode for each two plates). Due to short read length of Hiseq, the resulting reads were combined with 454 reads of shotgun metagenomic sequencing directly from environmental DNA to reach a better *de novo* assembly. BLASTp search of predicted ORFs from the resulting contigs against the CAZy database were applied for identification of carbohydrate-degrading genes. Obviously, sequence-based screening represented a much higher number of hits (cellulase, xylanase, amylase, protease and esterase/lipase) than function-based sequencing. It is because many genes in insert DNA are reluctant to be expressed in the heterologous expression host. However, there were several genes only identified from function-based screening, which was probably caused by their low homologies with known enzymes.

The sequence-based screening represented a low microbial diversity as expected in the oil reservoir sample, and alpha-diversity is 42.08 species. The domain Archaea was dominant and the *Euryarchaeota* is the most abundant phylum that contributed to most of the enzymes obtained in this study. Other three phylum with greater than 0.1% relative abundance were *Proteobacteria*, *Firmicutes* and *Thermotogae*. In addition, there is a considerable amount of sequences assigned to previously unknown microbes, indicating a huge unknown metagenomic diversity in the sampled environment. There were also some bias existed in abundance of some bacterial phylum presented by shogun sequencing and the fosmid library that was probably caused by difference in the

amplification and/or cloning of genomic DNA from these bacteria.

Identified clones with cellulase activities through both sequence- and function-based screening were then collected for further analysis due to their potential applications as biofuel. They were subcloned to achieve a higher copy number and five of them (S1C, S3C, S5C, S8C and F1C) produced significant signals of cellulase activities using a fluorescent substrate MUC in a quantitative assay. S5C, S8C and F1C also showed a strong thermal stability up to 80°C. Once looking into cellulase ORFs they encoded, F1C generated a cellulase with two distinct cellulase domains and it was perhaps resulted from the fusion of two archaeal cellulases. This is a unique and novel protein structure that may result in enhanced cellulase activity and thermo-stability. The thermo-stable cellulases of the three subclones were all predicted derived from thermophilic Archaea within the genus *Thermococcus*.

Currently they are all processed for codon-optimization to reach a better expression and purification using SDS-PAGE to visualize proteins. Finally, the proteins were expected to be extracted and then analysed using LC/MS.

Chapter 4. Soils harbor an extremely diverse microbial assemblage, with one gram of dry soil containing more than one billion microbial cells. Soil microbes produce an abundant and diverse array of secondary metabolites that play significant roles in microbial interactions with other microbes and with hosts, and have been exploited for their clinical utility. The traditional approach to study these metabolites is based on pure culture, isolating certain bacteria from a soil sample. However, most of microbes were reluctant to culture in lab condition. To over this limitation, metagenomics can directly

apply to study environmental DNA without culturing, providing insight into SM pathways derived from as-yet-uncultured microbes.

In this study a large-insert soil metagenomic clone library (~110kb and 19,200 clones) was constructed from an agricultural soil (Cullars Rotation, Auburn, AL) using a broad host range shuttle BAC vector, pSmartBAC-S. The insert size of this BAC library is critical to harbor complete SM pathways that are usually larger than 40kb. In addition, pSmartBAC-S is capable of supporting heterologous expression in both gram-positive and -negative bacteria, helping to access the metabolites encoded.

PKSs is composed of a multidomain architecture and requiring large genetic regions for a complete biosynthetic pathway, greatly contributing to medical industry such as antibacterial, antifungal and anticancer agents. Here we employed three different approaches for discovery of PKS/NRPS pathways from our BAC library. A survey of 16S rRNA genes contained within the pooled library clones revealed a very diverse assemblage of microbial genomes affiliated with nine bacterial phyla. The conserved KS domains of Type I PKS were PCR amplified using the same pooled library template DNAs, resulting in 110 unique KS domain amplicons with a range of only 32.3% - 82.7% amino acid identity with known KS domains in the GenBank nr/nt database. In the first approach, these unique KS domains were used to design radioactively labeled probes to identify 11 BAC clones that contain PKS pathways. In the second method, the universal primers were applied for screening each clone in the library, resulting in 110 PKS/NRPS clones. The final approach utilized the NGS technology to screening each BAC clone so that all kinds of SM pathways rather than

only PKS/NRPSs can be identified theoretically based on known databases. The applied pool strategy is the most creative part in this research, providing an efficient way to track each identified contigs back to the exact clone in the 384-well plate. BAC DNA from each column, row and plates were pooled together for unique barcoding, resulting in three dimensional barcodes for each clone. 358 exact identified clones were finally targeted using x, y and z axis. All the PKS/NRPS pathway-containing BAC clones were collected and then subjected to the NGS technology again to obtain complete or nearly complete pathways contained within each BAC clone, leading to discovery of limited homology to known PKS pathways. Obviously, the NGS screening is much more sensitive, producing a three time higher number of PKS/NRPS clones than PCR screening as well as avoiding bias on diversity of pathways due to specificity of PCR primers. As well, the NGS screening almost covered 46% clones identified by PCR screening. In addition, the NGS screening also discovered many other types of SM pathways (e.g. siderophore, terpene, type II and IV PKS), which will be collected and analyzed in the further research.

The pathway-containing BAC clones were transformed into an *E. coli* BTRA strain particularly engineered for PKS expression. BAC clones expressed in BTRA were screened for the synthesis of antibacterial compounds by bioassay against the pathogens Methicillin-resistant *Staphylococcus aureus* (MRSA), *Pseudomonas aeruginosa*, *Bacillus subtilis* and *Klebsiella pneumoniae*. Significant inhibitions have been observed from extracts of two clones P20G24 and P48L9 against MRSA #30, indicating potential polyketides of antibiotics. Clones expressing antimicrobial activity will be further

characterized by LC/MS analysis. These results illustrated the potential to obtain complete and novel SM pathways from a large-insert soil metagenomic library and express these pathways in a heterologous host.

In this chapter, we described a pipeline for identification and expression of PKS/NRPS pathways from soil microbes, followed by antibiotic assays against pathogens. Actually, it also can be applied to explore other SMs. Since these SMs might have potentials in different medicines, distinct assays will be employed to screen these clones for drug discovery in the further.

Reference

(1) Falkowski, P.; Scholes, R. J.; Boyle, E.; Canadell, J.; Canfield, D.; Elser, J.; Gruber, N.; Hibbard, K.; Hogberg, P.; Linder, S.; Mackenzie, F. T.; Moore, B., 3rd; Pedersen, T.; Rosenthal, Y.; Seitzinger, S.; Smetacek, V.; Steffen, W. *Science* **2000**, *290*, 291.

(2) Kirchman, D. L. *Proceedings of the National Academy of Sciences of the United States of America* **2012**, *109*, 17732.

(3) Kertesz, M. A. *FEMS microbiology reviews* **2000**, *24*, 135.

(4) Elser, J.; Bennett, E. *Nature* **2011**, *478*, 29.

(5) Rothschild, L. J.; Mancinelli, R. L. *Nature* **2001**, *409*, 1092.

(6) Szewzyk, U.; Szewzyk, R.; Stenstrom, T. A. *Proceedings of the National Academy of Sciences of the United States of America* **1994**, *91*, 1810.

(7) Torsvik, V.; Goksoyr, J.; Daae, F. L. *Applied and environmental microbiology* **1990**, *56*, 782.

(8) Roesch, L. F.; Fulthorpe, R. R.; Riva, A.; Casella, G.; Hadwin, A. K.; Kent, A. D.; Daroub, S. H.; Camargo, F. A.; Farmerie, W. G.; Triplett, E. W. *The ISME journal* **2007**, *1*, 283.

(9) Whitman, W. B.; Coleman, D. C.; Wiebe, W. J. *Proceedings of the National Academy of Sciences of the United States of America* **1998**, *95*, 6578.

(10) Pisciotta, J. M.; Zou, Y.; Baskakov, I. V. *PloS one* **2010**, *5*, e10821.

(11) Tripp, H. J.; Bench, S. R.; Turk, K. A.; Foster, R. A.; Desany, B. A.; Niazi, F.; Affourtit, J. P.; Zehr, J. P. *Nature* **2010**, *464*, 90.

(12) Eckburg, P. B.; Bik, E. M.; Bernstein, C. N.; Purdom, E.; Dethlefsen, L.; Sargent, M.; Gill, S. R.; Nelson, K. E.; Relman, D. A. *Science* **2005**, *308*, 1635.

(13) Plotnikoff, G. A.; Riley, D. *Global advances in health and medicine : improving healthcare outcomes worldwide* **2014**, *3*, 4.

(14) Houbraken, J.; Frisvad, J. C.; Samson, R. A. *IMA fungus* **2011**, *2*, 87.

(15) Pegler, S.; Healy, B. *Bmj* **2007**, *335*, 991.

(16) Chopra, I.; Roberts, M. *Microbiology and molecular biology reviews : MMBR* **2001**, *65*, 232.

(17) Fosso, M. Y.; Li, Y.; Garneau-Tsodikova, S. *MedChemComm* **2014**, *5*, 1075.

(18) Garey, K. W.; Salazar, M.; Shah, D.; Rodrigue, R.; DuPont, H. L. *The Annals of pharmacotherapy* **2008**, *42*, 827.

(19) Wipf, P.; Reeves, J. T.; Day, B. W. *Current pharmaceutical design* **2004**, *10*, 1417.

(20) Chien, A.; Edgar, D. B.; Trela, J. M. *Journal of bacteriology* **1976**, *127*, 1550.

(21) Fuchs, A. *Antonie van Leeuwenhoek* **1984**, *50*, 425.

(22) Pringsheim, E. G. *Medizinische Monatsschrift* **1971**, *25*, 118.

(23) Jokl, D. H. *Documenta ophthalmologica. Advances in ophthalmology* **1999**, *99*, 285.

(24) Ellis, H. *British journal of hospital medicine* **2010**, *71*, 223.

(25) Handelsman, J. *Microbiology and molecular biology reviews : MMBR* **2004**, *68*, 669.

(26) Vartoukian, S. R.; Palmer, R. M.; Wade, W. G. *FEMS microbiology letters* **2010**, *309*, 1.

(27) Weisburg, W. G.; Barns, S. M.; Pelletier, D. A.; Lane, D. J. *Journal of bacteriology* **1991**, *173*, 697.

(28) Woese, C. R.; Kandler, O.; Wheelis, M. L. *Proceedings of the National Academy of Sciences of the United States of America* **1990**, *87*, 4576.

(29) Bartlett, J. M.; Stirling, D. *Methods in molecular biology* **2003**, *226*, 3.

(30) Dunbar, J.; Ticknor, L. O.; Kuske, C. R. *Applied and environmental microbiology* **2001**, *67*, 190.

(31) Peterson, C. *Journal of visualized experiments : JoVE* **2007**, 164.

(32) Sumida, M.; Kato, Y.; Kurabayashi, A. *Genes & genetic systems* **2004**, *79*, 105.

(33) Rondon, M. R.; August, P. R.; Bettermann, A. D.; Brady, S. F.; Grossman, T. H.; Liles, M. R.; Loiacono, K. A.; Lynch, B. A.; MacNeil, I. A.; Minor, C.; Tiong, C. L.; Gilman, M.; Osburne, M. S.; Clardy, J.; Handelsman, J.; Goodman, R. M. *Appl Environ Microbiol* **2000**, *66*, 2541.

(34) Ni, J.; Yan, Q.; Yu, Y. *Scientific reports* **2013**, *3*, 1968.

(35) Grada, A.; Weinbrecht, K. *The Journal of investigative dermatology* **2013**, *133*, e11.

(36) Degaspari, J. *Healthcare informatics : the business magazine for information and communication systems* **2013**, *30*, 15.

(37) Koonin, E. V. *Nature biotechnology* **2007**, *25*, 540.

(38) Lee, D. G.; Jeon, J. H.; Jang, M. K.; Kim, N. Y.; Lee, J. H.; Lee, J. H.; Kim, S. J.; Kim, G. D.; Lee, S. H. *Biotechnology letters* **2007**, *29*, 465.

(39) Hohn, B.; Koukolikova-Nicola, Z.; Lindenmaier, W.; Collins, J. *Biotechnology*

1988, *10*, 113.

(40) Evans, G. A.; Snider, K.; Hermanson, G. G. *Methods in enzymology* **1992**, *216*, 530.

(41) Kim, U. J.; Shizuya, H.; de Jong, P. J.; Birren, B.; Simon, M. I. *Nucleic acids research* **1992**, *20*, 1083.

(42) Shizuya, H.; Birren, B.; Kim, U. J.; Mancino, V.; Slepak, T.; Tachiiri, Y.; Simon, M. *Proceedings of the National Academy of Sciences of the United States of America* **1992**, *89*, 8794.

(43) Shizuya, H.; Kouros-Mehr, H. *The Keio journal of medicine* **2001**, *50*, 26.

(44) O'Connor, M.; Peifer, M.; Bender, W. *Science* **1989**, *244*, 1307.

(45) Kakirde, K. S.; Wild, J.; Godiska, R.; Mead, D. A.; Wiggins, A. G.; Goodman, R. M.; Szybalski, W.; Liles, M. R. *Gene* **2011**, *475*, 57.

(46) Liles, M. R.; Williamson, L. L.; Rodbumrer, J.; Torsvik, V.; Goodman, R. M.; Handelsman, J. *Applied and environmental microbiology* **2008**, *74*, 3302.

(47) Liles, M. R.; Williamson, L. L.; Rodbumrer, J.; Torsvik, V.; Parsley, L. C.; Goodman, R. M.; Handelsman, J. *Cold Spring Harb Protoc* **2009**, *2009*, pdb prot5271.

(48) Delmont, T. O.; Robe, P.; Clark, I.; Simonet, P.; Vogel, T. M. *Journal of microbiological methods* **2011**, *86*, 397.

(49) Navarro-Noya, Y.; Hernandez-Rodriguez, C.; Zenteno, J. C.; Buentello-Volante, B.; Cancino-Diaz, M. E.; Jan-Roblero, J.; Cancino-Diaz, J. C. *Brazilian journal of microbiology : [publication of the Brazilian Society for Microbiology]* **2012**, *43*, 283.

(50) Beja, O.; Aravind, L.; Koonin, E. V.; Suzuki, M. T.; Hadd, A.; Nguyen, L. P.;

Jovanovich, S. B.; Gates, C. M.; Feldman, R. A.; Spudich, J. L.; Spudich, E. N.; DeLong, E. F. *Science* **2000**, *289*, 1902.

(51) Zhang, W.; Chen, J.; Yang, Y.; Tang, Y.; Shang, J.; Shen, B. *PloS one* **2011**, *6*, e17915.

(52) Kearse, M.; Moir, R.; Wilson, A.; Stones-Havas, S.; Cheung, M.; Sturrock, S.; Buxton, S.; Cooper, A.; Markowitz, S.; Duran, C.; Thierer, T.; Ashton, B.; Meintjes, P.; Drummond, A. *Bioinformatics* **2012**, *28*, 1647.

(53) Sun, S.; Chen, J.; Li, W.; Altintas, I.; Lin, A.; Peltier, S.; Stocks, K.; Allen, E. E.; Ellisman, M.; Grethe, J.; Wooley, J. *Nucleic acids research* **2011**, *39*, D546.

(54) Pagani, I.; Liolios, K.; Jansson, J.; Chen, I. M.; Smirnova, T.; Nosrat, B.; Markowitz, V. M.; Kyrpides, N. C. *Nucleic acids research* **2012**, *40*, D571.

(55) Glass, E. M.; Wilkening, J.; Wilke, A.; Antonopoulos, D.; Meyer, F. *Cold Spring Harbor protocols* **2010**, *2010*, pdb prot5368.

(56) Goecks, J.; Nekrutenko, A.; Taylor, J. *Genome biology* **2010**, *11*, R86.

(57) Suenaga, H.; Ohnuki, T.; Miyazaki, K. *Environmental microbiology* **2007**, *9*, 2289.

(58) Schmitz, J. E.; Daniel, A.; Collin, M.; Schuch, R.; Fischetti, V. A. *Appl Environ Microbiol* **2008**, *74*, 1649.

(59) Craig, J. W.; Chang, F. Y.; Kim, J. H.; Obiajulu, S. C.; Brady, S. F. *Applied and environmental microbiology* **2010**, *76*, 1633.

(60) Brideau, C.; Gunter, B.; Pikounis, B.; Liaw, A. *Journal of biomolecular screening* **2003**, *8*, 634.

(61) Hann, M. M.; Oprea, T. I. *Current opinion in chemical biology* **2004**, *8*, 255.

(62) Taupp, M.; Mewis, K.; Hallam, S. J. *Current opinion in biotechnology* **2011**, *22*, 465.

(63) Mewis, K.; Taupp, M.; Hallam, S. J. *Journal of visualized experiments : JoVE* **2011**.

(64) Mus-Veteau, I. *Comparative and functional genomics* **2002**, *3*, 511.

(65) Olins, P. O.; Lee, S. C. *Current opinion in biotechnology* **1993**, *4*, 520.

(66) Brady, S. F.; Clardy, J. *Angewandte Chemie* **2005**, *44*, 7063.

(67) Gomez-Escribano, J. P.; Bibb, M. J. *Methods in enzymology* **2012**, *517*, 279.

(68) Larsson, S.; Cassland, P.; Jonsson, L. J. *Applied and environmental microbiology* **2001**, *67*, 1163.

(69) Cronin, C. N.; McIntire, W. S. *Protein expression and purification* **2000**, *19*, 74.

(70) Sanger, F.; Nicklen, S.; Coulson, A. R. *Proceedings of the National Academy of Sciences of the United States of America* **1977**, *74*, 5463.

(71) Sanger, F.; Coulson, A. R. *Journal of molecular biology* **1975**, *94*, 441.

(72) Huse, S. M.; Huber, J. A.; Morrison, H. G.; Sogin, M. L.; Welch, D. M. *Genome biology* **2007**, *8*, R143.

(73) Caporaso, J. G.; Lauber, C. L.; Walters, W. A.; Berg-Lyons, D.; Huntley, J.; Fierer, N.; Owens, S. M.; Betley, J.; Fraser, L.; Bauer, M.; Gormley, N.; Gilbert, J. A.; Smith, G.; Knight, R. *The ISME journal* **2012**, *6*, 1621.

(74) McElhoe, J. A.; Holland, M. M.; Makova, K. D.; Su, M. S.; Paul, I. M.; Baker, C. H.; Faith, S. A.; Young, B. *Forensic science international. Genetics* **2014**, *13C*, 20.

(75) Voelkerding, K. V.; Dames, S. A.; Durtschi, J. D. *Clinical chemistry* **2009**, *55*, 641.

(76) van Dijk, E. L.; Auger, H.; Jaszczyszyn, Y.; Thermes, C. *Trends in genetics : TIG* **2014**.

(77) Roberts, R. J.; Carneiro, M. O.; Schatz, M. C. *Genome biology* **2013**, *14*, 405.

(78) Craig, J. P.; Bekal, S.; Niblack, T.; Domier, L.; Lambert, K. N. *Journal of nematology* **2009**, *41*, 281.

(79) Craig, C. L.; Cameron, C.; Griffiths, J.; Bauman, A.; Tudor-Locke, C.; Andersen, R. E. *BMC public health* **2009**, *9*, 425.

(80) Craig, S. B.; Graham, G. C.; Burns, M. A.; Dohnt, M. F.; Smythe, L. D.; McKay, D. B. *Annals of tropical medicine and parasitology* **2009**, *103*, 467.

(81) Greenleaf, W. J.; Sidow, A. *Genome biology* **2014**, *15*, 303.

(82) Craig, S. B.; Smythe, L. D.; Graham, G. C.; McKay, D. B. *The American journal of tropical medicine and hygiene* **2009**, *80*, 1067.

(83) Suenaga, K.; Wakabayashi, H.; Koshino, M.; Sato, Y.; Urita, K.; Iijima, S. *Nature nanotechnology* **2007**, *2*, 358.

(84) Suenaga, R.; Tomonaga, S.; Yamane, H.; Kurauchi, I.; Tsuneyoshi, Y.; Sato, H.; Denbow, D. M.; Furuse, M. *Amino acids* **2008**, *35*, 139.

(85) Suenaga, T.; Arase, H.; Yamasaki, S.; Kohno, M.; Yokosuka, T.; Takeuchi, A.; Hattori, T.; Saito, T. *European journal of immunology* **2007**, *37*, 3197.

(86) Suenaga, M.; Kawai, Y.; Watanabe, H.; Atsuta, N.; Ito, M.; Tanaka, F.; Katsuno, M.; Fukatsu, H.; Naganawa, S.; Sobue, G. *Journal of neurology, neurosurgery, and*

psychiatry **2008**, *79*, 496.

(87) Suenaga, K.; Higashihara, S.; Ohashi, M.; Oomi, G.; Hedo, M.; Uwatoko, Y.; Saito, K.; Mitani, S.; Takanashi, K. *Physical review letters* **2007**, *98*, 207202.

(88) Suenaga, A.; Narumi, T.; Futatsugi, N.; Yanai, R.; Ohno, Y.; Okimoto, N.; Taiji, M. *Chemistry, an Asian journal* **2007**, *2*, 591.

(89) Craig, J. *Diabetes self-management* **2009**, *26*, 8.

(90) Ziemert, N.; Podell, S.; Penn, K.; Badger, J. H.; Allen, E.; Jensen, P. R. *PloS one* **2012**, *7*, e34064.

(91) Craig, S. M.; Yu, F.; Curtis, J. R.; Alarcon, G. S.; Conn, D. L.; Jonas, B.; Callahan, L. F.; Smith, E. A.; Moreland, L. W.; Bridges, S. L., Jr.; Mikuls, T. R. *The Journal of rheumatology* **2010**, *37*, 275.

(92) Staunton, J.; Weissman, K. J. *Natural product reports* **2001**, *18*, 380.

(93) Chan, Y. A.; Podevels, A. M.; Kevany, B. M.; Thomas, M. G. *Natural product reports* **2009**, *26*, 90.

(94) Cheng, Y. Q.; Tang, G. L.; Shen, B. *Proceedings of the National Academy of Sciences of the United States of America* **2003**, *100*, 3149.

(95) Weissman, K. J. *Chembiochem : a European journal of chemical biology* **2006**, *7*, 485.

(96) Donadio, S.; Staver, M. J.; McAlpine, J. B.; Swanson, S. J.; Katz, L. *Science* **1991**, *252*, 675.

(97) Donadio, S.; Sosio, M. *Combinatorial chemistry & high throughput screening* **2003**, *6*, 489.

(98) Craig, A.; Mai, J.; Cai, S.; Jeyaseelan, S. *Infection and immunity* **2009**, *77*, 568.

(99) Craig, D. H.; Gayer, C. P.; Schaubert, K. L.; Wei, Y.; Li, J.; Laouar, Y.; Basson, M. D. *American journal of physiology. Cell physiology* **2009**, *296*, C193.

(100) Narasingarao, P.; Podell, S.; Ugalde, J. A.; Brochier-Armanet, C.; Emerson, J. B.; Brocks, J. J.; Heidelberg, K. B.; Banfield, J. F.; Allen, E. E. *The ISME journal* **2012**, *6*, 81.

(101) Craig, L. A.; Hong, N. S.; Kopp, J.; McDonald, R. J. *Experimental brain research* **2009**, *193*, 29.

(102) Smits, S. L.; Bodewes, R.; Ruiz-Gonzalez, A.; Baumgartner, W.; Koopmans, M. P.; Osterhaus, A. D.; Schurch, A. C. *Frontiers in microbiology* **2015**, *6*, 1069.

(103) Ruby, J. G.; Bellare, P.; Derisi, J. L. *G3* **2013**, *3*, 865.

(104) Hunt, M.; Gall, A.; Ong, S. H.; Brener, J.; Ferns, B.; Goulder, P.; Nastouli, E.; Keane, J. A.; Kellam, P.; Otto, T. D. *Bioinformatics* **2015**, *31*, 2374.

(105) Grard, G.; Fair, J. N.; Lee, D.; Slikas, E.; Steffen, I.; Muyembe, J. J.; Sittler, T.; Veeraraghavan, N.; Ruby, J. G.; Wang, C.; Makuwa, M.; Mulembakani, P.; Tesh, R. B.; Mazet, J.; Rimoin, A. W.; Taylor, T.; Schneider, B. S.; Simmons, G.; Delwart, E.; Wolfe, N. D.; Chiu, C. Y.; Leroy, E. M. *PLoS pathogens* **2012**, *8*, e1002924.

(106) Siegers, J. Y.; van de Bildt, M. W.; van Elk, C. E.; Schurch, A. C.; Tordo, N.; Kuiken, T.; Bodewes, R.; Osterhaus, A. D. *Emerging infectious diseases* **2014**, *20*, 1081.

(107) Zhou, J.; Zhang, W.; Yan, S.; Xiao, J.; Zhang, Y.; Li, B.; Pan, Y.; Wang, Y. *Journal of virology* **2013**, *87*, 4225.

(108) La Scola, B.; Desnues, C.; Pagnier, I.; Robert, C.; Barrassi, L.; Fournous, G.;

Merchat, M.; Suzan-Monti, M.; Forterre, P.; Koonin, E.; Raoult, D. *Nature* **2008**, *455*, 100.

(109) Yau, S.; Lauro, F. M.; DeMaere, M. Z.; Brown, M. V.; Thomas, T.; Raftery, M. J.; Andrews-Pfannkoch, C.; Lewis, M.; Hoffman, J. M.; Gibson, J. A.; Cavicchioli, R. *Proceedings of the National Academy of Sciences of the United States of America* **2011**, *108*, 6163.

(110) Fischer, M. G.; Suttle, C. A. *Science* **2011**, *332*, 231.

(111) Gaia, M.; Pagnier, I.; Campocasso, A.; Fournous, G.; Raoult, D.; La Scola, B. *PloS one* **2013**, *8*, e61912.

(112) Desnues, C.; La Scola, B.; Yutin, N.; Fournous, G.; Robert, C.; Azza, S.; Jardot, P.; Monteil, S.; Campocasso, A.; Koonin, E. V.; Raoult, D. *Proceedings of the National Academy of Sciences of the United States of America* **2012**, *109*, 18078.

(113) Zhou, J. L.; Zhang, W. J.; Yan, S. L.; Xiao, J. Z.; Zhang, Y. Y.; Li, B. L.; Pan, Y. J.; Wang, Y. J. *Journal of virology* **2013**, *87*, 4225.

(114) Gaia, M.; Benamar, S.; Boughalmi, M.; Pagnier, I.; Croce, O.; Colson, P.; Raoult, D.; La Scola, B. *PloS one* **2014**, *9*, e94923.

(115) Raoult, D.; Boyer, M. *Intervirology* **2010**, *53*, 321.

(116) Morgan, L. A.; Shanks, W. C.; Lovalvo, D. A.; Johnson, S. Y.; Stephenson, W. J.; Pierce, K. L.; Harlan, S. S.; Finn, C. A.; Lee, G.; Webring, M.; Schulze, B.; Duhn, J.; Sweeney, R.; Balistrieri, L. *J Volcanol Geoth Res* **2003**, *122*, 221.

(117) Balistrieri, L. S.; Shanks, W. C. I.; Cuhel, R. L.; Aguilar, C.; Klump, J. V. *U.S. Geological Survey Professional Paper 1717* **2007**, 173.

(118) Morgan, L. A.; Shanks, W. C., III; Pierce, K. L.; Lovalvo, D. A.; Lee, G. K.; Webring, M. W.; Stephenson, W. J.; Johnson, S. Y.; Harlan, S. S.; Schulze, B.; Finn, C. A. *U.S. Geological Survey Professional Paper 1717* **2007**, 95.

(119) Kan, J.; Clingenpeel, S.; Macur, R. E.; Inskeep, W. P.; Lovalvo, D.; Varley, J.; Gorby, Y.; McDermott, T. R.; Nealson, K. *The ISME journal* **2011**, *5*, 1784.

(120) Clingenpeel, S.; Macur, R. E.; Kan, J.; Inskeep, W. P.; Lovalvo, D.; Varley, J.; Mathur, E.; Nealson, K.; Gorby, Y.; Jiang, H.; LaFracois, T.; McDermott, T. R. *Environmental microbiology* **2011**, *13*, 2172.

(121) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. *Nucleic acids research* **1997**, *25*, 3389.

(122) Altschul, S. F.; Wootton, J. C.; Gertz, E. M.; Agarwala, R.; Morgulis, A.; Schaffer, A. A.; Yu, Y. K. *The FEBS journal* **2005**, *272*, 5101.

(123) Marchler-Bauer, A.; Lu, S.; Anderson, J. B.; Chitsaz, F.; Derbyshire, M. K.; DeWeese-Scott, C.; Fong, J. H.; Geer, L. Y.; Geer, R. C.; Gonzales, N. R.; Gwadz, M.; Hurwitz, D. I.; Jackson, J. D.; Ke, Z.; Lanczycki, C. J.; Lu, F.; Marchler, G. H.; Mullokandov, M.; Omelchenko, M. V.; Robertson, C. L.; Song, J. S.; Thanki, N.; Yamashita, R. A.; Zhang, D.; Zhang, N.; Zheng, C.; Bryant, S. H. *Nucleic acids research* **2011**, *39*, D225.

(124) Jones, P.; Binns, D.; Chang, H. Y.; Fraser, M.; Li, W.; McAnulla, C.; McWilliam, H.; Maslen, J.; Mitchell, A.; Nuka, G.; Pesseat, S.; Quinn, A. F.; Sangrador-Vegas, A.; Scheremetjew, M.; Yong, S. Y.; Lopez, R.; Hunter, S. *Bioinformatics* **2014**, *30*, 1236.

(125) Katoh, K.; Standley, D. M. *Molecular biology and evolution* **2013**, *30*, 772.

(126) Stamatakis, A. *Bioinformatics* **2014**, *30*, 1312.

(127) Yutin, N.; Raoult, D.; Koonin, E. V. *Virology journal* **2013**, *10*, 158.

(128) Ilyina, T. V.; Gorbalenya, A. E.; Koonin, E. V. *Journal of molecular evolution* **1992**, *34*, 351.

(129) Iyer, L. M.; Koonin, E. V.; Leipe, D. D.; Aravind, L. *Nucleic acids research* **2005**, *33*, 3875.

(130) Amann, R. I.; Ludwig, W.; Schleifer, K. H. *Microbiological reviews* **1995**, *59*, 143.

(131) Ferrer, M.; Golyshina, O.; Beloqui, A.; Golyshin, P. N. *Current opinion in microbiology* **2007**, *10*, 207.

(132) Lewin, A.; Wentzel, A.; Valla, S. *Current opinion in biotechnology* **2013**, *24*, 516.

(133) Daniel, R. *Current opinion in biotechnology* **2004**, *15*, 199.

(134) Banik, J. J.; Brady, S. F. *Current opinion in microbiology* **2010**, *13*, 603.

(135) Milshteyn, A.; Schneider, J. S.; Brady, S. F. *Chemistry & biology* **2014**, *21*, 1211.

(136) Delmont, T. O.; Malandain, C.; Prestat, E.; Larose, C.; Monier, J. M.; Simonet, P.; Vogel, T. M. *The ISME journal* **2011**, *5*, 1837.

(137) Wilson, M. C.; Piel, J. *Chemistry & biology* **2013**, *20*, 636.

(138) Healy, F. G.; Ray, R. M.; Aldrich, H. C.; Wilkie, A. C.; Ingram, L. O.; Shanmugam, K. T. *Applied microbiology and biotechnology* **1995**, *43*, 667.

(139) Nacke, H.; Engelhaupt, M.; Brady, S.; Fischer, C.; Tautzt, J.; Daniel, R. *Biotechnology letters* **2012**, *34*, 663.

(140) Jiang, C.; Ma, G.; Li, S.; Hu, T.; Che, Z.; Shen, P.; Yan, B.; Wu, B. *J Microbiol* **2009**, *47*, 542.

(141) Pope, P. B.; Denman, S. E.; Jones, M.; Tringe, S. G.; Barry, K.; Malfatti, S. A.; McHardy, A. C.; Cheng, J. F.; Hugenholtz, P.; McSweeney, C. S.; Morrison, M. *Proceedings of the National Academy of Sciences of the United States of America* **2010**, *107*, 14793.

(142) Klippel, B.; Sahm, K.; Basner, A.; Wiebusch, S.; John, P.; Lorenz, U.; Peters, A.; Abe, F.; Takahashi, K.; Kaiser, O.; Goesmann, A.; Jaenicke, S.; Grote, R.; Horikoshi, K.; Antranikian, G. *Extremophiles : life under extreme conditions* **2014**, *18*, 853.

(143) Mewis, K.; Armstrong, Z.; Song, Y. C.; Baldwin, S. A.; Withers, S. G.; Hallam, S. J. *Journal of biotechnology* **2013**, *167*, 462.

(144) Martin, M.; Biver, S.; Steels, S.; Barbeyron, T.; Jam, M.; Portetelle, D.; Michel, G.; Vandenbol, M. *Applied and environmental microbiology* **2014**, *80*, 4958.

(145) Rebuffet, E.; Groisillier, A.; Thompson, A.; Jeudy, A.; Barbeyron, T.; Czjzek, M.; Michel, G. *Environmental microbiology* **2011**, *13*, 1253.

(146) Lewin, A.; Johansen, J.; Wentzel, A.; Kotlar, H. K.; Drablos, F.; Valla, S. *Environmental microbiology* **2014**, *16*, 545.

(147) Kotlar, H. K.; Lewin, A.; Johansen, J.; Throne-Holst, M.; Haverkamp, T.; Markussen, S.; Winnberg, A.; Ringrose, P.; Aakvik, T.; Ryeng, E.; Jakobsen, K.; Drablos, F.; Valla, S. *Environmental microbiology reports* **2011**, *3*, 674.

(148) Magot, M.; Ollivier, B.; Patel, B. K. *Antonie van Leeuwenhoek* **2000**, *77*, 103.

(149) Lombard, V.; Golaconda Ramulu, H.; Drula, E.; Coutinho, P. M.; Henrissat, B.

Nucleic acids research **2014**, *42*, D490.

(150) Brethauer, S.; Wyman, C. E. *Bioresource technology* **2010**, *101*, 4862.

(151) Lynd, L. R.; Weimer, P. J.; van Zyl, W. H.; Pretorius, I. S. *Microbiology and molecular biology reviews : MMBR* **2002**, *66*, 506.

(152) Tissot, B. P.; Welte, D. H. **1984**.

(153) Xu, Z.; Zhang, R.; Wang, D.; Qiu, M.; Feng, H.; Zhang, N.; Shen, Q. *Appl Environ Microbiol* **2014**, *80*, 2941.

(154) Yin, Y.; Mao, X.; Yang, J.; Chen, X.; Mao, F.; Xu, Y. *Nucleic acids research* **2012**, *40*, W445.

(155) Meyer, F.; Paarmann, D.; D'Souza, M.; Olson, R.; Glass, E. M.; Kubal, M.; Paczian, T.; Rodriguez, A.; Stevens, R.; Wilke, A.; Wilkening, J.; Edwards, R. A. *BMC bioinformatics* **2008**, *9*, 386.

(156) Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A. A.; Dvorkin, M.; Kulikov, A. S.; Lesin, V. M.; Nikolenko, S. I.; Pham, S.; Prjibelski, A. D.; Pyshkin, A. V.; Sirotkin, A. V.; Vyahhi, N.; Tesler, G.; Alekseyev, M. A.; Pevzner, P. A. *Journal of computational biology : a journal of computational molecular cell biology* **2012**, *19*, 455.

(157) Hyatt, D.; Chen, G. L.; Locascio, P. F.; Land, M. L.; Larimer, F. W.; Hauser, L. J. *BMC bioinformatics* **2010**, *11*, 119.

(158) Kasana, R. C.; Salwan, R.; Dhar, H.; Dutt, S.; Gulati, A. *Current microbiology* **2008**, *57*, 503.

(159) Krishnan, P.; Bhat, R.; Kush, A.; Ravikumar, P. *Journal of applied microbiology*

2012, *113*, 308.

(160) Peltier, G. L.; Beckord, L. D. *Journal of bacteriology* **1945**, *50*, 711.

(161) Sokol, P. A.; Ohman, D. E.; Iglewski, B. H. *Journal of clinical microbiology* **1979**, *9*, 538.

(162) Ertugrul, S.; Donmez, G.; Takac, S. *Journal of hazardous materials* **2007**, *149*, 720.

(163) Lehmann, C.; Sibilla, F.; Maugeri, Z.; Streit, W. R.; de Maria, P. D.; Martinez, R.; Schwaneberg, U. *Green Chem* **2012**, *14*, 2719.

(164) Chernoglazov, V. M.; Jafarova, A. N.; Klyosov, A. A. *Analytical biochemistry* **1989**, *179*, 186.

(165) Chen, H. L.; Chen, Y. C.; Lu, M. Y. J.; Chang, J. J.; Wang, H. T. C.; Ke, H. M.; Wang, T. Y.; Ruan, S. K.; Wang, T. Y.; Hung, K. Y.; Cho, H. Y.; Lin, W. T.; Shih, M. C.; Li, W. H. *Biotechnology for biofuels* **2012**, *5*.

(166) Kim, S. J.; Lee, C. M.; Kim, M. Y.; Yeo, Y. S.; Yoon, S. H.; Kang, H. C.; Koo, B. S. *Journal of microbiology and biotechnology* **2007**, *17*, 905.

(167) Akcapinar, G. B.; Gul, O.; Sezerman, U. *Biotechnology progress* **2011**, *27*, 1257.

(168) Ko, K. C.; Han, Y.; Cheong, D. E.; Choi, J. H.; Song, J. J. *Journal of microbiological methods* **2013**, *94*, 311.

(169) Overbeek, R.; Begley, T.; Butler, R. M.; Choudhuri, J. V.; Chuang, H. Y.; Cohoon, M.; de Crecy-Lagard, V.; Diaz, N.; Disz, T.; Edwards, R.; Fonstein, M.; Frank, E. D.; Gerdes, S.; Glass, E. M.; Goesmann, A.; Hanson, A.; Iwata-Reuyl, D.; Jensen, R.; Jamshidi, N.; Krause, L.; Kubal, M.; Larsen, N.; Linke, B.; McHardy, A. C.; Meyer,

F.; Neuweger, H.; Olsen, G.; Olson, R.; Osterman, A.; Portnoy, V.; Pusch, G. D.; Rodionov, D. A.; Ruckert, C.; Steiner, J.; Stevens, R.; Thiele, I.; Vassieva, O.; Ye, Y.; Zagnitko, O.; Vonstein, V. *Nucleic acids research* **2005**, *33*, 5691.

(170) Schwede, T.; Kopp, J.; Guex, N.; Peitsch, M. C. *Nucleic acids research* **2003**, *31*, 3381.

(171) Rinker, K. D.; Kelly, R. M. *Applied and environmental microbiology* **1996**, *62*, 4478.

(172) Rossello-Mora, R.; Amann, R. *FEMS microbiology reviews* **2001**, *25*, 39.

(173) Kang, H. S.; Brady, S. F. *Journal of the American Chemical Society* **2014**, *136*, 18111.

(174) Cohen, L. J.; Kang, H. S.; Chu, J.; Huang, Y. H.; Gordon, E. A.; Reddy, B. V. B.; Ternei, M. A.; Craig, J. W.; Brady, S. F. *Proceedings of the National Academy of Sciences of the United States of America* **2015**, *112*, E4825.

(175) Chang, F. Y.; Brady, S. F. *Chembiochem : a European journal of chemical biology* **2014**, *15*, 815.

(176) Feng, Z. Y.; Chakraborty, D.; Dewell, S. B.; Reddy, B. V. B.; Brady, S. F. *Journal of the American Chemical Society* **2012**, *134*, 2981.

(177) Shen, B. *Current opinion in chemical biology* **2003**, *7*, 285.

(178) Hopwood, D. A.; Sherman, D. H. *Annual review of genetics* **1990**, *24*, 37.

(179) Fischbach, M. A.; Walsh, C. T. *Chemical reviews* **2006**, *106*, 3468.

(180) Parsley, L. C.; Linneman, J.; Goode, A. M.; Becklund, K.; George, I.; Goodman, R. M.; Lopanik, N. B.; Liles, M. R. *FEMS Microbiology Ecology* **2011**, *78*, 176.

(181) Foerstner, K. U.; Doerks, T.; Creevey, C. J.; Doerks, A.; Bork, P. *PloS one* **2008**, *3*.

(182) Schirmer, A.; Gadkari, R.; Reeves, C. D.; Ibrahim, F.; DeLong, E. F.; Hutchinson, C. R. *Applied and Environmental Microbiology* **2005**, *71*, 4840.

(183) Zhao, J.; Yang, N.; Zeng, R. Y. *Extremophiles : life under extreme conditions* **2008**, *12*, 97.

(184) Gillespie, D. E.; Brady, S. F.; Bettermann, A. D.; Cianciotto, N. P.; Liles, M. R.; Rondon, M. R.; Clardy, J.; Goodman, R. M.; Handelsman, J. *Applied and Environmental Microbiology* **2002**, *68*, 4301.

(185) Cecchini, D. A.; Laville, E.; Laguerre, S.; Robe, P.; Leclerc, M.; Dore, J.; Henrissat, B.; Remaud-Simeon, M.; Monsan, P.; Potocki-Veronese, G. *PloS one* **2013**, *8*.

(186) Culligan, E. P.; Sleator, R. D.; Marchesi, J. R.; Hill, C. *Isme Journal* **2012**, *6*, 1916.

(187) Gonzalez-Pastor, J. E.; Mirete, S. *Methods in molecular biology* **2010**, *668*, 273.

(188) Rabausch, U.; Juergensen, J.; Ilmberger, N.; Bohnke, S.; Fischer, S.; Schubach, B.; Schulte, M.; Streit, W. R. *Applied and environmental microbiology* **2013**, *79*, 4551.

(189) Gabor, E. M.; Alkema, W. B.; Janssen, D. B. *Environ Microbiol* **2004**, *6*, 879.

(190) Warren, R. L.; Freeman, J. D.; Levesque, R. C.; Smailus, D. E.; Flibotte, S.; Holt, R. A. *Genome research* **2008**, *18*, 1798.

(191) Wawrik, B.; Kerkhof, L.; Zylstra, G. J.; Kukor, J. J. *Applied and Environmental Microbiology* **2005**, *71*, 2232.

(192) Owen, J. G.; Reddy, B. V. B.; Ternei, M. A.; Charlop-Powers, Z.; Calle, P. Y.; Kim, J. H.; Brady, S. F. *Proceedings of the National Academy of Sciences of the United States of America* **2013**, *110*, 11797.

(193) Weber, T.; Blin, K.; Duddela, S.; Krug, D.; Kim, H. U.; Bruccoleri, R.; Lee, S. Y.; Fischbach, M. A.; Muller, R.; Wohlleben, W.; Breitling, R.; Takano, E.; Medema, M. H. *Nucleic acids research* **2015**, *43*, W237.

(194) Hadjithomas, M.; Chen, I. M. A.; Chu, K.; Ratner, A.; Palaniappan, K.; Szeto, E.; Huang, J. H.; Reddy, T. B. K.; Cimermancic, P.; Fischbach, M. A.; Ivanova, N. N.; Markowitz, V. M.; Kyrpides, N. C.; Pati, A. *mBio* **2015**, *6*.

(195) Caporaso, J. G.; Lauber, C. L.; Walters, W. A.; Berg-Lyons, D.; Lozupone, C. A.; Turnbaugh, P. J.; Fierer, N.; Knight, R. *Proceedings of the National Academy of Sciences of the United States of America* **2011**, *108 Suppl 1*, 4516.

(196) DeSantis, T. Z.; Hugenholtz, P.; Larsen, N.; Rojas, M.; Brodie, E. L.; Keller, K.; Huber, T.; Dalevi, D.; Hu, P.; Andersen, G. L. *Applied and environmental microbiology* **2006**, *72*, 5069.

(197) Wu, C.; Asakawa, S.; Shimizu, N.; Kawasaki, S.; Yasukochi, Y. *Molecular & general genetics : MGG* **1999**, *261*, 698.

(198) Borodovsky, M.; Mills, R.; Besemer, J.; Lomsadze, A. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **2003**, *Chapter 4*, Unit4 5.

(199) Guindon, S.; Dufayard, J. F.; Lefort, V.; Anisimova, M.; Hordijk, W.; Gascuel, O. *Systematic biology* **2010**, *59*, 307.

(200) Tamura, K.; Stecher, G.; Peterson, D.; Filipski, A.; Kumar, S. *Molecular biology and evolution* **2013**, *30*, 2725.

(201) Owen, J. G.; Reddy, B. V.; Ternei, M. A.; Charlop-Powers, Z.; Calle, P. Y.; Kim, J. H.; Brady, S. F. *Proceedings of the National Academy of Sciences of the United States of America* **2013**, *110*, 11797.

(202) Hiratsuka, T.; Suzuki, H.; Kariya, R.; Seo, T.; Minami, A.; Oikawa, H. *Angewandte Chemie* **2014**, *53*, 5423.

(203) George, I. F.; Hartmann, M.; Liles, M. R.; Agathos, S. N. *Applied and environmental microbiology* **2011**, *77*, 8184.

(204) Brady, S. F.; Simmons, L.; Kim, J. H.; Schmidt, E. W. *Natural product reports* **2009**, *26*, 1488.