Functional Genomics of Soil Bacteria using a Metagenomics Approach

by

Kavita S. Kakirde

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 4, 2012

Keywords: soil, metagenomics, shuttle BAC vector, metagenomic library construction, high molecular weight DNA, antibacterial compounds,

Approved by

Mark R. Liles, Chair, Associate Professor of Biological Sciences
Paul A. Cobine, Assistant Professor of Biological Sciences
Eduardus Duin, Associate Professor of Biochemistry
Omar A. Oyarzabal, Associate Professor of Biological Sciences

Abstract

Soil microbial communities are an abundant resource for natural product discovery. Traditional methods such as cultivation of soil microorganisms from soil under laboratory conditions have lead to discovery of new compounds but the vast majority of microorganisms are as yet unculturable and hence many prokaryotic phyla have yet to be explored for bioactive secondary metabolites. One of the significant breakthroughs to overcome this limitation is the application of metagenomics to investigate the genetic and functional diversity of as-yet-uncultured microorganisms from natural environments. Metagenomic analyses can provide extensive information on the structure, composition, and predicted gene functions of diverse environmental microbial assemblages. Our studies used a metagenomic approach to identify large-insert clones that express an antimicrobial activity. Bacterial artificial chromosome (BAC) vectors have been used to clone and express DNA fragments from single genomes and from entire microbial communities. Cloning and expression of large insert DNA in different host organisms can be of significance in the functional analysis and is facilitated by shuttle BAC vectors which permit the transfer and replication of BAC genomic libraries in the host organism of choice.

In the first study, we designed and constructed a novel Gram negative shuttle BAC vector that enables enables stable replication of cloned DNA in diverse Gram-negative species. This vector possesses an inducible copy system to increase the number of plasmids per cell. Thus, the vector that is maintained as a single copy can be induced by addition of arabinose thereby getting

a ~100-fold amplification of the DNA and potentially better expression of the cloned DNA due to a gene dosage affect. The pGNS-BAC vector can be used for high efficiency cloning of large fragments of genomic DNA transferred from Escherichia coli to other Gram-negative bacteria.

The second study describes screening a soil metagenomic library to identify recombinant clones producing an antimicrobial activity. Here we used a culture-independent and function based method to characterize the soil "metagenome" to access novel antibiotics of potential medical importance. Three different libraries were screened using various tester strains. After multiple rounds of screening and validation tests we identified several clones with antimicrobial activity. Clones of interest were further characterized using preliminary biochemical studies and genetic analysis.

The third study focused on detailed characterization of one of the clones (clone P6L4) identified from the screening of the large-insert library. The anti-MRSA activity derived from this clone was consistent and reproducible in all the bioassays that were performed. Basic biochemical and genetic analysis revealed that the anti-MRSA activity is likely due to the esterase produced by this clone which counteracts the action of the chloramphenicol acetyl transferase which in turn leads to growth inhibition of the MRSA by chloramphenicol.

Acknowledgments

I would like to give a heartfelt thanks to Dr. Mark Liles, who has been an excellent mentor during my research work over the past five years. This dissertation work would not be possible without his continued guidance and encouragement. I highly value his support and advice that have been vital to my development as a graduate student. Special thanks to my advisory committee members, Drs. Evert Duin, Paul Cobine, and Omar Oyzarbal for sharing their wisdom, expertise, and advice throughout the years. I would also like to thank Dr. Peter Panizzi for reviewing my dissertation. I would like to acknowledge Nancy Capps, Andrew Wiggins, Paul Bergen, Shamima Nasrin, Molly Staley, Dr. Larissa Parsley, Dr. Molli Newman, Dr. Abel Carrias, Jahangir Hossain, Chao Ran, Malachi Williams, Katherine Vest and Ann Marie Goode for their contribution to the research within this dissertation and for helping me to wade through tough times. I am thankful to all our colleagues at Lucigen Corp. for their collaborative work in these research projects. I am forever grateful to my parents Shodhan and Gauravi Kakirde, my grandmother Savita Kakirde, and my sister Namrata for their unconditional love and patience; and for the sacrifices they made to enable me to pursue my Doctor of Philosophy degree. Most importantly, I would like to thank my husband Sree Menon, who has been my strength, support and motivation through it all and has always driven me to give my best. I will always be indebted to God for giving me an opportunity to achieve my desired goals.

Table of Contents

List of Tables

List of Figures

CHAPTER I

LITERATURE REVIEW

A. Metagenomics for Characterization of Soil Microbial Communities

Metagenomic analyses can provide extensive information on the structure, composition, and predicted gene functions of diverse environmental microbial assemblages. Each environment presents its own unique challenges to metagenomic investigation and requires a specifically designed approach to accommodate physicochemical and biotic factors unique to each environment that can pose technical hurdles and/or bias the metagenomic analyses. In particular, soils harbor an exceptional diversity of prokaryotes that are largely undescribed beyond the level of ribotype and are a potentially vast resource for natural product discovery. The successful application of a soil metagenomic approach depends on selecting the appropriate DNA extraction, purification, and if necessary, cloning methods for the intended downstream analyses. The most important technical considerations in a metagenomic study include obtaining a sufficient yield of high-purity DNA representing the targeted microorganisms within an environmental sample or enrichment and (if required) constructing a metagenomic library in a suitable vector and host. Size does matter in the context of the average insert size within a clone library or the sequence read length for a high-throughput sequencing approach. It is also imperative to select the appropriate metagenomic screening strategy

to address the specific question(s) of interest, which should drive the selection of methods used in the earlier stages of a metagenomic project (e.g., DNA size, to clone or not to clone). Here, we present both the promising and problematic nature of soil metagenomics and discuss the factors that should be considered when selecting soil sampling, DNA extraction, purification, and cloning methods to implement based on the ultimate study objectives.

1. Introduction

Previous cultivation-based studies have proven soils to be an excellent resource for the discovery of novel microbial natural products (Schatz and Waksman, 1944). The discrepancy between the numbers of microorganisms visible via microscopy and the colonies obtained from laboratory cultivation is several orders of magnitude for most soils, and overcoming the "great plate count anomaly" (Staley and Konopka, 1985) in order to access a greater diversity of bacteria has become one of the most significant challenges and opportunities in the field of molecular microbial ecology. Many studies have demonstrated that the phylogenetic and functional diversity of microorganisms in various habitats, including soil, vastly exceeds the diversity of prokaryotic phyla known from cultivation (Ward et al., 1990; Hugenholtz et al., 1998; Rondon et al., 2000; Breitbart et al., 2003; Dinsdale et al., 2008). Fortunately, the recent development of metagenomic and other culture-independent approaches has enabled investigation of the functional genetic diversity of soil microorganisms without the inherent biases of cultivation.

Metagenomics can be defined as the genomic analysis of the collective microbial assemblage found in an environmental sample (Handelsman et al., 1998). There are many

variants on metagenomic approaches, which initially were dependent upon cloning of DNA from an environmental sample (Healy et al., 1995; Stein et al., 1996), but more recently many metagenomic approaches have relied upon high-throughput sequencing (Edwards et al., 2006). One of the main advantages of functional metagenomics is its ability to identify gene products from as-yet-uncultured microbes, many with no significant homolog within the GenBank database. Studies have applied a metagenomic approach to a number of different environments, such as soils (Rondon et al., 2000; Voget et al., 2003; Tringe et al., 2005), the complex microbiome of the rumen (Brulc et al., 2009), planktonic marine microbial assemblages (Beja et al., 2000a; Breitbart et al., 2002), deep sea microbiota (Sogin et al., 2006), an acid mine site (Tyson et al., 2004), arctic sediments (Jeon et al., 2009) and the Sargasso Sea (Venter et al., 2004).

This review focuses on metagenomic approaches for exploring the phylogenetic and functional diversity of soil microorganisms. Despite the promise of metagenomics as a strategy for the identification of novel natural bioactive products, xenobiotic pathways, and other metabolic processes, soils present a unique set of technical challenges for the successful isolation and analysis of metagenomic DNA. Many of the methods are labor- and cost-intensive, and the full extent of the project should be considered before embarking on a metagenomic study of a soil sample(s). A key strategic decision will be whether to adopt a sequence-only strategy or one that involves cloning of metagenomic DNA. This will be dependent upon the nature of the gene(s) or gene product(s) that are targeted, the degree of knowledge concerning these genetic loci within extant microorganisms and sequence databases, and the interest in identifying biological functions that may not be recognized from a purely sequence-driven approach.

2. Exploring the soil environment

Soil is the major component of most terrestrial environments and is considered to be the most diverse ecosystem on Earth, with respect to its native microbial populations. One gram of soil is estimated to contain millions of bacteria, archaea, viruses, and eukaryotic microorganisms (Torsvik and Ovreas, 2002; Fierer et al. 2007; Wommack et al., 2008), of which only a small percentage has been cultivated in the laboratory (Hugenholtz et al., 1998; Curtis and Sloan, 2005). From phylogenetic surveys of soil ecosystems it is known that the number of prokaryotic species in a single soil sample exceeds known cultured prokaryotes. The soil environment is an abundant yet under-characterized source of genetic diversity that has great potential to enrich our understanding of soil microbial ecology and provide enzymes and bioactive compounds useful to human society.

2.1.1 Soil composition affects microbial diversity

Soils are dynamic and heterogeneous environments in which bacteria, fungi, protozoa, and other eukaryotes compete for nutrients and space. Often, this competition leads to the production of secondary metabolites with antimicrobial activity, which may explain why the majority of previously characterized antibiotics originated from soil microbes (Burgess et al., 1999; Garbeva and de Boer, 2009). Microbes are subjected to both biotic stress (e.g., competition, parasitism) and abiotic stress (e.g., fluctuations in temperature, moisture levels, etc.), leading to a dynamic ecosystem that fosters a variety of microbial interactions and functions.

Microbial activity and growth in soils is affected by its physical, chemical, and biological properties, and as a result of microbial processes, the soil environment is dramatically

transformed in terms of its structure and chemistry (e.g., nitrogen fixation, organic matter decomposition). The physical composition (e.g., loamy, sandy, clay) of the soil will greatly influence its microbial population, as will its chemical characteristics, such as organic matter content and pH (Hassink et al., 1993). For example, the extent of bacterial diversity and number of bacteria present has been observed to be inversely related to the soil's particle size (Sessitsch et al., 2001). Although many soil microbes are mesophilic, more extreme environmental conditions or the presence of unusual contaminants may select for a distinct group of organisms, thus altering the overall community structure of that particular soil sample (George et al., 2009). In addition, the geographic location of the soil will affect phylogenetic composition and microbial growth, as temperature and moisture content will vary widely among different regions. Selecting a sampling site and method(s) is an important factor to consider when beginning a metagenomic analysis of soil microorganisms.

2.1.2 Soil sampling considerations

The depth of the soil sample will affect the number and types of microbes that are collected, as cell density is generally greater in surface soils when compared to subsurface soils. In addition, surface soils will contain phototrophic microorganisms (e.g., from the division Cyanobacteria) that will not be present at lower soil horizons (Veluci et al., 2006). In consideration of these variations, it is advisable to take multiple samples and pool the samples prior to analysis. Pooling is beneficial when a representative sample that encompasses diverse microorganisms is desired, but can be a disadvantage if the objective is to target a specific microbial population. In the latter case

it is important that the sampling site and method be selected accordingly, and to assess the presence of the targeted population by cultivation or via specific molecular probes. The depth of sampling and cross-contamination are also factors that should be considered. Soil augers are well suited for sampling because of their precision over using a shovel. Since sampling equipment may become contaminated with microbes from other layers before reaching the targeted depth, the top and/or outer layer of the sampled soil may be discarded. To prevent contamination between sampling runs, utilize a separate auger for each sample type, or the equipment can be treated with ethanol, bleach, and sterile water. After sampling it is critical to freeze or place samples on ice and process them as quickly as possible or store them at -80ºC. Soil samples that have been stored desiccated are not recommended for use, due to lower yields of cells and/or DNA.

2.2 Extraction and purification of soil microbial metagenomic DNA

When extracting metagenomic DNA from a soil sample, the first consideration is DNA size. If the goal of the study is high-throughput sequencing, PCR amplification, or small-insert clone libraries, then a harsh extraction method that results in substantially sheared, yet highly purified metagenomic DNA will be sufficient (Sections 4.3 and 4.4). Large-insert clone libraries will require adoption of an alternative DNA extraction protocol to provide sufficiently intact metagenomic DNA (Section 4.5). For any application, it is critical to isolate DNA from diverse microorganisms that are representative of the microbial assemblage; otherwise, downstream analyses may be biased against or in favor of a particular group of microorganisms (Liles et al., 2003; Feinstein et al., 2009). However, biased metagenomic libraries may be preferred, if one is targeting a consortium of microorganisms enriched for a specific functional activity

(Healy et al., 1995), in which case the relative abundance of targeted microbial taxa during the enrichment and metagenomic library construction process may be monitored. Two general approaches exist for environmental metagenomic DNA extraction, 1) DNA is directly extracted from the environmental sample; or 2) microbial cells are recovered from the environmental sample prior to lysis and DNA purification (i.e., "indirect extraction"). Direct extraction of metagenomic DNA has many advantages, including its decreased processing time and that it provides a greater DNA yield compared to other methods (Ogram et al., 1987). Unfortunately, this method often results in the isolation of a higher percentage of non-bacterial DNA (Ogram et al., 1987; Tsai and Olson, 1991; Tebbe and Vahjen, 1993). Indirect DNA extraction overcomes some limitations of the direct extraction method because it results in less non-bacterial DNA (Osborn and Smith, 2005) and, like direct extraction methods, can yield DNA from phylogenetically diverse origins (Gabor et al., 2003). However, indirect extraction methods are more time-consuming, in general provide lower DNA yields, and may bias against microorganisms that are not easily dissociated from the environmental matrix or lysed via chemical and enzymatic treatment. Selecting which extraction method to adopt depends greatly on the desired downstream application. The decrease in genomic DNA fragment size resulting from harsh direct extraction and purification methods is typically not a problem in PCR-based or pyrosequencing studies since the targeted genetic loci are of relatively small size (e.g., less than a few kilobase pairs). Conversely, the indirect extraction method is generally used when the size of extracted DNA fragments must be maintained for use in constructing large-insert metagenomic libraries, and/or a high proportion of bacterial DNA template is desired prior to the molecular application.

Because of soil's physical and chemical heterogeneity, DNA isolated from soils is often co-isolated with organic compounds that can inhibit downstream applications such as PCR and metagenomic library construction. Depending on the composition of the soil, these contaminants may include humic acids, polyphenols, polysaccharides, and nucleases, which can also degrade DNA (Tebbe and Vahjen, 1993; Zhou et al., 1996; Frostegard et al., 1999; Sylvia, 2005). The removal of these co-isolated contaminants is critical to successful DNA manipulation, and extraction and purification methods should be selected to yield DNA suitable for the ultimate metagenomic application.

3. Metagenomic Applications

Microorganisms in natural environments may contain genes that encode and express biosynthetic or biodegradative pathways of interest that have never been identified using culture-dependent methods. One strength of the metagenomic approach is in enabling researchers to investigate the phylogenetic and functional diversity of microorganisms at the community level, independent from cultivation-associated biases (Schloss and Handelsman, 2003; Cowan et al., 2005).

3.1. Natural product discovery: Enzymes

Enzymes expressed from cultured soil microorganisms have been harvested and used commercially for many decades. High-throughput screening of environmental metagenomic DNA libraries has led to the discovery of many novel enzymes that are of great use in industrial applications. Indeed, the very first metagenomic study involved the identification of cellulases from a bioreactor "zoolibrary" (Healy et al., 1995). There are many examples of enzymes discovered via a metagenomic approach, such as a multifunctional glycosyl hydrolase identified from a rumen metagenomic library

(Palackal et al., 2007), low pH, thermostable α-amylases discovered from deep sea and acidic soil environments (Richardson et al., 2002), pectinolytic lyases from soil samples containing decaying plant material (Solbak et al., 2005), agarases from soil (Voget et al., 2003) and lipolytic enzymes such as esterases and lipases (Rondon et al., 2000; Voget et al., 2003; Lee et al., 2004; Ferrer et al., 2005). In another study, 137 unique nitrilases were discovered from screening environmental (terrestrial and aquatic) DNA libraries using high-throughput and culture-independent methods (Robertson et al., 2004). A novel β-glucosidase gene isolated by screening a metagenomic library derived from alkaline polluted soil was found to be a first member of a novel family of β-glucosidase genes (Jiang et al., 2009). The discovery of a diverse set of genes that encode enzymes for cellulose and xylan hydrolysis from the resident bacterial flora of the hindgut paunch of a wood-feeding 'higher' termite (Nasutitermes sp.) and from moths was a result of metagenomic analysis (Brennan et al., 2004; Warnecke et al., 2007). In each of these studies, it should be noted that the rate of discovery is generally less than one clone with activity per 1,000 clones screened; therefore, the anticipated "hit rate" for any enzymatic activity should be considered prior to initiating metagenomic library screening. These are just a sampling of the many enzymatic activities discovered from metagenomes, providing ample evidence of the potential of this approach for the discovery of novel biocatalysts from the environment.

Mining for biocatalysts from metagenomic libraries usually involves three different strategies: 1) homology-driven metagenome mining based on high throughput sequencing, 2) substrate-induced gene expression (see Section 3.4), or 3) function-based screening. Unlike chemical synthesis, biocatalysis does not include the use of toxic

chemical reagents. The discovery of novel enzymes through these approaches is an economical and potentially environmentally responsible way to decrease the use of toxic chemicals traditionally used in many industries. This approach for enzyme discovery can help improve the efficiency of existing techniques and also enable novel processes for the production of various chemicals that serve as precursors in the synthesis of pharmaceuticals, insecticides, fertilizers, herbicides, etc.

3.2. Natural product discovery: Antibiotics

As-yet-uncultured microorganisms are an untapped reservoir for the discovery of secondary metabolites such as antibiotics (Gillespie et al., 2002). The biosynthetic pathways encoding the secondary metabolites can be captured by cloning large fragments of contiguous metagenomic DNA into heterologous hosts that are easier to manipulate in vitro, such as *E. coli* (Rondon et al., 2000; Gillespie et al., 2002; Liles et al., 2004). Many low molecular weight molecules are produced during specific growth phases such as during developmental stages or starvation (Clardy and Walsh, 2004) and exhibit bioactive properties. For many years, screening environmental microbial communities for natural products has led to the discovery proteolytic systems (Beja et al., 2000) from a variety of environmental metageomes. A diverse class of secondary metabolites is the polyketides (Moffitt and Neilan, 2003; Ginolhac et al., 2004; Schirmer et al., 2005; Wawrik et al., 2005), produced by modular enzymatic pathways with phenomenal structural heterogeneity and yet with some conserved DNA sequences that allow their identification via nucleic acid probes (see Section 5.2). Cloning and heterologous expression of environmental DNA into easily cultured bactrerial hosts has been shown to help in isolation of novel natural products and identification of the biosynthetic genes and

the mechanism of biosynthesis. This approach has been used for characrterization of isocyanide containing natural-product antibiotic 1 and identification of the first isocyanide synthase, IsnA (Brady and Clardy, 2005).

The adoption of heterologous hosts besides *E. coli* permits expression of cloned DNA from diverse sources. Streptomyces species and other Actinobacteria have been used as screening hosts for soil DNA libraries because of their ability to express diverse polyketide and other bioactive secondary metabolites and their relative ease of genetic manipulation (Martinez et al., 2005). For example, the antibiotic terragine with anti-Mycobacterium activity was discovered via heterologous expression of metagenomic clones within a Streptomyces lividans host (Wang et al., 2000). Another study introduced Type II PKS pathways, recovered from a metagenomic library, into S. livdans and S. albus hosts, resulting in the production of clone-specific metabolites (King et al., 2009). Beyond the well-characterized metabolites of Actinobacteria, many other bacterial divisions may also prove to be prodigious producers of antibiotics, and serve as alternative hosts. In a study that expressed metagenomic libraries in multiple Proteobacteria hosts, the antimicrobial products detected in each host were distinct, supporting the contention that each heterologous host may yield a novel range of expressed metabolites from a given metagenomic library (Craig et al., 2010). In another study, a PCR-based screening approach was used to analyze DNA extracted from desert soil for identifying sequences related to OxyC, which is an oxidative coupling enzyme involved in the synthesis of glycopeptide antibiotics (Banik and Brady, 2008). The same group also discovered eDNA clones producing long-chain N-acyltyrosine antibiotics after screening seven libraries constructed from different environmental samples that were

geographically distinct (Brady et al., 2004). In another recent study a soil DNA derived PKS system on functional analysis using *Streptomyces albus* as host was shown to encode unique derivative with activity against MRSA and vancomycin-resistant Enterococcus faecalis (Feng et al., 2011). Additional studies have investigated other metagenomes and have identified pathways involved in the biosynthesis of various antimicrobial compounds such as beta-lactamases (Rashamuse et al., 2009, Williamson et al., 2005) and antifungal agents (Chung et al., 2008). Other studies have discovered metagenomic clones producing triaryl cation antibiotics turbomycin A and B (Gillespie et al., 2002). This study, while using *E. coli* expression, is an example of the unique chemistry that may be derived from the combination of host metabolites (i.e., *E. coli* produced indole) and metagenomic clone chemistry (i.e., melanin pigment production).

3.3. Bioremediation

Xenobiotics include compounds such as antibiotics, pesticides, hormones, and other foreign biological or chemical contaminants that can affect a microbial community. Other examples of xenobiotics include aromatic compounds and their derivatives, and polychlorinated biphenyls (PCBs), anthropogenic chemical pollutants that persist in the environment and are recalcitrant to complete removal. Xenobiotic degradation can be achieved by biotic and/or abiotic reactions, and may be accelerated by harnessing microbial degradative activities to biostimulate or bioaugment the natural attenuation of environmental contaminants (Vogel, 1996; Cosgrove et al., 2010). The application of metagenomics may aid in the isolation of novel catabolic pathways for degradation of xenobiotic compounds, indicating the functional genetic capacity for contaminant

degradation and providing molecular tools useful for identification of the microbial taxa encoding the biodegradative gene(s).

A combined approach using metagenomics and other molecular techniques is commonly used to study microorganisms useful for bioremediation of environmental contaminants. Labeled substrates have been used to target and recover genes from populations involved in the degradation process (Sul et al., 2009). This group used [13C]-labeled biphenyl to identify biphenyl dioxygenase genes from bacteria capable of growing in PCB-contaminated river sediments. Other metagenomic studies have identified catabolic pathways that encode nitrilases, which play an important role in both biosynthetic and catabolic reactions (Robertson et al., 2004), as well as enzymes with catalytic properties that degrade organic contaminants (Kim et al., 2007).

3.4. Strategies to improve the isolation of biosynthetic and catabolic pathways

Extraction of total metagenomic DNA and cloning to construct libraries requires extensive labor, time, and resources. The number of positive clones obtained from screening these libraries for the presence or expression of a specific gene or function is often very low because the target pathways comprise a small percentage of the total cloned DNA, and only a subset of the cloned genes may be expressed in a given heterologous host. There are various strategies that can be employed prior to library construction and/or screening that can improve the frequency of biosynthetic or catabolic pathway isolation. Although these methods may result in a loss of considerable diversity from the environmental sample, they also have the power to select for a particular population or function(s) of interest. The loss of diversity can be mitigated by altering the degree of the selective pressure criteria used.

A commonly applied strategy is to enrich the environmental sample for microbial populations capable of growth on certain substrates or for survival under different physico-chemical conditions. Use of a selective medium will result in favorable growth and enrichment of the targeted population due to specific substrate utilization, as well as potentially other metabolically co-dependent microbial populations. Direct cloning from enrichment cultures enables studying metabolic activities of microbial assemblages and selection for specific microbes that produce an enzyme or compound of interest (Healy et al., 1995). This approach has also been used in the identification of biotin synthesis genes by isolation of clones carrying the biotin biosynthesis operon (Entcheva et al., 2001). Stable isotope probing (SIP) is an approach that enriches the DNA (or RNA) of microorganisms that can utilize a stable isotope (e.g., ^{13}C-glucose) and incorporate the isotope into newly synthesized nucleic acids (Radajewski et al., 2000; Dumont et al., 2006). The isolated "heavy" DNA from the treated environmental sample is subjected to density gradient centrifugation to separate the ^{13}C-labeled DNA for analysis, which may then serve as DNA template within a PCR to identify the microorganisms that have incorporated the labeled substrate (Radajewski et al., 2000). Metagenomic analysis in conjunction with SIP can access a multitude of functional genes since the labeled DNA is enriched for the genomes of microbial populations with specific metabolic capabilities (Wellington et al., 2003). DNA-SIP has also been used to retrieve genomic fragments of an active population by cloning the ^{13}C-labeled DNA without initial PCR amplification (Dumont et al., 2006). However, SIP has its limits and biases, such as dilution of the labeled substrate with unlabeled substrates and cross-feeding of ^{13}C-labeled metabolic intermediates by other organisms (Radajewski et al., 2000). When using DNA-SIP for

metagenomic analyses, the small amount of heavy DNA available can also be a hurdle to successful library construction. To overcome this challenge, methods such as multiple-displacement amplification (Dumont et al., 2006; Chen et al., 2008) and community growth enrichment by sediment slurries (Kalyuzhnaya et al., 2008) have been used to increase the amount of heavy DNA available for analysis. Despite these challenges, SIP coupled with metagenomics is an excellent culture-independent strategy to identify functional genes involved in the utilization of a variety of compounds or in degradation of environmental pollutants.

Another promising approach for identification of catabolic pathways has been described as substrate-induced gene expression screening (SIGEX). SIGEX identifies clones from an operon-trap metagenomic library that are induced in the presence of a specific substrate, resulting in green fluorescence protein expression that can be detected using fluorescence-activated cell sorting (Uchiyama et al., 2005). There are limitations of the SIGEX approach due to its dependence on in cis regulatory factors that are active within E. coli (de Lorenzo, 2005), so this rapid screen should not be considered an exhaustive survey, as is the case with any metagenomic analysis.

3.5. Metagenomics in quorum sensing regulation studies

Quorum sensing (QS)-mediated bacterial responses to cell density are specific to each bacterial species, and are important in understanding bacterial pathogenesis and other bacterial phenotypes in natural environments (e.g., bioluminescence of Vibrio fischeri within the light organ of the Euprymna squid). The use of a metagenomic approach to study QS regulation in the soil environment was pioneered by Williamson et al., wherein they identified clones producing unknown molecules that activated QS-

regulated genes (Williamson et al., 2005). Clones of interest were identified using a high throughput intracellular screen, i.e. the metagenomic DNA is within a host cell that contains a biosensor responsive to compounds inducing QS. Another study identified a clone that degraded N-acylhomoserine lactone (NAHL) from screening a pasture soil metagenomic library (Riaz et al., 2008). The identified gene was shown to encode a lactonase with NAHL degrading ability and the gene product efficiently quenched quorum-sensing-regulated pathogenic functions when expressed in Pectobacterium carotovorum. Metagenome-derived clones isolated in another study were found to encode novel lactonase family proteins interfering with QS (Schipper et al., 2008), that when expressed in *Pseudomonas aeruginosa* successfully inhibited motility and biofilm formation. Lastly, metagenomic libraries constructed with DNA isolated from activated sludge and soil have been screened using an *Agrobacterium* biosensor strain, resulting in the isolation of three unique clones with novel QS synthase genes (Hao et al., 2010).

4. Analyzing the soil metagenome

A variety of approaches may be employed for analyzing the soil metagenome, depending on the specific aims of the study. The ultimate downstream application should dictate the methods used for soil sampling, DNA extraction and purification, and library construction and screening (if necessary). The biologically, chemically, and physically heterogeneous nature of soils presents many challenges to the successful characterization of its microbial metagenome. Representative coverage of the soil microbial community requires isolation and cloning of a large amount of DNA from a small sample, and depends on insert size and the number of clones. It has been estimated that the number of plasmid clones (5 kb average insert size) and BAC clones (100 kb

average insert size) required for representative coverage of the diverse soil microbial community in one gram of soil is 10^7 and 10^6 respectively (Handelsman et al., 1998). This is of course based on the assumption that all species in a soil environment are equally abundant. Since members of a community are rarely equally represented, the metagenomic library with minimum coverage is more likely to represent only the abundant species. In order to achieve substantial representation of the genomes from rare members of the soil community, a 100- to 1000- fold coverage of the metagenome is needed in library construction (Riesenfeld et al., 2004). Since this translates to about 10,000 Gb of soil DNA, or 1011 BAC clones, it is not reasonable to suppose that bacterial taxa present in lower abundance will be represented within a metagenomic library unless an enrichment method is used. Also, when working with soil samples that have not been well-characterized, it is advisable to utilize a variety of different methods for DNA extraction and purification to empirically determine the ideal combination that will yield high-quality and high-diversity metagenomic DNA. Here, we discuss many of the metagenomic-based approaches used to study soil microbiology, as well as the approach-specific factors to consider when performing such analyses.

4.1. Sequencing

The use of PCR has become routine for molecular phylogenetic analysis based on ribotype diversity (Woese, 1987), often used in combination with community analysis methods such as denaturing gradient gel electrophoresis (e.g., Muyzer et al., 1993), 16S rRNA gene clone libraries (e.g., Chandler, 1997), or more recently microarrays (DeSantis et al., 2007; Liles et al., 2010). Although in many cases such studies are described as "metagenomic", since indeed the template DNA used is derived from diverse genomes,

such phylogenetic surveys of a single evolutionarily conserved gene are not truly metagenomic in nature and will not be further considered in this review.

Pyrosequencing and other next-generation approaches offer the capacity for massively parallel sequencing of metagenomic samples (Ronaghi, 2001). The accuracy of pyrosequencing is comparable to that achieved via Sanger sequencing (Huse et al., 2007), but it is more cost- and time-effective per sequenced nucleotide (Hugenholtz and Tyson, 2008), and sequencing read length has been gradually increasing with each iteration of sequencing technologies (Margulies et al., 2005). The increased availability of high-throughput sequencing technologies has made it possible for scientists to gain access to the genetic diversity within environmental communities (Sogin et al., 2006). Pyrosequencing has been used in the investigations of microbial diversity in soil (Roesch LF, 2007), deep sea ecosystems (Sogin et al., 2006) and phage populations from various environments (Dinsdale et al., 2008).

Because pyrosequencing relies on an amplification process, the same environmental contamination challenges that apply to PCR-based applications also apply to pyrosequencing. However, since pyrosequencing currently generates reads only 300-500 bp in length, obtaining intact, larger DNA is not critical (Metzker, 2005). Most commercially available soil DNA extraction methods yield DNA suitable for pyrosequencing, and if needed, further DNA purification methods can be employed (see section 4.2.1). Alternatively, the DNA template can be diluted (along with contaminants) to permit PCR amplification (Altshuler, 2006), or bovine serum albumin can be added to the reaction mixture to prevent humic acid-mediated inhibition (Kreader, 1996).

Read length is a critical factor in the probability that a metagenomic sequence will have a significant hit within GenBank or other database (Wommack et al., 2008). Even for a pure bacterial culture, it is not uncommon for a completely sequenced bacterial genome to have 35% to 45% of predicted open reading frames (ORFs) with no significant homolog in GenBank (Schwartz, 2000). This problem is only exacerbated with metagenomic sequences, with an even larger proportion of metagenomic sequences from soil and other environments having no significant BLAST homolog (Venter et al., 2004; Tringe et al., 2005; Pignatelli et al., 2008). Even with the difficulty in interpretation of much of the sequences within metagenomic datasets, substantial information related to the genomic composition, and predicted functions and metabolic pathways, of microbial communities has been unearthed from deep-sequencing approaches (Breitbart et al., 2002; Tyson et al., 2004).

4.2 Small-insert libraries

The construction and analysis of small-insert metagenomic libraries (less than ~10 kb average insert size) is a useful approach to identify gene product(s) encoded by a relatively small genetic locus, such as most enzymes, or genetic determinants of antibiotic resistance (Reisenfeld et al., 2004a; Parsley et al., 2010). Biases in cell lysis and cloning techniques may select against some prokaryotic taxa or gene products that are toxic to the host cell; therefore, it is important to select DNA extraction and cloning methods designed to yield a high proportion of DNA from the microorganisms of interest. Refer to Figure 1 for a schematic representation of the overall steps involved in metagenomic library construction.

4.2.1 Selection of vectors and host organisms

Vectors used for the construction of small-insert libraries often possess a promoter for transcription of the cloned gene inserts and should be compatible with the host selected for screening. A vector with two promoter sites flanking the multi-cloning site facilitates gene expression that is independent of gene orientation and the promoters associated with inserts (Lammle et al., 2007). With the possibility of the expressed gene product having toxic effects on the host organism, it is important to regulate the expression levels of the cloned genes, which can be achieved by using vectors with inducible control over gene expression of the insert or plasmid copy number (Sukchawalit et al., 1999; Saida et al., 2006).

An additional issue to consider when selecting a vector is its ability to replicate in multiple hosts to enable heterologous expression of specific gene(s) of interest. Although the utility of using *E. coli* as a heterologous host for metagenomic library construction has been well-established (Rondon et al., 2000; Pfeifer and Khosla, 2001; Gillespie et al., 2002; Liles et al., 2004), other bacterial hosts may be more suitable for some applications, particularly if the percent G+C content of the cloned gene(s) are significantly different from that of *E. coli*, or if the regulatory factors required for expression or the biosynthetic capacity may be enhanced within another prokaryote.

4.2.2 Preparation of DNA for cloning

The preparation of DNA for small-insert libraries is similar to that used for PCR- or pyrosequencing-based applications. A sufficient yield of DNA is necessary for successful library construction, and soil contaminants co-isolated with the DNA such as humic acids can interfere with efficient cloning. DNA extraction and purification

conditions should be harsh enough to lyse a variety of microbes and remove the majority of contaminants, while the degree of DNA fragmentation that is permissible will depend on the desired average insert size of the library. If the desired average insert size is less than 20 kb, a commercial kit (e.g., MoBio Laboratories, Qiagen) may provide a useful method for obtaining DNA of sufficient size, purity, and yield for small-insert cloning. One study using Antarctic top soil used two separate commercial kits to further purify the DNA after cell lysis for construction of small-insert libraries (Cieslinski et al., 2009). In cases when commercial kits are not suitable, such as soils with high clay content, it may be advisable to adopt cell-based ("indirect extraction") methods such as sucrose/Percoll density gradient centrifugation or Nycodenz treatment, which have been shown to generate DNA appropriate for small-insert cloning (Bakken and Lindahl, 1995).

Regardless of which DNA extraction method is used, it is possible that further purification will be required for efficient cloning. Many DNA purification methods may be effective in yielding DNA suitable for cloning, such as phenol and chloroform extraction, and/or treatment with hexadecyltrimethylammonium bromide (CTAB) or polyvinylpolypyrrolidone (PVPP), which may be combined with CsCl density centrifugation or hydroxyapatite column chromatographic purification (Holben et al., 1988; Selenska and Klingmuller, 1991; Knaebel and Crawford, 1995; Roose-Amsaleg et al., 2001; Lee et al., 2004). However, it has been shown that many of these methods (i.e., PVPP addition, CsCl density centrifugation, and hydroxyapatite column chromatographic purification) resulted in a decreased DNA yield (Steffan et al., 1988). In the case of indirect extraction methods, some studies have found that a washing step prior to cell lysis is useful for the removal of soluble inhibitors and extracellular DNA (Xia et al.,

1995; Harry et al., 1999). Unfortunately, many soil samples require a combination of these purification steps, which significantly increases processing time and can lead to an even greater loss of DNA. For example, one study compared DNA extracted from five different soils with various organic matter contents and found that the samples with the highest organic matter content required five purification steps to yield sufficiently pure DNA (Van Elsas et al., 1997). Following extraction and purification of the DNA, it may be physically sheared or partially restriction digested and then size-selected by extracting the DNA in the desired size range from an agarose gel (Riesenfeld et al., 2004; Lammle et al., 2007). Because the size-selected DNA will likely be less than 20 kb, it can be column-purified, the gel slices may be treated with GELase enzyme (Epicentre), or the DNA may be electroeluted from the gel prior to cloning (Osoegawa et al., 1998).

4.3 Large-insert libraries

Large-insert metagenomic libraries contain large, contiguous DNA fragments that have the potential to contain intact biosynthetic pathways involved in the synthesis of antimicrobial compounds, multiple enzymes with catabolic activity, or operons encoding other complex metabolic functions. However, along with potential advantages for some applications, large-insert cloning from soil microorganisms also presents many technical challenges in order to obtain and screen high-quality metagenomic libraries containing DNA from representative microorganisms.

4.3.1 Selection of vectors and host organisms

Because the applications appropriate for large-insert metagenomic libraries depend on their ability to capture large, intact genetic pathways, the selection of an appropriate cloning vector is critical to the maintenance and expression of the cloned

pathways. Several vector options exist for cloning HMW DNA from environmental samples, such as cosmids, fosmids, and BACs. The cosmid, a hybrid plasmid that contains cos sequences from the λ phage genome; was one of the first vectors used for cloning (Collins and Hohn, 1978). The packaging capacity of cosmids varies depending on the size of the vector itself but usually lies around 40-45 kb. While typical plasmids can maintain inserts of 1-20 kb, cosmids are capable of containing DNA inserts of about 30 kb up to 40 kb. The size limits ensure that vector self-ligation resulting in empty clones is not a problem. Both broad host range cosmids and shuttle cosmids are available (Craig et al., 2009). Cosmids can replicate like plasmids when they contain a suitable origin of replication and they commonly possess selective genes such as antibiotic resistance to facilitate screening of transfected cells. Fosmid vectors, which are similar to cosmids but are based on the *E. coli* F-factor replicon, were developed for constructing stable libraries from complex genomes (Kim et al., 1992). The low copy number of fosmid vectors offers higher stability than comparable high-copy number cosmids. A low copy number is optimal for long term survival of the plasmid in a host. Also, plasmid copy number determines gene dosage. Recombinant clones from large-insert libraries may express gene products that are toxic to the host and hence it is important to maintain libraries in single copy until screening for a function. Fosmid copy number is tightly regulated in E. coli to 1-2 copies per cell, and fosmids can typically accommodate cloned inserts between 40 and 50 kb. BAC vectors are based on the same F-factor replicon but have the capacity to maintain large inserts in excess of 100 kb (Shizuya et al., 1992). Along with the long-term stability conferred by the F-factor for maintenance, a modified BAC vector also containing an RK2 origin of replication is capable of inducible copy-

number, alternating between single-copy and high-copy BAC maintenance (Wild et al., 2002). The inducible-copy phenotype can have significant advantages for the yield of DNA from metagenomic clones, and potentially for expression of cloned genes.

Although fosmid vectors are limited in insert size compared to BAC vectors, their significantly higher cloning efficiency enables construction of metagenomic libraries with many thousands of transductants. Conversely, BAC vectors even though capable of accommodating higher insert sizes have lower cloning efficiency than that of fosmid vectors. As mentioned previously, HMW DNA for fosmid-based cloning may be treated with harsher extraction and purification methods, which could yield a higher concentration of DNA from more diverse microorganisms than that of DNA isolated for BAC-based cloning. However, because BAC vectors can stably maintain cloned inserts hundreds of kilobases in size, they offer a greater chance of isolating intact pathways or of linking phylogenetic and functional genetic information (Stein et al., 1996). Therefore, the predicted size of the pathway of interest, its native level of activity, and its relative abundance within the community must be considered when choosing a suitable cloning vector for large-insert metagenomic library construction.

As with small-insert libraries, *E. coli* is the preferred host for the construction of large-insert metagenomic libraries due to its high cloning efficiency. This host has been successfully used to express many bioactive enzymes and compounds in metagenomic studies (Handelsman et al., 1998; Heath et al., 2009). In addition, *Streptomyces lividans* has been used as a heterologous host for library screening, and it has more stringent promoter recognition and regulation properties when compared to *E. coli* (Martinez et al., 2005). Because large-insert libraries may contain clones that express gene products that

are toxic to the library host, it is important to maintain libraries in single copy until screening for a function and to consider the use of multiple hosts to increase the probability of identifying and characterizing the function(s) of interest. It has been shown that clones positive for a specific activity detected using one host may not be detected in a different host and vice-versa (Li and Qin, 2005; Wang et al., 2006; Craig et al., 2009; Craig et al., 2010). A range of Gram-positive and -negative bacteria can be used as hosts for heterologous expression, and the corresponding vectors selected should be compatible with those hosts (Sosio et al., 2000; Martinez et al., 2004; Hain et al., 2008). Vector systems such as pRS44 enable shuttling into other Gram-negative hosts and have higher potential for function-based screening across species barriers and heterologous gene expression (Aakvik et al., 2009). Several other factors are necessary for successful expression of the cloned pathways (e.g., co-factors, post-translation modification enzymes, inducers, chaperones etc.), which may be provided by the vector or the host organism.

4.3.2 Preparation of DNA for cloning

Large-insert metagenomic libraries are the most challenging to construct, but also can provide significant advantages for some applications since they enable identification and characterization of intact functional pathways encoded on large, contiguous DNA fragments (Stein et al., 1996; Beja et al., 2000b; Rondon et al., 2000; Courtois et al., 2003). All of the considerations discussed previously regarding the selection of DNA extraction and purification methods apply to large-insert cloning, along with an additional critical issue: the construction of large-insert metagenomic libraries depends on obtaining sufficiently pure DNA of high molecular weight (in excess of ~100 kb). However, most

extraction and purification methods result in DNA significantly smaller than this size (Tien et al., 1999; Wellington et al., 2003; Miller and Day, 2004). Although a few methods can yield DNA from soil greater than 1 Mbp in size (Berry et al., 2003; Liles et al., 2004), it has been demonstrated that these indirect extraction methods can result in inefficient cloning due to contaminants that may be co-isolated with the metagenomic DNA and require further purification.

The successful recovery of high molecular weight (HMW) metagenomic DNA from soil microorganisms presents many extraction and purification challenges. A primary goal is to obtain DNA from an assemblage of diverse bacterial cells that are representative of the soil microbial community DNA. However, the harsh extraction methods (i.e., bead-beat lysis) typically employed for PCR or small-insert cloning applications will result in substantially fragmented DNA that is much too small for large-insert cloning. The use of indirect DNA extraction methods can somewhat alleviate this dilemma by first separating the cells from the soil sample, embedding them in an agarose plug, and then carefully lysing the cells and purifying the resulting DNA rather than performing the extraction *in situ*. Repeated homogenization and differential centrifugation are often sufficient to separate the cells from the soil sample (Faegri et al., 1977; Hopkins et al., 1991), although other dispersion methods include the use of cation-exchange resin (Macdonald, 1986; Jacobsen and Rasmussen, 1992) and incubating the soil with sodium deoxycholate or polyethylene glycol (Liles et al., 2008). Another novel method that is capable of selectively concentrating DNA within a gel while rejecting high concentrations of contaminants is SCODA (Pel et al., 2009), but the quantities of DNA

capable of being extracted may not be sufficient for cloning without further amplification.

The choice of extraction and purification method also depends on which cloning vector will be employed, such as a fosmid or bacterial artificial chromosome (BAC). Metagenomic libraries constructed in a fosmid vector are introduced into their heterologous host using a λ phage-based packaging system, which limits the clone insert size to 40-50 kb. Although DNA isolated for fosmid libraries must be treated carefully to prevent excessive shearing of DNA, using a fosmid vector does allow the use of harsher extraction and purification methods than those that may be used for BAC cloning. Also, during fosmid library construction, the DNA is typically size-selected by physically shearing the DNA into fragments of a desired length rather than by restriction digestion. This "direct size-selection" method eliminates the need for gel extraction (which can lead to DNA loss) and the possibility of DNA degradation due to over-digestion. An alternative to the physical shearing method was proposed by Quaiser and colleagues, who constructed fosmid libraries containing soil metagenomic DNA contaminated with humic and fulvic acids by embedding the DNA in agarose, electrophoresing the DNA through agarose containing PVP, and then combining the subsequent removal of the PVP with the size-selection step which resulted in purified, "clonable" DNA in the 30-100 kb size range (Quaiser et al., 2002). In combination with other purification steps, the inclusion of a formamide plus NaCl treatment was shown to significantly increase the efficiency of cloning of large DNA fragments into fosmid or BAC vectors (Liles et al., 2008). Factors that have been demonstrated to affect the size of recovered DNA include not only the DNA extraction method used but also the microbial growth status and chemical

composition of the soil (Bertrand et al., 2005). In general, DNA extracted from bacterial cells is significantly larger than DNA directly extracted from soil but is also found in lower yields (Liles et al., 2008); however, this loss can be reduced by using wide-bore pipette tips to prevent shearing of DNA, performing multiple rounds of indirect extraction on each soil sample, minimizing the amount of agarose that is retained during size selection, or using electroelution as an alternative to extraction of DNA from the agarose gel (Osoegawa et al., 1998).

5. Metagenomic library screening

The analysis of metagenomic libraries involves two main strategies, function-based or sequence-based screening. The choice of screening method depends on many factors, including the type of library constructed, the genetic loci or functional activity of interest, and the time and resources available to characterize the library. Both approaches offer advantages and disadvantages, which will be discussed here.

5.1 Function-based screening

Function-based methods involve screening a metagenomic library to detect the expression of a particular phenotype conferred on the host by cloned DNA (Henne et al., 1999). Because the frequency of discovering active pathways from metagenomic libraries is often low, high-throughput screening of library clones is the most efficient approach for function-based detection of activity. By screening on indicator media, *E. coli* recombinant clones that express a novel phenotype (not already encoded on the *E. coli* genome) may be recognized. As opposed to high-throughput screening methods, a direct selection for a positive clone that has acquired resistance to an antibiotic or heavy metal can be performed by excluding microorganisms that are unable to grow in the

presence of these selective compounds (Riesenfeld et al., 2004; Mirete et al., 2007; Parsley et al., 2010).

Another approach for functional screening of metagenomic libraries is to use host strains or mutants of host strains that require heterologous complementation for growth under selective conditions (Simon and Daniel, 2009). Growth is exclusively observed in the case of recombinant clones that possess the gene of interest and produce an active product. This strategy has been applied for the detection of enzymes involved in poly-3-hydroxybutyrate metabolism (Wang et al., 2006), DNA polymerase I (Simon et al., 2009), operons for biotin biosynthesis (Entcheva et al., 2001), lysine racemases (Chen et al., 2009), glycerol dehydratases (Knietsch et al., 2003) and naphthalene dioxygenase (Ono et al., 2007).

Screening can also be performed by detecting a specific phenotypic characteristic, in which individual clones are assayed for a particular trait. Incorporation of specific substrates in the growth medium will allow the identification of the corresponding enzymatic activity encoded by a metagenomic clone(s). Examples include the identification of esterases (Elend et al., 2006; Chu et al., 2008) by formation of a clear halo around a colony on the indicator medium and the identification of extradiol dioxygenases by the production of a yellow compound (Suenaga et al., 2007). Metagenomic clones expressing an antimicrobial activity may be detected by growth inhibition assays of a suitable tester organism using soft agar overlays over the clone colonies or a microtiter plate assay using the supernatant extracts from the clone cultures (Rondon et al., 2000; Courtois et al., 2003; Brady et al., 2004; Craig et al., 2009). As

discussed previously, SIGEX is an additional functional screening approach in order to identify genes for substrate catabolism.

Although function-based screening is a powerful tool to identify novel natural products or metabolic activities from as-yet-uncultured organisms, it is often limited by a number of obstacles that may be difficult to overcome. Detecting a recombinant clone that expresses a gene product will depend upon successful gene transcription, translation, protein folding, and secretion from the host organism. By adopting high-throughput screening protocols and multiple heterologous expression hosts, the probability of discovering the function(s) of interest may be improved.

5.2 Sequence- based screening

Sequence-based screening involves direct sequencing of metagenomic DNA, either with or without cloning prior to sequencing and then subjecting the sequences to bioinformatic analyses (Kunin et al., 2008; Sleator et al., 2008). Practically speaking, a sequence-only approach to metagenomics involves significantly less laboratory bench work, relative to cloning-based approaches. Recent developments in next generation sequencing (NGS) technologies have made a available a number of methods that can be used for sequencing, although with varying costs and capabilities. GS20, the first instrument based on the 454 pyrosequencing technology was shown to sequence up to 25 million bases of a bacterial genome in a four hour run, with average read lengths of 110 bp and 96% raw read accuracy (Margulies et al., 2005). A current model 454 GS-FLX sequencer using Titanium chemistry can achieve read lengths of up to 500 bp, with future improvements in read length expected. By comparison, the Illumina Solexa platform based on fluorescently labeled sequencing by synthesis generates 35 to 76 bp on average.

The latest version of the short read sequencer from Applied Biosystems, called the SOLiD4, generates 100 Gb per run with read length of 50 bp. Though NGS technologies provide good overall coverage for single genomes, the short read lengths can be a serious limitation for efficient assembly of metagenomic sequences. The cost per megabase is highest for 454 sequencing at approximately $10 followed by Solexa and SOLiD at about $5 and $2, respectively (Rothberg and Leamon 2008). With the rapid growth and developments in this field it is very likely that the cost and read estimates will keep changing as NGS technology advances. The selection of a metagenomic strategy should be informed by the degree to which the gene(s) of interest are expected to be identified from a sequence-only approach; the interest (or lack thereof) in obtaining functional cloned genes, and the availability of time and resources for the project. As the cost per base pair of sequence has dropped dramatically through adoption of NGS technology, this has enabled large scale sequencing efforts accessible to individual academic researchers. Still, sequence data analysis can consume more time and resources than are initially anticipated. Fortunately, bioinformatics approaches to analyze metagenomic datasets have been developed that allow rapid comparative analyses.

With the enormous amount of sequence data generated by these different approaches, it is very important to have bioinformatics tools for such high-throughput sequence pipelines. Metagenomic studies must first curate the sequence reads to obtain data of sufficient quality, eliminating ambiguous base pairs and any vector or adaptor sequences. The edited sequences can then be used for gene prediction, and if desired, contig assembly. Given the non-exhaustive nature of most metagenomic sequence datasets, especially for analysis of soil communities, it is expected that contig assembly

will be of limited benefit. For very diverse microbial assemblages, contig sizes will be relatively short, and chimeric contigs will likely be present at a high frequency. Once high-quality metagenomic sequences are available, they can be deposited within sequence databases (e.g, GenBank env) and compared against other environmental metagenomic datasets. A useful tool for accessing metagenomic information is CAMERA (Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis), developed to serve the needs of the microbial ecology research community by creating a data repository and a bioinformatics resource to facilitate metagenomic sequence data storage, access, analysis, and synthesis (Smarr, 2006). A freely available open source system that can process metagenome sequence data is the metagenomics RAST server (MG-RAST)(Meyer et al., 2008). The MG-RAST server compares protein as well as nucleotide databases for functional assignments of sequences in the metagenome accompanied by a phylogenetic summary. Just like next generation sequencing technology enabled generation of vast amount of sequence data, tools like MG-RAST have enabled high-performance computing for annotation and analysis of metagenomes.

There are available bioinformatics tools for gene prediction, such as MEGAN (MEtaGenome ANalyzer), a program that compares a set of DNA reads (or contigs) against databases of known sequences using comparative tools such as BLAST algorithms. MEGAN can then be used to compute and interactively explore the taxonomical content of the dataset by using NCBI taxonomy to summarize and order the results (Huson et al., 2007). Once a dataset of metagenomic sequences with significant GenBank hits has been assembled, these sequences can then be categorized by a

subsystems approach using SEED to organize predicted gene functions according to related biological processes (Overbeek et al., 2005). SEED enables rapid annotation of metagenomic sequences according to similarity to previously known gene products. The predicted genes may also be assigned a phylogenetic classification using *Treephyler* for rapid taxonomic profiling of metagenomic sequences (Schreiber et al., 2010).

With each of these bioinformatics tools and approaches, it should be acknowledged that the predictive power of the sequence analysis is limited by the previously described gene functions available in public databases and that many putative functions may be inaccurately annotated. While this potential source of bias does affect the utility of a sequence-based approach to metagenomics, such intensive sequence-driven surveys of natural environments have profoundly affected our collective view of prokaryotic diversity and the extent of functional genetic diversity that has yet to be understood in terms of biological functionality (Venter et al., 2004).

6. Conclusions

The development of metagenomic approaches has provided an unprecedented level of access to microbial genomes from many different environments, making it possible to characterize the phylogenetic and functional diversity of as-yet-uncultured microorganisms from various biomes of interest. Because of its complex and dynamic nature, soil presents unique challenges for metagenomic applications. Selecting the most suitable combination of soil sampling, DNA extraction and purification, cloning and/or sequencing method that is most appropriate for the metagenomic study should begin with consideration of the ultimate desired outcome, for an application-driven approach to soil metagenomics.

The use of cutting-edge metagenomic-based technologies to access soil microbial communities has led to a remarkable increase in the discovery of pathways that encode diverse gene products, such as enzymes and antimicrobial compounds. Soils are expected to be a continuing rich resource of novel genetic and functional pathways of use and interest to academia and industry.

B. Anbtibiotics: Modes of Action

1. Gentamicin and Kanamycin

Gentamicin and Kanamycin belong to the class of antibiotics referred to as aminoglycosides that interfere with bacterial protein synthesis. Aminoglycosides bind to the 30S ribosomal subunit inducing a significant increase in misreading of messenger RNA (Davies and Davis, 1968), resulting in the bacterial inability to synthesize proteins vital for growth. They are known to inhibit ribosomal translocation where the peptidyl-tRNA moves from the A-site to the P-site (Davies *et al*., 1965; Cabanas *et al*., 1978; Misumi *et al*., 1978). Both gentamicin and kanamycin irreversibly bind to specific 30S subunit proteins and 16S rRNA leading to interference with the initiation complex and misreading of mRNA. The induced mistranslation results in insertion of incorrect amino acids into the polypeptide making it toxic or non-functional.

2. Nalidixic acid

Nalidixic acid is the first of quinolone antibiotics, a family of synthetic antibacterial drugs. It is a broad spectrum antibiotic that is bacteriostatic at lower concentrations and bacteriocidal at higher concentrations. It binds to the A subunit of

DNA gyrase preventing unwinding of the bacterial DNA and interfering with DNA replication and transcription (Crumplin and Smith, 1976; Gellert *et al*., 1976).

3. Chloramphenicol

Chloramphenicol is a broad spectrum antibiotic and is mainly bacteriostatic. It may be bacteriocidal when used at high concentrations or against highly susceptible organisms. It binds to the 23S rRNA of the 50S subunit of bacterial ribosomes preventing peptide bond formation (Wisseman et al., 1953, 1954;). It suppresses the activity of peptidyl transferase thus preventing transfer of amino acids to growing peptide chains and inhibiting protein synthesis (Gale and Folkes, 1953).

4. Vancomycin

Vancomycin is a glycopeptide antibiotic used to treat infections caused by Gram-positive bacteria. Vancomycin inhibits cell wall biosynthesis and assembly (Jordan and Reynolds, 1967). It prevents N-acetylmuramic acid (NAM) - and N-acetylglucosamine (NAG)-peptide subunits from being incorporated into the peptidoglycan matrix by binding to the terminal D-alanyl-D-alanine moieties of the nascent peptidoglycan chains via hydrogen bonds (Nieto and Perkins, 1971). The result is prevention of cell wall synthesis in two different ways, by blocking polymerization and preventing the backbone polymers already formed from cross linking with each other (Perkins and Nieto, 1972; 1973; 1974). The cell wall falls apart due to the inability of the peptide chain to interact properly with the cross linking enzyme. Vancomycin also leads to alteration in bacterial-cell-membrane permeability and RNA synthesis.

5. Ampicillin and Methicillin

Ampicillin and Methicillin belong to the penicillin group of beta-lactam antibiotics, a broad class of bacteriocidal antibiotics that includes all antibiotic agents containing a beta-lactam ring in their molecular structures. Beta-lactam antibiotics inhibit the synthesis of the peptidoglycan layer of bacterial cell walls (Blumberg and Strominger, 1974). The last step of peptidoglycan synthesis is facilitated by transpeptidases known as penicillin-binding proteins (PBPs) that are located inside the cell wall. Beta-lactam antibiotics irreversibly bind to the active site of PBPs and competitively inhibit them thus preventing the final crosslinking of the peptidoglycan layer and disrupting cell wall synthesis (Izaki *et al*., 1968; Waxman and Strominger, 1983). The resulting build up of peptidoglycan precursors triggers the activation of cell wall autolytic enzymes (hydrolases) that mediate cell lysis. Ampicillin is a broad-spectrum antibiotic and Methicllin is a narrow-spectrum antibiotic of the penicillin class. Methicillin is no longer clinically used, being replaced by more stable penicillins and the term methicillin-resistant *Staphylococcus aureus* (MRSA) may refer to *S. aureus* strains resistant to all penicillins. The two main modes of resistance mechanisms are enzymatic hydrolysis of the beta-lactam ring by the enzyme beta-lactamase or penicillinase and alteration in PBPs (Bycroft and Shute, 1985). The latter is the premise of resistance in MRSA. Beta-lactams are unable to bind the altered PBPs and are not as effective in disrupting cell wall synthesis. The altered PBPs differ from other PBPs in that the active site does not bind to methicillin or other beta-lactams. The transpeptidation reaction proceeds normally enabling cell wall synthesis in the presence of antibiotics (Lowy, 2003).

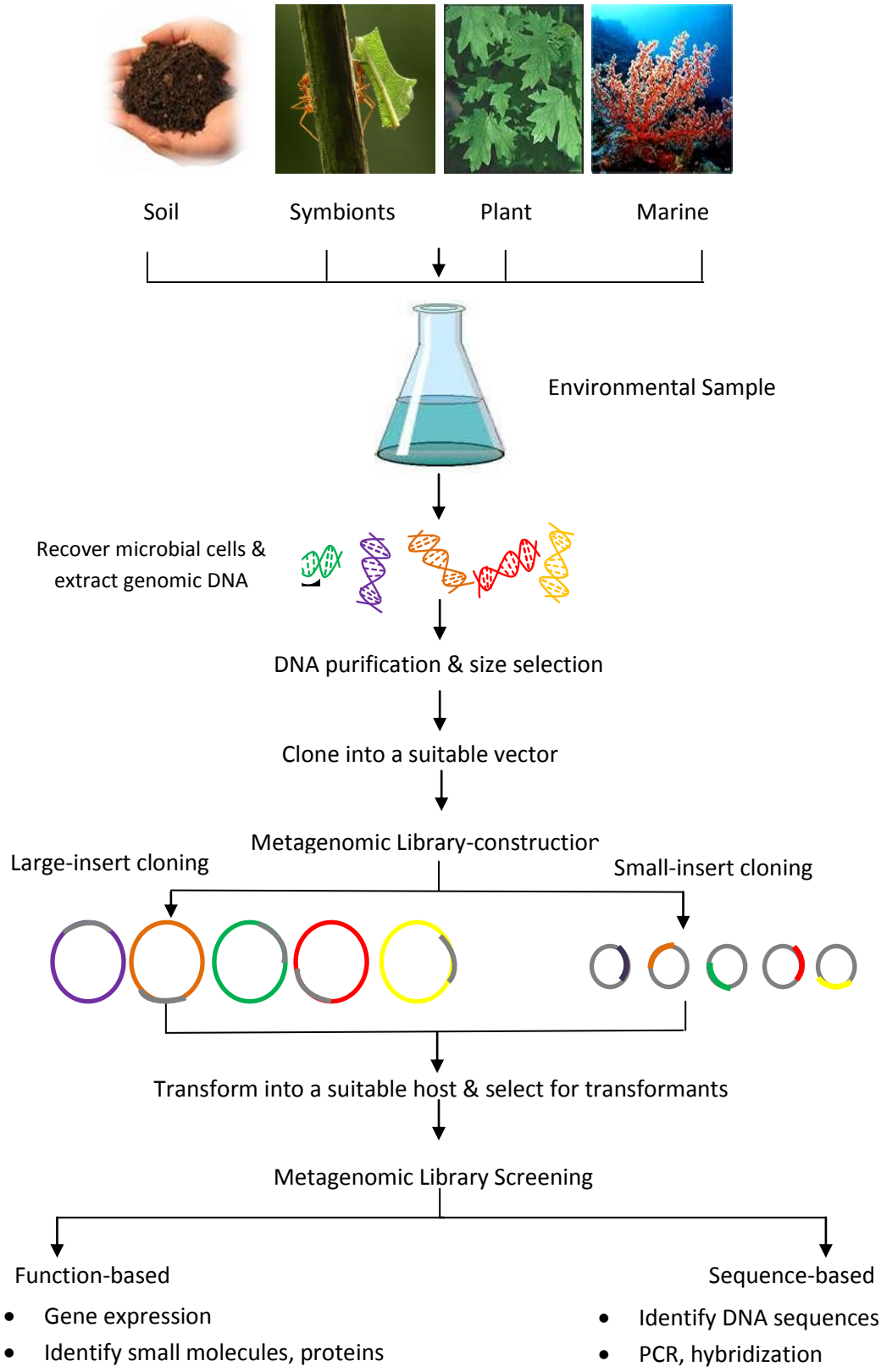


Figure 1. Schematic of microbial metagenomic library construction and screening.

CHAPTER II

GRAM-NEGATIVE SHUTTLE BAC VECTOR FOR HETEROLOGOUS EXPRESSION OF METAGENOMIC LIBRARIES

A. ABSTRACT

Bacterial artificial chromosome (BAC) vectors enable stable cloning of large DNA fragments from single genomes or microbial assemblages. A novel shuttle BAC vector was constructed that permits replication of BAC clones in diverse Gram-negative species. The "Gram-negative shuttle BAC" vector (pGNS-BAC), uses the F replicon for stable single-copy replication in *E. coli* and the broad-host-range RK2 mini-replicon for high-copy replication in diverse Gram-negative bacteria. As with other BAC vectors containing the *oriV* origin, this vector is capable of an arabinose-inducible increase in plasmid copy number. Resistance to both gentamicin and chloramphenicol is encoded on pGNS-BAC, permitting selection for the plasmid in diverse bacterial species. The *oriT* from an IncP plasmid was cloned into pGNS-BAC to enable conjugal transfer, thereby allowing both electroporation and conjugation of pGNS-BAC DNA into bacterial hosts. A soil metagenomic library was constructed in pGNS-BAC-1 (the first version of the vector, lacking gentamicin resistance and *oriT*), and recombinant clones were demonstrated to replicate in diverse Gram-negative hosts, including *Escherichia coli, Pseudomonas* spp., *Salmonella enterica, Serratia marcescens, and Enterobacter*

nimipressuralis. This shuttle BAC vector can be utilized to clone genomic DNA from diverse sources, and then transfer it into diverse Gram-negative bacterial species to facilitate heterologous expression of recombinant pathways.

B. INTRODUCTION

BAC vectors using the modified F plasmid are commonly used for construction and analysis of genomic libraries; greatly facilitating research that relies upon the stable maintenance of very large DNA inserts (Shizuya et al., 1992). Early versions of BAC vectors (e.g., pBELOBAC11) provided excellent stability of recombinant clones, but due to the single copy number of the BAC replicon processing for library construction or clone analysis required large culture volumes to achieve sufficient vector DNA (Kim et al., 1996). The introduction of an additional origin of replication (oriV) from a broad host-range RK2 plasmid permitted a stable but inducible copy system, wherein the copy number was controlled by an arabinose-inducible replicator protein (TrfA) inserted into the *E. coli* chromosome (Wild et al., 2002, Wild and Szybalski, 2004a). With inducible copy number, BAC clones can be maintained at single copy under control of the F replicon and then induced to multiple copies (50- to 150-fold induction) by the addition of 0.01% arabinose to the culture medium. These vectors were further developed by Szybalski's lab to a new class of pBAC/oriV "copy-control tightly regulated expression vectors" (Wild and Szybalski, 2004b).

While providing significant advantages, these commercially available inducible BAC vectors (e.g., CopyRight v2.0 BAC, Lucigen Corp., Middleton, WI) were described as limited to replication within an *E. coli* host. For sequence-based mapping and molecular analysis, maintenance in *E. coli* is sufficient. However, for construction and

functional screening of metagenomic libraries it is advantageous to transfer recombinant clones into multiple bacterial expression hosts to improve heterologous expression of cloned metagenomic DNA (Craig et al., 2010; Handelsman et al., 1998; Rondon et al., 1998). Various shuttle vectors have been used to transfer recombinant clones into alternative heterologous hosts, such as *Streptomyces* and *Pseudomonas* spp. (Martinez et al., 2004; Wang et al., 2000). By incorporating into a BAC vector both oriV and trfA, which is the mini-replicon necessary for RK2 plasmid replication (Thomas et al., 1981), a much greater host range, can be achieved, including most Gram-negative bacterial species.

A previous phylogenetic analysis of a soil metagenomic library indicated the very low prevalence of 16S rRNA genes from Gram-positive phyla (Liles et al., 2003), reflecting the poor lysis of Gram-positive cells when attempting to clone large DNA fragments. Therefore, a Gram-negative shuttle BAC vector would be particularly advantageous in cloning DNA derived from the diverse Gram-negative bacteria represented within metagenomic libraries, allowing for conjugal transfer and heterologous expression of metagenomic cloned DNA in multiple bacterial hosts. One such example is pRS44, a RK2-based broad-host-range cloning vector (Aakvik et al., 2009). Unlike the pGNS-BAC vector, which has the complete RK2 mini-replicon contained within the vector (i.e., oriV and trfA), the pRS44 vector system requires transposon-mediated insertion of the trfA gene within the desired host species. Increased expression of cloned DNA due to copy number induction can be very important in functional screening of metagenomic libraries; thus, the pGNS-BAC vector increases the

probability of identifying clones with specific functions by expanding the range of genomic library hosts for expression.

C. MATERIALS AND METHODS

1. Bacterial strains and media. *E. coli* strain DH10B was used as the primary host for transformations. Cultures were grown at 37^{o}C in Luria-Bertani broth or agar plates supplemented with the appropriate antibiotics. Concentrations of antibiotics were 12.5 µg/ml chloramphenicol (Cm) and 30 µg/ml gentamicin sulfate (Gm). *Pseudomonas putida, Pseudomonas aeruginosa, Pseudomonas stutzeri, Pseudomonas fluorescens, Salmonella enterica, Serratia marcescens, Vibrio vulinificus,* and *Enterobacter nimipressuralis* were used as recipients to test the host range of pGNS-BAC-1 (Table 1).

2. Construction of pGNS-BAC-1 Bacterial culture collection. The P_{ara}BAD promoter of plasmid pJW544 drives expression of the *trfA* gene, and the TrfA replication initiation protein then binds to *oriV* iterons (Perri et al., 1991). A BamHI restriction site within the promoter was destroyed by restriction digestion and subsequent fill-in with Klenow DNA polymerase and dNTPs. Plasmid DNAs were extracted from *E. coli* cultures using a Promega Wizard Plus SV Minipreps kit (Madison, WI). Restriction and DNA sequence analysis was conducted to confirm loss of the BamHI site, and induction of plasmid copy number with 0.01% arabinose was performed to confirm that the copy-inducible phenotype was still functional. The resulting plasmid was named pGNS-BAC-1.

3. Soil metagenomic library construction. To determine if recombinant BAC clones in the vector pGNS-BAC-1 were capable of replication within Gram-negative bacterial hosts, a small-insert BAC library was constructed from bacterial cells that were first

extracted from the soil prior to DNA isolation (Liles *et al*., 2008). Briefly, the extracted and washed bacterial cells were incorporated into agarose plugs, lysed, and then high molecular weight (HMW) metagenomic DNA was electrophoresed from the plug. Purification by a formamide denaturation step (70% final concentration) resulted in removal of associated nuclease activity from the HMW DNA and improved cloning efficiency (Liles *et al*., 2008). The formamide-treated metagenomic DNA was partially restriction digested with HindIII, electroeluted from an agarose gel, and ligated into a Hind-III digested and dephosphorylated pGNS-BAC-1 vector. The ligated vector and insert DNA was transformed into *E. coli* strain DH10B, and transformants were selected on LB containing 12.5 µg/ml Cm. Transformants were robotically picked into a 96-well format and stored in 10% glycerol at -80^{o}C.

4. Electroporation of BAC DNA into bacterial strains. Random clones were selected from the soil metagenomic library in pGNS-BAC-1. Plasmid DNAs were extracted using a manual alkaline lysis protocol and characterized by restriction fragment length polymorphism (RFLP) analysis using HindIII (Promega). Clones with insert DNA were transformed into electrocompetent *Serratia marcescens, V. vulnificus*, and *Pseudomonas putida* (1 mm gap cuvette, 1.8 kV, 600 Ohms, 10 µF). Cells were grown in SOC recovery medium for 1 hour at 37^{o}C and plated on LB agar supplemented with Cm. Plasmid DNAs were isolated and subjected to RFLP analysis as above to test for the presence of the recombinant BAC DNA in each bacterial host.

5. Construction of pGNS-BAC. A Gm resistance cassette was obtained from plasmid pBSL141 (Alexeyev et al., 1995) as a NheI restriction fragment and ligated into an Eco47III site of the vector pGNS-BAC-1. Transformants were selected on LB containing

both Cm and Gm. Restriction digests with EcoRV established the presence of the Gm-resistance cassette, resulting in the plasmid pGNS-BAC-2.

A cloning region from the vector pSMART BAC v2.0 (Lucigen Corporation, Middleton, WI) containing the counter-selectable *sacB* gene and pUC19 origin of replication was cloned into pGNS-BAC-2 to reduce the background of transformants without inserts (i.e., by *sacB*-mediated counter-selection) and to provide very high copy number for preparation of empty vector DNA (pUC19 origin). The cloning region was PCR amplified using flanking primers, purified, blunt-ended, and ligated to a filled-in HindIII restriction site of the pGNS-BAC-2 vector. The ligation was transformed into electrocompetent *E. coli* strain DH10B and plated onto LB containing Cm and Gm. Transformants were screened for sucrose sensitivity. Plasmid DNA was extracted from sucrose-sensitive clones and restriction digested with HindIII to confirm the addition of the cloning region to pGNS-BAC-2. The resulting BAC vector was designated as pGNS-BAC-3.

To introduce the ability to conjugally transfer the BAC vector, the *oriT* (*mob*) gene from pLOF-Km was PCR amplified using the primers *mob*F (5´ GATC**CTCGAG**GGATCCTTTTTGTCCG) and *mob*R (5´ GATC**CTCGAG**CAGCCGACCAGGCT) (Herrero et al., 1990). The PCR primers include 5´ XhoI restriction sites (bold). After amplification and XhoI digestion, the amplicon was ligated into the XhoI site of pGNS-BAC-3, transformed into *E. coli* strain DH10B, and selected on LB containing Cm and Gm. Clones containing the *oriT* were identified via PCR using the mobF and mobR primers, and the resultant plasmid was verified by restriction digestion with Sau3AI. The final vector construct, pGNS-BAC-4,

also referred to as the pGNS-BAC vector, was stored as a glycerol stock at -80^{o}C. The pGNS-BAC vector was sequenced completely, and the sequence was deposited within the GenBank database (accession number HQ245711).

6. Conjugal transfer of BAC vector DNA into bacterial strains. The pGNS-BAC vector was electroporated (1 mm gap cuvette, 1.8 kV, 600 Ohms, 10 µF) into *E.coli* strain SM10, which permits conjugal transfer of *oriT*-containing plasmids (Simon et al., 1983). Cells were grown in SOC recovery medium for 1 hour at 37^{o}C and plated on LB agar supplemented with Cm (12.5 µg/ml) and Gm (30 µg/ml). *E. coli* strain SM10 having the pGNS-BAC vector was used as the donor for conjugation experiments. Confirmation of the ability of *oriT* to mediate conjugal transfer was performed using *S. marcescens* as the recipient. LB broth supplemented with Cm and Gm was used to grow the donor, and LB broth without antibiotics was used to grow the recipient. Cultures were grown overnight at 37^{o}C with aeration. Donor and recipient were mixed in a ratio of 1:4 (50 µl and 200 µl respectively) and treated with 1 ml of 10 mM $MgSO_4$. After mixing thoroughly, centrifugation was carried out at 15,000xg for 10 minutes. One ml of supernatant was discarded, and the cells were resuspended in the remaining liquid and spread on a nitrocellulose membrane placed on the surface of an LB agar plate. Following incubation at 37^{o}C for 4 hours the membrane was transferred to an LB agar plate containing 1 mM IPTG and incubated at 37^{o}C for another 12 hours. Cells were then washed off the membrane with 3 ml of 10 mM $MgSO_4$ and collected in a tube. Different dilutions of this cell mixture were then spread on LB agar containing Cm and Gm (to select against the recipient) and colistin (10 µg/ml, to select against the donor). This procedure allows exclusive selection of the *S. marcescens* transconjugants.

Transconjugants were selected and screened for the presence of pGNS-BAC vector DNA by isolating plasmid DNA from the recipient hosts after varying times of cultivation, in the presence and absence of Cm and/or Gm and/or 0.01% arabinose. The presence of plasmid DNAs was confirmed by restriction analysis. The cell counts of donor, recipient, and transconjugants were estimated by plating a range of serial dilutions on suitable media.

7. Increase in MIC of Cm and Gm conferred by pGNS-BAC. Minimum inhibitory concentration (MIC) testing using the macrodilution method was carried out to test the degree of resistance to gentamicin or chloramphenicol conferred by pGNS-BAC on *E. coli* or *S. marcescens*. Both bacterial species were tested in the presence and absence of the BAC vector and with or without addition of 0.01% arabinose to the cation-adjusted Mueller-Hinton broth (CAMHB) (Table 2). Antibiotic stock solutions of Gm (960 µg/ml), Cm (400 µg/ml), and arabinose (0.01% and 0.02%) were made using CAMHB. The final concentration range tested for Gm was from 7.5 µg/ml with a twofold consecutive increase up to 960 µg/ml and likewise for Cm from 3.125 µg/ml up to 400 µg/ml. The experiment was conducted in triplicate, with inclusion of the controls: 1) bacterial strains without vector DNA, 2) bacterial growth without any antibiotics added, and 3) media only. Tubes were incubated overnight at 37ºC and turbidity was measured the next day to determine the MIC of the antibiotic.

D. RESULTS

1. pGNS-BAC-1 construction and analysis. pGNS-BAC-1 was tested as a shuttle vector under control of either of its two origins of replication (i.e., F and RK2) (Table 1). The pGNS-BAC-1 vector is maintained in *E. coli* as a single-copy plasmid by repressing

the RK2 origin of replication with the addition of 0.1% glucose to the growth medium. Induction of plasmid copy number in *E. coli* was achieved by supplementation with 0.01% arabinose (Figure 1, panel A). The pGNS-BAC-1 vector was electroporated into *P. putida, P. aeruginosa, P. stutzeri, P. fluorescens, S. enterica, S. marcescens, V. vulinificus, and E. nimipressuralis* (Table 1). Isolated colonies from each transformation were used to inoculate LB broth cultures containing Cm, and the plasmid DNAs extracted from each host revealed a banding pattern identical to the pGNS-BAC-1 plasmid (data not shown). In some cases, the DNA isolated from non-*E. coli* hosts (e.g., *S. marcescens*) was retransformed into *E. coli*, yielding Cm-resistant clones with a pGNS-BAC-1 restriction profile (data not shown).

2. Construction of a soil metagenomic library and lateral transfer of recombinant clones. To determine if recombinant pGNS-BAC-1 clones can also stably replicate in different bacterial hosts, a metagenomic library was constructed within pGNS-BAC-1. Metagenomic DNA was extracted from soil at the Bonanza Creek Experimental Forest near Fairbanks, AK, and was partially restriction digested and ligated into pGNS-BAC-1. *E. coli* transformants were picked into 96-well plates, and random clones were analyzed by RFLP to identify large-insert containing clones. Random clones containing DNA inserts of approximately 75.0 kb, 79.8 kb, 83.9 kb, and 86.0 kb were electroporated into *S. marcescens, V. cholerae, and E. nimipressuralis*. Cm-resistant transformants were successfully isolated for each of the clones in each of the bacterial hosts. The range of transformation efficiencies for the clones containing inserts relative to the empty pGNS-BAC-1 vector was 94.9% to 259% for *S. marcescens*, 45.7% to 76.2% for *V. cholerae*, and 55.9% to 104.9% for *E. nimipressuralis*.

3. pGNS-BAC construction and analysis. Although pGNS-BAC-1 was maintained in multiple Gram-negative bacterial hosts, its utility as a shuttle vector was limited due to the presence of only a single antibiotic resistance gene and an inability to be conjugally transferred to recipient hosts. Therefore, a Gm resistance cassette and an *oriT* were added to pGNS-BAC-1.

An improved multiple cloning region with a removable counter-selectable marker was also added to pGNS-BAC-1 to provide much lower background during transformations, resulting in the final pGNS-BAC vector construct. Cells containing intact pGNS-BAC vector are sucrose-sensitive due to the presence of the *sacB* gene within the cloning region. This region is removed as a restriction fragment during preparation of the vector for cloning. The final vector size is 11.9 Kb, and recombinant clones are sucrose-resistant (data not shown). The complete sequence of pGNS-BAC was determined and annotated and submitted to GenBank (accession number HQ245711).

E. coli strain SM10 containing pGNS-BAC was mixed with *S. marcescens* to test its ability to be conjugally transferred and to replicate within a bacterial host other than *E. coli*. Transconjugants that were CmR and GmR were readily obtained (> 1 x 105 transconjugants μg^{-1} DNA). Representative transconjugants were inoculated into broth cultures with and without antibiotic selection, and after 12 to 16 hours of growth, plasmid DNAs were isolated and restriction digested to determine plasmid yield and stability. Plasmid DNAs corresponding to the pGNS-BAC restriction profile were observed by RFLP (Figure 1, panel B).

In the absence of arabinose copy-induction, *E. coli* (pGNS-BAC) had an MIC for Cm of 25 µg/ml and an MIC for Gm of 60 µg/ml, whereas S. marcescens (pGNS-BAC) had an MIC for Cm of 12.5 µg/ml and an MIC for Gm of 30 µg/ml (Table 2). In the presence of arabinose, *E. coli* (pGNS-BAC) had an MIC for Cm of 200 µg/ml and 480 µg/ml for Gm, and *S. marcescens* (pGNS-BAC) had an MIC for Cm of 200 µg/ml and an MIC for Gm of 240 µg/ml (Table 2). Thus, both *E. coli* and *S. marcescens* harboring pGNS-BAC had a 32-fold increase in resistance to Cm as a result of arabinose-mediated copy-induction, and a similar increase in resistance to Gm in the presence of arabinose (32-fold for *E. coli*, and 16-fold for *S. marcescens*; Fig. 3). However, in the absence of the pGNS-BAC vector no arabinose-induced changes in MIC levels were observed (Table 2).

E. DISCUSSION

The pGNS-BAC vector provides the ability to clone DNA inserts and maintain recombinant clones at single copy in *E. coli*, utilizing the well-described stability of the F plasmid. The addition of arabinose results in induction of pGNS-BAC copy number mediated by *trfA* located on the plasmid. Copy-induction greatly increases plasmid DNA yield and could improve heterologous expression of cloned DNA via a gene-dosage mechanism (Rine *et al*., 1983). The RK2 mini-replicon that affords the copy-inducible phenotype in *E. coli* also permits replication in a broad range of Gram-negative bacterial hosts. Large-insert clones within the first version of the shuttle vector pGNS-BAC-1 were capable of transfer and replication within phylogenetically diverse bacterial species. The final pGNS-BAC vector construct has a significantly expanded host range compared

to pGNS-BAC-1 due to the addition of genes for Gm resistance and its ability to be conjugally transferred.

This inducible-copy and Gram-negative shuttle vector can be employed for metagenomic analysis of diverse environments, most of which contain abundant Gram-negative species, as well as to heterologously express specific genetic pathways. Construction of a soil metagenomic library in the pGNS-BAC vector provides the ability to transfer entire libraries, or specific recombinant clones, into bacterial hosts that may be more closely related to the bacterial taxa from which the cloned DNA was derived. Ideally, metagenomic libraries from a given source DNA could be constructed in both pGNS-BAC and a Gram-positive shuttle vector, thereby providing the widest possible range of heterologous expression hosts.

The rapidly advancing science of metagenomics requires molecular tools to enhance the heterologous expression of cloned DNAs. The metagenomic libraries constructed in pGNS-BAC will have all of the properties valued in previous libraries, such as stable maintenance of large inserts, with added features that could greatly facilitate manipulation and expression of recombinant clones in a variety of different Gram negative hosts.

Table 2.1. Bacterial strains and plasmids.

Bacteria	Plasmids of interest	Plasmid-encoded antibiotic resistance or other characteristic	Source
E. coli strain DH10B	pJW544	BAC vector, Cm^R, *oriV*	Wild et al., 2004
E. coli strain DH10B	pGNS-BAC1	BAC vector, Cm^R, *oriV*, BamHI site-minus	This study
E. coli strain SM10	pGNS-BAC	BAC vector, Cm^R, Gm^R, *oriV*, *oriT*, *sacB*	This study
Pseudomonas putida	pGNS-BAC1 or pGNS-BAC	Cm^R, or Cm^R and Gm^R	This study
Pseudomonas aeruginosa	pGNS-BAC1 or pGNS-BAC	Cm^R, or Cm^R and Gm^R	This study
Pseudomonas stutzeri	pGNS-BAC1 or pGNS-BAC	Cm^R, or Cm^R and Gm^R	This study
Pseudomonas fluorescens	pGNS-BAC1 or pGNS-BAC	Cm^R, or Cm^R and Gm^R	This study
Salmonella enterica	pGNS-BAC1 or pGNS-BAC	Cm^R, or Cm^R and Gm^R	This study
Vibrio vulinificus	pGNS-BAC1 or pGNS-BAC	Cm^R, or Cm^R and Gm^R	This study
Enterobacter nimipressuralis	pGNS-BAC1 or pGNS-BAC	Cm^R, or Cm^R and Gm^R	This study

Table 2.2. MIC for Cm and Gm conferred by pGNS-BAC.

Bacterial strain	pGNS-BAC	Arabinose	MIC (µg/ml)	
			Cm	Gm
E. coli	+	-	25	60
E. coli	+	+	200	480
E. coli	-	-	6.25	15
E. coli	-	+	6.25	15
S. marcescens	+	-	12.5	30
S. marcescens	+	+	200	240
S. marcescens	-	-	6.25	15
S. marcescens	-	+	6.25	15

Figure 2.1. Isolation of BAC vector DNA from *E. coli* and *S. marcescens.*

(Panel A) Lane 1, molecular weight marker (1 kb Plus, Promega); Lanes 2-4, *E. coli* containing pGNS-BAC (minus the stuffer fragment containing *sacB*) with 0.2% glucose added to the medium (lane 2), with no arabinose or glucose added to the medium (lane 3), or with 0.01% arabinose added to the growth medium (lane 4). (Panel B) Lane 1, molecular weight ladder; Lane 2, pGNS-BAC isolated from *E. coli;* and Lane 3, pGNS-BAC isolated from *S. marcescens*. DNAs in lanes 2 and 3 were restriction digested with BsrGI.

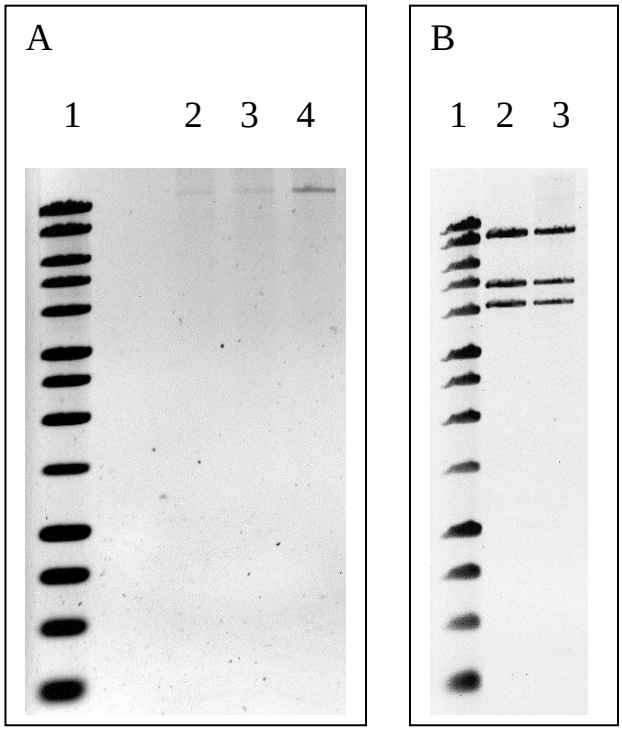


Figure 2.2. Annotated plasmid map for pGNS-BAC-1 (Panel A) and pGNS-BAC (Panel B).

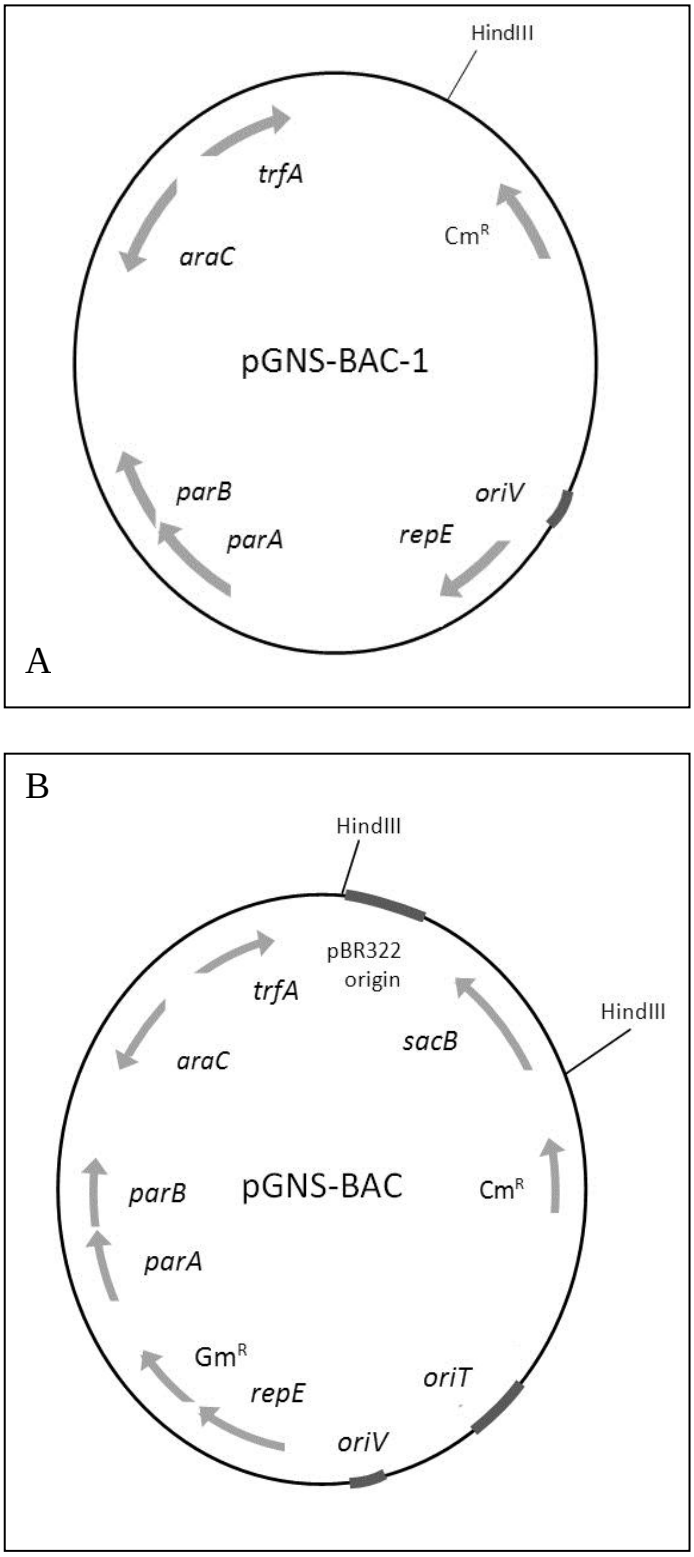
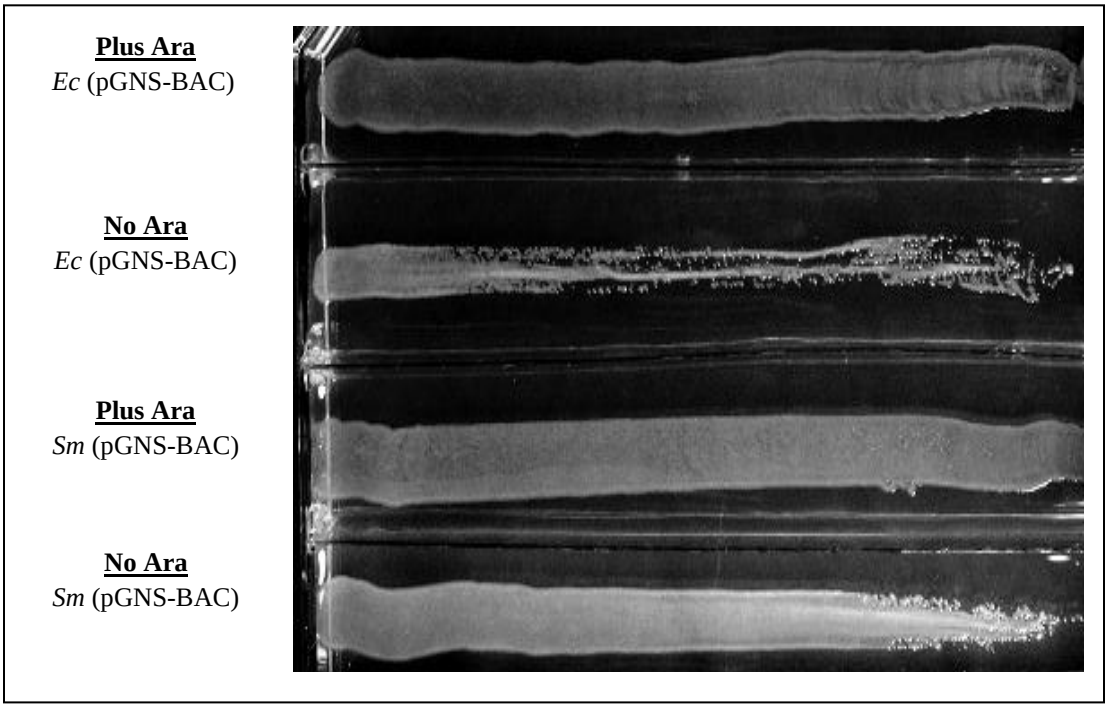


Figure 2.3. Growth pattern of *E. coli* and *S. marcescens* on a Cm gradient agar plate with and without arabinose.

Growth of *E. coli* with the pGNS-BAC vector (*Ec* (pGNS-BAC)) and *S. marcescens* with the pGNS-BAC vector (*Sm* (pGNS-BAC)) in the presence (Plus Ara) and absence (No Ara) of arabinose, and, on a Cm gradient agar plate ranging from no added Cm (Left) to 500 µg/ml Cm (right).



CHAPTER III

SCREENING SOIL METAGENOMIC LIBRARIES TO IDENTIFY RECOMBINANT CLONES PRODUCING AN ANTIMICROBIAL ACTIVITY

A. ABSTRACT

The emergence of multidrug-resistant (MDR) pathogens has led to an increased need for new antibiotic classes. Significant numbers of clinically-used antibiotics are natural products derived from cultured microorganisms. Studies have shown that the diverse microbial communities in soils are potentially a great resource for novel natural products but due to limitations of culturing methods a majority of them are unexplored. To tap into this vast resource, we combined randomly sheared, large-insert cloning with the ability to express clones in multiple heterologous hosts. High molecular weight DNA was isolated from diverse soil microorganisms, sheared and cloned into a bacterial artificial chromosome (BAC) shuttle vector. Three different metagenomic libraries were constructed that had an average insert size of 68kb, 42kb and 113 kb, respectively. Functional screening of clones in *E. coli* was used to identify top candidates with inhibitory activity against tester strains like *Pseudomonas aeruginosa, Alcaligenes faecalis,* methicillin-resistant *Staphyloccocus aureus* (MRSA). The library was screened in 96-well microtitre plate format with an *in situ* lysis method for detecting both intra- and extracellular compounds. A negative control (empty vector) was used in all

bioassays. These candidates were further evaluated for consistency of results and validated in multiple bioassays. Transformation of naïve *E. coli* with BAC DNA isolated from these clones confirmed presence of an antibacterial activity against the respective tester strains used.

B. INTRODUCTION

The impact of bacterial infections on the society is significant and has been a cause of concern with the emergence of hypervirulent and MDR strains (Levy et al., 2004). Discovery of novel antibiotics is of great importance to combat these pathogens. Different approaches that have been used traditionally for discovering antibiotics include cultivation of bacteria that were previously "unculturable" (Kaeberlein et al., 2002), genetically modifying existing pathways (Pfeifer et al., 2001), direct cloning and expression of metagenomic DNA from natural environments (Rondon et al., 2000), and screening natural products (Singh 2006) against a target bacterial culture. The huge costs and a high rate of antibiotic rediscovery have limited the investments of pharmaceutical industries putting the onus of antibiotic discovery on academic researchers in the field.

One of the very effective and common means of identifying novel antibiotics is isolation of secondary metabolites produced by soil microorganisms. In this study a culture-independent and function-based method to characterize the soil "metagenome" was used to access novel antibiotics of potential medical importance. As opposed to small-insert metagenomic libraries the large-insert metagenomic libraries screened in this study have a higher probability of containing intact biosynthetic pathways necessary for the synthesis of the new chemical entities (NCEs). Enhancing the expression of cloned genes may lead to subsequent increase in the concentration of antibiotic products and the

likelihood of their detection. Arabinose-induction of the BAC vector copy number (under control of PBAD:trfA) was used for amplification from single copy to ~50 copies per *E. coli* cell.

An advantage of expression in *E. coli* (host) is easy and safe scale-up. With depletion of nutrients a complex cascade of regulatory signals leads to a change of expression from primary metabolism genes to those responsible for secondary metabolite synthesis. Thus, prolonged incubation of cultures in stationary phase is another way to increase expression of cloned genes. The innovations used in this study were using high molecular weight metagenomic DNA from soil microbial communities for construction of large-insert Random Shear Shuttle BAC libraries, improving heterologous gene expression and production of recombinant proteins by using a newly developed shuttle BAC vector and developing new methods for efficient screening of large soil libraries to accelerate the speed of discovery of the recovered antimicrobial activities. Clones identified on the basis of antimicrobial activity against tester strains may be promising candidates for potential therapeutics.

C. MATERIALS AND METHODS

1. Construction and Screening of Libraries. Three different metagenomic libraries were constructed and used for functional screening (Table 3.1). HMW metagenomic DNA was isolated from soils reperesenting diverse microbial communites. The isolation and purification of this HMW DNA from soils was done with some modifications to the published protocol (Liles et al., 2008). The bacterial cells in the soil sample were recovered by soil homogenization followed by differential centrifugation, thus separating

them from most eukaryotic cells and the soil particles. The cells were washed several times and embedded in an agarose plug. *In situ* lysis was carried out in the agarose plug by enzymatic treatment. Metagenomic DNA from the agarose plugs was electrophoresed into an agarose gel at 70V for 4-5 hrs followed by gel extraction, concentration and purification. Restriction endonuclease digestion or random shearing was used as applicable to get the desired size range of fragments for BAC cloning. The DNA was then blunt ended, ligated into the vector and transformed into the host *E. coli* strain of choice. For efficient high-throughput screening of these libraries various screening protocols were used:

Library SL 1: The library in the form of a 384 well plate was replica plated onto a 96 well plate containing LB media with Chloramphenicol (Cm) 12.5 µg/ml and arabinose 0.01%. The plates were grown at 37°C for 48 hrs with shaking at 200rpm. After 48 hrs the cultures were subjected to lysis by exposure to $CHCl_3$ vapors and then spun down to pellet the cell material at 4000 rpm. The supernatant collected from each plate was then spotted onto an LB media plate and overlayed with soft LB agar containing a 1:1000 diluted log phase culture of *P. aeruginosa*. Plates were incubated overnight at 37°C and then observed for zones of inhibition.

Library SL 3: A modified assay protocol was used for the screening of this library. As before the library in the form of a 384 well plate was replica plated onto a 96 deep well plate containing 150 µl LB media with Cm 12.5 µg/ml and arabinose 0.01%. The plates were grown at 37°C for 48 hrs with shaking at 200rpm. After 48 hrs 50 µl LB containing Ampicillin 200µg/ml and 0.4% SDS was added to each to each well using a robotic liquid dispenser. Plates were incubated again at 37^{o}C for 12 hours and then 50µl of a

1:1000 diluted log-phase culture of *P. aeruginosa* was added to each well followed by incubation at 37^{o}C for 24 hrs. The OD_{600} was recorded for each well as well as OD_{520} for the *Pseudomonas* cultures (as indicator of pyocyanin pigment production). Each plate was also examined visually to determine wells with impaired growth of the tester strain.

Library SL 5: Using a pin replicator the BAC library containing *E. coli* cells was inolcuted into deep 96-well plates containing about 1.0 ml of LB containing chloramphenicol 12.5 µg/ml plus arabinose 0.01% and plates were incubated at 37^{o}C for 48 hrs. The *E. coli* cells were lysed by freezing at -80^{o}C, followed by rapid thawing at 55^{o}C. 100 µl of 1:1000 diluted log phase tester (MRSA) culture with Nalidixic acid 30 µg/ml was added to each well followed by incubation at 37^{o}C for 24 hrs. Finally 165 µl of the viability indicator resazurin solution (0.02%) was added to each well and the plates were incubated at 37^{o}C till a color change from blue to pink was observed.

2. Validation of positive antibiotic producing hits. Each recombinant clone identified as inhibiting the growth and/or viability of the tester strain was grown from the library 384-well plate into LB broth culture containing Cm 12.5 µg/ml, incubated overnight at 37^{o}C and a separate glycerol stock was stored at -80^{o}C. Each positive clone was then inoculated into replicate wells (n=4) of a 96-well plate and grown as above to retest the clone for inhibitory activity against its respective tester strain that showed sensitivity previously. Every positive clone that demonstrated reproducible antibiotic activity was tested for its ability to inhibit growth of the tester strain by removal of *E. coli* cells by centrifugation rather than cell lysis, and transfer of supernatants to another microtitre plate. Results for each positive clone were noted according to the degree and consistency of inhibitory activity observed thus helping narrow down the the list to top candidates of

choice for the next stage. In the resazurin based bioassays, fluorescence readings of reduced resazurin (resorufin) were recorded (530 nm excitation and 590 nm emission wavelengths) and used for calculating the % growth inhibition of the tester culture in comparison with the empty vector negative control.

3. Retransformation of antibiotic-producing clones. The validated antibiotic producing clones were grown in 3 ml of LB containing Cm and after 24 hours of growth at 37^{o}C plasmid DNA was extracted by alkaline lysis method. A restriction digestion of each BAC clone with BamHI (or EcoRI) was resolved by PFGE using conditions suitable for plasmid RFLP analysis (i.e., 6V/cm, 1 sec to 15 sec switch time, for 12 hours at 15oC). The insert size for each BAC clone was estimated. BAC DNA was transformed into a naïve *E. coli* strain and selected on LB containing Cm for the presence of the plasmid. Transformation was done by electroporation (1 mm gap cuvette, 1.8 kV, 600 Ohms, 10 µF) into commercially available electrocompetent BAC replicator V2.0 cells (Lucigen Corp.) for Library SL1 and SL5 and into *E. cloni* 10G cells (Luciegen Corp) for Library SL3. Each re-transformed clone was re-tested as above for antibiotic activity. Only recombinant clones showing evidence of a metagenomic insert and consistent and re-transformable antibiotic activity were selected to be studied further.

4. Testing for the effect of arabinose induction. The final shortlist of clones was used to study the difference in the antibiotic activity when grown in the presence and absence of arabinose induction. These clones were grown in LB containing Cm but in two sets, one with arabinose 0.01% and the other without arabinose. The remaining procedure was the same as that in the validation assays. Results were recorded and data was analyzed to

determine whether arabinose was essential for copy induction and significant antibiotic activity against the tester strains.

D. RESULTS

1. Screening libraries for identification of antibiotic-producing clones. *E. coli* clones and target organisms were cultured separately to get optimized culture conditions for clone expression and antibiotic detection. Also to enhance expression of cloned genes clone cultures were incubated for prolonged time in stationary phase. Incubation in stationary phase is a standard method for inducing secondary metabolite synthesis in industrial production of pharmaceuticals (Strobel and Sullivan, 1999). Depletion of the growth medium leads to a complex cascade of regulatory signals which shut down expression of primary metabolism (growth) genes, and turn on secondary metabolism (survival) genes (Nystrom, 2004). For some organisms, secondary metabolism includes secreted products that inhibit growth of competitor bacteria, i.e., antibiotics. The induction of secondary metabolism in the host likely increases the probability of expressing cloned genes and the likelihood of detecting antibiotic activity. Each *E. coli* clone was therefore grown for 48 hrs at 37^{o}C prior to assaying clone supernatants. The use of Ampicillin and SDS and nalidixic acid in the later screening protocols coupled with the freeze-thaw process eliminated the hassle of using $CHCl_3$ for cell lysis. The concnetrations of these used in the screening were such that they inhibited the growth of any remaining *E. coli* cells, while not interfering with growth of the tester strains. After the Amp and SDS treatment some cell debris does remain in the well, but is clumped at the bottom of each well resulting in a clear supernatant that may contain bioactive compound(s) synthesized and by the *E. coli* recombinant clone. With resazurin only

viable cells result in color change from blue to pink since the dye is a viability indicator which fluoresces bright pink upon reduction by metabolically active cells. In the screening of Library SL1, clones exhibiting a zone of inhibition were considered as positive hits, with library SL3, clones showing very little or no growth in the well were considered as positive for activity and for library SL5 clones that resulted in less fluorescence than untreated MRSA controls were selected as positive (Figure 3.1 A and B).

2. Validation of positive antibiotic producing hits. Multiple bioassays that were carried out as outlined in the methods yielded a set of clones that were positive for activity against the respective tester strains (Figure 3.2). The progression from growing clone clutures in 96 well plates to growth in cultures tubes gave similar results. The number of clones selected after the validation rounds was narrowed down to 3 from Library SL1, 2 from Library SL3 and 28 from Library SL5 (Table 3.2).

3. Retransformation of antibiotic-producing clones. All the antibacterial clones that were analyzed using PFGE had large and unique cloned DNA (Figure 3.3) which was successfully transformed into naïve *E. coli*. The resulting transformants on testing for antibacterial activity as described before had significant activity against the respective tester strains (Figure 3.4), demonstrating that the clone DNA was necessary and sufficient to confer the activity on the *E. coli* host. These clones were the top candidates selected for 454 sequencing and subsequent analysis.

4. The effect of arabinose induction. As discussed before, clone amplification leads to enhanced expression of secondary metabolites from cloned genes and may increase downstream concentrations of antibiotic products, and therefore the likelihood of their

detection. Copy-control cloning vectors used in these libaries can be amplified from single-copy to ~50 copies per *E. coli* cell by addition of arabinose to the growth medium. Clone amplification-increased expression is likely the result of increased gene dosage. Enhanced expression of cloned genes due to clone amplification may increase downstream concentrations of antibiotic products, and therefore the likelihood of their detection. This effect was clearly demonstrated when clone cultures grown in the absence of arabinose showed very little or almost no inhibition of the tester strains as compared to good activity when grown in presence of arabinose (Figure 3.4).

E. DISCUSSION

Metagenomic analysis of uncultured microorganisms is a recent strategy that has been used in the discovery of novel antibiotics. Although it is a more inclusive method to capture the vast majority of microorganisms that are as yet uncultured under laboratory conditions, functional metagenomics can be riddled with challenges that can limit matural product discovery. Some of these are the isolation of HMW DNA with high quality and purity and an efficient screening methodology. Both these limitations were tackled in this study by using protocols that enabled isolation of high quality HMW DNA and *in situ* lysis of the host *E. coli* cells for high throughput library screening. These screening methods were more sensitive, faster and detected both extra- and intracellular compounds. In the latest library that was constructed an average insert size of over 100kb was obtained using randomly sheared DNA. This is an important breakthrough in the field as large insert sizes greatly increase the probability of containing an entire biosynthetic pathway in the cloned genes.

Another innovation in this study was the use of inducible-copy number BAC vectors. The advantage of using BAC vectors is the high stability of both the vector and the insert when maintained at a single or low copy and the ability to be induced to give a high copy number when required, e.g., when high DNA yields are needed in constructing libraries or to induce copy number and potentially achieve better expression of cloned DNA thereby leading to a significant increase ein the drug yield for screening. The pSMART-BAC-S vector used in contsrcution of Library SL5 allows high-throughput conjugation-based transfer of large-insert BAC clones into both Gram-negative and Gram-positive hosts, with chromosomal integration or stable episomal maintenance for heterologous expression.

The results have shown that eDNA can be cloned into BAC libraries and stably maintained in *E. coli*. Function-based analysis of metagenomic libraries was employed as a useful tool for identifying soil derived recombinant eDNA clones that showed growth inhibition of various tester bacterial cultures. As a proof of concept for the antibacterial activity from the cloned DNA, the DNA after cloning into a naïve *E. coli* host showed a very similar pattern of growth inhibition of tester strains. A combination of innovative methods in this study led to identification of various metagenomic clones that are good candidates for further characterization.

Table 3.1. Details of different Metagenomic libraries used for screening.

Library	Vector	Soil Source	Avg. insert size	# Clones
SL1	pSmartBAC	Hancock, WI	68kb	9216
SL3	pGNSBAC-1	Fairbanks, AK	42kp	27648
SL5	pSmart-BAC-S	Cullars Rotation, AL	113kb	19000

Table 3.2. Shortlisted clones after validation experiments

Library	# clones	Clone IDs
SL1	3	P11K11, P15G24, P17L5
SL3	2	P29E3, P30A5
SL5	28	P2P12, P2A13, P5A4, P5C24, P6B5, P6L4, P6L5, P9L21, P14O1, P18N22, P20I6, P22C4, P22E10, P23K15, P27K16, P27M10, P28H1, P31G24, P35B14, P36M1, P37A9, P37A11, P37O10, P43A3, P46O24, P49M4

Figure 3.1 Examples of metagenomic clones exhibiting inhibition of tester strain growth

A. Examples of *P. aeruginosa* growth inhibitory activity in supernatants isolated from *E. coli* expressing BAC cloned DNA. Note complete inhibition in well D3, and partial inhibition in well F3.

B. Metagenomic clone (blue) that inhibited MRSA viability.

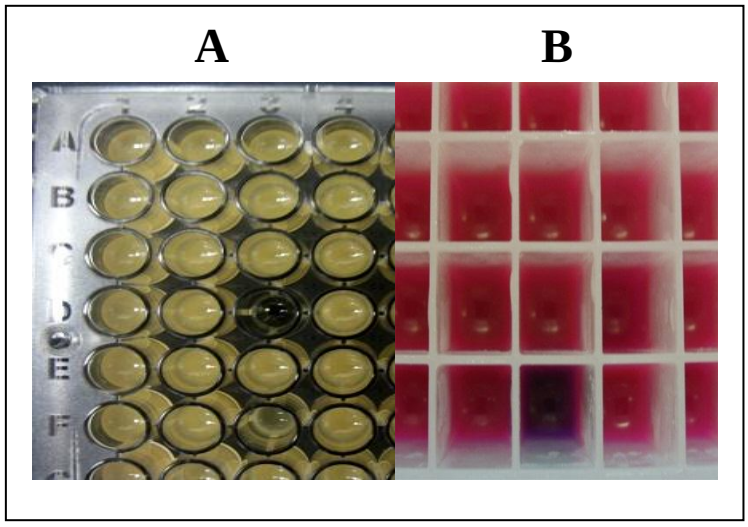


Figure 3.2 Antibacterial activity exhibited by the shortlisted metagenomic clones. The graph represents the % growth inhibition (Y axis) of the tester culture by the metagenomic clones (X axis) relative to the empty vector negative control, considered to have no inhibitory effect and calculated by measuring the fluorescence of reduced resazurin.

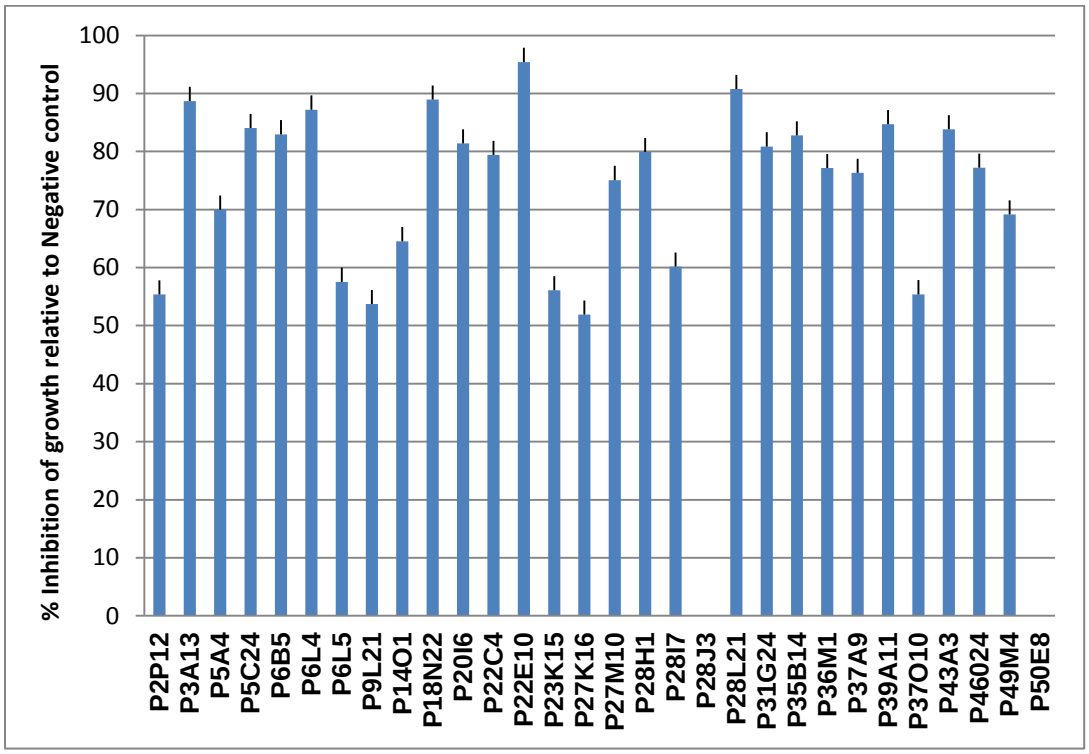
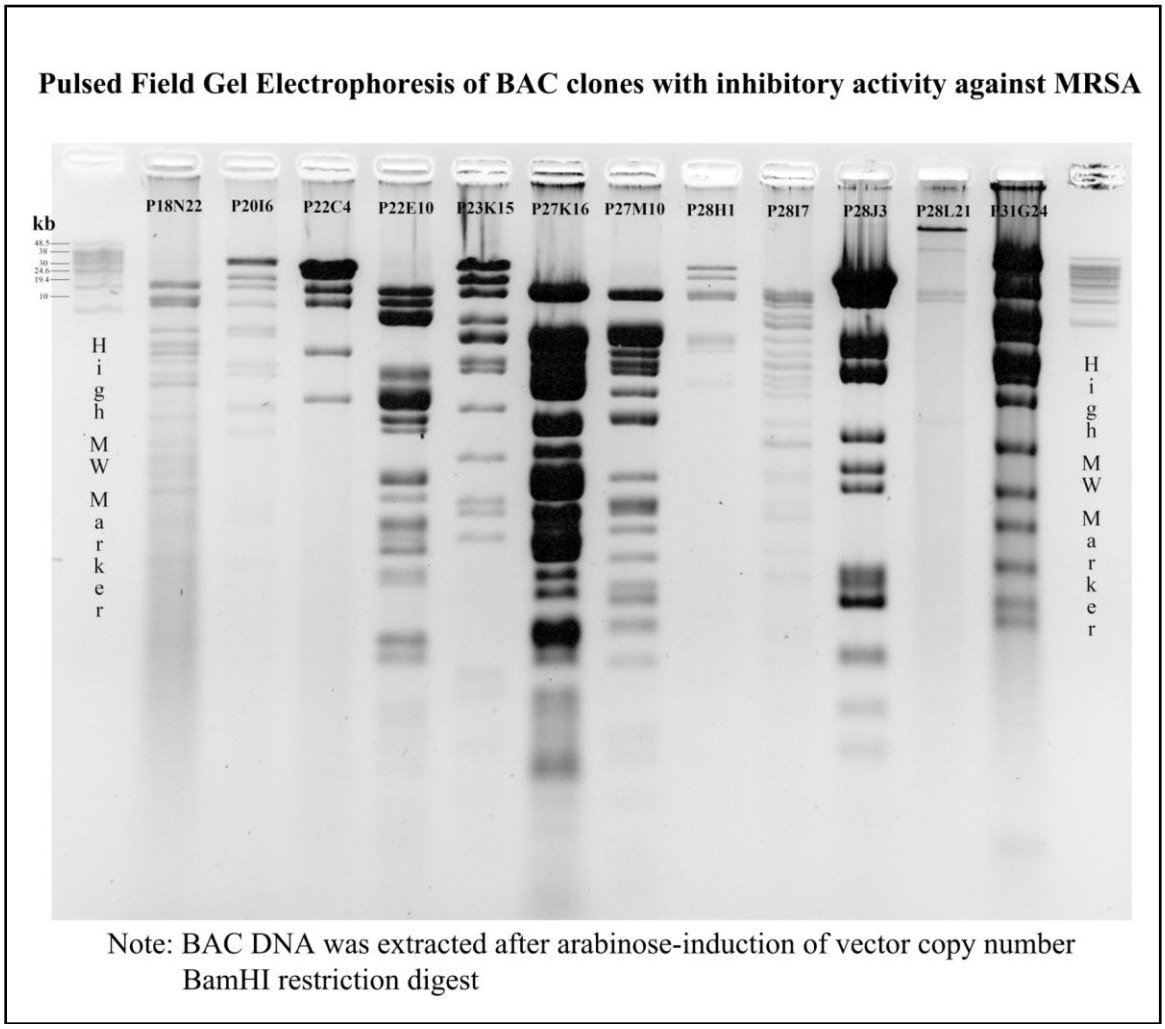


Figure 3.3 RFLP pattern of BAC DNA isolated from active metagenomic clones



Note: BAC DNA was extracted after arabinose-induction of vector copy number
BamHI restriction digest

Figure 3.4 Antibacterial actvity exhibited after transforming the cloned DNA into a naive *E. coli* host.

The graph represents the % growth inhibition (Y axis) of the tester culture by the metagenomic clones (X axis) relative to the empty vector negative control, considered to have no inhibitory effect and calculated by measuring the fluorescence of reduced resazurin.

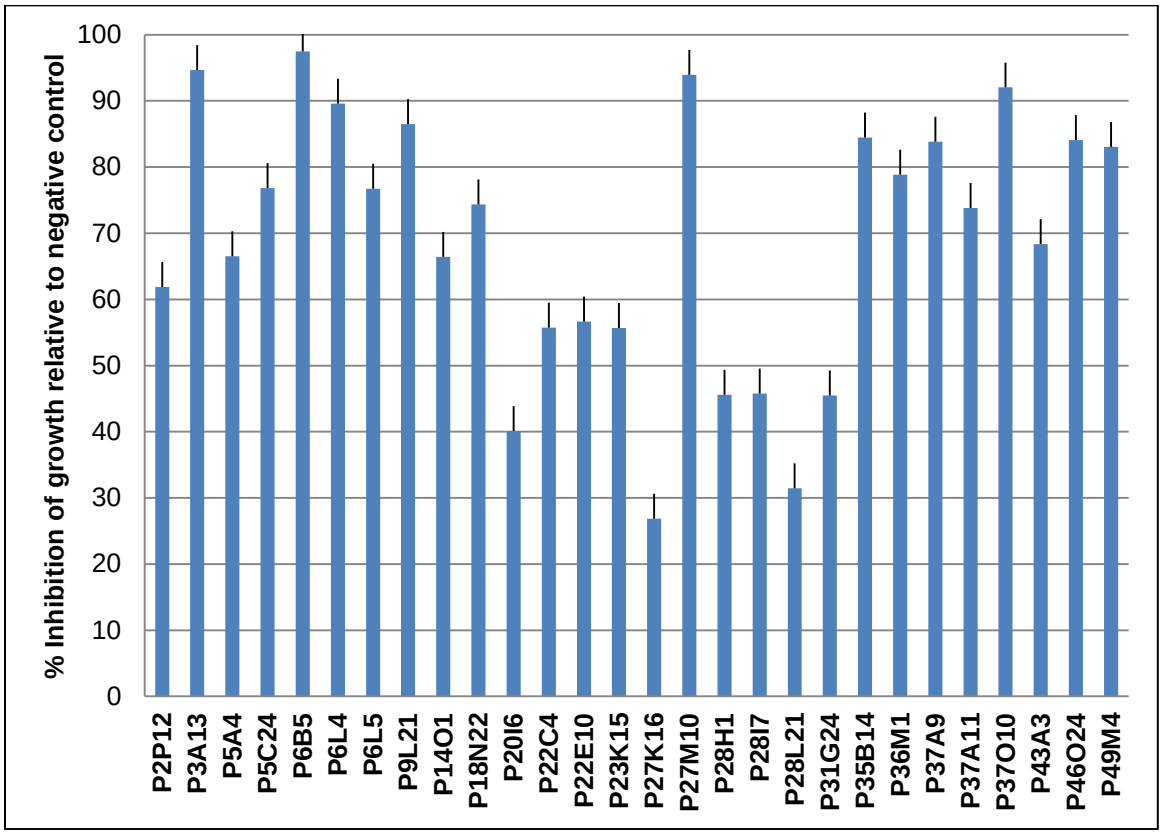
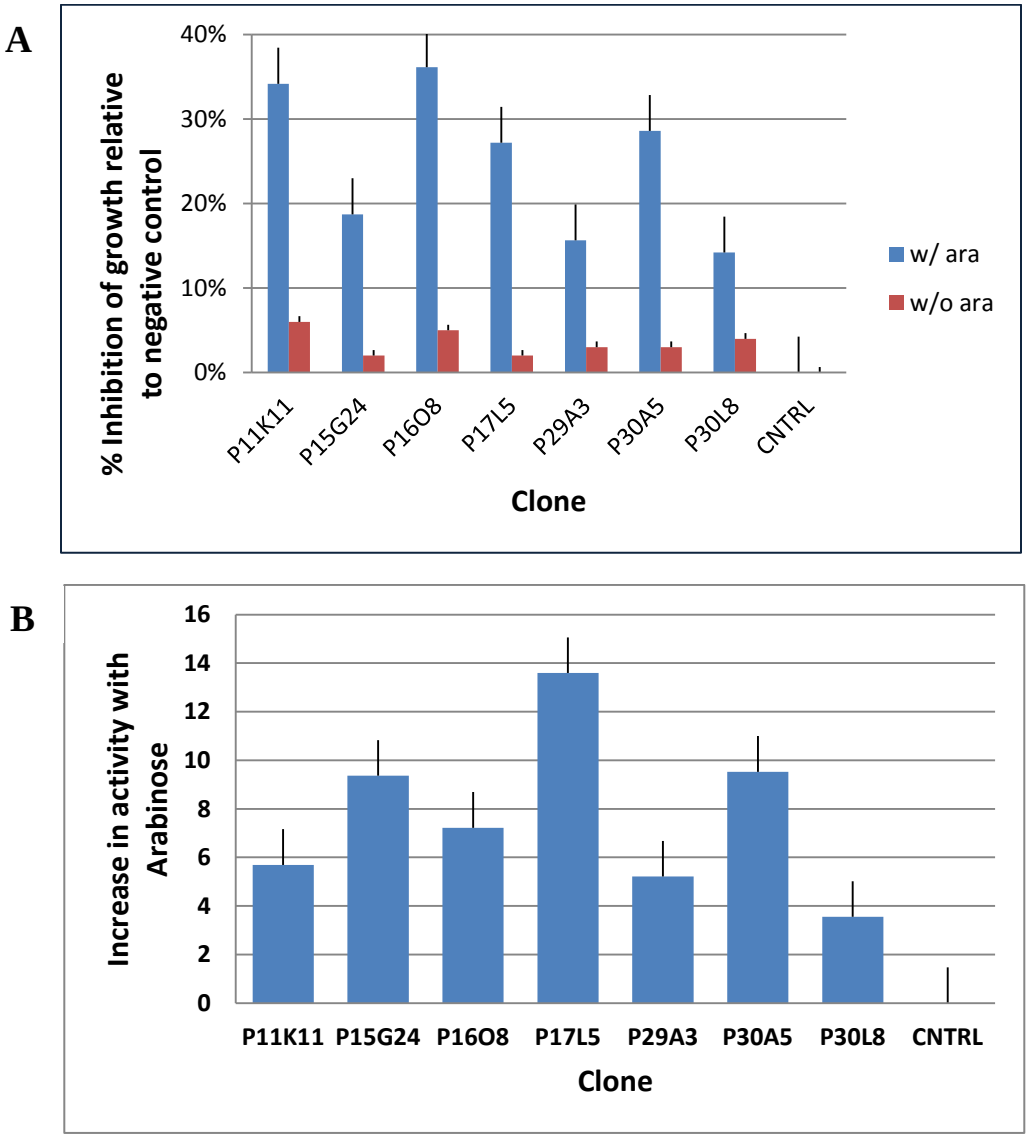


Figure 3.5 Effect of arabinose induction on the antimicrobial activity against respective tester strains.

Shown here is comparison of cultures grown in the presence and absence of arabinose. Graph A represents the % growth inhibition (Y axis) of the tester culture by the metagenomic clones (X axis) for two different treatments and relative to the Empty vector negative control, considered to have no inhibitory effect. Difference between the values for the two treatments groups was used to calculate the corresponding fold increase in activity for each clone as shown in Graph B.



CHAPTER IV

CHARACTERIZATION OF ANTIBIOTIC-EXPRESSING METAGENOMIC CLONES

A. ABSTRACT

Screening of multiple metagenomic libraries for recombinant clones that expressed an antibacterial activity resulted in a collection of validated clones that were subjected to preliminary biochemical characterization. Most assays were performed in a 96-well format using MRSA as the bacterial pathogen for bioassay-guided fractionation. Clone cultures were processed and analyzed to determine if the active compound(s) was extra- or intracellular, for heat stability and fractionation using a 3KD MWCO membrane. Based on the biochemical results, a smaller subset of clones that expressed a non-proteinaceous, small molecular weight antibacterial product(s) were selected for DNA sequence analysis. One clone candidate (P6L4) was chosen for further biochemical and genetic studies to predict the gene(s) present in the cloned insert. Biochemical characterization was done by LC-MS and then subcloning was used to determine the gene(s) responsible for the antibacterial activity. The anti-MRSA activity derived from clone P6L4 is most likely the result of an esterase that reactivates the endogenous chloramphenicol (added to the culture media) from its acetylated form. Results from the amplification, cloning and expression of the esterase encoding gene support this

observation and sequence analysis suggets a probable origin from the phylum *Acidobacteria*.

B. INTRODUCTION

Uncultured bacteria are a significant source for the discovery of novel small molecules with antimicrobial properties (Handelsman et al., 1998; Rondon et al., 2000). Function-based metagenomic analysis is a powerful approach to access the biosynthetic machinery of these uncultured bacteria to identify natural products such as antibiotics or enzymes. Screening a metagenomic library for clones that express a particular phenotype or function is the first step in identification of the genes that encode the biosynthesis of the antimicrobial compound.

Isolation of terragine E and other related compounds from recombinant clones from combinatorial biosynthetic libraries was one of the first examples of functional metagenomic analysis for discovery of novel compounds (Wang et al., 2000). Other examples include the long-chain N-acyl antibiotics isolated from environmental DNA (Brady and Clardy, 2000), characterization of the antibiotic palmitoylputrescine and its biosynthetic gene (Brady and Clardy, 2004), and the identication of the triaryl cations, designated as turbomycin A and turbomycin B, respectively, with antibiotic activity against gram-negative and gram-positive organisms (Gillespie et al., 2002). A metagenomic approach has been used in studies to search for novel biocatalysts such as lipases or esterases (Henne et al., 2000).

Based on the results of screening large-insert soil metagenomic libraries for antibacterial activity, a total of 33 recombinant clones were selected for further biochemical and genetic characterization. A preliminary characterization for these clones

was performed to determine the properties of the active compound(s) such as intra/extracellular secretion from the host, approximate estimation of molecular weight, heat stability and activity against multiple MRSA strains. The best lead candidates were selected for DNA sequence analysis and comparison of predicted genes against the NCBI GenBank database. The most complete biochemical and genetic analysis was conducted to characterize clone P6L4, as it had the greatest activity that was determined to be due to a small molecular weight compound, and this clone was determined to express an esterase that reactivated the acetylated Cm present in *E. coli* cultures and is likely of *Aciobacterial* origin.

C. MATERIALS AND METHODS

1. Preliminary characterization of active clones. The 33 clones selected in this study were tested to determine if the active compound(s) was extra- or intracellular. For each clone, 2 ml LB broth containing 12.5µg/ml Cm and 0.01% arabinose was inoculated with the *E. coli* glycerol stock stored at -80°C. The culture tubes were incubated at 37°C for 48 hrs at 200 rpm. Cultures were then divided into two sets (1 ml each), one set was subjected to a freeze-thaw process described previously and the other set was processed for cell-free supernatants. These samples were then tested in a bioassay in 96-well microtiter plates with three replicates (200 µl in each well) for each clone and treatment. Appropriate negative controls (empty vector with no insert) were used in the bioassay and 20µl of diluted log phase MRSA strain EAMC30 was added to each well. Nalidixic acid 30µg/ml was used to inhibit any residual *E. coli* cells in the cell lysates from the set of freeze-thaw treatment. Plates were incubated at 37°C for 24 hrs at 200 rpm. 30µl of a

0.02% resazurin stock solution was added to each well and plates were again incubated at 37°C for 4-5 hrs at 200 rpm. Fluorescence readings were recorded using a microtiter plate reader (excitation at 530nm and emission at 590nm) and the percent reduction of resazurin fluorescence of the MRSA strain EAMC 30 for each clone was determined by comparison with the respective negative controls.

To evaluate the heat stability and molecular weight of the clone-expressed activity, cultures were inoculated for the same set of 33 clones and the negative controls and incubated as before. At the end of 48 hrs, the cultures were processed using either the freeze-thaw or cell-free supernatant methods chosen on the basis of which method yielded the most antibacterial activity. For two of the clones (P6B5 and P37O10) the freeze-thaw treatment was used and all of the remaining cell-free supernatants were divided into three sets to test the heat stability, estimate the molecular weight of the active fractions and test against an additional MRSA strain Xen 31. For the test of heat stability to indicate activity due to proteinaceous products, the cell-free supernatant or the lysate was transferred to a microcentrifuge tube and placed in a boiling water bath for 10 minutes. Tubes were cooled to RT and a bioassay was set up as described before. To determine if the activity was due to a compound less than 3 kDa in size, 1 ml of the sample was fractionated using a centrifugal filter (VWR) with a modified polyethersulfone (PES) membrane with 3 kDa molecular weight cutoff (MWCO). The spin time was 15 minutes at 14,000 x g. The concentrate (reconstituted with half strength LB broth) and the filtrate were tested in the bioassay format as described before. To determine activity against a bioluminescent MRSA strain, a standard bioassay was used against a 1:1000 diluted log-phase culture of the Xen 31 strain. Cell-free supernatants

from a few select clones from library SL5 were also tested against the Xen 31 strain by the standard bioassay format. Bioluminescence was recorded after 24 hours of addition of the tester culture using Promega Glomax luminometer.

2. DNA sequence generation and analysis. The five clones from libraries SL1 and SL3 and the 12 top candidates from library SL5 were selected for complete insert sequencing using 454 pyrosequencing. BAC DNA was isolated from 100 mlcultures of each of the clones as described in Molecular Cloning, a laboratory manual. The purified DNA was sent to the Luciegen Corporation (Middleton, WI) to generate bar-coded shotgun subclone libraries that were sequenced at the EnGenCore Center at the Univeristy of South Carolina (Columbia, SC) using a Genome Sequencer FLX system (Roche) as per the manufacturer's instructions. The sequences were trimmed for quality (using a quality score cutoff of 0.01) and assembled into contiguous fragments (contigs) using the CLC genomics workbench (Cambridge, MA) *de novo* assembler. The contig that represented the complete (or nearly complete) clone insert DNA was exported in FASTA format. ORFs were identified within the complete insert sequence using a GeneMark heuristic model for prediction of prokaryotic genes (http://exon.gatech.edu/gmhmm2_prok.cgi). The ORF sequences were compared against the GenBank nr/nt database using BLASTx for predicting gene products.

3. HPLC and MS studies of clone P6L4. Clone P6L4 was selected for detailed biochemical and genetic characterization. A culture tube containing 10 ml of LB broth with 12.5µg/ml Cm and 0.01% arabinose was inoculated from a glycerol stocks clone P6L4 and the empty vector negative control, and the cultures were grown for 48 hours and then filtered to prepare a cell-free supernatant. A portion (5 ml) of the supernatant

was treated with 500µl of glacial acetic acid and the remainder (5 ml) was treated with 500µl of ammonium hydroxide to produce acidic and basic conditions, respectively. After mixing thoroughly, an equal volume of ethyl acetate was added to each sample and shaken vigorously to mix the two layers. The samples were allowed to separate into two distinct layers and the aqueous and organic phases were collected followed by drying at 60°C. The extracts were resuspended in 100 µl of sterile water thus achieving a 50-fold concentration that was tested in a bioassay against MRSA strain EAMC30 after 10-fold serial dilutions in LB broth.

Mid-scale cultures (500 ml) for clone P6L4 and the negative control were grown using LB broth containing 12.5µg/ml Cm and 0.01% arabinose. After 48 hrs, cultures were subjected to centrifugation at 10,000 x g for 10 min to pellet the cells and then the supernatant was filtered using a 0.2 µm bottle top filter to produce the cell-free supernatant. After extraction with ethyl acetate the samples were concentrated using a rotovap and resuspended in sterile water. After confirmation of activity against MRSA strain EAMC30, the extract was analysed by High Performance Liquid Chromatography (LC). Samples for both P6L4 and control were loaded on to a reverse phase C-18 column and a linear (0-100%) methanol gradient was used as the mobile phase. Fractions were collected at 1 ml/min for a 30 minute run, dried at 60°C using a centrivap, resuspended in sterile water and tested against MRSA strain EAMC30. A Cm control was also used for comparison where the empty vector clone was grown under the same conditions in LB containing arabinose but without addition of Cm. Cm was added at a concentration of 12.5µg/ml to the cell-free supernatant and then subjected to the same procedure as P6L4 and the negative control. Active fractions were then analyzed by LC-MS.

4. Detection of fluorescently labeled chloramphenicol analogue by TLC. A 2 ml LB broth culture containing 12.5µg/ml Cm and 0.01% arabinose of clone P6L4 and the negative control were grown at 37°C for 12 hrs at 200 rpm. A portion (600 µl) of this culture was then transferred to fresh medium (6ml LB broth containing 0.01% arabinose) and 30 µl of the BODIPY FL chloramphenicol (FAST CAT kit from Molecular PROBES) was added to each tube. Tubes with covered with aluminum foil and incubated in the dark (to protect from light) at 37°C for 48 hrs at 200 rpm. Then 1 ml aliquots were withdrawn every 12 hours (including T_0 until T_{48}) and extracted with an equal volume of ethyl acetate. The organic layer was transfered and stored in tubes for light sensitive material at -20°C. At the end of 48 hrs all of the extracts were dried in the centrivap at 60°C and resuspended in 15ul of ethyl acetate thereby concentrating the samples. A small portion (5 µl) of each sample was spotted on a silica gel TLC plate (1 cm from the base) and a mixture of chloroform and methanol (87:13) was used as the mobile phase. After the run the plate was visualized under UV light.

5. Amplification, cloning and expression of esterase genes from clone P6L4. This was done by using the Expresso Rhamnose SUMO Cloning and expression System (Lucigen Corporation, Middleton, WI). Custom primers were designed for amplification of the genes encoding three different esterases as follows:

<u>Putative esterase</u>

Forward primer –

5'- **CGCGAACAGATTGGAGGT**GCG CGATGGTCTTCTTTTAGT

Reverse primer –

5'- **GTGGCGGCCGCTCTATTA**TTAAGCGAAAGCGTCGCCGGG

Metallophosphoesterase

Forward primer –

5’- **CGCGAACAGATTGGAGGT**ATCTATGGCGTCAAAAAGGTA

Reverse primer –

5’- **GTGGCGGCCGCTCTATTA**CGGTGCCGCCCGCAGCGTAAT

Phospholipase/carboxylesterase family

Forward primer –

5’- **CGCGAACAGATTGGAGGT**CCTTTGCTGCATCAGTTCTAC

Reverse primer –

5’- **GTGGCGGCCGCTCTATTA**TTCATGGTGCAGCCCTCGGAA

For both the forward and the reverse primers in each case the 18 bases shown in bold correspond to the two ends of the pRham vector sequence adjoining the insertion site. The following sequence is that of the target coding region and in the case of the reverse primers it represents the reverse complement of the last 7 codons of the target coding region.

The design of the amplification protocol used was as follows:

A 50 µl reaction included 2.5 µl of the 10 µM stock of each primer (F’ & R’), 25 µl of the EconoTaq PLUS 2x Master Mix (Lucigen Corp.), 1 µl of the DNA template (~5 ng) and 19 µl of Nuclease free water.

The cycling conditions used were as follows:

The thermocycler was pre-heated to 94°C and for initial denaturation of the target DNA template the reactions were incubated at 94°C for 2 min. A total of 25 amplification cycles were performed with denaturation at 94°C for 15 sec, annealing at 55°C for 15 sec,

and extension at 72°C for one min. The final extension step was at 72°C for 10 min. Then 10 µl of the PCR product was loaded onto an agarose gel (SB gel run for 2 hrs at 165 V) for analysis. A PCR product was obtained for the putative esterase (E) and the carboxylesterase (Ce).

The PCR amplicon was cloned into the pRham vector using *E. cloni* 10G chemically competent cells (Lucigen Corp.). The cells and the vector DNA were thawed on ice, with 2 µl (25 ng) of vector DNA mixed with 1 µl of the PCR product and added to 40µl of the cells. The mixture was stired gently with a pipet tip so as to avoid any air bubbles and warm the cells. The mixture was then transferred to a pre-chilled 15 ml polypropylene culture tube and placed on ice for 30 min. The cells were heat shocked by placing the tube in a 42°C bead bath for 45 sec followed by 2 min on ice. Then 960 µl of recovery medium was added to the cells in the culture tube and incubated at 37°C for 1 hr at 200 rpm. Then 100 µl of the transformed cells were plated on YT agar plates containing 30µg/ml kanamycin (Kan) and plates were incubated overnight at 37°C.

Transformants from each plate (E and Ce) were then grown in YT broth with Kanat 37°C for 16 hrs at 200 rpm followed by extraction of plasmid DNA. PCR (design and conditions as before) was used to confirm the presence of the cloned insert by using two primer sets, the custom-designed primers and the SUMO forward and pETite reverse primers provided in the kit. The PCR product was analyzed as before by agarose gel electrophoresis and purified by the Wizard PCR Clean up System (Promega, Madison, WI). The amplified and purified DNA was sequenced and compared with the original sequence using the CLC Genomics Workbench and also analyzed by BLASTx.

To verify induction of protein expression a standard induction protocol was used. A 5 ml LB broth culture containing 30 µg/ml Kan was inoculated with the subclones (pRham-e and pRham-Ce) containing the respective pRham expression construct and incubated at 37°C at 200 rpm until the cultures reached an optical density of 0.4 at 600nm (OD_{600}). A 1 ml aliquot of the cultures was withdrawn and these uninduced cells were collected by centrifugation at 12,000 x g for 5 min. The pelleted cells were resuspended in 50 µl of the SDS-PAGE loading buffer and stored at -20 to be used as the uninduced control. To the remaining cultures rhamnose at a final concentration of 0.2% was added to induce expression and incubation was continued for 6 hrs. The OD_{600} was recorded and a 1 ml aliquot of each culture was processed as described previously. The induced samples were diluted appropriately to match the OD units of the uninduced samples. Samples added to the SDS-PAGE laoding buffer were heated to 95°C for 5 minutes followed by centrifugation for 1 minute at 12,000 x g. Evaluation of expression was done by SDS-PAGE analysis. The preparation of buffers, separating gel, stacking gel, staining/destaining solutions, and the electrophoresis was carried out as per the protocol in Molecular Cloning: A Laboratory Manual (Sambrook and Russell, 2001).

6. Confirmation of anti-MRSA activity of the P6L4 subclones. The esterase subclones pRham-e and pRham-Cet can be selected using Kan, but in order to evaluate their respective ability to modify Cm it was necessary to introduce an additional vector that confers Cm resistance. Both of the *E. coli* strains containing the pRham-e and pRham-Ce constructs were made electrocompetent by chilling log phase cultures to 4°C, pelleting the cells, washing the cells multiple times with cold 10% glycerol and then resuspending the cell pellet in cold 10% glycerol.and the competent cells were transformed by

electroporation (using conditions as described previously) with the pGNSBAC vector DNA. An aliquot (100µl) of the transformed cells were plated on YT agar plates containing 12.5µg/ml Cm and 30µg/ml Gm and plates were incubated overnight at 37°C. An appropriate negative control was also designed by using an *E. coli* strain containing the empty pRham vector and processing it similarly for electroporation with the pGNSBAC vector.

Each of the transformants were inoculated into 5 ml of LB broth containing 12.5µg/ml Cm, 30µg/ml Kan, 0.2% rhamnose, and additional inoculations were made into 5ml LB broth containing 12.5µg/ml Cm and 30µg/ml Kan without any added rhamnose. Cultures were incubated at 37°C for 48 hrs at 200 rpm and cell-free supernatants were collected followed by testing against MRSA strain EAMC30 using the 96-well microtiter plate bioassay. Results for induced expression with addition of rhamnose and in the absence of rhamnose were compared for the inhibiton of MRSA growth.

7. Comparative codon usage for clone P6L4. The frequency of codon usage was calculated for the complete insert of clone P6L4 using EMBOSS, The European Molecular Biology Open Software Suite (Rice et al., 2000) and compared against that for the complete genome of *Esherichia coli strain K12 substrain DH10B* and *Candidatus Solibacter usitatus Ellin6076*, a soil bacterial species in subgroup 3 of the *Acidobacteria* phylum.

D. RESULTS

1. Preliminary characterization of active clones. The results from all the tests used for

the preliminary characterization of the 33 clones are summarized in Table 4.1. Among the 33 clones tested, for 31 clones it was observed that the cell-free supernatant had slightly higher or equivalent activity as compared to the cell lysate (Figure 4.1). This suggests that the active compound(s) is most likely extracellular or readily secreted out of the cell for most of the clones. For two clones, namely P6B5 and P37O10 the activity in the cell lysate was significantly higher than the minimal activity detected in the cell free supernatant. This presumably shows that the active compound(s) is intracellular or membrane-associated, not readily secreted out of the cell and/or freely soluble in the supernatant. For all of the subsequent tests for clones P6B5 and P37O10 the cell lysate was used in bioassays, whereas the cell-free supernatant was used for the other 31 clones. The activity for each clone or clone fraction was calculated as % reduction of resazurin fluorescence as compared to the respective negative control. For the negative control, both the cell free-supernatant and the cell lysate were tested so that a true comparison could be made with the respective clone samples.

Cell-free supernantants or lysates after heating for 10 minutes in a boiling water bath and bioassay against MRSA strain EAMC30 showed varied results for the different metagenomic clones. Half of the clones showed no significant loss in activity indicating that the active compound(s) is heat stable and is likely to be non-proteinaceous. Among the remaining clones there was a significant drop in activity and four of the clones showed complete lost of activity as a result of the boiling treatment suggesting that the active compound(s) is heat sensitive and is likely proteinaceous. In the fractionation assay, filtrate obtained after separation using a 3 kDa MWCO membrane was tested. For 11 of the 33 clones the filtrate (<3 kDa fraction) did not show a loss of activity. For 14

clones, even though there was a decrease in the % activity the overall inhibition of MRSA strain EAMC30 resazurin-derived fluorescence was still more than 50%. For 8 clones there was a significant drop in the activity with less than 50% inhibition. For a subset of the clones that were tested against the bioluminescent MRSA strain, along with recording luminescence an image showing the light output from the cells growing in the microtiter plate was captured (Figure 4.2). Actively growing MRSA gave a strong light output, which was color-coded for intensity analysis whereas MRSA that was inhibited did not produce light.

In the bioassay testing with MRSA strain Xen 31, more than 50% inhibition of tester culture growth was seen by clones P29E3, P6L4, P18N22, P28I7, and P28L21. These clones are among the ones that were top candidates selected for the 454 sequencing analysis, thus indicating the presence of genes involved in production of antibacterial compound(s). The remaining clones that were also tested showed less than 50% or no activity. This is not entirely surprising considering different strains of tester bacteria may have a different growth inhibition pattern when tested against secondary metabolites from clone cultures.

2. Sequence analysis. Good quality DNA sequences were obtained for 17 clones using 454 pyrosequencing. Sequence reads were trimmed and CLC Genomics Workbench 4.9 was used for *de novo* assembly. Contigs that represented the entire clone insert, with high number of reads and typically > 50x coverage were selected for analysis. For the five clones from metagenomic libraries SL1 and SL3 and two of the clones from SL5 multiple contigs were generated for the clone insert whereas for the remaining 10 clones the entire insert was obtained after *de novo* assembly. Since the insert size of clones from Library

SL5 is much larger than that of the other library clones from SL1 or SL3, the number of predicted ORFs in each of the SL5-derived clones was much greater with more than 100 ORFs per most clones. Annotation of clones from Libraries SL1 & SL3 showed the presence of many hypothetical proteins and no significant hits in GenBank or as-yet-unassigned functions suggesting the probability of the presence of novel genes contained within the cloned DNA. The annotations for the 12 clones from SL5 are summarized in Table 4.2. ORF maps for all the inserts and contigs were generated by importing the annotations from NCBI BLASTx into the CLC Genomics workbench. A plot of the % G+C for each of these contigs/insert sequences was obtained by using the program cpgplot by the European Bioinformatics Institute. The %G+C plot is depicted together with the cloneannotations in Figure 4.3A to 4.3Y.

3. Characterization of clone P6L4. Although each of the 33 clones was a promising candidate for further characterization, clone P6L4 was selected first for biochemical and genetic studies. It had shown the best and consistent results over the entire preliminary and validation tests in bioassays against all MRSA strains (Figure 4.2). Also, clone P6L4 was easy to work with since the cell-free supernatant was shown to almost completely inhibit the growth of tester strains. Organic extraction with ethyl acetate under basic conditions was found to be the best suited approach with retention of activity in the dried and resuspended large-scale extracts. The concentrated extracts were then subjected to LC analysis for separation of the active fraction from other components in the extract. The negative control extract was processed in parallel for comparison. Among the 0-30 minute fractions collected by HPLC, fractions with elution time between 20-21 minutes always showed complete inhibition of the MRSA strain EAMC30. Curiously, even the

negative control fractions in the same range inhibited the growth of the MRSA strain. The chromatograms for the negative control and P6L4 showed similar peaks for the 20-21 minute elution time although they had different intensities. A Cm reference sample also showed a very similar chromatogram pattern. Chromatograms for all three are shown in Figure 4.4A.

To investigate this further, a Cm control was used for comparion with P6L4, wherein 12.5µg/ml Cm was added to the culture after 48 hr incubation. This provided a similar background profile to the spent supernatant from the P6L4 culture extracts and also served as the Cm reference standard. LC fractions shown to be active by bioassay were analyzed by LC-MS the results for which can be seen in Figure 4.4 B. The most dominant peak in the LC-MS for P6L4 and Cm control had the same position. The mass spectra of these peaks under the negative ion mode gave identical results for P6L4 and the Cm control. The highest abundance ratio in each case was a compound with 321 mass ion and elemental composition similar to Cm. These results indicate the presence of a similar compound in P6L4 and the Cm control, which is most likely Cm. Comparison of the retention time in LC, the absorption maxima and the identical LC-MS profiles suggests that the active compound from P6L4 was Cm. Ideally the Cm added in the cultures should be inactivated by the chroramphenicol acetyl transferase (CAT) encoded on the BAC vector. This is true in case of the negative control in which Cm was still detected after 48 hrs albeit at a very low concentration that was sub-inhibitory for MRSA. But clearly in the case of P6L4 the concentration was high enough for inhibition, indicating that CAT activity is counteracted in P6L4 thus reactivating the Cm activity.

Enzymatic reactivation of chloramphenicol by chloramphenicol acetate esterases that counteract the CAT activity has been reported previously (Nakano et al., 1977; Sohaskey and Barbour 1999; Sohaskey and Barbour 2000; Sohaskey 2004). A similar mechanism is probably responsible for the activity of P6L4 since the insert sequence had multiple ORFs with esterases as the predicted gene product. Three different esterases were predicted, including a putative esterase (E), a carboxylesterase (Ce) and a metallophosphoesterase (MPe).

4. **Detection of fluorescently labeled chloramphenicol analogue by TLC.** To determine the esterase activity of P6L4, a fluorescent BODIPY FL Cm substrate (BCAM) was added to the cultures. TLC results (Figure 4.5) showed that the concentration of BCAM in the negative control decreased (T_0-T_{48}) and that of the acetylated forms increased over time. Two different acetylated forms of BCAM were observed in the negative control whereas only one of these was seen in P6L4. A reverse trend was seen in P6L4 wherein the BCAM concentration increased and the acetylated form decreased over time suggesting reactivation of Cm by CAE activity. The genes encoding three different esterases in P6L4 were subcloned to investigate this and ascertain the role of the esterase.

5. Amplification, cloning and expression of esterase genes from clone P6L4. Two of the esterase encoding genes (e and ce) were successfully amplified (Figure 4.6 A) using the custom designed primers. Multiple rounds of PCR using a gradient, touchdown, and varying concentrations of primers and template did not yield an amplicon for mpe and therefore only e and ce were used for cloning into the pRham vector. Transformants from both subclone e and subclone ce were used for DNA extraction and upon amplification

showed a strong PCR product of expected size (figure 4.6 B). The sequences of the amplicons from the esterase subclones aligned with the original gene sequences from the P6L4 insert. Further validation of the successful cloning of the esterases was seen from the results of SDS-PAGE analysis (Figure 4.6 C). In case of both E and Ce the induced sample clearly showed greater concentration of the protein at an approximate molecular mass predicted for each respective esterase as compared that of the uninduced sample.

6. Confirmation of anti-MRSA activity of the P6L4 subclones. After confirming that the two esterases had been cloned, the evaluation of their role in the activity of clone P6L4 was conducted by testing the subclones in a bioassay against the MRSA strain EAMC30. Transformation of the subclones (and the empty vector control) with pGNSBAC enabled growth in LB broth containing Cm. Inhibition of MRSA growth was observed for subclone pRham-e but not for pRham-ce. Also, as expected the inhibition was much higher when rhamnose was added for induced expression (Figure 4.7). These results indicate that the putative esterase is most likely responsible for the anti-MRSA activity observed in clone P6L4 supernatants.

7. Comparative codon usage for clone P6L4. The top BLASTx hit for the putative esterase from clone P6L4 was *Candidatus Solibacter usitatus Ellin6076*, a soil bacterial species in subgroup 3 of the phylum *Acidobacteria*. Hence, it is very likely that the DNA in the P6L4 insert is of Acidobacteria origin. Codon usage was calculated to gain a better idea of similarity between the insert sequence of P6L4 and a known *Acidobacteria* genome sequence. A codon is a triplet (three nucleotide series) that encodes a specific amino acid residue (61 codons) in the polypeptide chain or terminates translation (3 stop codons). Thus there are 64 codons but only 20 amino acids leading to many amino acids

being encoded by more than one specific codon and the genetic code is said to be degenerate. However, different organisms show preference for a particular codon over other codons encoding the same amino acid. The frequency of use of this codon is greater than that expected by chance. The codon usage analysis of the P6L4 insert sequence showed a preference for particular synonymous codons. Codon usage in *Escherichia coli strain K12 substrain DH10B* (the host organism) was also determined. A compasrison of the frequency of each codon was done for all three, the P6L4 complete insert and the complete genome sequences of *S. usitatus* and the *E. coli* strain (Figure 4.8). As seen in the graphical representation, there are many similarities in the codon usage pattern of *Solibacter* and *E. coli DH10B.* Differences in codon usage preference among organisms lead to a variety of problems concerning heterologous gene expression but the fact that there were no significant differences in codon usage might explain the likely success of heterologous expression of the esterase in *E. coli.* The %G+C content of the putative esterase gene from P6L4 was 60.42%, and that of the *Solibacter* complete genome is 61.90% providing additional support for the *Acidobacteria* origin of clone P6L4.

E. DISCUSSION

Function-based metagenomics has enabled the annotation of many proteins previously listed as hypothetical proteins in the GenBank database. It is a useful methodological approach that complements sequence-driven metagenomic analysis of microbial communities for discovery of novel genes and gene products. Recombinant clones identified from the screening of metagenomic libraries were characterized to gain information about the basic properties of the active compound(s). The clones characterized in this study showed the presence of growth inhibitory activity against

multiple MRSA strains. For many of the clones, the active compound(s) are heat resistant (non-proteinaceous compounds) and easily secreted out of the host cell. Active supernatants on passage through a size exclusion membrane showed that the estimated molecular weight for most of the active compunds was less than 3kDa. DNA sequence analysis was conducted for prediction of genes in the cloned metagenomic DNA. The cloned DNA from some active clones is predicted to have genes involved in polyketide enterocin synthesis and isoprenoid biosynthesis. For more than one recombinant clone the top BLAST hit for many ORFs was from a member of the phylum *Acidobacteria*, members of which are likely to contain PKS-related genes according to a recent genome sequencing study (Ward et al., 2009; Parsley et al., 2011) and are known to be involved in the synthesis of polyketides (Staunton & Weissman, 2001; Stinear et al., 2004). At least two clones contain a predicted gene product that is likely a Radical SAM (S-adenosylmethionine) domain protein. Proteins belonging to this superfamily function in antibiotic and herbicide biosynthesis pathways. Many of the gene products from these metagenomic clones are hypothetical proteins of unknown function and may be indentified by further experiments as used for describing the genes and gene product from clone P6L4.

The putative esterase from clone P6L4 is responsible for reactivation of Cm as supported by the results from the biochemical and genetic studies. A common mechanism of Cm resistance is the inactivation of Cm by chloramphenicol acetyltransferase (CAT), by addition of an acetyl group to C3 of Cm resulting in 3-acetyl Cm which is then converted to 1-acetyl Cm and may also be acetylated at both C1 and C3 by CAT to form 1,3-diacetyl Cm (Nakagawa et al., 1979). The putative esterase activity counteracts the

CAT mechanism encoded by the *cat* gene on the cloning vectors used. An important objective of metagenomic studies is to gain access to the genomes of as yet unculturable microorganisms. The BAC libraries in this study have given an insight into the genetic composition of the cloned metagenomic DNA that is representative of the microbial assemblage of the sampled soil. With any metagenomic analysis there is always the possibility of discovering housekeeping genes along with discovery of novel genes encoding the function of interest. Various screens can be designed for detecting other functions from the BAC libraries and these may provide more information about the cloned gene inserts. Metagenomic studies are multi-faceted and can be used not only for gene discovery but also for mining information about the regulatory processes, codon usage, gene organization and gene expression in the uncultured microorganisms that constitute the majority of microbial communities in any natural environment.

As this study illustrates, each metagenomic clone can contain a unique combination of genetic elements and biochemical products, such that each clone requires separate analyses. In this study P6L4 was selected on the basis of rational criteria for targeting the best drug-like antibiotic candidate from this metagenomic library. Many other clones identified in this study await further investigation. The progress made in these studies toward generation of large-insert metagenomic libraries in shuttle BAC vectors will be applied in the future for generation of larger-scale libraries that can encompass a greater diversity of soil microbial metagenomes and be expressed in multiple hosts. The progress made toward development of novel screening methods in this study will be very necessary in evaluating the larger-scale libraries that are produced. In total, this thesis research represents a proof-of-concept for application of a functional

metagenomic approach in identifying antimicrobial-expressing recombinant clones from a large-insert soil metagenomic library. Future research will mine the unique functions unearthed from these efforts.

E. FUTURE WORK

The BAC vectors used in the construction of these metagenomic libraries allow transfer and stable maintenance of the cloned insert DNA into different hosts (Gram negative and or Gram positive). An increase in the number of active clones and in the diversity of the antimicrobial compound(s) may be achieved by using multiple heterologous hosts for screening. Based on the sequence information available currently for the validated clones, an alternative bacterial expression host may be selected as the best matched host for a specific clone. For example clones with insert DNA that is possibly of *Acidobacterial* origin can be transferred to this host which may lead to an increased expression from the native promoters.

Further studies also include testing validated clones against a broader panel of bacterial tester strains including certain pathogenic strains. *Aeromonas hydrophila, Bacillis spp, Legionella pneumophila, Campylobacter jejuni and Mycobacetria* are some of the pathogens that will be tested for susceptibility to the clones. Fungal and yeast species will also be used as tester strains in similar bioassays as described earlier. This will increase the probability of discovering a broad spectrum antimicrobial and results from the bioassays will be helpful in determining the potency of the expressed bioactive compound(s) against the different tester species. Large scale cultures for each clone of interest will be grown for the production of bioactive compound(s) in higher quantity which will help in the elucidation of chemical structure. For the most promising lead

candidates, characterization of active compounds will include determination of the chemical structure and testing for potency, toxicity and efficacy in an animal model.

Table 4.1. Preliminary characterization of active metagenomic clones. MRSA strain EAMC30 was used as the tester strain in all bioassays, unless otherwise indicated. Percent inhibition values were calculated in comparison with the corresponding empty vector negative control by measuring the fluorescence of reduced resazurin.

Clone ID	%Reduction of Resazurin fluorescence	% Inhibition of MRSA viability relative to the empty vector negative control…				
		againstMRSA Xen 31	in the cell lysate	in the cell-free supernatant	in the lysate*/ supernatant boiled at 10 min	in the < 3 kDa fraction
P11K11	39.6	0.0	14.6	60.4	65.7	41.2
P15G24	9.7	33.1	50.1	90.3	80.2	57.8
P17L5	45.2	20.2	53.8	54.8	75.1	54.6
P29E3	21.5	58.8	61.6	78.5	79.6	46.9
P30A5	26.1	46.0	0.0	73.9	67.9	44.5
P2P12	52.0	0.0	34.6	48.0	56.7	59.6
P3A13	8.1	1.3	89.0	91.9	86.4	38.7
P5A4	63.7	1.6	3.6	36.3	24.7	52.2
P5C24	17.8	24.2	84.2	82.2	68.7	65.0
P6B5	43.1	41.3	56.9	19.3	92.2*	76.5
P6L4	7.5	78.1	88.3	92.5	89.1	85.4
P6L5	25.7	45.1	6.8	74.3	14.9	54.4
P9L21	7.8	13.9	5.7	92.2	58.4	71.8
P14O1	75.2	0.0	12.6	24.8	0.0	36.7
P18N22	10.8	54.3	5.9	89.2	60.8	72.3
P20I6	11.4	2.9	55.1	88.6	78.5	22.9
P22C4	60.5	0.0	24.0	39.5	0.0	39.1
P22E10	53.4	0.0	7.6	46.6	71.7	54.9
P23K15	7.7	0.0	86.7	92.3	86.5	81.0
P27K16	11.6	0.0	15.2	88.4	8.3	52.8
P27M10	7.7	10.7	87.0	92.3	88.3	82.7
P28H1	8.6	22.4	86.8	91.4	80.8	76.5
P28I7	8.2	61.1	72.8	91.8	84.4	81.3
P28L21	8.4	57.8	70.7	91.6	75.6	78.7
P31G24	8.3	0.0	87.9	91.7	86.3	73.3
P35B14	39.6	31.2	34.9	60.4	0.0	48.2
P36M1	56.0	0.0	5.5	44.0	8.9	55.8

P37A9	54.0	0.0	2.8	46.0	7.0	3.5
P37A11	59.5	0.0	15.6	40.5	0.0	53.7
P37O10	41.2	0.0	58.8	29.1	90.6*	16.0
P43A3	66.5	0.0	3.8	33.5	28.5	46.2
P46O24	8.6	16.5	50.1	91.4	55.0	65.7
P49M4	10.0	14.9	29.7	90.0	7.4	55.7

Figure 4.1. Clone activity in cell lysates and cell-free supernatants.

Comparison of the two different treatments used for processing 48 hour old metagenomic clone cultures prior to bioassay. Lysates and cell free supernatants for each were tested in parallel against MRSA strain EAMC30. The % growth inhibition of MRSA (Yaxis) by clones (X axis) relative to the empty vector negative control (considered to have no inhibitory effect) was calculated by measuring the fluorescence of reduced resazurin.

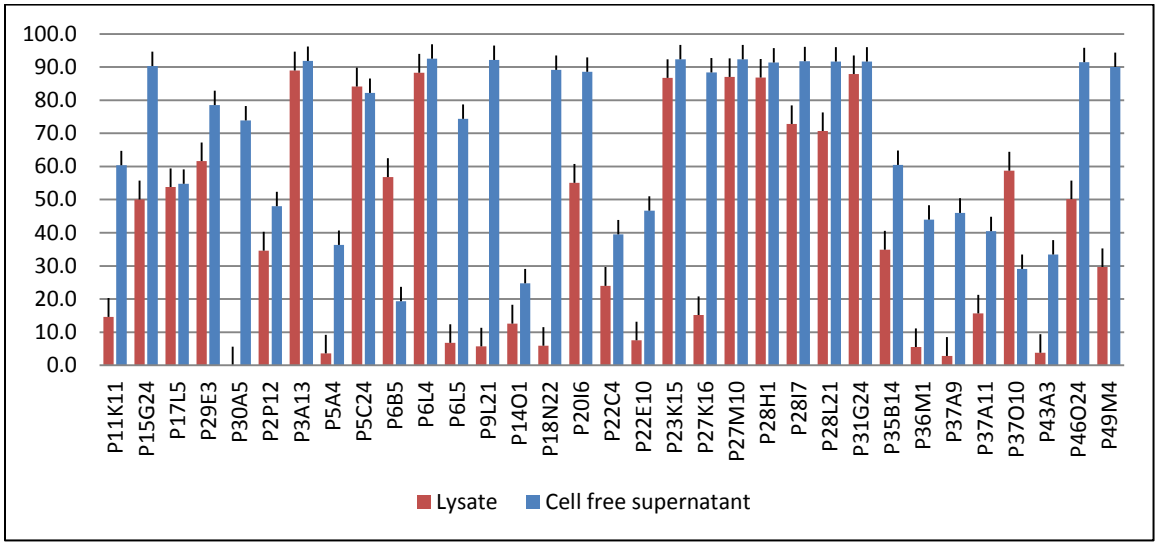


Figure 4.2. Inhibition of the growth of a bioluminescent MDR *S. aureus* strain by supernatants from metagenomic clones.

Actively growing MRSA was observed to have strong bioluminescent emission, which was color-coded for intensity analysis. MRSA that did not grow did not emit bioluminescence.

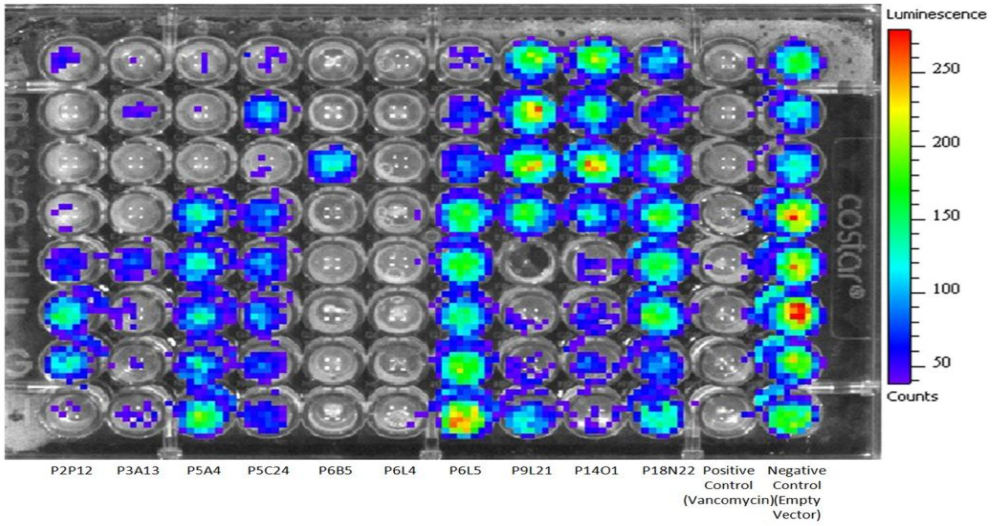
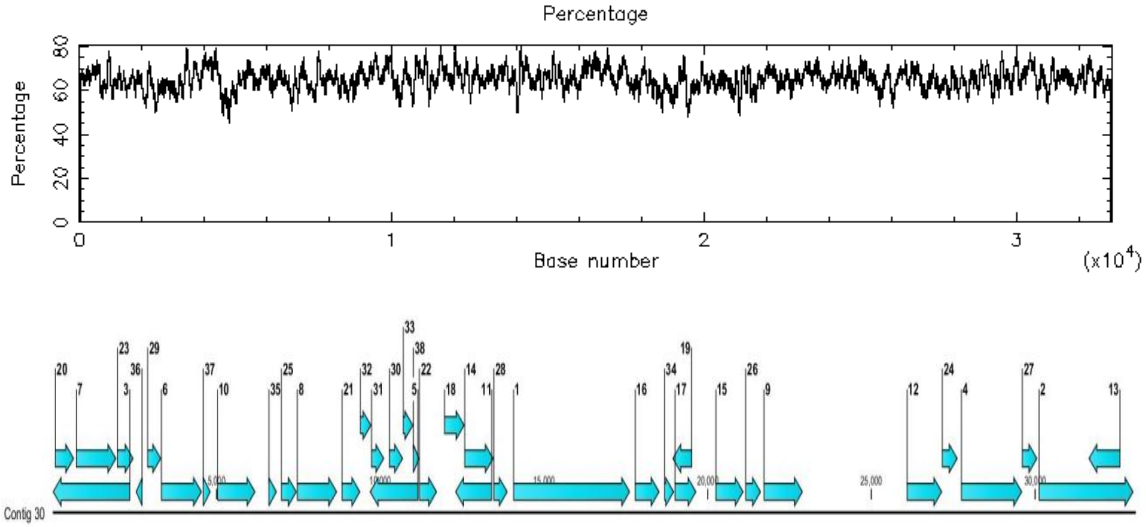


Figure 4.3A to 4.3Y. DNA sequence and annotation for 17 antibacterial metagenomic BAC clones.

For each clone a plot of % G+C content and an ORF map for the contiguous sequence contained within the recombinant clone insert is depicted.

A. P11K11 contig 30

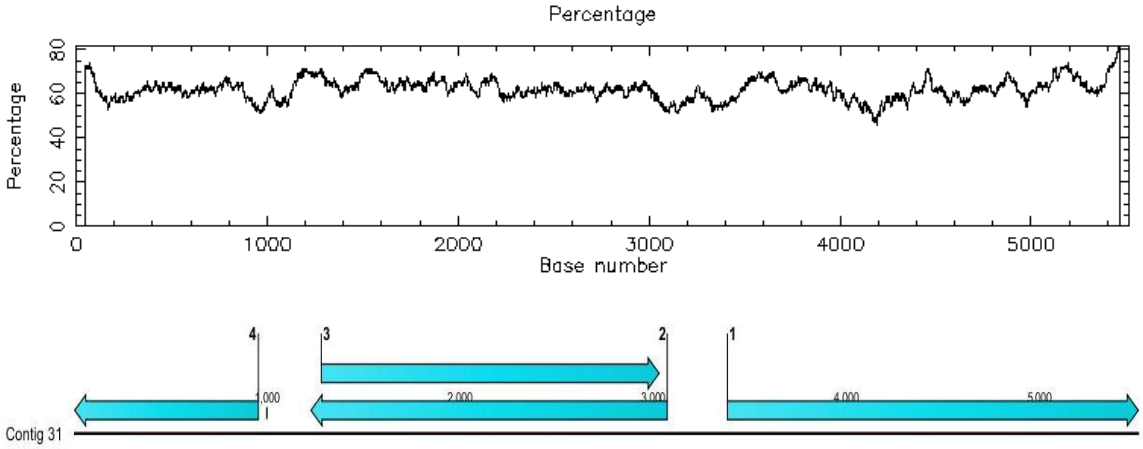


ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	3564	hypothetical protein Atu0967	*[Agrobacterium tumefaciens str. C58]*	0	48
2	2892	glutamate-ammonia-ligase adenylyltransferase	*[Brucella suis ATCC 23445]*	0	47
3	2364	sarcosine oxidase alpha subunit protein	*[Agrobacterium radiobacter K84]*	2E -24	27
4	1863	cytochrome c-type biogenesis protein CcmF	*[Stappia aggregata IAM 12614]*	0	63
5	1482	hypothetical protein Atu0961	*[Agrobacterium tumefaciens str. C58]*	1E-21	40
6	1245	hypothetical protein AZC_1105	*[Azorhizobium caulinodans ORS 571]*	2.00E-115	57

7	1230	conserved hypothetical protein	*[Serratia odorifera DSM 4582]*	2.00E-75	39
8	1221	Phage major capsid protein, HK97	*[Nitrobacter winogradskyi Nb-255]*	1.00E-137	62
9	1194	hypothetical protein OCAR_6577	*[Oligotropha carboxidovorans OM5]*	2.00E-08	24
10	1164	phage portal protein, HK97 family	*[Starkeya novella DSM 506]*	1.00E-122	60
11	1125	hypothetical protein Bru83_04273	*[Brucella sp. 83/13]*	1.00E-02	33
12	1080	cytochrome c-type biogenesis protein, putative	*[Ochrobactrum anthropi ATCC 49188]*	2E-39	37
13	972	hypothetical protein NB311A_12644	*[Nitrobacter sp. Nb-311A]*	5.00E-07	27
14	891	hypothetical protein METDI2079	*[Methylobacterium extorquens DM4]*	3.00E-65	48
15	855	ATPase, histidine kinase-, DNA gyrase B-, and HSP90-like domain protein	*[Labrenzia alexandrii DFL-11]*	2.00E-62	50
16	750	hypothetical protein BAZG_01351	*[Brucella sp. NVSL 07-0026]*	7.00E-44	51
17	660	two component response regulator	*[Agrobacterium vitis S4]*	3.00E-101	82
18	621	hypothetical protein RPB_3461	*[Rhodopseudomonas palustris HaA2]*	5.00E-64	61
19	597	family S13 unassigned peptidase	*[Burkholderia pseudomallei BCC215]*	6.00E-14	33
20	579	dnaK-type molecular chaperone dnaK	*[Mesorhizobium loti MAFF303099]*	8.00E-35	44
21	567	phiE125 gp8 hypothetical protein	*Hyphomicrobium denitrificans ATCC 51888]*	5.00E-33	41
22	549	hypothetical protein RPA1902	*[Rhodopseudomonas palustris CGA009]*	2.00E-33	47
23	501	hypothetical protein Oant_1708	*[Ochrobactrum anthropi ATCC 49188]*	1.00E-53	65

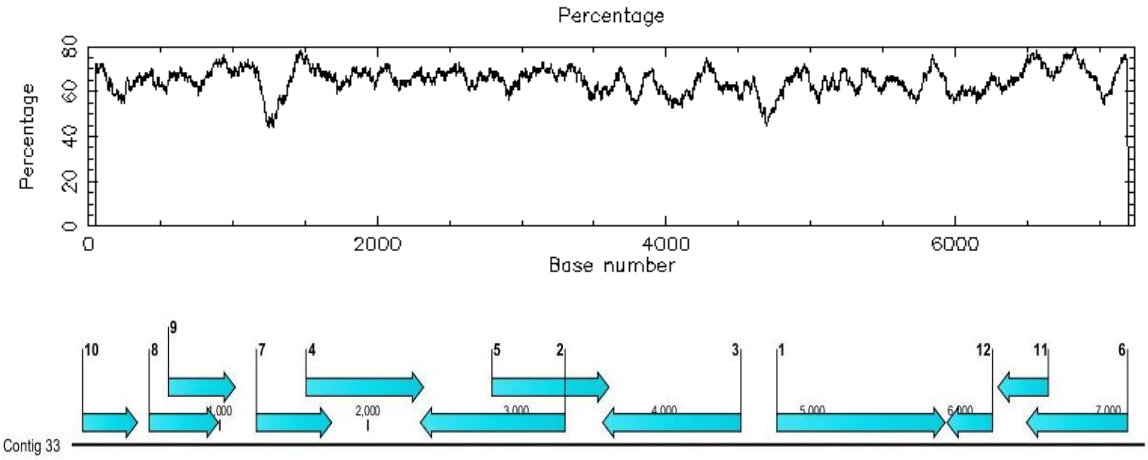
24	477	CcmE/CycJ protein	*[Methylocella silvestris BL2]*	5.00E-41	59
25	477	phage prohead protease, HK97 family	*phage prohead protease, HK97 family*	2.00E-35	53
26	474	putative LipA protein	*[Azorhizobium caulinodans ORS 571]*	7.00E-15	36
27	468	cytoChrome c-type biogenesis protein CcmH	*[Roseibium sp. TrichSKD4]*	3.00E-42	62
28	432	hypothetical protein SPO2251	*[Ruegeria pomeroyi DSS-3]*	2.00E-31	55
29	423	conserved hypothetical protein	*[Enhydrobacter aerosaccus SK60]*	6.00E-06	33
30	414	TP901-1 family phage major tail protein	*[Parvibaculum lavamentivorans DS-1]*	2.00E-47	65
31	396	putative phage tail protein p028	*[Bacillus phage SPP1] [Bacillus amyloliquefaciens DSM7]*	1.00E-08	33
32	342	phage head-tail adaptor	*[Desulfarculus baarsii DSM 2075]*	2.00E-08	33
33	315	gene transfer agent (GTA) like protein	*[Nitrobacter hamburgensis X14]*	6.00E-29	69
34	300	hypothetical protein SADFL11_311	*[Labrenzia alexandrii DFL-11]*	8.00E-03	37
35	249	hypothetical protein Nwi_1162	*[Nitrobacter winogradskyi Nb-255]*	4E-13	52
36	249	hypothetical protein mll6859	*[Mesorhizobium loti MAFF303099]*	2.00E-08	50
37	231	hypothetical protein Rru_A2704	*[Rhodospirillum rubrum ATCC 11170]*	3.00E-12	54
38	192	hypothetical protein RPA1900	*[Rhodopseudomonas palustris CGA009]*	7.00E-16	66

B. P11K11 CONTIG 31



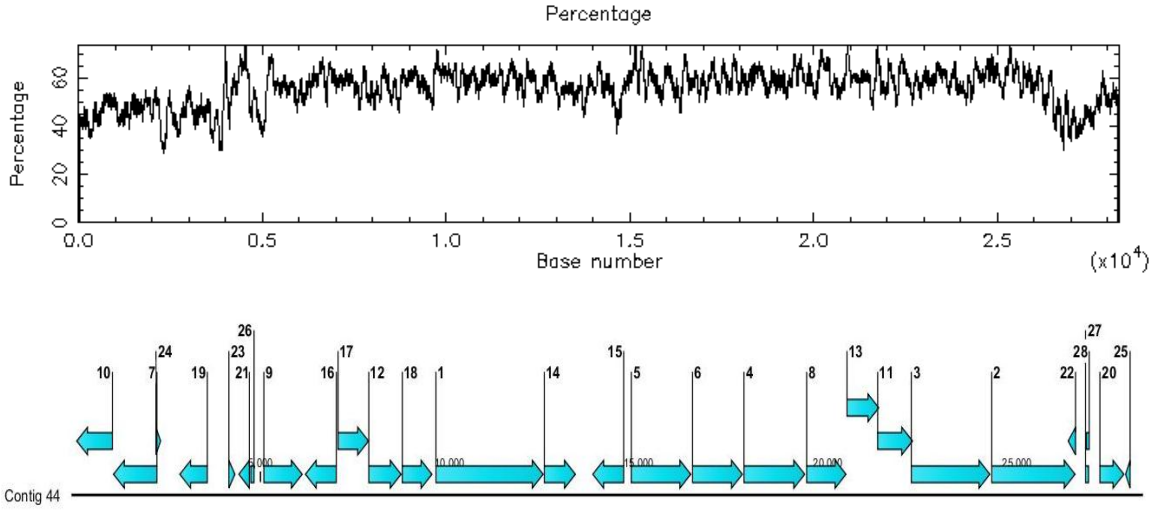
ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	2133	aminopeptidase N	*[Ochrobactrum intermedium LMG 3301]*	0	58
2	1871	ABC transporter related protein	*[Mesorhizobium opportunistum WSM2075]*	0	66
3	1755	cobyrinic Acid a,c-diamide synthase	*[Neisseria elongata subsp. glycolytica ATCC 29315]*	4.00E-36	33
4	957	probable ATP-binding/permease fusion ABC transporter	*[Stappia aggregata IAM 12614]*	1.00E-56	37

C. P11K11 CONTIG 33



ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	1143	hypothetical protein Arad_1482	*[Agrobacterium radiobacter K84]*	7.00E-04	26
2	984	chromate transport protein	*[Agrobacterium tumefaciens str. C58]*	1.00E-88	62
3	942	conserved hypothetical protein	*[Pseudovibrio sp. JE062]*	4.00E-27	37
4	801	hypothetical protein Avi_7458	*[Agrobacterium vitis S4]*	1.00E-64	65
5	798	A, transposase OrfB	*[Burkholderia mallei SAVP1]*	2.00E-05	31
6	690	ErfK/YbiS/YcfS/YnhG family protein	*[Desulfotomaculum acetoxidans DSM 771]*	2.00E-05	32
7	516	transcriptional regulator	*[Agrobacterium vitis S4]*	1.00E-50	69
8	474	conserved hypothetical protein	*[Pseudovibrio sp. JE062]*	2.00E-12	29
9	459	hypothetical protein	*[Podospora anserina S mat+]*	0.002	31
10	378	hypothetical protein Oant_2678	*[Ochrobactrum anthropi ATCC 49188]*	3.00E-18	41
11	348	Peptidoglycan-binding domain 1 protein	*[Sinorhizobium meliloti BL225C]*	7.00E-06	42
12	315	hypothetical protein RPC_1781	*Rhodopseudomonas palustris BisB18]*	3.00E-10	38

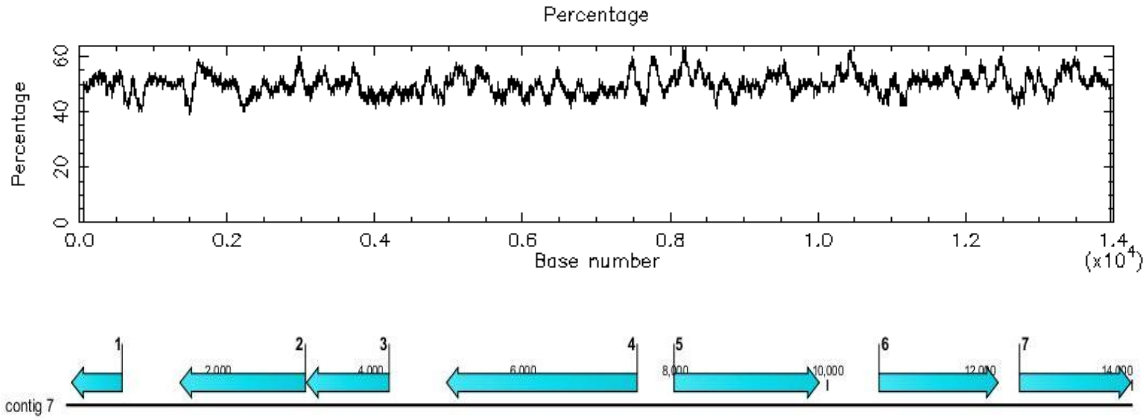
D. P11K11 CONTIG 44



ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	2862	hypothetical protein	*[Paramecium tetraurelia strain d4-2]*	8e-122,	34
2	2232	NAD(P) transhydrogenase beta subunit	*Ruminococcus albus 8*	2.00E-26	89
3	2103	hypothetical protein	*[Candidatus Kuenenia stuttgartiensis]*	3.00E-31	26
4	1635	probable calmodulin	*[Planctomyces maris DSM 8797]*	1.00E-27	27
5	1599	hypothetical protein GobsU_09848	*[Gemmata obscuriglobus UQM 2246]*	5.00E-69	37
6	1344	hypothetical protein PM8797T_09209	*[Planctomyces maris DSM 8797]*	1e-119,	56
7	1176	protoporphyrinogen oxidase	*[Vibrio cholerae V52]*	0	99
8	1068	methanol dehydrogenase regulator (moxR)-like protein	*[Rhodopirellula baltica SH 1]*	9.00E-106	55
9	1041	WD40 repeat, subgroup	*[Ktedonobacter racemifer DSM 44963]*	2.00E-49	39
10	972	plasmid-partitioning protein	*[Plasmid F]*	0	99

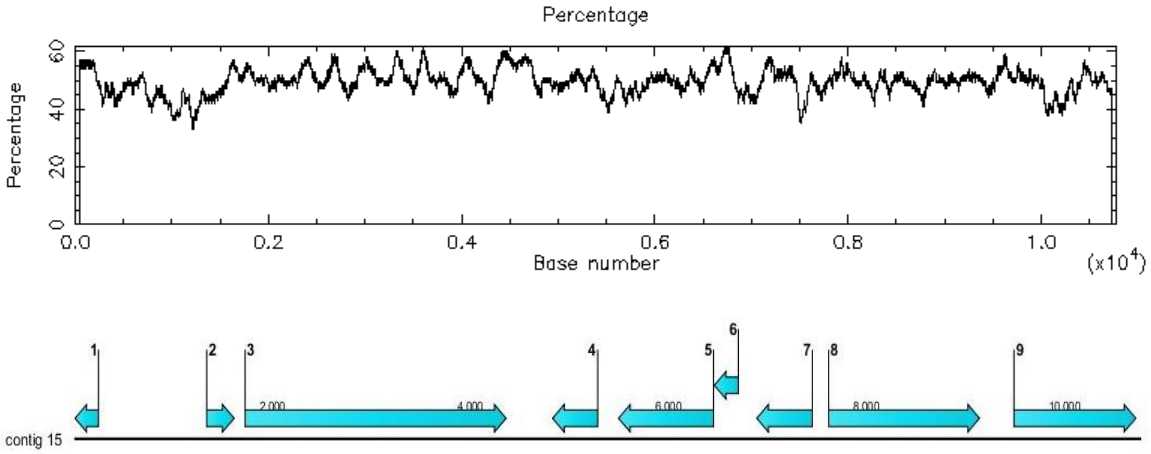
11	948	hypothetical protein DSM3645_19608 [Blastopirellula marina DSM 3645]		4.00E-63	43
12	882	hypothetical protein PM8797T_23941	*[Planctomyces maris DSM 8797]*	3.00E-07	26
13	867	hypothetical protein PM8797T_23941	*[Planctomyces maris DSM 8797]*	4.00E-10	29
14	849	hypothetical protein PM8797T_23941	*[Planctomyces maris DSM 8797]*	2.00E-11	27
15	846	hypothetical protein PM8797T_23941	*[Planctomyces maris DSM 8797]*	1.00E-15	31
16	846	hypothetical protein PM8797T_23941	*[Planctomyces maris DSM 8797]*	8.00E-13	29
17	834	hypothetical protein PM8797T_23941	*[Planctomyces maris DSM 8797]*	1.00E-15	30
18	810	hypothetical protein PM8797T_23941	*[Planctomyces maris DSM 8797]*	4.00E-05	34
19	756	replication protein	*[Plasmid F]*	1.00E-147	100
20	600	hypothetical protein CLOSCI_03331	*[Clostridium scindens ATCC 35704]*	3.00E-130	100
21	306	hypothetical protein pU302L_094	*[Salmonella enterica subsp. enterica serovar Typhimurium]*	8.00E-10	100
22	222	LacOPZ-alpha peptide from pUC9; putative	*[unidentified cloning vector]*	3.00E-20	90
23	189	conserved hypothetical protein	*[Escherichia coli MS 196-1]*	4.00E-27	100
24	153	hypothetical protein EcE24377A_E0023	*[Escherichia coli E24377A]*	4.00E-20	98
25	147	hypothetical protein ECH7EC4501_6204	*[Escherichia coli O157:H7 str. EC4501]*	2.00E-10	100
26	111	putative reverse transcriptase	*[Platanus x acerifolia]*	7.00E-05	83
27	108	conserved domain protein	*[Escherichia coli MS 84-1]*	1.00E-11	98
28	102	conserved hypothetical protein [Enterococcus faecalis AR01/DG]	*[Enterococcus faecalis AR01/DG]*	3.00E-10	100

E. P15G24 CONTIG 7



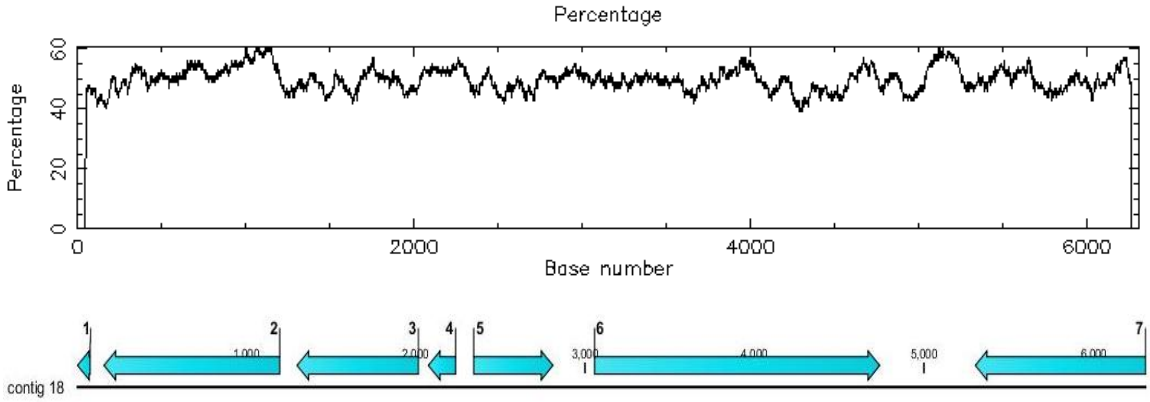
ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	678	Ankyrin	*Sulfolobus islandicus Y.N.15.51*	4.00E-14	32
2	1662	no significant hit			
3	1101	no significant hit			
4	2520	hypothetical protein CLOSTMETH_01752	*Clostridium methylpentosum DSM 5476*	3.00E-18	36
5	1923	pre-neck appendage preprotein	*Bacillus phage Nf*	3.00E-15	25
6	1581	hypothetical protein NEUTE2DRAFT_148577	*Neurospora tetrasperma FGSC 2509*	2.00E-18	47
7	1479	outer membrane pathogenesis protein	*Agrobacterium radiobacter K84*	1.00E-04	24

F. P15G24 CONTIG 15



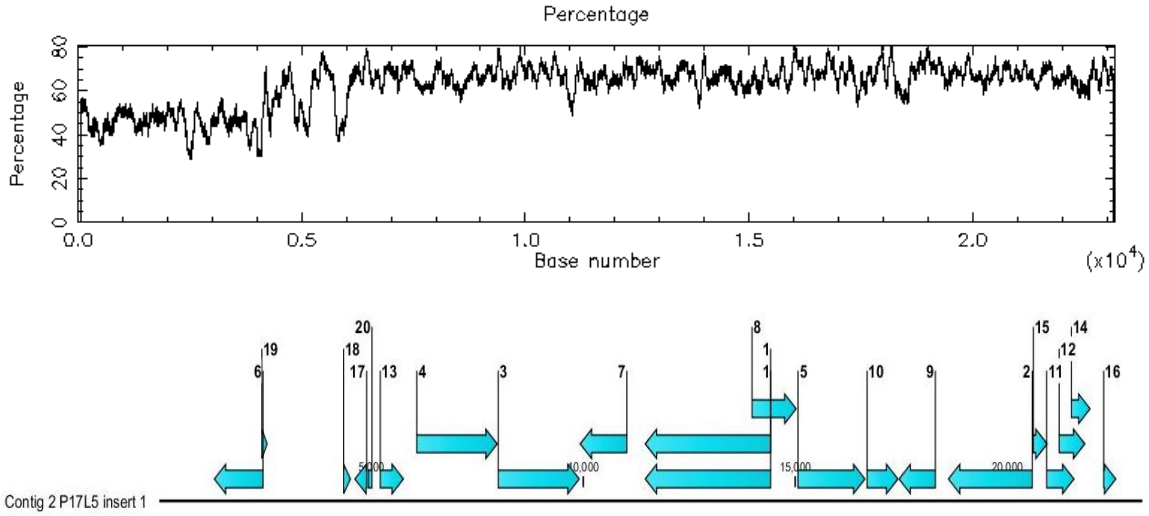
ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	249	No significant hit			
2	288	predicted protein	*Nematostella vectensis*	7.00E-15	30
3	2649	predicted protein	*Naegleria gruberi*	1.00E-21	29
4	468	No significant hit			
5	972	No significant hit			
6	255	No significant hit			
7	573	Ankyrin	*Thiocapsa marina 5811*	5.00E-18	34
8	1536	No significant hit			30
9	1242	No significant hit			

G. P15G24 CONTIG 18



ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	81	no significant hit			
2	1044	Bacillolysin	*Niabella soli DSM 19437*	7.00E-73	40
3	723	Ankyrin	*Sulfolobus islandicus Y.N.15.51*	2.00E-25	37
4	168	hypothetical protein CPAR2_403190	*Candida parapsilosis*	2.00E-06	40
5	474	26S proteasome non-ATPase regulatory subunit, putative	*Phytophthora infestans T30-4*	5.00E-06	39
6	1689	no significant hit			
7	1011	no significant hit			

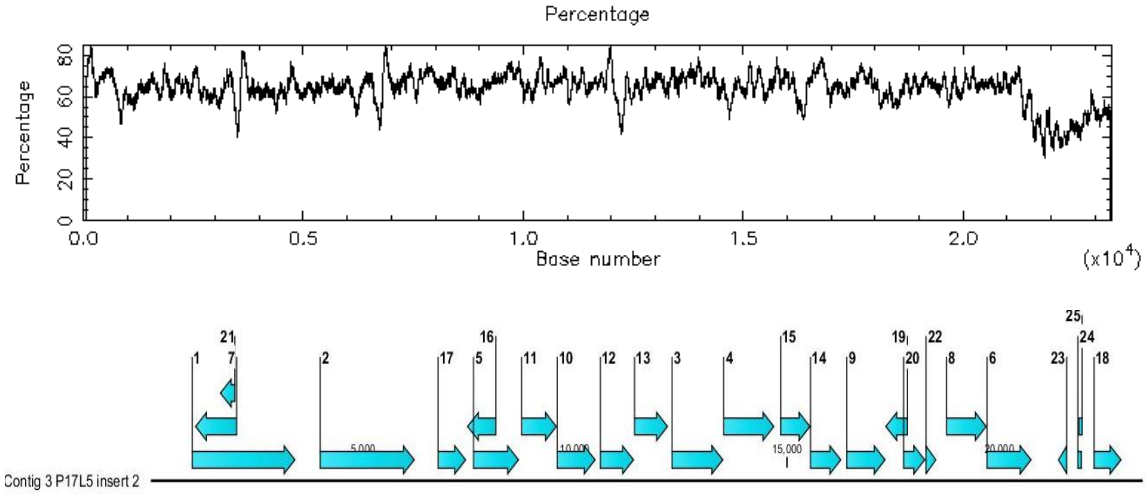
H. P17L5 CONTIG 2



ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	2982	spermine synthase	*[Solibacter usitatus Ellin6076]*	0	47
2	2004	hypothetical protein Acid_6976	*[Solibacter usitatus Ellin6076]*	3.00E-124	43
3	1938	putative metal-dependent phosphohydrolase	*[uncultured Acidobacteria bacterium cosmid p2H8]*	3.00E-78	36
4	1926	peptidase S8 and S53 subtilisin kexin sedolisin	*[Acidobacterium sp. MP5ACTX9]*	4.00E-49	36
5	1605	tetratricopeptide TPR_4	*[Methylobacterium nodulans ORS 2060]*	1.00E-129	49
6	1176	protoporphyrinogen oxidase	*[Vibrio cholerae V52]*	0	99
7	1128	hypothetical protein PM8797T_08574	*[Planctomyces maris DSM 8797]*	6.00E-40	36
8	1065	N-methyltryptophan oxidase	*[Chloroflexus aurantiacus J-10-fl]*	6.00E-106	57
9	885	2-hydroxy-3-oxopropionate reductase	*[Thermobaculum terrenum ATCC BAA-798]*	1.00E-72	50

10	753	acid phosphatase, HAD superfamily protein	*[Rickettsiella grylli]*	3.00E-29	35
11	669	serine/threonine protein kinase	*[Solibacter usitatus Ellin6076]*	9.00E-14	47
12	630	hypothetical protein PRABACTJOHN_04411	*[Parabacteroides johnsonii DSM 18315]*	3.00E-06	31
13	573	ABC transporter-related protein	*[Geobacter metallireducens GS-15]*	2.00E-53	58
14	468	serine/threonine protein kinase	*[Haliangium ochraceum DSM 14365]*	7.00E-26	47
15	354	serine/threonine protein kinase	*[Solibacter usitatus Ellin6076]*	1.00E-18	77
16	312	serine/threonine protein kinase	*[Haliangium ochraceum DSM 14365]*	4.00E-08	41
17	306	hypothetical protein pU302L_094	*[Salmonella enterica subsp. enterica serovar Typhimurium]*	8.00E-10	100
18	189	conserved hypothetical protein	*[Escherichia coli MS 196-1]*	4.00E-27	100
19	153	hypothetical protein EcE24377A_E0023	*[Escherichia coli E24377A]*	4.00E-20	98
20	111	putative reverse transcriptase	*[Platanus x acerifolia]*	7.00E-05	83

I. P17L5 CONTIG 3

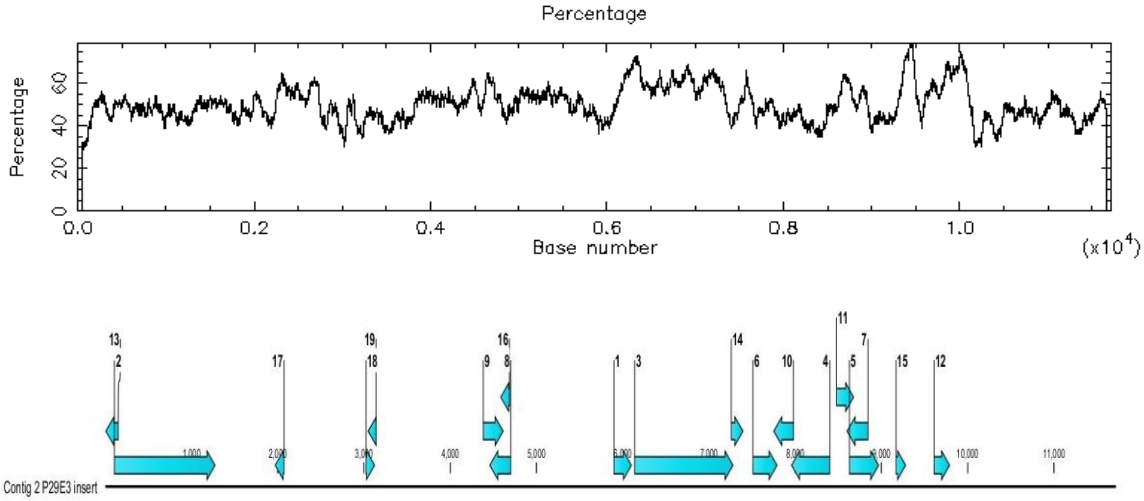


ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	2445	ATPase	*[Solibacter usitatus Ellin6076]*	0	69
2	2241	surface antigen (D15)	*[Candidatus Koribacter versatilis Ellin345]*	1.00E-103	32
3	1221	7,8-didemethyl-8-hydroxy-5-deazariboflavin synthase, CofH subunit	*[Thermincola sp. JR]*	3.00E-114	56
4	1209	probable chlorohydrolase	*[Blastopirellula marina DSM 3645]*	2.00E-55	39
5	1086	acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamine O-acyltransferase	*[Solibacter usitatus Ellin6076]*	3.00E-68	49
6	1059	deoxyguanosinetriphosphate triphosphohydrolase	*[Desulfurivibrio alkaliphilus AHT2]*	7.00E-105	56
7	993	hypothetical protein 1100011001330_R2601_13514	*[Pelagibaca bermudensis HTCC2601]*	6.00E-37	55

8	960	phosphopentomutase	*[Thermosediminibacter oceani DSM 16646]*	2.00E-105	57
9	927	TonB-like protein	*[Candidatus Koribacter versatilis Ellin345]*	5.00E-07	29
10	918	Oxidoreductase-like [	*delta proteobacterium MLMS-1]*	1.00E-71	52
11	843	protein of unknown function DUF1009	*[Acidobacterium sp. MP5ACTX8]*	3e-71,	53
12	807	Radical SAM domain protein	*[Acetohalobium arabaticum DSM 5501]*	6.00E-12	28
13	804	hypothetical protein DSM3645_04470 [	*Blastopirellula marina DSM 3645]*	3.00E-47	44
14	738	ubiquinone/menaquinone biosynthesis methyltransferase	*[Rhodothermus marinus DSM 4252]*	2.00E-52	49
15	714	prenyltransferase	*[Geobacter sulfurreducens PCA]*	2.00E-51	50
16	693	hypothetical protein STIAU_5450	*[Stigmatella aurantiaca DW4/3-1]*	3.00E-06	31
17	675	Outer membrane chaperone Skp (OmpH) [	*Geobacter metallireducens GS-15]*	3.00E-13	30
18	660	hypothetical protein CLOSCI_03331	*[Clostridium scindens ATCC 35704]*	3.00E-130	100
19	534	signal transduction histidine kinase	*[Rothia mucilaginosa DY-18]*	2.00E-08	33
20	522	tRNA isopentenyltransferase	*[delta proteobacterium MLMS-1]*	1.00E-25	42
21	366	hypothetical protein CaO19.13746	*[Candida albicans SC5314]*	4.00E-09	41
22	258	RNA-binding protein Hfq	*[Solibacter usitatus Ellin6076]*	5.00E-13	46

23	222	LacOPZ-alpha peptide from pUC9; putative	*[unidentified cloning vector]*	3.00E-20	90
24	108	conserved domain protein	*[Escherichia coli MS 84-1]*	1e-11,	98
25	102	conserved hypothetical protein	*[Enterococcus faecalis AR01/DG]*	3.00E-10	100

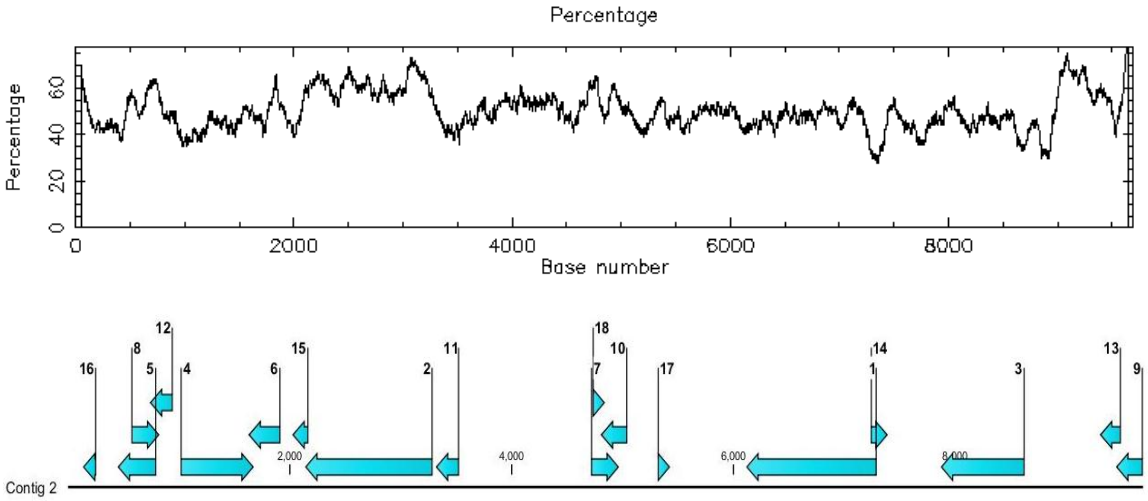
J. P29E3



ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	210	hypothetical protein HMPREF9552_04933	*Escherichia coli MS 198-1*	3.00E-14	97
2	1176	protoporphyrinogen oxidase	*Vibrio cholerae V52*	0	100
3	1149	transcriptional repressor protein	*uncultured bacterium*	0	99
4	453	hypothetical protein CLOSCI_03331	*Clostridium scindens ATCC 35704*	1.00E-63	99
5	348	site-specific recombinase, phage integrase family	*Escherichia coli MS 119-7*	1.00E-61	100
6	288	conserved hypothetical protein	*Streptomyces ghanaensis ATCC 14672*	1.00E-10	100
7	252	orf681	*Escherichia coli*	2.00E-37	98
8	252	ybl209	*Escherichia coli BL21(DE3)*	2.00E-34	98
9	240	hypothetical protein ECH7EC4501_6204	*Escherichia coli O157:H7 str. EC4501*	3.00E-11	100
10	237	hypothetical protein CLOSCI_03331	*Clostridium scindens ATCC 35704*	7.00E-41	100
11	207	restriction endonuclease	*Photobacterium damselae subsp. piscicida*	6.00E-07	76

12	189	conserved hypothetical protein	*Escherichia coli MS 196-1*	4e-27,	100
13	153	hypothetical protein EcE24377A_E0023	*Escherichia coli E24377A*	2.00E-20	100
14	144	GCN5-related N-acetyltransferase	*Birmingham IncP-alpha plasmid*	7.00E-06	100
15	120	hypothetical protein ECH7EC4196_4052	*Escherichia coli O157:H7 str. EC4196*	2.00E-14	100
16	111	hypothetical protein SeSPA_A3240	*Salmonella enterica subsp. enterica serovar Saintpaul str. SARA23*	2.00E-11	97
17	111	putative reverse transcriptase	*Platanus x acerifolia*	7.00E-05	82
18	108	conserved domain protein	*Escherichia coli MS 84-1*	1.00E-11	97
19	102	conserved hypothetical protein	*Enterococcus faecalis AR01/DG*	3.00E-10	100

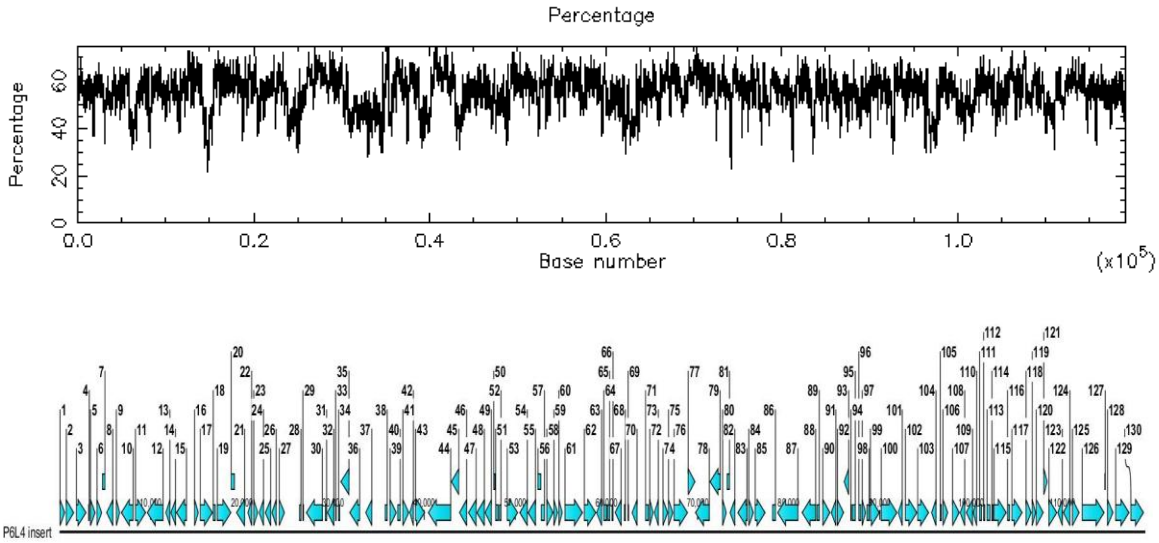
K. P30A5



ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	1176	protoporphyrinogen oxidase [Vibrio cholerae V52]	*[Vibrio cholerae V52]*	0	100
2	1149	transcriptional repressor protein	*[uncultured bacterium]*	0	99
3	756	replication protein	*Plasmid F*	1.00E-147	100
4	660	hypothetical protein CLOSCI_03331	*[Clostridium scindens ATCC 35704]*	3.00E-130	100
5	348	site-specific recombinase, phage integrase family x	*[Escherichia coli MS 119-7]*	1.00E-61	100
6	288	conserved hypothetical protein	*[Streptomyces ghanaensis ATCC 14672]*	1.00E-10	100
7	252	ybl209	*[Escherichia coli BL21(DE3)]*	2.00E-34	98
8	252	orf681	*[Escherichia coli]*	2e-37,	98
9	243	hypothetical protein EcolH2_00650	*[Escherichia coli H299]*	3.00E-38	100
10	240	hypothetical protein ECH7EC4501_6204	*[Escherichia coli O157:H7 str. EC4501]*	3.00E-11	100
11	210	hypothetical protein HMPREF9552_04933	*[Escherichia coli MS 198-1]*	3.00E-14	97

12	207	restriction endonuclease	*[Photobacterium damselae subsp. piscicida]*	6.00E-07	76
13	189	conserved hypothetical protein	*[Escherichia coli MS 196-1]*	4.00E-27	100
14	153	hypothetical protein EcE24377A_E0023	*[Escherichia coli E24377A]*	2.00E-20	100
15	144	GCN5-related N-acetyltransferase	*[Birmingham IncP-alpha plasmid]*	7.00E-06	100
16	120	hypothetical protein ECH7EC4196_4052	*[Escherichia coli O157:H7 str. EC4196]*	2.00E-14	100
17	111	putative reverse transcriptase	*[Platanus x acerifolia]*	7.00E-05	82
18	111	hypothetical protein SeSPA_A3240	*[Salmonella enterica subsp. enterica serovar Saintpaul str. SARA23]*	2.00E-11	97

L. P6L4



ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	585	regulatory protein ArsR	*Micromonospora aurantiaca ATCC 27029*	5.00E -42	49
2	897	activator of Hsp90 ATPase 1 family protein	*Micromonospora aurantiaca ATCC 27029*	8.00E -26	29
3	1119	hypothetical protein RHA1_ro00504	*Rhodococcus jostii RHA1*	3.00E -16	45
4	150	transcriptional regulator, AraC family	*Ktedonobacter racemifer DSM 44963*	5.00E -04	40
5	585	transcriptional regulator, AraC family	*Ktedonobacter racemifer DSM 44963*	2.00E -43	48
6	606	bifunctional deaminase-reductase domain protein	*Ktedonobacter racemifer DSM 44963*	1.00E -63	60
7	357	conserved hypothetical protein	*Paenibacillus sp. oral taxon 786 str. D14*	3.00E -31	60
8	762	dienelactone hydrolase	*Methanoculleus marisnigri JR1*	4.00E -71	57
9	447	hypothetical protein Acid345_4585	*Candidatus Koribacter versatilis Ellin345*	5.00E -49	72

10	1362	beta-galactosidase	*Thermobaculum terrenum ATCC BAA-798*	1.00E-155	58
11	1194	putative S1B family peptidase	*Anaerolinea thermophila UNI-1*	1.00E-69	43
12	1818	PAS/PAC sensor signal transduction histidine kinase	*Chthoniobacter flavus Ellin428*	7.00E-48	36
13	540	acetyltransferase	*Bacillus coahuilensis m4-4*	7.00E-43	48
14	606	No significant hit			
15	1185	putative esterase	*Candidatus Solibacter usitatus Ellin607*	9.00E-22	28
16	489	hypothetical protein Xaut_2234	*Xanthobacter autotrophicus Py2*	1.00E-09	34
17	1407	leucine aminopeptidase-related protein	*Erythrobacter sp. NAP1*	4.00E-43	32
18	318	No significant hit			
19	1566	hypothetical protein sce6585	*Sorangium cellulosum 'So ce 56'*	3.00E-13	36
20	435	probable N-acetylglutamate synthase	*Planctomyces maris DSM 8797*	5.00E-31	54
21	981	hypothetical protein sce6608	*Sorangium cellulosum 'So ce 56'*	2.00E-45	39
22	498	DinB family protein	*Herpetosiphon aurantiacus ATCC 23779*	1.00E-47	57
23	516	hypothetical protein Haur_4171	*Herpetosiphon aurantiacus ATCC 23779*	5.00E-38	45
24	561	bifunctional deaminase-reductase domain protein	*Ktedonobacter racemifer DSM 44963*	2.00E-54	57
25	681	AraC family transcriptional regulator	*Rubrobacter xylanophilus DSM 9941*	1.00E-59	52
26	555	hypothetical protein RoseRS_3033	*Roseiflexus sp. RS-1*	9.00E-53	57
27	660	CmR			
28	237	gp29			
29	159	no significant hit			

30	1842	p68			
31	222	hypothetical protein			
32	780	apramycin acetyl transferase			
33	144	int			
34	225	hypothetical protein EfaeDRAFT_1157			
35	972	plasmid-partitioning protein			
36	1167	protoporphyrinogen oxidase			
37	756	replication protein			
38	294	resolvase			
39	759	beta-lactamase	*Ktedonobacter racemifer DSM 44963*	9.00E-69	52
40	354	no significant hit			
41	732	hypothetical protein Xcel_2577	*Xylanimonas cellulosilytica DSM 15894*	1.00E-21	36
42	468	no significant hit			
43	1062	no significant hit			
44	2538	hypothetical protein OSCT_2889	*Oscillochloris trichoides DG6*	6.00E-173	43
45	885	no significant hit			
46	873	no significant hit			
47	921	hypothetical protein SS1G_07480	*Sclerotinia sclerotiorum 1980*	9.00E-05	27
48	906	no significant hit			
49	762	no significant hit			
50	219	no significant hit			
51	348	PemK-like protein	*Microcoleus chthonoplastes PCC 7420*	1.00E-45	77
52	246	hypothetical protein alr7074	*Nostoc sp. PCC 7120*	3.00E-09	44
53	1224	hypothetical protein BBR47_37950	*Brevibacillus brevis NBRC 100599*	2.00E-37	29
54	894	no significant hit			
55	906	no significant hit			

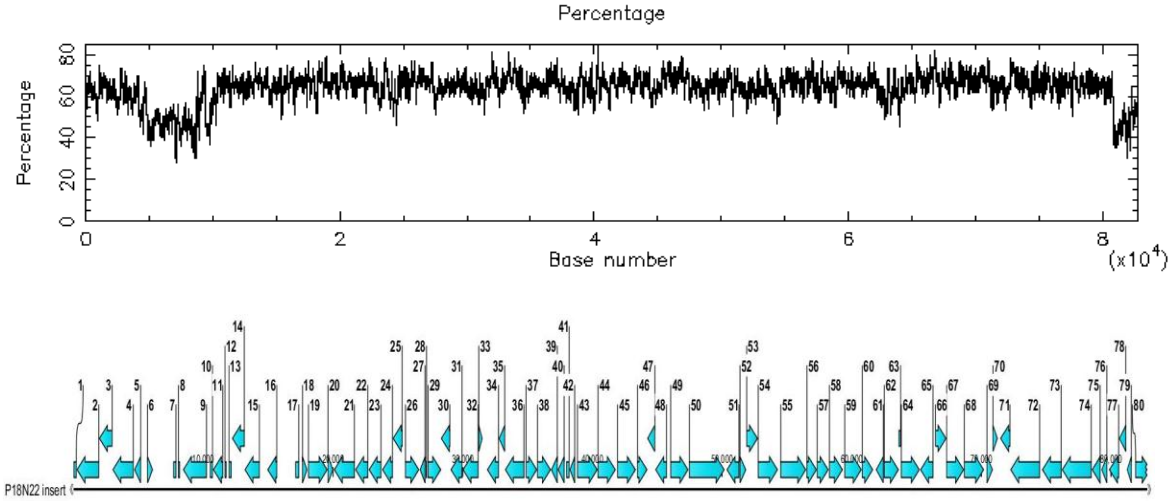
56	402	two component transcriptional regulator, winged helix family	*Dethiosulfovibrio peptidovorans DSM 11002*	6.00E-06	32
57	420	no significant hit			
58	759	Short-chain dehydrogenase/reductase SDR	*NC10 bacterium 'Dutch sediment'*	8.00E-43	43
59	492	putative integron gene cassette protein	*uncultured bacterium]*	3.00E-07	30
60	453	hypothetical protein Kfla_3931	*Kribbella flavida DSM 17836*	9.00E-26	45
61	1998	glycosyl transferase family protein	*Candidatus Methanoregula boonei 6A8*	8.00E-11	27
62	1335	major facilitator superfamily MFS_1	*Anaeromyxobacter dehalogenans 2CP-1*	3.00E-53	36
63	639	carbonic anhydrase	*Aeromonas hydrophila subsp. hydrophila ATCC 7966*	4.00E-84	69
64	390	hypothetical protein Avi_0533	*Agrobacterium vitis S4*	2.00E-45	68
65	300	no significant hit			
66	195	no significant hit			
67	735	no significant hit			
68	198	no significant hit			
69	150	no significant hit			
70	627	hypothetical protein sce1838	*Sorangium cellulosum 'So ce 56*	4.00E-13	29
71	405	no significant hit			
72	444	hydrolases of the alpha/beta superfamily	*Microscilla marina ATCC 23134*	7.00E-23	46
73	519	transposase IS4 family protein	*Herpetosiphon aurantiacus ATCC 23779*	9.00E-20	38
74	603	no significant hit			
75	549	2',5' RNA ligase	*Geobacter metallireducens GS-15*	1.00E-34	45
76	1551	no significant hit			

77	822	metallophosphoesterase	*Methanobacterium sp. AL-21*	2.00E-45	42
78	1605	serine/threonine protein kinase	*Herpetosiphon aurantiacus ATCC 23779*	7.00E-52	41
79	1203	response regulator receiver protein	*Anaerolinea thermophila UNI-1*	3.00E-06	22
80	483	hypothetical protein SeryN2_34165	*Saccharopolyspora erythraea NRRL 2338*	4.00E-21	43
81	342	no significant hit			
82	600	ECF subfamily RNA polymerase sigma-24 factor	*Herpetosiphon aurantiacus ATCC 23779*	2.00E-38	48
83	1107	putative outer membrane adhesin like protein	*Shewanella sp. MR-7*	3.00E-04	26
84	579	DNA-3-methyladenine glycosylase I	*Geobacter lovleyi SZ*	5.00E-69	67
85	1173	hypothetical protein SrosN15_03733	*Streptomyces roseosporus NRRL 15998*	5.00E-07	34
86	402	possible bacteriophage envelope protein	*Sphingobacterium spiritivorum ATCC 33300*	7.00E-25	44
87	2304	Hypothetical protein CBG23651	*Caenorhabditis briggsae*	6.00E-41	34
88	1575	M23 family metalloendopeptidase	*Leptospira interrogans serovar Lai str. 56601*	1.00E-08	32
89	342	no significant hit			
90	855	hypothetical protein Hoch_4337	*Haliangium ochraceum DSM 14365*	3.00E-36	40
91	588	conserved hypothetical protein	*Microscilla marina ATCC 23134*	9.00E-47	50
92	768	hypothetical protein MXAN_7068	*Myxococcus xanthus DK 1622*	3.00E-62	52
93	582	GCN5-related N-acetyltransferase	*Ktedonobacter racemifer DSM 44963*	2.00E-48	48
94	144	no significant hit			
95	405	no significant hit			

96	309	no significant hit			
97	456	no significant hit			
98	324	no significant hit			
99	1062	hypothetical protein BACI_c18230	*Bacillus anthracis CI*	2.00E-30	32
100	1803	oligoendopeptidase F	*Ktedonobacter racemifer DSM 44963*	1.00E-173	52
101	510	no significant hit			
102	1338	adenosine deaminase	*Burkholderia pseudomallei 1710b*	2.00E-04	31
103	1290	no significant hit			
104	561	deaminase-reductase domain-containing protein	*Candidatus Solibacter usitatus Ellin6076*	5.00E-61	63
105	228	no significant hit			
106	603	hypothetical protein sce1838	*Sorangium cellulosum 'So ce 56'*	2.00E-11	28
107	795	protein of unknown function DUF899 thioredoxin family protein	*Ktedonobacter racemifer DSM 44963*	5.00E-109	75
108	582	MIP family channel protein	*Ktedonobacter racemifer DSM 44963*	1.00E-64	72
109	771	transcriptional regulator, ArsR family	*Ktedonobacter racemifer DSM 44963*	4.00E-57	46
110	447	phosphotyrosine protein phosphatase	*Syntrophus aciditrophicus SB*	1.00E-43	58
111	288	no significant hit			
112	216	no significant hit			
113	390	no significant hit			
114	210	no significant hit			
115	1419	type I secretion target GGXGXDXXX repeat protein domain protein	*Synechococcus sp. PCC 7335*	4.00E-122	51
116	300	no significant hit			
117	1212	peptidase C14 caspase catalytic subunit p20	*Methylobacterium nodulans ORS 2060*	8.00E-41	44

118	654	alpha/beta hydrolase fold protein	*Rhodomicrobium vannielii ATCC 17100*	2.00E-32	39
119	444	cyclase/dehydrase	*Prosthecochloris aestuarii DSM 271*	3.00E-13	34
120	837	phospholipase/carboxylesterase family	*Aciduliprofundum boonei T469*	3.00E-47	37
121	438	hypothetical protein BCAS0686	*Burkholderia cenocepacia J2315*	8.00E-14	39
122	948	WD40-like Beta Propeller	*Bacillus cereus R309803*	2.00E-20	29
123	567	hypothetical protein Adeh_2296	*Anaeromyxobacter dehalogenans 2CP-C*	3.00E-34	43
124	783	Methyltransferase type 11	*bacterium Ellin514*	3.00E-08	31
125	822	Glycoside hydrolase family 25	*Oscillatoria sp. PCC 6506*	4.00E-32	38
126	2478	ATP-dependent Clp protease ATP-binding subunit	*Anaerolinea thermophila UNI-1*	0	71
127	135	no significant hit			
128	684	hypothetical protein RHA1_ro00504	*Rhodococcus jostii RHA1*	4.00E-19	49
129	1542	N-acetylmuramoyl-L-alanine amidase	*Mobiluncus curtisii ATCC 43063*	1.00E-15	39
130	1470	beta-lactamase	*Herpetosiphon aurantiacus ATCC 23779*	6.00E-166	61

M. P18N22



ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	240	no significant hit			
2	1731	putative type II secretion protein	*Pseudomonas aeruginosa PAb1*	2.00E-163	56
3	1032	GDP-mannose 4,6-dehydratase	*Thermomicrobium roseum DSM 5159*	2.00E-90	50
4	1617	glycosyl transferase family protein	*Opitutus terrae PB90-1*	6.00E-72	34
5	525	glycosyl transferase, group 1	*Methylobacillus flagellatus KT*	3.00E-32	45
6	438	CmR			
7	237	gp29			
8	159	no significant hits			
9	1842	p68			
10	222	hypothetical protein			
11	780	apramycin acetyl transferase			
12	144	int			
13	225	hypothetical protein EfaeDRAFT_1157			
14	972	plasmid-partitioning protein			
15	1167	plasmid-partitioning protein SopA			
16	756	replication protein			
17	294	resolvase			
18	399	no significant hits			

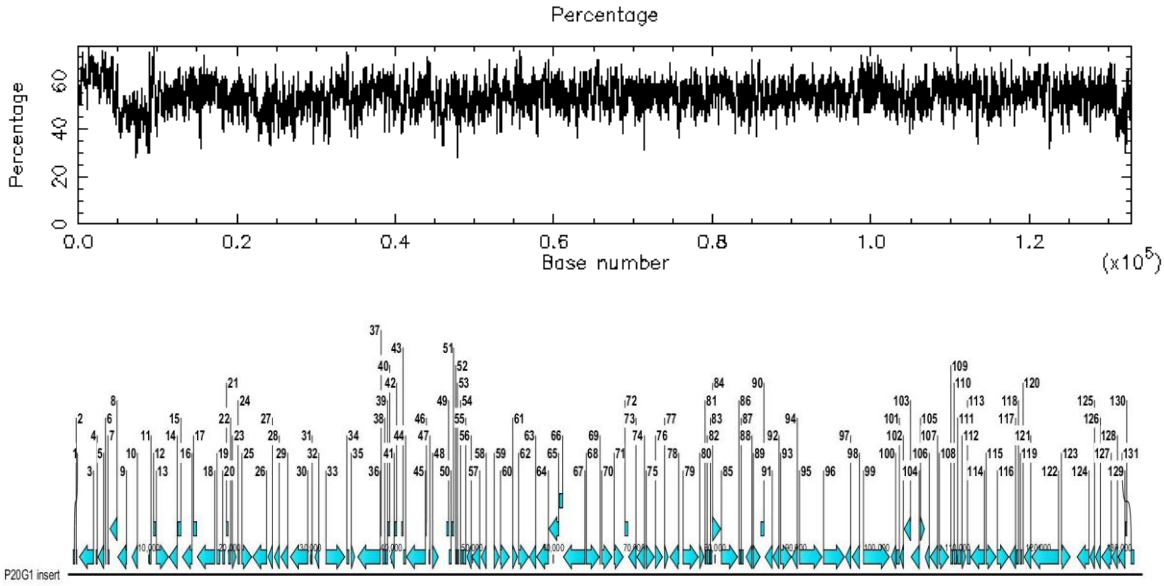
19	1497	AMP nucleosidase	*Polaromonas sp. JS666*	0	74
20	432	protein of unknown function UPF0047	*Thioalkalivibrio sp. HL-EbGR7*	9.00E-44	68
21	1608	Glucose-methanol-choline oxidoreductase	*alpha proteobacterium BAL199*	0	68
22	999	extracellular solute-binding protein family 3	*Variovorax paradoxus S110*	3.00E-120	67
23	969	putative secreted protein	*Bordetella petrii DSM 12804*	5.00E-92	55
24	816	ATP-binding component of ABC transporter	*Bordetella parapertussis*	7.00E-110	73
25	753	taurine ABC transporter, permease protein	*Bordetella petrii DSM 12804*	4.00E-92	81
26	1086	OmpA/MotB domain-containing protein	*Polaromonas naphthalenivorans CJ2*	2.00E-35	60
27	438	no significant hits			
28	93	no significant hits			
29	1041	hypothetical protein Bpro_3480	*Polaromonas sp. JS666*	7.00E-81	60
30	717	transmembrane protein	*Sideroxydans lithotrophicus ES-1*	2.00E-15	44
31	909	hypothetical protein ebA3896	*Aromatoleum aromaticum EbN1*	2.00E-55	56
32	1221	rtcB protein	*Azoarcus sp. BH72*	3.00E-159	70
33	318	no significant hits			
34	903	Transporter, drug/metabolite exporter family	*Ralstonia solanacearum UW551*	4.00E-68	51
35	489	methylated-DNA/protein-cysteine methyltransferase	*Desulfovibrio fructosovorans JJ*	9.00E-36	54
36	1488	transcriptional regulator, AraC family	*Desulfovibrio sp. FW1012B*	1.00E-153	59
37	864	hypothetical protein Daci_5147	*Delftia acidovorans SPH-1*	4.00E-59	48
38	1053	selenophosphate synthase	*Cupriavidus metallidurans CH34*	2.00E-128	66

39	513	hypothetical protein NE2209	*[Nitrosomonas europaea ATCC 19718*	3.00E-25	53
40	516	preprotein translocase subunit SecB	*Laribacter hongkongensis HLHK9*	9.00E-41	62
41	258	glutaredoxin	*Azoarcus sp. BH72*	4.00E-29	75
42	423	rhodanese-like protein	*Thiobacillus denitrificans ATCC 25259*	2.00E-20	40
43	1575	phosphoglycerate mutase, 2,3-bisphosphoglycerate-independent	*Sideroxydans lithotrophicus ES-1*	6.00E-178	61
44	1392	Peptidase M23	*Methylotenera mobilis JLW8*	1.00E-46	36
45	1428	carboxyl-terminal protease	*Sideroxydans lithotrophicus ES-1*	9.00E-131	58
46	795	adenylyltransferase	*Variovorax paradoxus S110*	3.00E-80	61
47	603	TetR family transcriptional regulator	*Sideroxydans lithotrophicus ES-1*	6.00E-60	69
48	885	acetylglutamate kinase	*Sideroxydans lithotrophicus ES-1*	5.00E-102	69
49	1389	hypothetical protein BB1357	*Bordetella bronchiseptica RB50*	0.00E+00	72
50	2733	hypothetical protein Mpe_A2083	*Methylibium petroleiphilum PM1*	0.00E+00	45
51	978	PhoH family protein	*Candidatus Accumulibacter phosphatis clade IIA str. UW-1*	9.00E-110	69
52	501	hypothetical protein ebA1336	*Aromatoleum aromaticum EbN1*	6.00E-37	55
53	888	hypothetical protein Tbd_2703		9.00E-102	70
54	1542	apolipoprotein N-acyltransferase	*Thiobacillus denitrificans ATCC*	7.00E-106	50
55	1986	AMP-dependent synthetase and ligase	*Dechloromonas aromatica RCB*	0.00E+00	71
56	792	ABC transporter related	*Dechloromonas aromatica RCB*	2.00E-107	79
57	894	ABC transporter permease	*Azoarcus sp. BH72*	1.00E-109	73

58	1059	putative branched-chain amino acid transport permease	*Azoarcus sp. BH72*	2.00E-129	72
59	1326	branched chain amino acid ABC transporter periplasmic protein	*Azoarcus sp. BH72*	7.00E-136	59
60	864	ABC transporter ATP-binding protein	*Azoarcus sp. BH73*	5.00E-116	79
61	603	Glyoxalase/bleomycin resistance protein/dioxygenase	*Burkholderia sp. H160*	2.00E-57	74
62	1143	phenylacetate--CoA ligase	*Azoarcus sp. BH72*	5.00E-127	68
63	180	no significant hits			
64	1440	hypothetical protein ebA4929	*Aromatoleum aromaticum EbN1*	6.00E-33	50
65	1017	hypothetical protein Mfla_2478	*Methylobacillus flagellatus KT*	1.00E-62	52
66	864	hypothetical protein Bpro_4887	*Polaromonas sp. JS666*	3.00E-61	52
67	1362	chromate transporter chromate ion transporter (CHR) family	*Variovorax paradoxus S110*	1.00E-146	74
68	1536	hypothetical protein NE1839	*Nitrosomonas europaea ATCC 19718*	0.00E+00	77
69	471	sugar oxidoreductase	*Sorangium cellulosum 'So ce 56'*	4.00E-41	55
70	390	hypothetical protein H16_A3407	*Ralstonia eutropha H16*	6.00E-26	48
71	780	two component transcriptional regulator, LytTR family	*Pseudoxanthomonas suwonensis 11-1*	5.00E-57	49
72	2286	signal transduction histidine kinase, LytS	*Pseudoxanthomonas suwonensis 11-1*	7.00E-51	50
73	1512	D-alanyl-D-alanine carboxypeptidase/D-alanyl-D-alanine-endopeptidase	*Candidatus Accumulibacter phosphatis clade IIA str. UW-1*	3.00E-113	52
74	2283	patatin-like phospholipase	*Methylibium petroleiphilum PM1*	2.00E-152	43

75	660	hydrogenase cytochrome b-type subunit	*Azoarcus sp. BH72*	3.00E-42	45
76	447	cytochrome c, class II	*Acidovorax sp. JS42*	7.00E-32	56
77	774	transmembrane anti-sigma factor	*Variovorax paradoxus EPS*	1.00E-48	48
78	555	sigma-24 (FecI-like)	*Rhodoferax ferrireducens T118*	2.00E-40	51
79	366	hypothetical protein Daci_3003	*Delftia acidovorans SPH-1*	1.00E-34	69
80	1086	no significant hit			

N. P20G1



ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	237	gp29			
2	159	no significant hits			
3	1842	p68			
4	222	hypothetical protein			
5	780	apramycin acetyl transferase			
6	144	int			
7	225	hypothetical protein EfaeDRAFT_1157			
8	972	plasmid-partitioning protein			
9	1167	protoporphyrinogen oxidase			
10	756	replication protein			
11	294	resolvase			
12	333	ResB family protein	*Alkalilimnicola ehrlichii MLHE-1*	8.00E-05	39
13	1596	cytochrome c assembly protein	*Candidatus Koribacter versatilis Ellin345*	4.00E-34	45
14	1074	PWWP domain protein	*Aspergillus clavatus NRRL 1*	4.00E-05	33
15	456	GCN5-related N-acetyltransferase	*Meiothermus ruber DSM 1279*	1.00E-25	45

16	1287	carboxyl-terminal protease	*Candidatus Solibacter usitatus Ellin6076*	4.00E-50	37
17	480	Holliday junction resolvase YqgF	*Anaeromyxobacter sp. K*	7.00E-17	39
18	2244	penicillin amidase	*Gloeobacter violaceus PCC 7421*		
19	444	cytochrome c family protein	*Stigmatella aurantiaca DW4/3-1*	8.00E-06	35
20	318	no significant hits			
21	279	no significant hits			
22	336	no significant hits			
23	543	Micrococcal nuclease	*Thermobaculum terrenum ATCC BAA-798*	3.00E-32	51
24	258	no significant hits			
25	1272	Thiol-disulfide oxidoreductase resA	*Lysinibacillus fusiformis ZC1*	4.00E-21	35
26	1803	penicillin-binding protein	*Microscilla marina ATCC 23134*	2.00E-66	37
27	633	hypothetical protein Deipr_1715	*Deinococcus proteolyticus MRP*	7.00E-33	36
28	696	hypothetical protein ACP_1760	*Acidobacterium capsulatum ATCC 51196*	4.00E-16	34
29	939	no significant hits			
30	2232	serine/threonine protein kinase	*Candidatus Koribacter versatilis Ellin345*	2.00E-41	26
31	249	no significant hits			
32	564	no significant hits			
33	2430	hypothetical protein RoseRS_2802	*Roseiflexus sp. RS-1*	9.00E-84	35
34	327	no significant hits			
35	501	no significant hits			
36	2871	phage tail tape measure protein, TP901 family	*Gordonia bronchialis DSM 43247*	4.00E-47	41
37	117	no significant hits			

38	348	no significant hits			
39	324	phage protein, HK97 gp10 family	*Xylanimonas cellulosilytica DSM 15894*	9.00E-07	33
40	276	no significant hits			
41	549	no significant hits			
42	315	no significant hits			
43	243	no significant hits			
44	348	no significant hits			
45	2430	hypothetical protein Sthe_2835	*Sphaerobacter thermophilus DSM 20745*	4.00E-98	42
46	141	no significant hits			
47	201	no significant hits			
48	795	no significant hits			
49	312	no significant hits			
50	318	no significant hits			
51	306	no significant hits			
52	204	no significant hits			
53	231	no significant hits			
54	279	no significant hits			
55	438	no significant hits			
56	594	phage protein, HK97 gp10 family	*Xylanimonas cellulosilytica DSM 15894*	9.00E-07	33
57	1059	no significant hits			
58	705	hypothetical protein Bcav_2656	*Beutenbergia cavernae DSM 12333*	1.00E-15	27
59	705	hypothetical protein AM202_03510	*Actinobacillus minor 202*	4.00E-19	47
60	1182	phage integrase family site specific recombinase	*Azoarcus sp. BH72*	3.00E-20	32
61	660	hypothetical protein XCC3211	*Xanthomonas campestris pv. campestris str. ATCC 33913*	6.00E-41	47
62	1314	hypothetical protein	*uncultured Acidobacterium*	7.00E-28	37
63	891	hypothetical protein Ava_3538	*Anabaena variabilis ATCC 29413*	2.00E-14	28

64	1536	histidine ammonia-lyase	*Chloroflexus aggregans DSM 9485*	0.00E+00	67
65	1311	imidazolonepropionase	*Chloroflexus aurantiacus J-10-fl*	6.00E-104	48
66	489	YCII domain-containing protein	*Hyphomonas neptunium ATCC 15444*	8.00E-32	49
67	2757	hypothetical protein P700755_19977	*Psychroflexus torquis ATCC 700755*	2.00E-67	44
68	1683	urocanate hydratase	*Anaerolinea thermophila UNI-1*	0.00E+00	72
69	126	no significant hits			
70	1347	circadian clock protein, kaic	*Variovorax paradoxus EPS*	1.00E-167	68
71	1230	signal transduction histidine kinase with CheB and CheR activity	*Gemmata obscuriglobus UQM 2246*	4.00E-45	43
72	447	no significant hits			
73	1023	Tetratricopeptide repeat family	*Microcoleus chthonoplastes PCC 7420*	7.00E-12	36
74	1023	TPR domain-containing protein	*Carboxydothermus hydrogenoformans Z-2901*	1.00E-12	28
75	1113	hypothetical protein Acid_5877	*Candidatus Solibacter usitatus Ellin6076*	2.00E-121	63
76	954	arginase/agmatinase/formiminoglutamase	*Chloroflexus aurantiacus*	2.00E-61	43
77	519	phage SPO1 DNA polymerase-related protein	*Opitutus terrae PB90-1*	3.00E-55	63
78	1176	von Willebrand factor, type A	*Candidatus Koribacter versatilis Ellin345*	5.00E-28	32
79	1956	threonyl-tRNA synthetase	*Candidatus Solibacter usitatus Ellin6076*	0.00E+00	51

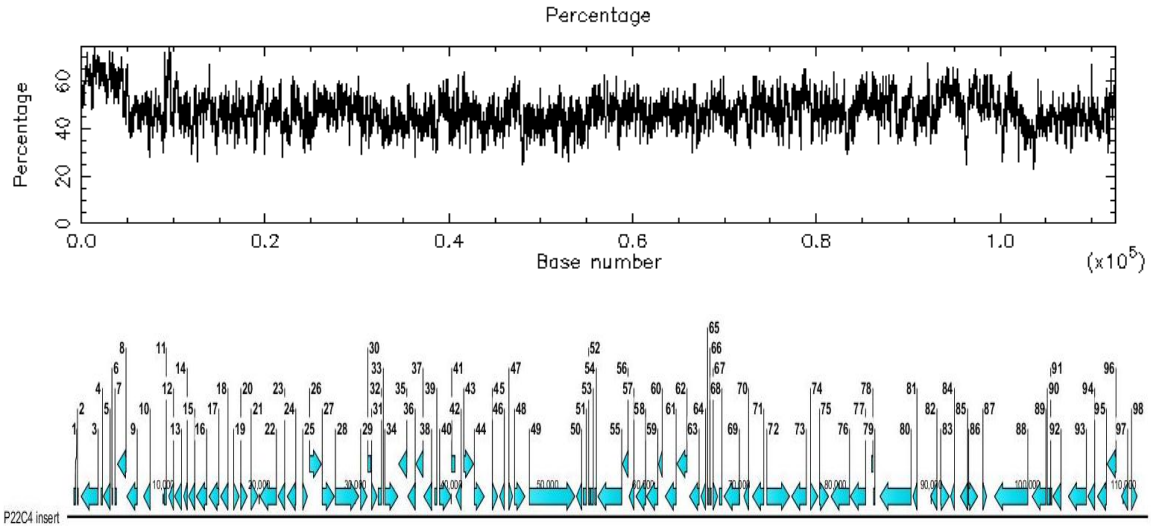
80	597	translation initiation factor 3	*Cupriavidus metallidurans CH34*	1.00E-38	56
81	204	ribosomal protein L35	*Acidobacterium sp. MP5ACTX9*	7.00E-14	64
82	381	ribosomal protein L20	*Acidobacterium capsulatum ATCC 51196*	8.00E-31	69
83	273	hypothetical protein BBta_5434	*Bradyrhizobium sp. BTAi1*	2.00E-19	56
84	1092	hypothetical protein CLOLEP_02088	*Clostridium leptum DSM 753*	5.00E-101	54
85	2079	phenylalanyl-tRNA synthetase subunit beta	*Carboxydothermus hydrogenoformans Z-2901*	1.00E-122	51
86	282	no significant hits			
87	324	hypothetical protein Acid345_0720	*Candidatus Koribacter versatilis*	1.00E-11	43
88	801	no significant hits			
89	948	hypothetical protein AciX9_1316	*Acidobacterium sp. MP5ACTX9]*	2.00E-22	28
90	474	no significant hits			
91	975	oxidoreductase domain protein	*Geobacillus sp. Y4.1MC1*	3.00E-71	43
92	849	myo-inositol catabolism protein	*Geobacillus kaustophilus HTA426*	7.00E-45	39
93	1380	probable soluble lytic transglycosylase	*Candidatus Chloracidobacterium thermophilum]*	3.00E-36	31
94	762	no significant hits			
95	2871	aconitate hydratase 1	*bacterium Ellin514]*	0.00E+00	66
96	2670	hypothetical protein kuste3266	*Candidatus Kuenenia stuttgartiensis*	4.00E-49	26
97	720	possible urease accessory protein	*Mariprofundus ferrooxydans PV-1*	3.00E-27	38

98	978	periplasmic binding protein/LacI transcriptional regulator	*bacterium Ellin514*	3.00E-92	59
99	3282	alpha-mannosidase	*Terriglobus saanensis SP1PR4*	0.00E+00	57
100	645	peptidyl-prolyl cis-trans isomerase, cyclophilin-type	*Oceanicola granulosus HTCC2516*	1.00E-23	41
101	450	hypothetical protein CLOHIR_02006 [	*Clostridium hiranonis DSM 13275*	9.00E-44	63
102	495	peptidylprolyl cis-trans isomerase, cyclophilin-type	*Synechococcus sp. JA-2-3B'a(2-13)*	4.00E-32	54
103	861	Cof protein	*[Candidatus Koribacter versatilis Ellin345]*	7.00E-34	32
104	1098	aminodeoxychorismate lyase	*Acidobacterium sp. MP5ACTX9*	1.00E-46	39
105	573	2'-5' RNA ligase	*Sphaerobacter thermophilus DSM 20745*	7.00E-26	35
106	600	peroxiredoxin	*Planctomyces brasiliensis DSM 5305*	1.00E-50	57
107	1017	band 7 protein	*Halothermothrix orenii H 168*	7.00E-71	49
108	1233	threonine dehydratase	*Meiothermus ruber DSM 1279*	4.00E-130	64
109	315	hypothetical protein DSM3645_05894	*Blastopirellula marina*	1.00E-20	66
110	375	lipoprotein	*Synechococcus sp. JA-3-3Ab*	2.00E-11	38
111	477	no significant hits			
112	546	no significant hits			
113	426	no significant hits			
114	1767	gamma-glutamyltransferase	*Candidatus Solibacter usitatus Ellin6076*	4.00E-127	50
115	1305	natural resistance-associated macrophage protein	*Thermobaculum terrenum ATCC BAA-798*	2.00E-121	56

116	1476	mechanosensitive ion channel/cyclic nucleotide-binding domain-containing protein	*[Myxococcus xanthus DK 1622]*	2.00E-49	29
117	849	no significant hits			
118	336	no significant hits			
119	261	FUR family transcriptional regulator	*Aquifex aeolicus VF5*	4.00E-13	50
120	132	no significant hits			
121	888	hypothetical protein CHY_2378	Carboxydothermus hydrogenoformans Z-2901	4.00E-34	36
122	3408	AAA ATPase	Acetohalobium arabaticum DSM 5501]	5.00E-13	20
123	1182	peptidase M48, Ste24p	*Candidatus Koribacter versatilis Ellin345*	4.00E-72	50
124	1542	outer membrane assembly lipoprotein YfiO	*Terriglobus saanensis SP1PR4*	8.00E-32	32
125	663	hypothetical protein CLOHIR_01101	*Clostridium hiranonis DSM 13275*	1.00E-54	56
126	669	serine/threonine-protein kinase PrkC	*Mitsuokella multacida DSM 20544*	3.00E-07	35
127	1287	16S rRNA (5-methyl-C967)-methyltransferase	*Geobacter bemidjiensis Bem*	2.00E-75	40
128	762	rhodanese sulfurtransferase	*Francisella philomiragia subsp. philomiragia ATCC 25017*	1.00E-57	46
129	930	methionyl-tRNA formyltransferase	*Thermoanaerobacter tengcongensis MB4*	4.00E-79	50
130	225	Alkaline phosphatase	*Azotobacter vinelandii DJ]*	2.00E-15	63

131	438	CmR			

O. P22C4



ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	237	gp29			
2	159	No significant hit			
3	1842	p68			
4	222	hypothetical protein			
5	780	apramycin acetyl transferase			
6	144	int			
7	225	hypothetical protein EfaeDRAFT_1157			
8	972	plasmid-partitioning protein			
9	1167	protoporphyrinogen oxidase			
10	756	replication protein			
11	294	resolvase			
12	570	aldo/keto reductase	aldo/keto reductase	7.00E -39	54
13	795	Uncharacterized oxidoreductase yvaG	*Oscillatoria sp.*	3.00E -100	70
14	450	hypothetical protein PAU_02593	*Photorhabdus asymbiotica subsp*	8.00E -27	46
15	726	glutamine amidotransferase, class-I	*Listeria ivanovii*	2.00E -32	36
16	1152	TPR repeat-containing protein	*Methanospirillum hungatei*	1.00E -04	22

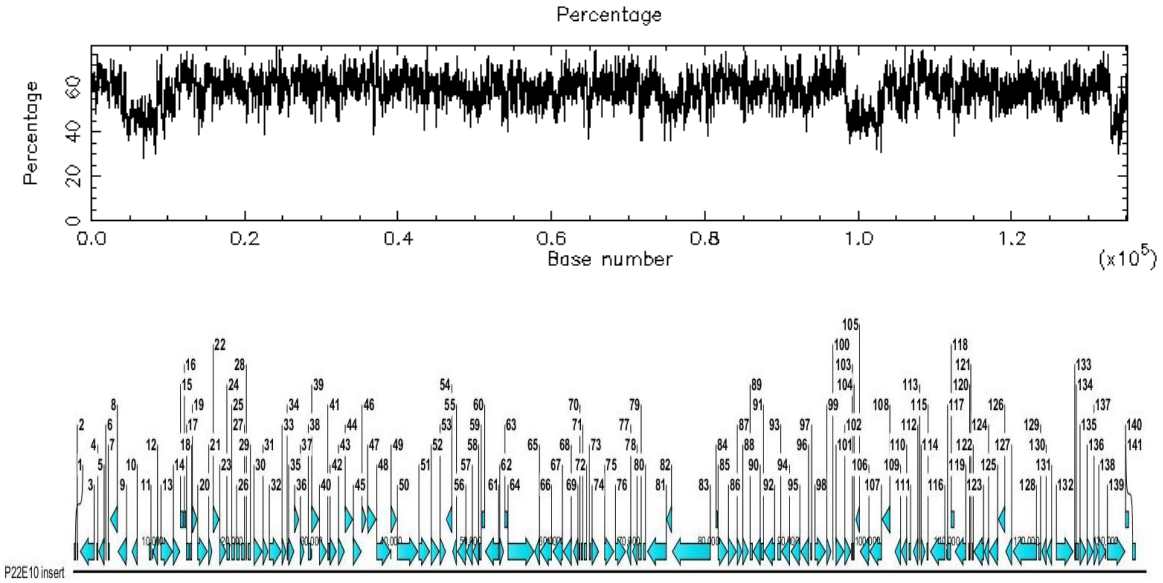
17	1104	putative hydrolase	*Myxococcus xanthus DK 1622*	2.00E-29	31
18	846	hypothetical protein ACP_2865	*Acidobacterium capsulatum*	1.00E-06	29
19	651	RNA polymerase sigma factor, sigma-70 family	*Verrucomicrobiae bacterium DG1235*	3.00E-27	41
20	750	transmembrane anti-sigma factor	*Bacillus cellulosilyticus DSM 2522*	1.00E-04	31
21	804	TonB-like protein	*Candidatus Koribacter versatilis Ellin345]*	4.00E-16	45
22	1803	GTP-binding protein LepA	*Candidatus Chloracidobacterium thermophilum*	0.00E+00	74
23	774	VWFA-related domain protein	*Acidobacterium*	1.00E-30	35
24	972	glucokinase	*Bacillus megaterium QM B1551*	8.00E-50	42
25	558	DedA family protein	*gamma proteobacterium*	6.00E-15	27
26	1314	serine/threonine protein kinase	*Candidatus Koribacter versatilis Ellin345]*	6.00E-36	54
27	1344	TPR repeat-containing serine/threonin protein kinase	*Candidatus Koribacter versatilis Ellin345]*	3.00E-82	43
28	2508	TPR repeat-containing serine/threonin protein kinase	*Candidatus Koribacter versatilis Ellin345]*	6.00E-147	39
29	756	hypothetical protein Npun_R4688	*Nostoc punctiforme PCC 73102*	5.00E-69	51
30	393	hypothetical protein MXAN_7426	*Myxococcus xanthus DK 1622*	1.00E-21	45
31	657	hypothetical protein Npun_R4687	*Nostoc punctiforme PCC 73102*	2.00E-51	48
32	414	putative cyclase	*NC10 bacterium 'Dutch sediment']*	6.00E-19	41
33	78	No significant hit			
34	1440	GH3 auxin-responsive promoter	*Nostoc punctiforme PCC 73102*	6.00E-102	45
35	933	hypothetical protein Npun_R4694	*Nostoc punctiforme PCC 73102*	7.00E-52	40

36	861	short-chain dehydrogenase	*Vibrio coralliilyticus ATCC BAA-450*	5.00E-43	42
37	816	conserved hypothetical protein	Bacteriovorax marinus SJ	6.00E-57	47
38	897	putative Rieske iron-sulphur domain protein	*Bacteriovorax marinus*	2.00E-100	63
39	288	No significant hit			
40	1293	DNA helicase-related protein	*Clostridium kluyveri DSM 555*	6.00E-35	24
41	414	glycogen synthase (ADP-glucose)	*Halanaerobium praevalens DSM*	6.00E-11	47
42	630	hypothetical protein NIDE1855	*Candidatus Nitrospira defluvii*	4.00E-06	50
43	1095	hypothetical protein PL1_0904	*Paenibacillus larvae*	1.00E-66	41
44	1125	class V aminotransferase	*Candidatus Solibacter usitatus*	4.00E-80	41
45	636	peroxidase	*Starkeya novella*	3.00E-94	78
46	567	hypothetical protein UBAL2_79310104a	*Leptospirillum rubarum*	8.00E-08	42
47	459	conserved hypothetical protein	*Prevotella marshii DSM 16973*	2.00E-07	33
48	1161	putative carboxypeptidase G2	*Thermomicrobium roseum*	6.00E-78	47
49	4818	hypothetical protein Anae109_3679	*Anaeromyxobacter sp*	0.00E+00	40
50	621	dephospho-CoA kinase	*Candidatus Solibacter usitatus*	2.00E-42	45
51	393	hypothetical protein L8106_17009	*Lyngbya sp.*	7.00E-15	42
52	228	hypotheical protein	*Nostoc punctiforme*	2.00E-17	61
53	339	hypothetical protein alr7075	*Nostoc sp.*	5.00E-14	46
54	249	hypothetical protein L8106_14325	*Lyngbya sp. PCC 8106*	1.00E-09	41
55	2571	carbamoyl-phosphate synthase large subunit	*Candidatus Koribacter versatilis Ellin345*	0.00E+00	64
56	660	carbamoyl-phosphate synthase, large subunit	*Geobacter sp. M21*	4.00E-82	70
57	594	No significant hit			

58	1089	ABC efflux pump, in membrane subunit	*Candidatus Koribacterner*	1.00E-38	28
59	1257	glycosyl transferase group 1 family protein	*Caulobacter crescentus*	3.00E-22	29
60	465	hypothetical protein AciX8DRAFT_4751	*Acidobacterium*	1.00E-14	34
61	1209	glycosyl transferase, group 1 family	*Acidobacterium capsulatum*	5.00E-98	49
62	1134	hypothetical protein ACP_2425	*Acidobacterium capsulatum*	5.00E-80	46
63	1083	ABC efflux pump, inner membrane subunit	*Candidatus Koribacterner*	5.00E-33	32
64	549	Phosphate acetyltransferase	*bacterium Ellin514*	7.00E-61	64
65	201	No significant hit			
66	207	No significant hit			
67	546	4-diphosphocytidyl-2C-methyl-D-erythritolsynthase	*Acetohalobium arabaticum*	1.00E-29	35
68	342	hypothetical protein gll3552	*Gloeobacter violaceus*	6.00E-41	71
69	1671	thiamine pyrophosphate protein domain protein TPP-binding protein	*Thermobaculum terrenum ATCC*	0.00E+00	66
70	576	hypothetical protein Cpin_1703	*Chitinophaga pinensis*	1.00E-07	22
71	1260	hypothetical protein Acid345_4436	*Candidatus Koribacter versatilis*	2.00E-15	29
72	2436	ABC efflux pump, inner membrane subunit	*Candidatus Koribacter versatilis*	0.00E+00	45
73	1605	response regulator receiver modulated serine phosphatase	*Candidatus Koribacter versatilis*	4.00E-54	50
74	840	hypothetical protein Tter_2345	*Thermobaculum terrenum ATCC*	2.00E-42	36
75	1014	alcohol dehydrogenase GroES domain-containing protein	*Burkholderia sp.*	7.00E-120	64
76	2004	hypothetical protein Cpin_1755	*Chitinophaga pinensis*	0.00E+00	67

77	1671	thiamine pyrophosphate protein domain protein TPP-binding	*Chitinophaga pinensis*	0.00E+00	75
78	201	No significant hit			
79	213	No significant hit			
80	3318	CnaB domain-containing protein	*Candidatus Solibacter usitatus*	0.00E+00	42
81	504	IS1 transposase B	*E Coli*	3.00E-94	100
82	732	beta-lactamase domain-containing protein	*Burkholderia phymatum*	7.00E-99	70
83	966	transposase, IS30 family	*Octadecabacter antarcticus*	8.00E-90	55
84	477	secreted protein	*Streptomyces sp*	8.00E-47	70
85	837	two-component system sensor protein	*uncultured bacterium BLR5*	1.00E-59	44
86	1023	hypothetical protein Fjoh_1645	*Flavobacterium johnsoniae*	6.00E-31	38
87	462	FG-GAP repeat/HVR domain-containing protein	*Stigmatella aurantiaca*	4.00E-08	58
88	3546	DNA/RNA non-specific endonuclease		1.00E-65	53
89	1521	monooxygenase FAD-binding	*Acidobacterium sp.*	0.00E+00	77
90	345	transposase IS4 family protein	*Herpetosiphon aurantiacus ATCC*	6.00E-09	41
91	195	No significant hit			
92	918	No significant hit			
93	1947	chaperone protein HtpG	*Rhodococcus erythropolis SK121*	0	49
94	780	thioesterase	*Anabaena variabilis ATCC*	3.00E-76	57
95	1014	ornithine cyclodeaminase	*Beggiatoa sp. PS*	1.00E-99	51
96	1005	Pyridoxal-phosphate dependent enzyme superfamily	*Microcoleus chthonoplastes PCC 7420*	1.00E-109	60
97	726	hypothetical protein ALOHA_HF1019P19.15c	*uncultured marine bacterium HF10_19P19*	1.00E-08	24
98	660	CmR			

P. P22E10



ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	237	gp29			
2	159	no significant hits			
3	1842	p68			
4	222	hypothetical protein			
5	780	apramycin acetyl transferase			
6	144	int			
7	225	hypothetical protein EfaeDRAFT_1157			100
8	972	plasmid-partitioning protein			
9	1167	protoporphyrinogen oxidase			100
10	756	replication protein			
11	294	resolvase			
12	660	putative cyclooxygenase	*Roseobacter litoralis Och 149*	2.00E -31	40
13	1521	xylulokinase	*Dictyoglomus thermophilum H-6-12*	8.00E -151	57
14	909	hypothetical protein PM8797T_05950	*hypothetical protein PM8797T_05950*	2.00E -46	38
15	441	no significant hits			
16	276	no significant hits			

17	414	PilT domain-containing protein	*Syntrophobacter fumaroxidans MPOB*	8.00E-16	37
18	297	pseudouridine synthase, RluA family	*Clostridium papyrosolvens DSM*	1.00E-17	50
19	675	pseudouridine synthase, RluA family protein	*Paenibacillus vortex V453*	1.00E-53	51
20	1314	conserved hypothetical protein	*Methylosinus trichosporium OB3b*	2.00E-10	26
21	501	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase	*Pedobacter heparinus DSM 2366*	9.00E-31	47
22	819	3-methyl-2-oxobutanoate hydroxymethyltransferase	*Thermoanaerobacterium thermosaccharolyticum DSM 571*	3.00E-80	53
23	858	pantoate--beta-alanine ligase	*Clostridium thermocellum ATCC 27405*	8.00E-75	51
24	429	aspartate 1-decarboxylase	*Fibrobacter succinogenes subsp. succinogenes S85*	2.00E-35	59
25	438	no significant hits			
26	453	no significant hits			
27	447	GatB/Yqey domain protein	*Capnocytophaga sputigena Capno*	1.00E-15	39
28	150	no significant hits			
29	387	hypothetical protein Bd1865	*Bdellovibrio bacteriovorus HD100*	3.00E-04	34
30	1122	TPR repeat-containing protein	*Anaeromyxobacter dehalogenans 2CP-C*	4.00E-04	33
31	720	uracil-DNA glycosylase superfamily	*Haliangium ochraceum DSM 14365*	1.00E-73	58
32	1590	hypothetical protein sce5057	*Sorangium cellulosum 'So ce 56'*	3.00E-51	33
33	576	type I phosphodiesterase/nucleotide pyrophosphatase	*Syntrophobacter fumaroxidans MPOB*	4.00E-40	51
34	177	no significant hits			

35	756	leucyl/phenylalanyl-tRNA--protein transferase	*Halomonas elongata DSM 2581*	8.00E-64	59
36	621	alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen	*Chloroherpeton thalassium ATCC 35110*	2.00E-23	35
37	555	putative Thioredoxin	*Thiomonas sp. 3As*	2.00E-16	35
38	402	no significant hits			
39	975	no significant hits			
40	987	ATPase associated with various cellular activities AAA_3	*bacterium Ellin514*	5e--105	63
41	255	no significant hits			
42	882	conserved hypothetical protein	*Candidatus Kuenenia stuttgartiensis*	2.00E-87	54
43	837	hypothetical protein Hoch_0521	*Haliangium ochraceum DSM 14365*	2.00E-17	41
44	1035	conserved hypothetical protein	*Candidatus Kuenenia stuttgartiensis*	3.00E-54	43
45	1092	von Willebrand factor type A domain protein	*delta proteobacterium NaphS2*	1.00E-34	31
46	618	von Willebrand factor type A domain protein	*delta proteobacterium NaphS3*	3.00E-05	29
47	1155	no significant hits			
48	1800	hypothetical protein PARMER_01892	*Parabacteroides merdae ATCC 43184*	1.00E-29	26
49	765	hypothetical protein BACCOPRO_01656	*Bacteroides coprophilus DSM 18228*	2.00E-12	26
50	2646	protein-P-II uridylyltransferase	*Geobacter sulfurreducens PCA*	1.00E-119	33
51	1443	protease Do	*Syntrophobacter fumaroxidans MPOB*	8.00E-78	40
52	1095	no significant hits			
53	807	formate dehydrogenase subunit D	*Polaribacter sp. MED152*	1.00E-41	40

54	789	no significant hits			
55	570	protein of unknown function DUF330	*Fibrobacter succinogenes subsp. succinogenes S85*	7.00E-05	27
56	1161	hypothetical protein HG1285_06365	*Hydrogenivirga sp. 128-5-R1-1*	7.00E-12	24
57	798	ABC-type transport system involved in resistance to organic solvents, ATPase component	*[uncultured bacterium]*	2.00E-69	50
58	792	hypothetical protein Sfum_2450	*Syntrophobacter fumaroxidans MPOB*	3.00E-53	46
59	318	COG1366: Anti-anti-sigma regulatory factor (antagonist of anti-sigma factor)	*[Magnetospirillum magnetotacticum MS-1]*	2.00E-13	43
60	429	Serine phosphatase RsbU	*Endoriftia persephone 'Hot96_1+Hot96_2'*	7.00E-13	33
61	1758	stage II sporulation E family protein	*Planctomyces limnophilus DSM 3776*	2.00E-57	30
62	570	hypothetical protein RoseRS_1412	*Roseiflexus sp. RS-1*	3.00E-17	34
63	396	transcriptional repressor, CopY family	*bacterium Ellin514*	3.00E-28	46
64	3348	peptidase M56 BlaR1	*bacterium Ellin514*	5.00E-46	23
65	702	Phospholipase/Carboxylesterase	*Chlorobium ferrooxidans DSM 13031*	2.00E-18	29
66	1518	hypothetical protein STH2009	*Symbiobacterium thermophilum IAM 14863*	1.00E-77	44
67	1269	no significant hits			
68	1080	oxygen-independent coproporphyrinogen III oxidase	*Thermincola sp. JR*	4.00E-55	40
69	669	ferrochelatase	*Haliangium ochraceum DSM 14365*	2.00E-55	48
70	213	ferrochelatase	*Myxococcus xanthus DK 1622*	4.00E-19	68

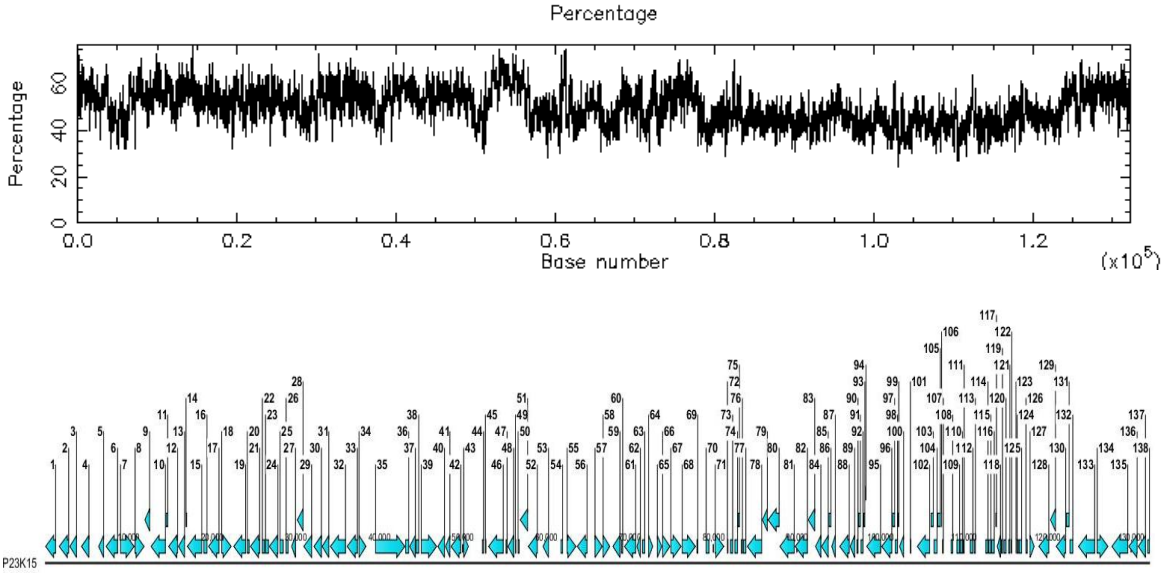
71	270	ribosomal protein S20	*Slackia exigua ATCC 700122*	4.00E-16	55
72	396	no significant hits			
73	261	glutaredoxin	*Terriglobus saanensis SP1PR4*	2.00E-06	39
74	915	hypothetical protein Plabr_3776	*Planctomyces brasiliensis DSM 5305*	8.00E-33	36
75	1248	glycoside hydrolase family protein	*Thermoanaerobacter pseudethano*	8.00E-62	34
76	1362	PBS lyase HEAT domain-containing protein repeat-containing protein	*Cyanothece sp. PCC 7822*	8.00E-06	25
77	558	no significant hits			
78	654	no significant hits			
79	426	pilin, type IV, putative	*Candidatus Koribacter versatilis Ellin345*	2.00E-13	47
80	360	no significant hits			
81	2466	hypothetical protein PM8797T_27557	*Planctomyces maris DSM 8797*	3.00E-41	30
82	693	Endo-1,4-beta-xylanase	*Spirochaeta thermophila DSM 6578*	2.00E-04	36
83	4875	secreted endo-1,4-beta-xylanase	*Microbispora corallina*	1.00E-07	24
84	270	no significant hits			
85	1275	homoserine dehydrogenase (HDH): ThrA, metL	*Thermodesulfovibrio yellowstonii DSM 11347*	2.00E-97	46
86	1101	threonine synthase	*Thermodesulfovibrio yellowstonii DSM 11347*	2.00E-114	65
87	669	PREDICTED: hypothetical protein	*Vitis vinifera*	5.00E-31	40
88	801	regulatory protein GntR HTH	*Haliangium ochraceum DSM 14365*	1.00E-21	35
89	378	no significant hits			
90	981	no significant hits			
91	405	conserved hypothetical protein	*gamma proteobacterium HTCC5015*	7.00E-12	33

92	1395	glucose-methanol-choline oxidoreductase	*Nocardioides sp. JS614*	2.00E-119	49
93	480	succinate-semialdehyde dehydrogenase	*gamma proteobacterium HTCC5015*	5.00E-36	49
94	1101	succinate-semialdehyde dehydrogenase	*gamma proteobacterium HTCC5015*	3.00E-78	48
95	1191	conserved hypothetical protein	*Microscilla marina ATCC 23134*	4.00E-115	50
96	1059	no significant hits			
97	222	no significant hits			
98	1425	PHP domain protein	*Syntrophothermus lipocalidus DSM 12680*	2.00E-15	25
99	612	hypothetical protein GM18_3082	*Geobacter sp. M18*	2.00E-14	39
100	141	50S ribosomal protein L32	*Magnetospirillum gryphiswaldense MSR-1*	6.00E-09	62
101	1062	phosphate acyltransferase	*Desulfurobacterium thermolithotrophum DSM 11699*	4.00E-79	46
102	843	malonyl CoA-acyl carrier protein transacylase	*Geobacillus sp. WCH70*	1.00E-44	44
103	189	no significant hits			
104	219	no significant hits			
105	540	no significant hits			
106	1185	no significant hits			
107	1563	no significant hits			
108	1068	hypothetical protein STAUR_5380	*Stigmatella aurantiaca DW4/3-1*	7.00E-13	30
109	789	hypothetical protein Deba_1821	*Desulfarculus baarsii DSM 2075*	1.00E-33	34
110	771	hypothetical protein Dde_1909	*Desulfovibrio desulfuricans subsp. desulfuricans str. G20*	8.00E-25	35
111	297	no significant hits			
112	651	no significant hits			
113	108	no significant hits			

114	579	putative acetyltransferase	*Sorangium cellulosum 'So ce 56'*	5.00E-23	42
115	240	HicB family protein	*Planctomyces brasiliensis DSM 5305*	3.00E-17	50
116	1902	deoxyxylulose-5-phosphate synthase	*uncultured bacterium*	0	60
117	471	cyclic nucleotide-binding protein	*Arthrospira platensis str. Paraca*	6.00E-12	31
118	453	putative transcriptional regulator, Crp/Fnr family	*Methylobacter tundripaludum SV96*	4.00E-10	35
119	1452	6-phosphogluconate dehydrogenase	*Cellvibrio japonicus Ueda107*	0.00E+00	72
120	204	Protein of unknown function UPF0150	*Crocosphaera watsonii WH 8501*	2.00E-16	58
121	132	no significant hits			
122	210	no significant hits			
123	1287	phosphofructokinase	*Pirellula staleyi DSM 6068*	9.00E-174	70
124	636	tetratricopeptide tpr_1 repeat-containing protein	*Micromonospora sp. L5*	3.00E-05	31
125	1143	erythronate-4-phosphate dehydrogenase	*Shigella flexneri 2a str. 301*	7.00E-84	47
126	894	Haloacid dehalogenase domain-containing protein hydrolase	*Isosphaera pallida ATCC 43644*	9.00E-87	57
127	831	short-chain dehydrogenase/reductase SDR	*Thermotoga lettingae TMO] >gb|ABV33390.1| short-chain dehydrogenase/reductase SDR [Thermotoga lettingae TMO]*	2.00E-72	60
128	3033	von Willebrand factor type A	*Clostridium thermocellum DSM 2360*	1.00E-102	48
129	342	nitrogen regulatory protein P-II	*Ketogulonicigenium vulgare Y25*	2.00E-29	75
130	639	Ammonium transporter	*Azotobacter vinelandii DJ*	2.00E-38	56

131	672	Ammonium transporter	*Candidatus Poribacteria sp. WGA-A3*	9.00E-73	63
132	2181	glutamine synthetase catalytic region	*Pirellula staleyi DSM 6068*	0.00E+00	77
133	195	hypothetical protein AFE_1579	*Acidithiobacillus ferrooxidans ATCC 23270*	2.00E-11	60
134	363	PilT domain-containing protein	*Geobacter uraniireducens Rf4*	9.00E-23	47
135	837	Xylose isomerase domain-containing protein TIM barrel	*Planctomyces brasiliensis DSM 5305*	1.00E-65	48
136	843	pseudouridine synthase, RluA family	*Chthoniobacter flavus Ellin428*	3.00E-55	42
137	597	RNA polymerase sigma-H factor	*Thermomicrobium roseum DSM 5159*	1.00E-26	44
138	1005	hypothetical protein DAPPUDRAFT_274329	*Daphnia pulex*	7.00E-14	56
139	2301	hypothetical protein PPSIR1_26408	*Plesiocystis pacifica SIR-1*	1.00E-52	30
140	471	hypothetical protein Plabr_4602	*Planctomyces brasiliensis DSM 5305*	5.00E-04	34
141	438	CmR	*BAC vector*		

Q. P23K15



ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	1284	DNA helicase-related protein	*Clostridium kluyveri DSM 555]*	2.00E-43	27
2	1218	TPR repeat-containing protein	*Candidatus Solibacter usitatus Ellin6076*	8.00E-32	26
3	894	diacylglycerol kinase catalytic region	*Symbiobacterium thermophilum IAM 14863*	6.00E-49	36
4	978	no significant hit			
5	690	putative methyltransferase	*Lactococcus lactis subsp. cremoris MG1363*	5.00E-19	30
6	1467	inosine-5'-monophosphate dehydrogenase	*Candidatus Solibacter usitatus Ellin6076*	3.00E-179	70
7	1719	prolyl-tRNA synthetase	*Anaeromyxobacter sp. Fw109-5*	0	61
8	1176	serine/threonine protein kinase	*Oscillochloris trichoides DG6*	2.00E-37	34
9	663	conserved hypothetical protein	*Oscillatoria sp. PCC 6506*	1.00E-43	44
10	1806	phosphoenolpyruvate -protein phosphotransferase	*Syntrophus aciditrophicus SB*	2.00E-120	41

11	291	phosphoadenosine phosphosulfate reductase	*Geobacter sulfurreducens PCA*	7.00E-15	51
12	1173	oligopeptide transport system permease protein AppC	*Persephonella marina EX-H1*	1.00E-90	53
13	819	binding-protein-dependent transport systems inner membrane component	*Halanaerobium sp. 'sapolanicus*	5.00E-51	47
14	177	no significant hit			
15	1818	extracellular solute-binding protein family 5	*bacterium Ellin514*	2.00E-108	38
16	465	UspA domain-containing protein	*Candidatus Korarchaeum cryptofilum OPF8*	9.00E-15	39
17	1341	sodium/calcium exchanger membrane region	*Micromonospora sp. L5*	1.00E-48	37
18	1215	no significant hit			
19	1500	amine oxidase	*Myxococcus xanthus DK 1622*	7.00E-57	35
20	312	no significant hit			
21	1200	hypothetical protein tlr1265	*Thermosynechococcus elongatus BP-1*	5.00E-98	52
22	363	no significant hit			
23	450	histidine kinase	*Burkholderia phymatum STM815*	2.00E-11	34
24	1119	beta-lactamase	*Teredinibacter turnerae T7901*	3.00E-121	59
25	468	transcriptional regulator, AsnC family protein	*alpha proteobacterium*	4.00E-34	45
26	336	no significant hit			
27	564	no significant hit			
28	786	N-acetylmannosaminyltransferase	*Clostridium perfringens D str. JGS1721*	6.00E-64	50
29	1026	glycosyl transferase family protein	*Geobacter metallireducens GS-15*	4.00E-100	58

30	987	NAD-dependent epimerase/dehydratase	*Carboxydibrachium pacificum DSM 12653*	2.00E-85	56
31	885	type 11 methyltransferase	*Chloroherpeton thalassium ATCC 35110*	1.00E-51	38
32	1899	asparagine synthase	*Chloroherpeton thalassium ATCC 35110*	3.00E-133	42
33	1320	hypothetical protein N47_G32660	*uncultured Desulfobacterium sp*	2.00E-143	62
34	915	HEPN domain-containing protein	*Dyadobacter fermentans DSM 1805*	1.00E-73	49
35	3600	WD repeat-containing protein	*Acaryochloris marina MBIC11017*	1.00E-115	40
36	465	secreted protein	*Xanthomonas fuscans subsp. aurantifolii str. ICPB 10535*	1.00E-12	31
37	774	hypothetical protein PPSIR1_13180	*Plesiocystis pacifica SIR-1*	1.00E-17	35
38	396	hypothetical protein Minf_0191	*Methylacidiphilum infernorum V4*	2.00E-07	41
39	1926	transcriptional regulator domain-containing protein	*Candidatus Solibacter usitatus Ellin6076*	3.00E-92	37
40	882	4-hydroxybenzoate polyprenyltransferase	*Candidatus Solibacter usitatus Ellin6076*	1.00E-84	56
41	612	3-octaprenyl-4-hydroxybenzoate carboxy-lyase	*Acidobacterium capsulatum ATCC 51196*	2.00E-37	45
42	1272	hypothetical protein VspiD_32170	*Verrucomicrobium spinosum DSM 4136]*	1.00E-53	36
43	660	CmR			
44	237	gp29			
45	159	no significant hit			
46	1842	p68			
47	222	hypothetical protein			
48	780	apramycin acetyl transferase			
49	144	int			
50	225	hypothetical protein EfaeDRAFT_1157			

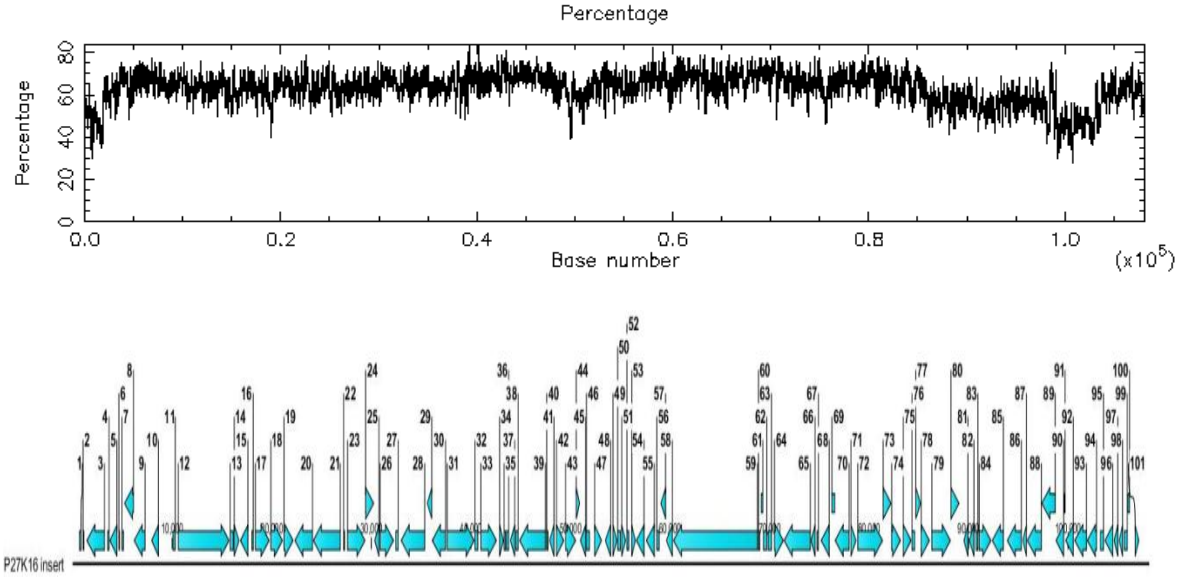
51	972	plasmid-partitioning protein			
52	1167	protoporphyrinogen oxidase			
53	756	replication protein			
54	294	resolvase			
55	1146	hypothetical protein Xcel_3215	*Xylanimonas cellulosilytica DSM 15894*	3.00E-07	29
56	1254	transposase IS111A/IS1328/IS1533	*Halothiobacillus neapolitanus c2*	9.00E-124	50
57	954	succinylglutamate desuccinylase/aspartoacylase family protein	*Subdoligranulum variabile DSM 15176*	3.00E-18	26
58	885	no significant hit			
59	933	Transposase	*Rhodobacterales bacterium HTCC2150*	1.00E-116	66
60	237	Pentapeptide repeat protein	*Microcoleus chthonoplastes PCC 7420*	7.00E-10	58
61	1425	APHP	*Methanospirillum hungatei JF-1*	2.00E-09	39
62	501	Rieske (2Fe-2S) domain-containing protein	*Candidatus Solibacter usitatus Ellin6076*	4.00E-55	65
63	366	no significant hit			
64	576	hypothetical protein NB231_04320	*Nitrococcus mobilis Nb-231*	4.00E-04	37
65	615	ECF subfamily RNA polymerase sigma-24 factor	*Candidatus Desulforudis audaxviator MP104C*	4.00E-27	38
66	963	conserved hypothetical protein	*Acidobacterium sp. MP5ACTX8*	0.34	26
67	1320	hypothetical protein Acid_7245	*Candidatus Solibacter usitatus Ellin6076*	2.00E-15	26
68	1674	hypothetical protein alr1903	*Nostoc sp. PCC 7120*	3.00E-25	28
69	198	Pentapeptide repeat protein	*Microcoleus chthonoplastes PCC 7420*	3.00E-07	54
70	432	no significant hit			

71	1119	phage integrase	*Candidatus Nitrospira defluvii*	1.00E-20	30
72	159	no significant hit			
73	351	no significant hit			
74	417	hypothetical protein AM1_5872	*Acaryochloris marina MBIC11017*	3.00E-19	50
75	231	no significant hit			
76	216	no significant hit			
77	381	hypothetical protein MettrDRAFT_1702	*Methylosinus trichosporium OB3b*	8.00E-08	30
78	1806	hypothetical protein BpV2_055 [Bathycoccus sp. RCC1105 virus BpV2]	*Bathycoccus sp. RCC1105 virus BpV2*	1.00E-05	34
79	681	no significant hit			
80	1407	hypothetical protein Nham_0315	*Nitrobacter hamburgensis X14*	4.00E-29	27
81	1809	spindle assembly 6 homolog	*Xenopus (Silurana) tropicalis*	2.00E-06	22
82	1551	no significant hit			
83	876	putative membrane-anchored cell surface protein	*Nitrobacter sp. Nb-311A*	9.00E-07	32
84	738	no significant hit			
85	879	hypothetical protein	*Trypanosoma brucei TREU927*	2.00E-16	30
86	327	no significant hit			0
87	525	no significant hit			0
88	1224	gp27	*Streptomyces phage phiSASD1*	7.00E-07	24
89	633	no significant hit			
90	222	no significant hit			
91	291	no significant hit			
92	327	no significant hit			
93	225	no significant hit			
94	144	no significant hit			
95	1755	no significant hit			
96	1296	hypothetical protein Dalk_3968	*Desulfatibacillum alkenivorans AK-01*	2.00E-90	43
97	336	no significant hit			
98	345	no significant hit			
99	168	no significant hit			

100	579	no significant hit			
101	153	no significant hit			
102	1575	no significant hit			
103	348	no significant hit			
104	459	no significant hit			
105	399	no significant hit			
106	117	no significant hit			
107	183	no significant hit			
108	294	Bacteriophage Lambda NinG	*Riemerella anatipestifer DSM 15868*	5.00E-08	35
109	408	no significant hit			
110	339	no significant hit			
111	201	no significant hit			
112	405	no significant hit			
113	228	no significant hit			
114	339	no significant hit			
115	354	no significant hit			
116	393	no significant hit			
117	174	no significant hit			
118	480	SpoU rRNA methylase family protein	*Polaribacter sp. MED152*	8.00E-27	45
119	231	no significant hit			
120	345	no significant hit			
121	243	no significant hit			
122	153	no significant hit			
123	111	no significant hit			
124	237	no significant hit			
125	378	no significant hit			
126	249	no significant hit			
127	594	DNA methylase	*Aeromonas salmonicida subsp. salmonicida A449*	2.00E-61	55
128	1278	conserved hypothetical protein	*uncultured archaeon*	2.00E-58	32
129	693	hypothetical protein KSE_52180	*Kitasatospora setae KM-6054*	4.00E-04	24
130	1230	ATPase	*Hahella chejuensis KCTC 2396*	5.00E-18	25
131	411	no significant hit			
132	429	no significant hit			

133	2028	peptidase M1, membrane alanine aminopeptidase	*Candidatus Solibacter usitatus Ellin6076*	3.00E-100	34
134	1380	DNA repair protein RadA	*Geobacter metallireducens GS-15*	4.00E-143	56
135	1995	Metal dependent amidohydrolase	*Erythrobacter sp. SD-21*	7.00E-31	27
136	1086	class V aminotransferase	*Candidatus Solibacter usitatus Ellin6076*	8.00E-69	40
137	957	L-asparaginase II	*Geobacillus sp. G11MC16*	2.00E-62	43
138	306	competence protein ComEA helix-hairpin-helix repeat protein	*Olsenella uli DSM 7084*	1.00E-09	44

R. P27K16



ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	237	gp29			
2	159	no significant hit			
3	1842	p68			
4	222	hypothetical protein			
5	780	apramycin acetyl transferase			
6	144	int			
7	225	hypothetical protein EfaeDRAFT_1157			
8	972	plasmid-partitioning protein			
9	1167	protoporphyrinogen oxidase			
10	756	replication protein			
11	294	resolvase			
12	5172	hypothetical protein Ftrac_0248	*Marivirga tractuosa DSM 4126*	5.00E-100	38
13	384	hypothetical protein Cpin_4641	*Chitinophaga pinensis DSM 2588*	3.00E-06	31
14	549	no significant hit			
15	822	no significant hit			
16	207	no significant hit			

17	1455	hypothetical protein BC1002_7102	*Burkholderia sp. CCGE1002*	3.00E-43	27
18	1236	hypothetical protein BC1002_7103	*Burkholderia sp. CCGE1002*	6.00E-38	29
19	987	hypothetical protein bll3582	*Bradyrhizobium japonicum USDA 110*	2.00E-07	50
20	1851	no significant hit			
21	2778	hypothetical protein NB231_12491	*Nitrococcus mobilis Nb-231*	4.00E-60	31
22	171	hypothetical protein Nham_2180	*Nitrobacter hamburgensis X14*	3.00E-14	70
23	1824	multicopper oxidase, type 3	*Candidatus Koribacter versatilis Ellin345*	1.00E-116	41
24	882	serine/threonine protein phosphatase	*Gemmatimonas aurantiaca T-27*	3.00E-67	50
25	504	no significant hit			
26	1428	hypothetical protein gll4242	*Gloeobacter violaceus PCC 7421*	2.00E-63	36
27	315	hypothetical protein MC7420_4440	*Microcoleus chthonoplastes PCC 7420*	3.00E-20	52
28	2436	rhamnulose-1-phosphate aldolase/alcohol dehydrogenase	*Thermobaculum terrenum ATCC BAA-798*	7.00E-80	34
29	528	no significant hit			
30	1404	Di-haem cytochrome c peroxidase	*Plesiocystis pacifica SIR-1*	3.00E-88	45
31	2775	TPR repeat-containing serine/threonin protein kinase	*Candidatus Koribacter versatilis Ellin345*	5.00E-73	38
32	366	sigma factor, ECF-like family protein	*Gemmatimonas aurantiaca T-27*	2.00E-08	42
33	1641	beta-lactamase domain protein	*Gloeobacter violaceus PCC 7421*	7.00E-32	28
34	441	hypothetical protein Sros_4960	*Streptosporangium roseum DSM 43021*	4.00E-14	35
35	231	conserved hypothetical protein	*Aspergillus terreus NIH2624*	4.00E-05	36

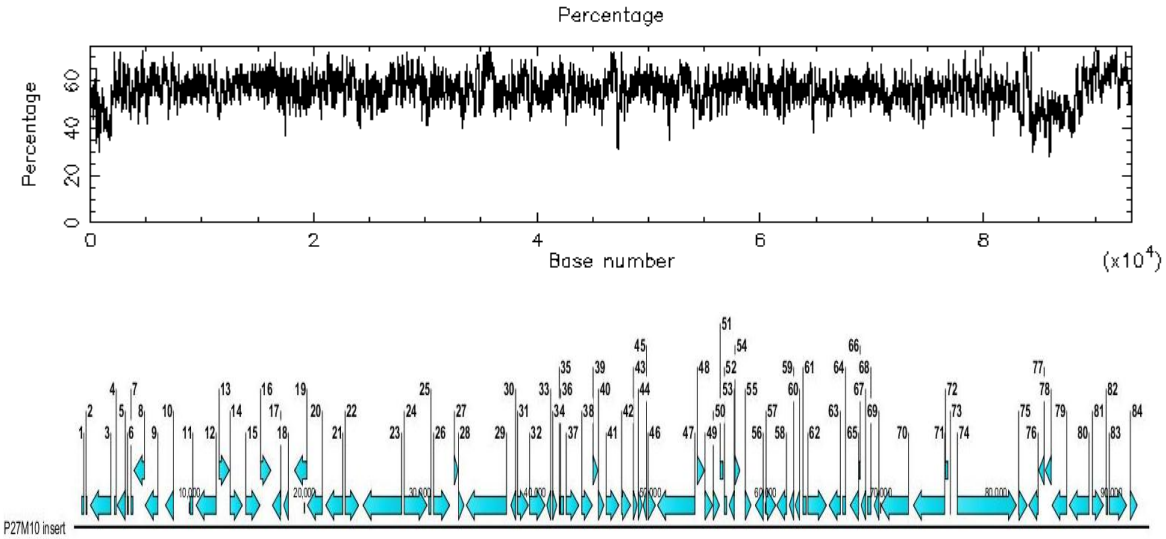
36	297	no significant hit			
37	516	no significant hit			
38	303	no significant hit			
39	2673	serine/threonine protein kinase	*Gemmatimonas aurantiaca T-27*	7.00E-85	35
40	198	serine/threonine protein kinase	*Gemmatimonas aurantiaca T-27*	9.00E-17	74
41	519	CHRD domain containing protein	*Nitrosococcus halophilus Nc4*	2.00E-05	36
42	804	AraC family transcriptional regulator	*Candidatus Solibacter usitatus Ellin6076*	2.00E-30	35
43	1077	blue (type 1) copper domain protein	*Ktedonobacter racemifer DSM 44963*	3.00E-13	43
44	411	plastocyanin-like protein	*Candidatus Koribacter versatilis Ellin345*	2.00E-14	42
45	603	beta-lactamase domain protein	*Opitutaceae bacterium TAV2*	1.00E-48	46
46	387	beta-lactamase domain-containing protein	*Shewanella woodyi ATCC 51908*	2.00E-31	69
47	765	no significant hit			
48	609	TetR family transcriptional regulator	*marine actinobacterium PHSC20C1*	1.00E-07	29
49	465	hypothetical protein GAU_1461	*Gemmatimonas aurantiaca T-27*	5.00E-12	37
50	366	no significant hit			
51	471	hypothetical protein CAP2UW1_2163	*Candidatus Accumulibacter phosphatis clade IIA str. UW-1*	5.00E-40	54
52	276	no significant hit			
53	447	no significant hit			
54	849	no significant hit			
55	876	XRE family transcriptional regulator	*Candidatus Solibacter usitatus Ellin6076*	6.00E-93	58
56	315	protein of unknown function	*NC10 bacterium*	2.00E-19	46
57	543	methionine-R-sulfoxide reductase	*Roseiflexus castenholzii DSM 13941*	8.00E-61	76

58	681	peptide methionine sulfoxide reductase	*Geobacter sp. M21*	9.00E-63	65
59	8499	glycosyltransferase 36	*Candidatus Solibacter usitatus Ellin6076*	0	44
60	204	no significant hit			
61	240	no significant hit			
62	399	MutT/NUDIX family protein	*Bifidobacterium breve DSM 20213*	2.00E-14	34
63	396	no significant hit			
64	954	cell surface protein	*Hydrogenivirga sp. 128-5-R1-1*	8.00E-11	31
65	2706	conserved membrane protein of unknown function	*NC10 bacterium*	0	47
66	471	conserved hypothetical protein	*bacterium Ellin514*	4.00E-24	53
67	216	cold shock protein	*Gemmatimonas aurantiaca T-27*	2.00E-27	87
68	882	hypothetical protein GAU_1047	*Gemmatimonas aurantiaca T-27*	7.00E-26	28
69	357	no significant hit			
70	1473	fumarate reductase/succinate dehydrogenase flavoprotein	*Ktedonobacter racemifer DSM 44963*	5.00E-128	56
71	579	formate dehydrogenase, alpha subunit	*Sphaerobacter thermophilus DSM 20745*	5.00E-64	74
72	2556	formate dehydrogenase, alpha subunit	*Sphaerobacter thermophilus DSM 20745*	0	65
73	861	4Fe-4S ferredoxin iron-sulfur binding domain-containing protein	*Sphaerobacter thermophilus DSM 20745*	4.00E-108	72
74	966	Polysulfide reductase NrfD	*Sphaerobacter thermophilus DSM 20745*	2.00E-49	52
75	864	formate dehydrogenase accessory protein	*Sphaerobacter thermophilus DSM 20745*	4.00E-55	41
76	351	hypothetical protein Sthe_3393	*Sphaerobacter thermophilus DSM 20745*	9.00E-25	55

77	555	hypothetical protein Krac_2611	*Ktedonobacter racemifer DSM 44963*	4.00E-41	55
78	885	6-phosphogluconate dehydrogenase, NAD-binding	*Stigmatella aurantiaca DW4/3-1*	8.00E-82	54
79	1929	high molecular weight glutenin subunit 15*y	*Aegilops kotschyi*	8.00E-06	25
80	876	no significant hit			
81	489	DoxX family protein	*Verrucomicrobiae bacterium DG1235*	1.00E-28	48
82	540	MarR family transcriptional regulator	*Gemmatimonas aurantiaca T-27*	9.00E-42	66
83	270	no significant hit			
84	1254	Xaa-Pro aminopeptidase	*Gemmatimonas aurantiaca T-27*	6.00E-90	47
85	1188	outer membrane protein	*Candidatus Koribacter versatilis Ellin345*	2.00E-22	34
86	1482	major facilitator transporter	*Polaromonas sp. JS666*	1.00E-100	66
87	429	DoxX	*Anaeromyxobacter dehalogenans 2CP-C*	5.00E-52	72
88	1485	MATE efflux family protein	*uncultured bacterium 66*	3.00E-153	67
89	1422	MATE efflux family protein	*Myxococcus xanthus DK 1622*	7.00E-153	61
90	846	5'-3' exonuclease	*Candidatus Solibacter usitatus Ellin6076*	8.00E-58	54
91	150	ATP-dependent DNA ligase	*Anaeromyxobacter sp. Fw109-5*	1.00E-10	63
92	834	ATP dependent DNA ligase	*Variovorax paradoxus EPS*	5.00E-96	65
93	1254	DNA primase small subunit	*Anaeromyxobacter sp. Fw109-5*	2.00E-144	66
94	1011	ECF subfamily RNA polymerase sigma factor	*Sorangium cellulosum*	1.00E-108	83

95	384	DGPFAETKE domain-containing protein	*Anaeromyxobacter dehalogenans 2CP-C*	8.00E-32	55
96	867	hypothetical protein Tbis_1787	*Thermobispora bispora DSM 43833*	2.00E-30	46
97	459	glyoxalase/bleomycin resistance protein/dioxygenase	*Nostoc punctiforme PCC 73102*	5.00E-44	59
98	429	DGPFAETKE	*Pseudomonas fluorescens Pf0-1]*	6.00E-51	74
99	381	glyoxalase/bleomycin resistance protein/dioxygenase	*Mesorhizobium loti MAFF303099*	1.00E-36	61
100	240	glyoxalase/bleomycin resistance protein	*uncultured archaeon*	7.00E-10	46
101	438	CmR			

S. P27M10



ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	237	gp29			
2	159	no significant hit			
3	1842	p68			
4	222	hypothetical protein			
5	780	apramycin acetyl transferase			
6	144	int			
7	225	hypothetical protein EfaeDRAFT_1157			
8	972	plasmid-partitioning protein			
9	1167	protoporphyrinogen oxidase			
10	756	replication protein			
11	294	resolvase			
12	1806	hypothetical protein Acid345_2913	*Candidatus Koribacter versatilis Ellin345*	4E-180	52
13	954	putative proline racemase	*Acidobacterium capsulatum ATCC 51196*	4E-123	64
14	1131	oxidoreductase, FAD-dependent	*Acidobacterium capsulatum ATCC 51196*	3E-116	61

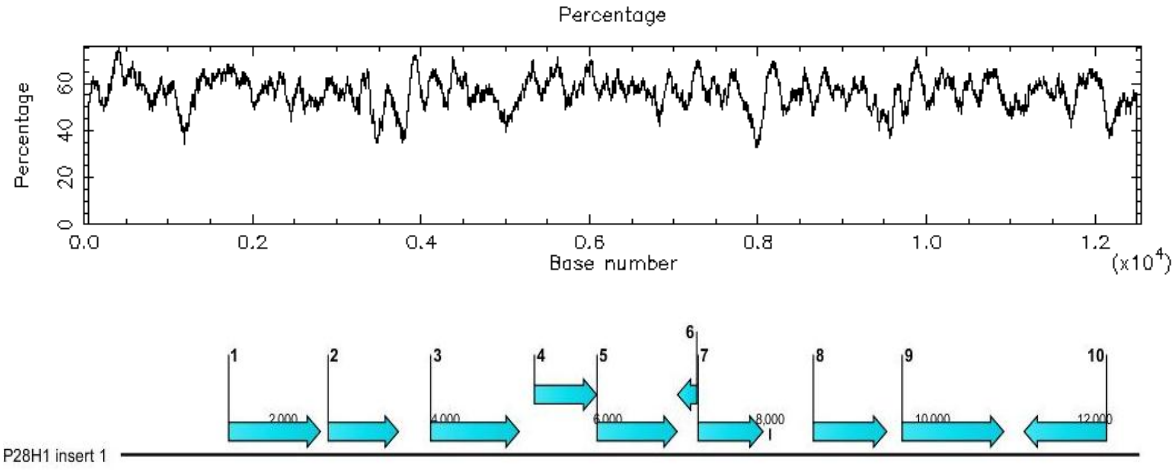
15	1293	pyridine nucleotide-disulfide oxidoreductase	*Acidobacterium capsulatum ATCC 51196*	1E-103	52
16	951	dihydrodipicolinate synthase	*Acidobacterium capsulatum ATCC 51196*	3E-102	61
17	795	hypothetical protein MkanA1_13575	*Mycobacterium kansasii ATCC 12478*	2E-16	43
18	471	ribonuclease H	*Myxococcus xanthus DK 1622*	2E-12	38
19	1143	geranylgeranyl reductase	*Planctomyces maris DSM 8797*	7E-103	48
20	1362	Cyclopropane-fatty-acyl-phospholipid synthase	*Planctomyces maris DSM 8797*	1E-115	53
21	1545	multicopper oxidase, type 2	*Candidatus Solibacter usitatus Ellin6076*	0	62
22	1239	mandelate racemase/muconate lactonizing protein	*Candidatus Solibacter usitatus Ellin6076*	3E-179	73
23	3414	TonB-dependent receptor	*Acidobacterium sp. MP5ACTX9*	0	38
24	2046	sulfatase	*Candidatus Solibacter usitatus Ellin6076*	1E-78	32
25	252	no significant hit			
26	1455	Phospholipase C	*Acidobacterium sp. MP5ACTX8*	0	67
27	378	NmrA family protein	*Desulfovibrio sp. FW1012B*	1E-10	40
28	546	NmrA family protein	*Anaeromyxobacter sp. Fw109-5*	2E-32	41
29	3555	hypothetical protein PARMER_00222	*Parabacteroides merdae ATCC 43184*	1E-135	29
30	480	3-demethylubiquinone-9 3-methyltransferase	*Opitutaceae bacterium TAV2*	5E-76	85
31	1002	hypothetical protein Acid345_3216	*Candidatus Koribacter versatilis Ellin345*	2E-23	33
32	1404	class V aminotransferase	*Terriglobus saanensis SP1PR4*	1E-136	53

33	408	transcriptional regulator, HxlR family	*Acidobacterium capsulatum ATCC 51196*	1E-31	50
34	417	Glyoxalase/bleomycin resistance protein/dioxygenase	*Blastopirellula marina DSM 3645*	4E-46	60
35	117	no significant hit			
36	318	no significant hit			
37	1161	Beta-lactamase	*Ktedonobacter racemifer DSM 44963*	5E-147	68
38	978	cyclase	*Bradyrhizobium japonicum USDA 110*	2E-58	42
39	471	hypothetical protein sce8962	*Sorangium cellulosum 'So ce 56*	4E-42	63
40	528	hypothetical protein Vapar_4411	*Variovorax paradoxus S110*	1E-56	58
41	1173	BNR/Asp-box repeat protein	*Acidobacterium capsulatum ATCC 51196*	0	78
42	843	alpha/beta hydrolase fold protein	*Geobacter sp. FRC-32*	6E-97	72
43	399	hypothetical protein Caul_2644	*Caulobacter sp. K31*	8E-24	51
44	417	glyoxalase/bleomycin resistance protein/dioxygenase	*Ralstonia eutropha JMP134*	5E-26	50
45	357	hypothetical protein Cyan7822_2938	*Cyanothece sp. PCC 7822*	8E-29	50
46	705	deiodinase, iodothyronine, type I	*Candidatus Koribacter versatilis Ellin345*	5E-29	68
47	3312	TonB-dependent receptor	*Candidatus Koribacter versatilis Ellin345*	0	68
48	696	amino acid transporter	*Candidatus Koribacter versatilis Ellin345*	5E-47	58
49	723	amino acid transporter	*Candidatus Koribacter versatilis Ellin345*	6E-47	54
50	594	alkylhydroperoxidase AhpD domain protein	*Acidobacterium capsulatum ATCC 51196*	1E-32	38

51	309	hypothetical protein Acid345_1680	*Candidatus Koribacter versatilis Ellin345*	1E-30	74
52	258	hypothetical protein Acid345_1681	*Candidatus Koribacter versatilis Ellin345*	8E-24	67
53	513	no significant hit			
54	495	no significant hit			
55	552	diguanylate cyclase (GGDEF) domain protein	*Acidobacterium capsulatum ATCC 51196*	4E-15	39
56	753	peptidase, T1A (proteasome) family	*Acidobacterium capsulatum ATCC 51196*	7E-79	59
57	888	transglutaminase	*Microcystis aeruginosa NIES-843*	1E-90	57
58	918	Xylose isomerase domain-containing protein TIM barrel	*Dyadobacter fermentans DSM 18053*	1E-60	43
59	444	hypothetical protein Acid345_3598	*Candidatus Koribacter versatilis Ellin345*	4E-37	63
60	447	hypothetical protein Acid345_1103	*Candidatus Koribacter versatilis Ellin345*	4E-26	46
61	342	Chaperonin Cpn10	*Acidobacterium capsulatum ATCC 51196*	7E-39	79
62	1668	chaperonin GroEL	*Candidatus Koribacter versatilis Ellin345*	0	83
63	1047	phenazine biosynthesis PhzC/PhzF protein	*Candidatus Koribacter versatilis Ellin345*	1E-78	55
64	303	Excinuclease ABC C subunit domain protein	*Acidobacterium sp. MP5ACTX8*	8E-16	52
65	786	protein of unknown function DUF899 thioredoxin family protein	*Chthoniobacter flavus Ellin428*	4E-78	60
66	144	no significant hit			

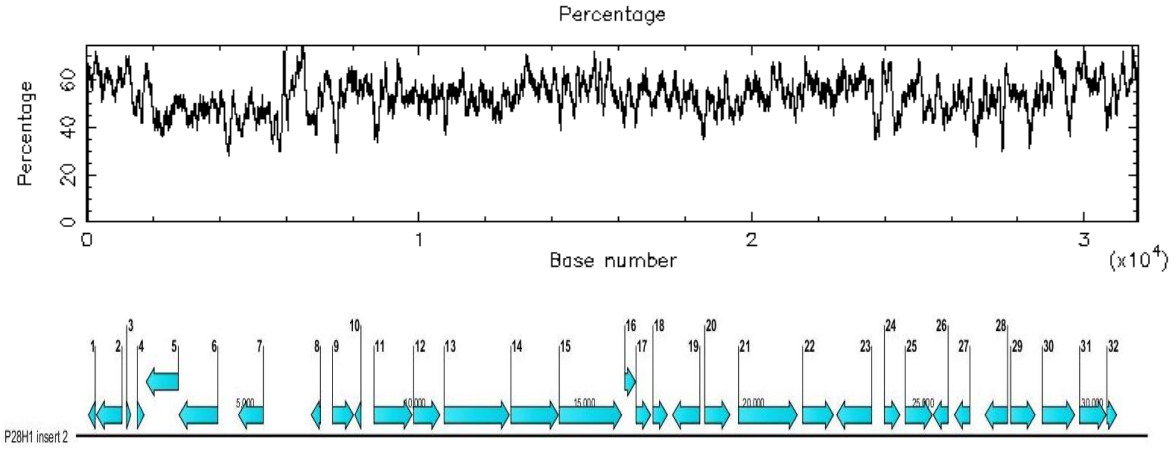
67	459	Activator of Hsp90 ATPase 1 family protein	*Acidobacterium sp. MP5ACTX8*	2E-35	57
68	342	ArsR family transcriptional regulator	*Streptomyces sp. AA4*	6E-18	61
69	474	conserved hypothetical protein	*Acidobacterium sp. MP5ACTX8*	5E-31	54
70	2442	hypothetical protein Acid345_1927	*Candidatus Koribacter versatilis Ellin345*	0	52
71	2784	cell surface receptor IPT/TIG domain-containing protein	*Delftia acidovorans SPH-1*	1E-39	36
72	282	no significant hit			
73	60	no significant hit			
74	5190	Ig family protein	*Roseiflexus castenholzii DSM 13941*	0	59
75	798	dienelactone hydrolase	*bacterium Ellin514*	1E-84	59
76	876	two component LuxR family transcriptional regulator	*Pseudomonas mendocina ymp*	2E-24	34
77	534	no significant hit			
78	576	no significant hit			
79	1350	no significant hit			
80	1785	polyvinyl-alcohol dehydrogenase	*Bradyrhizobium japonicum USDA 110*	7E-132	45
81	1020	gluconolactonase	*Planctomyces limnophilus DSM 3776*	3E-70	44
82	198	glutaredoxin 2	*Candidatus Solibacter usitatus Ellin6076*	0.0000006	40
83	1566	AMP-dependent synthetase and ligase	*Candidatus Koribacter versatilis Ellin345*	0	61
84	660	CmR			

T. P28H1 INSERT 1



ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	1146	two-component sensor histidine kinase-like protein	*Paenibacillus larvae subsp. larvae B-3650*	2E-50	36
2	882	2-dehydropantoate 2-reductase	*Verrucomicrobium spinosum DSM 4136*	6E-74	51
3	1107	sulfate ABC transporter, periplasmic sulfate-binding protein	*Geobacter uraniireducens Rf4*	1E-138	80
4	783	sulfate transporter permease	*Azoarcus sp. BH72*	3E-93	70
5	1002	sulfate ABC transporter permease protein CysW	*Cupriavidus metallidurans CH34*	2E-80	75
6	255	sulphate transport system permease protein 1	*Dechloromonas aromatica RCB*	1E-24	69
7	816	sulphate transport system permease protein 1	*Nitrosospira multiformis ATCC 25196*	6E-73	60
8	921	conserved exported protein of unknown function	*NC10 bacterium*	4E-08	33
9	1269	oxidoreductase domain-containing protein	*Shewanella sp. W3-18-1*	1E-131	54
10	1032	phosphate transport system regulatory protein PhoU	*Geobacter sulfurreducens PCA*	5E-80	55

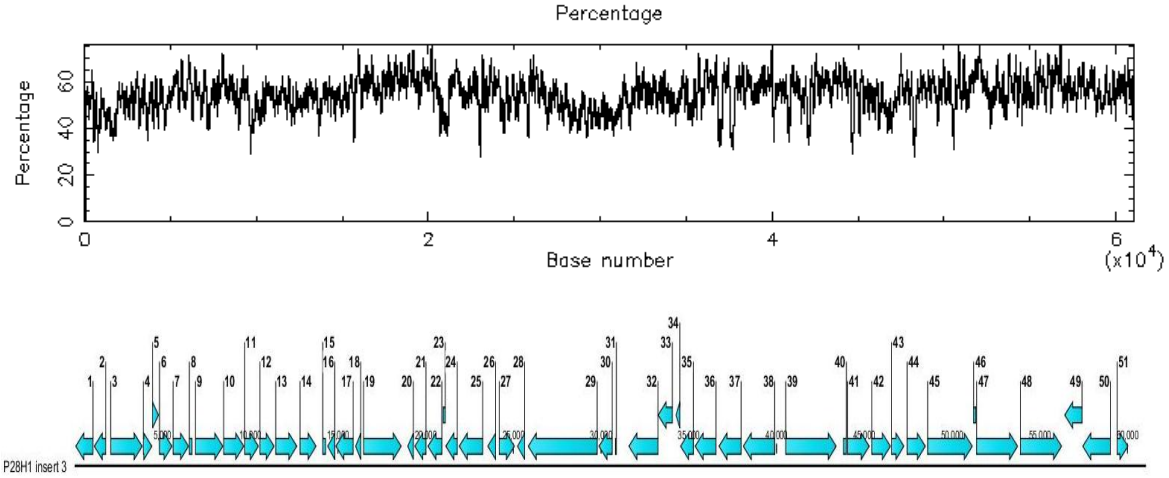
U. P28HI INSERT 2



ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	222	hypothetical protein			
2	780	apramycin acetyl transferase			
3	144	int [Cloning vector pTARa]			
4	225	hypothetical protein EfaeDRAFT_1157			
5	972	plasmid-partitioning protein			
6	1167	protoporphyrinogen oxidase			
7	756	replication protein			
8	294	resolvase			
9	642	TetR family transcriptional regulator	*Rhodopseudomonas palustris HaA2*	9E-55	54
10	201	glutathione-dependent formaldehyde-activating GFA	*Thiomonas intermedia K12 K12*	9E-10	49
11	1146	LacI family transcription regulator	*Acidothermus cellulolyticus 11B*	8E-16	32
12	798	hypothetical protein Cflav_PD3372	*bacterium Ellin514*	6E-08	27
13	1947	hypothetical protein Cwoe_0480	*Conexibacter woesei DSM 14684*	5E-100	35
14	1422	hypothetical protein PM8797T_22933	*Planctomyces maris DSM 8797*	2E-11	25
15	1863	pyridine nucleotide-disulphide oxidoreductase	*Blastopirellula marina DSM 3645*	3E-179	58

16	339	transcriptional regulator, ArsR family	*bacterium Ellin514*	7E-36	73
17	453	hypothetical protein blr7360	*Bradyrhizobium japonicum USDA 110*	5E-39	56
18	447	hypothetical protein blr7360	*Bradyrhizobium japonicum USDA 110*	9E-26	53
19	831	transcriptional regulator, AraC family	*Victivallis vadensis ATCC BAA-548*	4E-13	35
20	765	hypothetical protein ObacDRAFT_6920	*Opitutaceae bacterium TAV2*	0.00008	23
21	1758	hypothetical protein GYMC10_3424	*Paenibacillus sp. Y412MC10*	5E-61	31
22	933	hypothetical protein PM8797T_00392	*Planctomyces maris DSM 8797*	3E-51	38
23	1056	Dipeptidyl aminopeptidase/acylaminoacyl-peptidase-like protein	*Chthoniobacter flavus Ellin428*	4E-104	54
24	480	hypothetical protein PFL_2652	*Pseudomonas fluorescens Pf-5*	2E-31	50
25	819	hypothetical protein amb3252	*Magnetospirillum magneticum AMB-1*	0.00008	39
26	462	hypothetical protein Sulku_2674	*Sulfuricurvum kujiense DSM 16994*	8E-29	55
27	480	type I polyketide synthase	*Chlamydomonas reinhardtii*	0.25	27
28	687	hypothetical protein Mmwyl1_2453	*Marinomonas sp. MWYL1*	4E-36	36
29	744	pyridoxal phosphate enzyme, YggS family	*Persephonella marina EX-H1*	6E-57	49
30	981	hypothetical protein Hore_09230	*Halothermothrix orenii H 168*	1E-10	38
31	804	pyrroline-5-carboxylate reductase	*Stigmatella aurantiaca DW4/3-1*	1E-71	53
32	312	hypothetical protein DSM3645_27241	*Blastopirellula marina DSM 3645*	2E-13	42

V. P28H1 INSERT 3

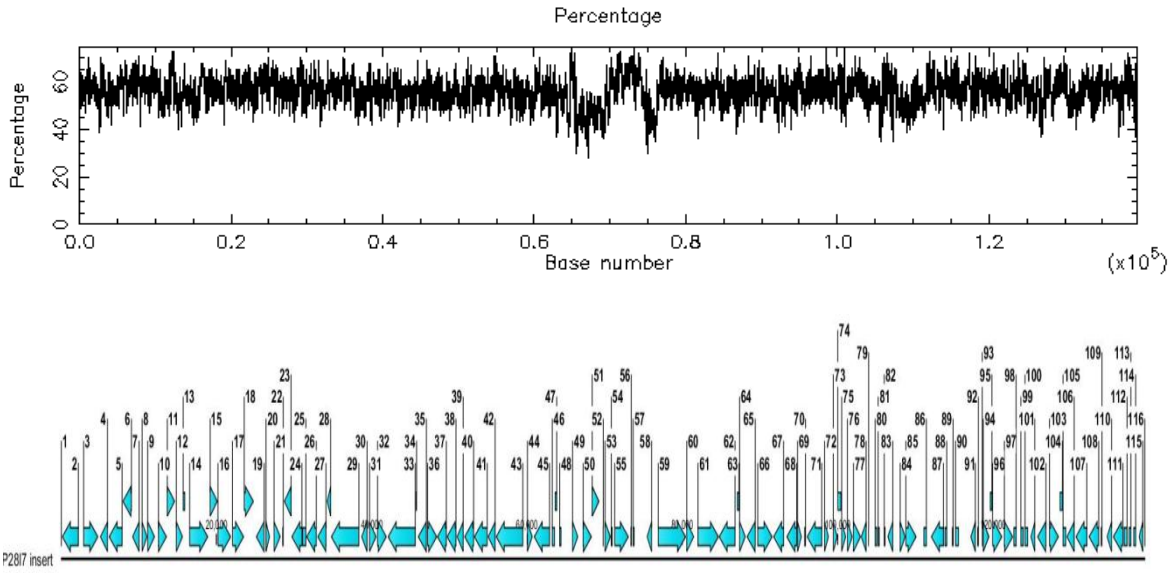


ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	1026	hypothetical protein Tter_2228	*Thermobaculum terrenum ATCC BAA-798*	1E-71	43
2	699	N-acetylglucosaminyltransferase	*Salinibacter ruber M8*	1E-26	37
3	1833	no significant hit			
4	528	2-oxoglutarate dehydrogenase, E2 subunit, dihydrolipoamide succinyltransferase	*Geobacillus sp. C56-T3*	1E-12	35
5	381	4'-phosphopantetheinyl transferase	*Candidatus Solibacter usitatus Ellin6076*	1E-30	51
6	753	3-oxoacyl-[acyl-carrier-protein] reductase	*Bacillus pumilus SAFR-032*	6E-54	49
7	924	PfkB domain-containing protein	*Pirellula staleyi DSM 6068*	4E-43	36
8	195	no significant hit			
9	1602	phospholipase/Carboxyl esterase	*Dyadobacter fermentans DSM 18053*	3E-39	27
10	1188	3-oxoacyl-(acyl-carrier-protein) synthase 2	*Thermoanaerobacter ethanolicus JW 200*	4E-77	42
11	825	Beta-ketoacyl synthase	*Frankia symbiont of Datisca glomerata*	5E-10	35

12	852	hypothetical protein AURANDRAFT_62341	*Aureococcus anophagefferens]*	0.00001	28
13	1248	secreted protein	*Streptomyces coelicolor A3(2)*	8E-72	40
14	954	hypothetical protein Plim_1627	*Planctomyces limnophilus DSM 3776*	0.00003	33
15	201	no significant hit			
16	450	hypothetical protein VspiD_07150	*Verrucomicrobium spinosum DSM 4136]*	3E-39	52
17	1032	arsenical-resistance protein	*Acidobacterium capsulatum ATCC 51196*	7E-147	74
18	336	transcriptional regulator, ArsR family	*Anaeromyxobacter dehalogenans 2CP-1*	2E-26	67
19	2190	hypothetical protein SNOG_08075	*Phaeosphaeria nodorum SN15*	0.00004	26
20	366	assimilatory nitrite reductase subunit	*Haladaptatus paucihalophilus DX253*	2E-16	42
21	681	enhancing lycopene biosynthesis protein 2	*Sulfurihydrogenibium yellowstonense SS-5*	3E-53	48
22	837	hypothetical protein SULAZ_1484	*Sulfurihydrogenibium azorense Az-Fu1*	2E-85	58
23	159	no significant hit			
24	696	NAD binding domain of 6-phosphogluconate dehydrogenase family	*Microcoleus chthonoplastes PCC 7420*	7E-54	53
25	1380	hypothetical protein Minf_0231	*Methylacidiphilum infernorum V4*	1E-33	39
26	489	Rieske (2Fe-2S) domain-containing protein	*Nostoc punctiforme PCC 73102*	3E-13	33
27	909	polysaccharide deacetylase	*Anabaena variabilis ATCC 29413*	3E-50	50
28	456	no significant hit			
29	3969	no significant hit			
30	795	no significant hit			
31	99	no significant hit			
32	1707	no significant hit			
33	855	no significant hit			

34	264	putative type-I PKS	*Streptomyces griseus subsp. griseus NBRC 13350*	8.1	35
35	780	Prepilin peptidase	*Desulfuromonas acetoxidans DSM 684*	3E-31	37
36	1233	putative type IV fimbrial assembly protein PilC	*delta proteobacterium NaphS2*	2E-80	43
37	1305	twitching motility protein	*Thermosinus carboxydivorans Nor1*	5E-95	55
38	1836	general secretory pathway protein E	*Thermosinus carboxydivorans Nor1*	9E-142	43
39	2916	DNA translocase FtsK	*Carboxydothermus hydrogenoformans Z-2901*	6E-126	54
40	207	no significant hit			
41	1278	metal dependent phosphohydrolase	*Acidobacterium sp. MP5ACTX9*	1E-98	47
42	1104	transcription regulator	*Lactobacillus plantarum subsp. plantarum ST-III*	7E-20	28
43	744	no significant hit			
44	1083	laminin G domain-containing protein	*Caulobacter segnis ATCC 21756*	0.00002	24
45	2586	conserved hypothetical protein	*Chthoniobacter flavus Ellin428*	2E-07	22
46	174	no significant hit			
47	2382	Beta-agarase	*Victivallis vadensis ATCC BAA-548*	8E-88	39
48	2394	hypothetical protein Sros_4284	*Streptosporangium roseum DSM 43021*	3E-46	29
49	1032	response regulator receiver modulated metal dependent phosphohydrolase	*Paenibacillus curdlanolyticus YK9*	7E-95	54
50	1620	sensory box histidine kinase/response regulator	*Pseudomonas syringae pv. tomato str. DC3000*	1E-81	42
51	660	CmR			

W. P28I7



ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	96	no significant hit			
2	2154	radical SAM domain protein	*Acidobacterium capsulatum ATCC 51196*	0	78
3	1932	PgPepO oligopeptidase	*Candidatus Koribacter versatilis Ellin345*	0	65
4	1002	zinc-binding alcohol dehydrogenase family protein	*Ammonifex degensii KC4*	5.00E -108	60
5	1773	chloride transporter, ClC family	*Acidobacterium capsulatum ATCC 51196*	1.00E -158	60
6	1194	hypothetical protein ACP_0083	*Acidobacterium capsulatum ATCC 51196*	6.00E -28	28
7	933	Ornithine cyclodeaminase	*Thermococcus sibiricus MM 739*	1.00E -45	38
8	714	putative esterase	*Acidobacterium sp. MP5ACTX8*	2.00E -78	57
9	969	glutathione synthase/ribosomal protein S6 modification glutaminyl transferase-like protein	*Acidobacterium sp. MP5ACTX9*	5.00E -120	68

10	1140	carboxylate-amine ligase	*Candidatus Solibacter usitatus Ellin6076*	7.00E -160	70
11	1047	aminopeptidase	*Candidatus Solibacter usitatus Ellin6076*	1.00E -129	66
12	885	hypothetical protein Acid_3924	*Candidatus Solibacter usitatus Ellin6076*	3.00E -78	53
13	300	hypothetical protein Acid_3924	*Candidatus Solibacter usitatus Ellin6076*	1.00E -30	64
14	2430	TPR repeat-containing serine/threonin protein kinase	*Candidatus Koribacter versatilis Ellin345*	3.00E -106	35
15	1017	hypothetical protein Acid345_3334	*Candidatus Koribacter versatilis Ellin345*	4.00E -100	62
16	1713	hypothetical protein Acid345_3334	*Candidatus Koribacter versatilis Ellin345*	3.00E -111	40
17	1512	amino acid transporter	*Candidatus Koribacter versatilis Ellin345*	3.00E -171	61
18	1272	hypothetical protein ZOD2009_13206	*Haladaptatus paucihalophilus DX253*	6.00E -80	49
19	1083	glycosyl hydrolase, BNR repeat-containing protein	*Gemmata obscuriglobus UQM 2246*	2.00E -61	44
20	561	hypothetical protein Acid345_3332	*Candidatus Koribacter versatilis Ellin345*	4.00E -54	58
21	888	conserved hypothetical protein	*Ktedonobacter racemifer DSM 44963*	4.00E -49	36
22	168	no significant hit			
23	966	homoserine kinase	*Terriglobus saanensis SP1PR4*	3.00E -60	51
24	1383	threonine synthase	*Terriglobus saanensis SP1PR4*	4.00E -144	58
25	504	hypothetical protein Acid345_1368	*Candidatus Koribacter versatilis Ellin345*	9.00E -35	57
26	1335	amidohydrolase 2	*Candidatus Solibacter usitatus Ellin6076*	8.00E -65	33
27	1158	aminotransferase	*Candidatus Solibacter usitatus Ellin6076*	2.00E -98	49
28	561	alkylhydroperoxidase like protein, AhpD family	*Acidobacterium sp. MP5ACTX8*	7.00E -19	35
29	3642	TonB-dependent receptor plug	*Acidobacterium sp. MP5ACTX9*	7.00E -125	32

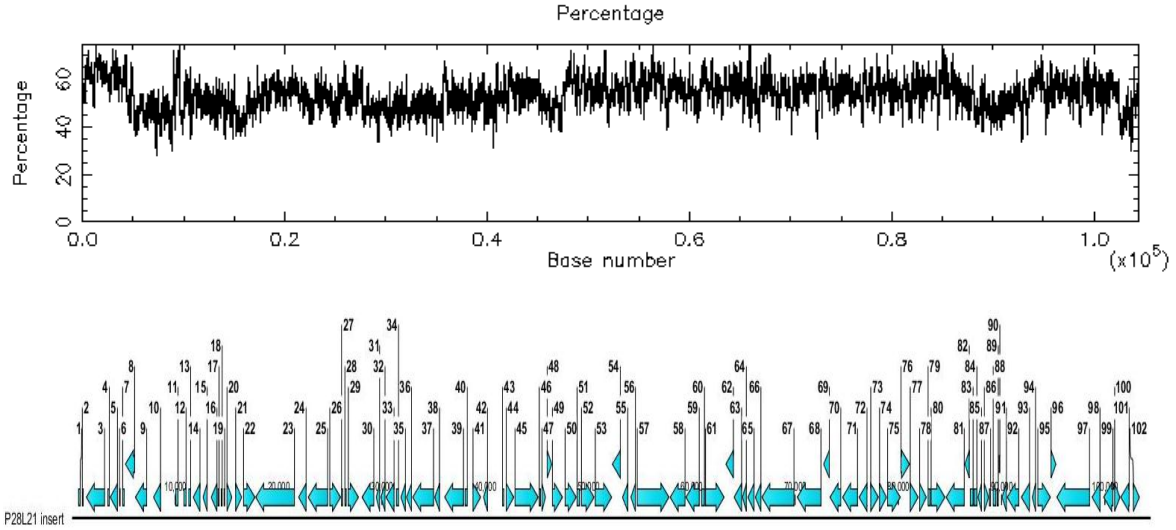
30	789	oxidoreductase, short chain dehydrogenase/reductase family	*Acidobacterium capsulatum ATCC 51196*	3.00E-72	52
31	942	branched-chain amino acid aminotransferase I	*Anaerolinea thermophila UNI-1*	2.00E-98	59
32	1209	cystathionine beta-lyase	*Microscilla marina ATCC 23134*	1.00E-90	43
33	3600	TonB-dependent receptor	*Candidatus Koribacter versatilis Ellin345*	0	54
34	159	no significant hit			
35	1137	LacI family transcription regulator	*Candidatus Koribacter versatilis Ellin345*	2.00E-125	61
36	1263	TPR repeat-containing protein	*Candidatus Koribacter versatilis Ellin345*	7.00E-37	30
37	1242	ABC efflux pump, inner membrane subunit	*Candidatus Koribacter versatilis Ellin345*	6.00E-126	60
38	1245	ABC efflux pump, inner membrane subunit	*Candidatus Koribacter versatilis Ellin345*	2.00E-120	57
39	870	hypothetical protein Tpen_1785	*Thermofilum pendens Hrk 5*	5.00E-36	33
40	1245	hypothetical protein Cpin_6276	*Chitinophaga pinensis DSM 2588*	4.00E-61	39
41	1737	hypothetical protein Acid345_0425	*Candidatus Koribacter versatilis Ellin345*	0	73
42	999	TPR repeat-containing protein	*Candidatus Koribacter versatilis Ellin345*	6.00E-83	53
43	3516	hypothetical protein Acid345_0423	*Candidatus Koribacter versatilis Ellin345*	0	50
44	723	GntR family transcriptional regulator	*Candidatus Koribacter versatilis Ellin345*	1.00E-85	66
45	2073	Beta-N-acetylhexosaminidase	*Candidatus Koribacter versatilis Ellin345*	0	56
46	384	no significant hit			
47	291	4-hydroxybenzoate polyprenyltransferase	*Candidatus Solibacter usitatus Ellin6076*	4.00E-11	67
48	294	resolvase			
49	756	replication protein [Plasmid F]			

50	1167	protoporphyrinogen oxidase			
51	972	plasmid-partitioning protein [Plasmid F]			
52	144	int			
53	780	apramycin acetyl transferase			
54	222	hypothetical protein			
55	1842	p68			
56	159	no significant hit			
57	237	gp29			
58	660	CmR			
59	3624	pyruvate:ferredoxin (flavodoxin) oxidoreductase	*Microcoleus chthonoplastes PCC 74200.0*	0	68
60	999	dihydroorotate dehydrogenase 2	*Cyanothece sp. PCC 7424*	2.00E -116	65
61	2727	diguanylate cyclase and metal dependent phosphohydrolase	*Acidobacterium sp. MP5ACTX9*	6.00E -180	46
62	2139	cellulase precursor	*Candidatus Koribacter versatilis Ellin345*	0	64
63	297	glycosyl transferase, group 1	*Candidatus Koribacter versatilis Ellin345*	2.00E -17	62
64	822	glycosyl transferase, group 1	*Candidatus Koribacter versatilis Ellin346*	1.00E -82	54
65	1158	no significant hit			
66	1938	peptidase S9, prolyl oligopeptidase	*Candidatus Koribacter versatilis Ellin345*	0	80
67	1392	radical SAM domain protein	*delta proteobacterium NaphS2*	2.00E -73	38
68	1362	phospholipase C	*Candidatus Koribacter versatilis Ellin345*	5.00E -95	47
69	585	hypothetical protein Glov_1457	*Geobacter lovleyi SZ*	1.00E -32	41
70	222	hypothetical protein Dehly_1129	*Dehalogenimonas lykanthroporepellens BL-DC-9*	1.00E -11	73
71	1971	acetyl-coenzyme A synthetase	*Candidatus Koribacter versatilis Ellin345*	0	76
72	606	hypothetical protein Acid345_2332	*Candidatus Koribacter versatilis Ellin345*	2.00E -25	51
73	585	ECF subfamily RNA polymerase sigma-24 factor	*Candidatus Solibacter usitatus Ellin6076*	2.00E -32	45

74	504	no significant hit			
75	612	hypothetical protein ACP_2749	*Acidobacterium capsulatum ATCC 51196*	1.00E-18	42
76	696	phosphoribosylformylglycinamidine synthase I	*Candidatus Koribacter versatilis Ellin345*	6.00E-107	75
77	1014	oxidoreductase	*Candidatus Koribacter versatilis Ellin345*	4.00E-99	57
78	726	putative lipoprotein	*Burkholderia pseudomallei 1106b*	3.00E-04	31
79	123	no significant hit			
80	306	no significant hit			
81	234	no significant hit			
82	219	two component LuxR family transcriptional regulator	*Candidatus Koribacter versatilis Ellin345*	5.00E-08	56
83	729	major intrinsic protein	*Candidatus Solibacter usitatus Ellin6076*	2.00E-81	70
84	744	hypothetical protein Noca_3089	*Nocardioides sp. JS614*	1.00E-45	51
85	1410	hypothetical protein Noca_3089	*Nocardioides sp. JS614*	8.00E-150	58
86	417	no significant hit			
87	1665	chaperonin GroEL	*Candidatus Koribacter versatilis Ellin345*	0	87
88	327	chaperonin, 10 kDa	*Acidobacterium capsulatum ATCC 51196*	1.00E-37	80
89	210	rhodanese-like protein	*Geobacter metallireducens GS-15*	6.00E-06	38
90	468	putative cytochrome c family protein	*Acidobacterium capsulatum ATCC 51196*	6.00E-23	42
91	675	no significant hit			
92	198	no significant hit			
93	117	no significant hit			
94	825	hypothetical protein Acid_4156	*Candidatus Solibacter usitatus Ellin6076*	8.00E-99	76
95	306	D-aminoacylase	*Gloeobacter violaceus PCC 7421*	6.00E-11	54
96	1284	D-aminoacylase	*Gloeobacter violaceus PCC 7421*	1.00E-139	59
97	1167	hypothetical protein Acid_5342	*Candidatus Solibacter usitatus Ellin6076*	2.00E-164	78

98	393	no significant hit			
99	402	no significant hit			
100	453	hypothetical protein Acid345_1103	*Candidatus Koribacter versatilis Ellin345*	7.00E-23	53
101	612	hypothetical protein FraEuI1c_6756	*Frankia sp. EuI1c*	2.00E-15	30
102	1161	phosphoesterase	*Candidatus Koribacter versatilis Ellin345*	4.00E-60	47
103	1251	beta-ketoacyl synthase	*Candidatus Koribacter versatilis Ellin345*	1.00E-159	69
104	405	conserved hypothetical protein	*Ktedonobacter racemifer DSM 44963*	2.00E-22	42
105	435	hypothetical protein MXAN_0913	*Myxococcus xanthus DK 1622*	2.00E-33	52
106	1026	GHMP kinase	*Candidatus Koribacter versatilis Ellin345*	3.00E-94	63
107	1500	glycosyl transferase family protein	*Candidatus Koribacter versatilis Ellin345*	7.00E-31	32
108	1350	Aromatic-L-amino-acid decarboxylase	*Mesorhizobium opportunistum WSM2075*	2.00E-128	57
109	192	no significant hit			
110	654	hypothetical protein Acid345_3645	*Candidatus Koribacter versatilis Ellin345*	1.00E-55	58
111	1317	major facilitator superfamily protein	*Cyanothece sp. PCC 7425*	1.00E-112	52
112	486	thermosensitive gluconokinase	*Stigmatella aurantiaca DW4/3-1*	8.00E-41	53
113	333	no significant hit			
114	351	hypothetical protein AciX9_1029	*Acidobacterium sp. MP5ACTX9*	5.00E-10	30
115	582	hypothetical protein AciPR4_2329	*Terriglobus saanensis SP1PR4*	6.00E-16	30
116	165	no significant hit			78

X. P28L21



ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	237	gp29			
2	159	no significant hit			
3	1842	p68			
4	222	hypothetical protein			
5	780	apramycin acetyl transferase			
6	144	int			
7	225	hypothetical protein EfaeDRAFT_1157			
8	972	plasmid-partitioning protein			
9	1167	protoporphyrinogen oxidase			
10	756	replication protein			
11	294	resolvase			
12	285	no significant hit			
13	255	no significant hit			
14	729	RNA polymerase, sigma-24 subunit, ECF subfamily	*Odoribacter splanchnicus DSM 20712*	3.00E-05	29
15	471	no significant hit			
16	330	no significant hit			
17	261	no significant hit			
18	186	no significant hit			
19	162	no significant hit			

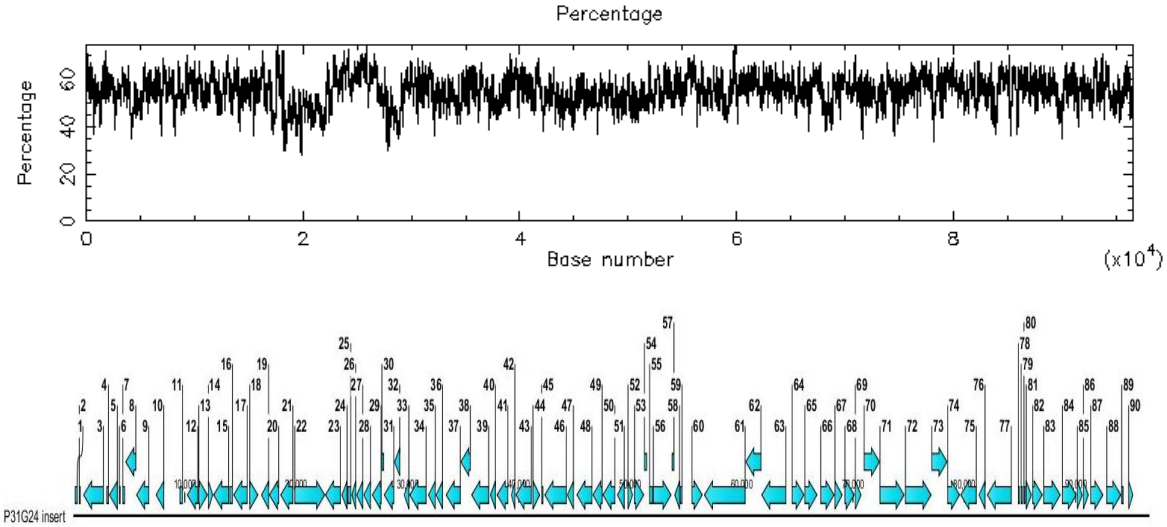
20	135	no significant hit			
21	543	no significant hit			
22	660	no significant hit			
23	1173	hypothetical protein RCCS2_04354	*Roseobacter sp. CCS2*	9.00E-09	25
24	3819	hypothetical protein PE36_18104	*Moritella sp. PE36*	1.00E-180	46
25	798	no significant hit			
26	1953	transcriptional regulator domain-containing protein	*Candidatus Solibacter usitatus Ellin6076*	1.00E-91	33
27	1113	no significant hit			
28	177	no significant hit			
29	201	no significant hit			
30	1053	acyltransferase 3	*Verrucomicrobium spinosum DSM 4136*	1.00E-31	34
31	1209	hypothetical protein Mmc1_2155	*Magnetococcus sp. MC-1*	7.00E-45	32
32	402	hypothetical protein LIC12054	*Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130*	4.00E-05	31
33	474	no significant hit			
34	852	no significant hit			
35	282	no significant hit			
36	510	no significant hit			
37	564	no significant hit			
38	2094	Hemolysin-type calcium-binding region	*Pelagibaca bermudensis HTCC2601*	1.00E-07	40
39	543	no significant hit			
40	1899	Hypothetical protein COLAER_01991	*Collinsella aerofaciens ATCC 25986*	5.00E-13	38
41	282	no significant hit			
42	810	conserved hypothetical protein	*Thiomonas sp. 3As*	2.00E-59	46
43	456	hypothetical protein Tint_2920	*Thiomonas intermedia K12*	1.00E-23	41
44	192	conjugative relaxase domain protein	*Acidobacterium sp. MP5ACTX8*	1.00E-06	69
45	798	conjugative relaxase domain protein	*Acidobacterium sp. MP5ACTX8*	3.00E-81	58

46	2319	conjugative relaxase domain protein	*Acidobacterium sp. MP5ACTX8*	3.00E-129	41
47	204	hypothetical protein NB311A_05880	*Nitrobacter sp. Nb-311A*	3.00E-05	59
48	411	no significant hit			
49	513	resolvase	*Nitrobacter sp. Nb-311A*	4.00E-53	69
50	1041	resolvase	*Nitrococcus mobilis Nb-231*	1.00E-100	49
51	1143	nicotinamide nucleotide transhydrogenase, subunit alpha	*Nodularia spumigena CCY9414*	5.00E-101	53
52	324	nicotinamide nucleotide transhydrogenase, subunit alpha2	*NC10 bacterium 'Dutch sediment'*	1.00E-28	72
53	1281	NAD(P)(+) transhydrogenase (AB-specific)	*Anaeromyxobacter sp. K*	1.00E-127	61
54	1665	peptidase M28	*Candidatus Solibacter usitatus Ellin6076*	3.00E-82	34
55	840	hypothetical protein Acid345_0035	*Candidatus Koribacter versatilis Ellin345*	6.00E-71	50
56	585	no significant hit			
57	519	no significant hit			
58	3138	peptidase S41	*Clostridium thermocellum ATCC 27405*	1.00E-12	26
59	1512	peptidase M16	*Algoriphagus sp. PR1*	4.00E-66	36
60	1368	peptidase S16B family protein	*Gemmatimonas aurantiaca T-27*	5.00E-78	41
61	354	no significant hit			
62	1875	peptidase M1, membrane alanine aminopeptidase	*Candidatus Solibacter usitatus Ellin6076*	5.00E-87	30
63	798	protein of unknown function DUF1009	*Acidobacterium sp. MP5ACTX9*	1.00E-79	57
64	798	acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamine O-acyltransferase	*Candidatus Solibacter usitatus Ellin6076*	8.00E-71	55
65	453	beta-hydroxyacyl-(acyl-carrier-protein) dehydratase FabZ	*Acidobacterium capsulatum ATCC 51196*	5.00E-40	54

66	663	outer membrane protein, putative	*Thermodesulfovibrio yellowstonii DSM 11347*	1.00E-07	26
67	615	outer membrane protein	*Acidobacterium capsulatum ATCC 51196*	2.00E-08	26
68	3180	surface antigen (D15)	*Candidatus Koribacter versatilis Ellin345*	5.00E-108	51
69	2367	ATP-dependent Clp protease, ATP-binding subunit ClpC	*Acidobacterium capsulatum ATCC 51196*	0	67
70	624	ABC transporter	*Mariprofundus ferrooxydans PV-1*	3.00E-40	49
71	1116	lipoprotein releasing system, transmembrane protein, LolC/E family	*delta proteobacterium NaphS2*	5.00E-38	33
72	1518	no significant hit			
73	849	response regulator receiver protein	*Desulfarculus baarsii DSM 2075*	3.00E-29	32
74	813	lipid A biosynthesis acyltransferase	*Candidatus Koribacter versatilis Ellin345*	2.00E-50	41
75	750	protein of unknown function DUF374	*Nitrosococcus halophilus Nc4*	4.00E-40	44
76	1308	peptidase T	*Oligotropha carboxidovorans OM5*	3.00E-104	51
77	894	GTP-binding protein Era	*Ammonifex degensii KC4*	6.00E-72	49
78	840	50S ribosomal protein L11 methyltransferase	*Caldicellulosiruptor owensensis OL*	1.00E-30	34
79	726	protein of unknown function DUF558	*Desulfurivibrio alkaliphilus AHT2*	2.00E-36	38
80	276	hypothetical protein MCP_1658	*Methanocella paludicola SANAE*	3.00E-04	34
81	1407	secreted serine protease MCP-01	*Candidatus Chloracidobacterium thermophilum*	4.00E-93	51
82	1821	TPR repeat-containing protein	*Candidatus Solibacter usitatus Ellin6076*	2.00E-18	22
83	495	hypothetical protein NIDE3048	*Candidatus Nitrospira defluvii*	9.00E-21	40
84	348	putative anti-sigma factor antagonist	*uncultured Acidobacteria bacterium*	6.00E-30	57

85	348	anti-anti-sigma factor	*Terriglobus saanensis SP1PR4*	1.00E-23	61
86	414	putative anti-sigma regulatory factor, serine/threonine protein kinase	*Candidatus Koribacter versatilis Ellin345*	9.00E-23	43
87	486	peptidase S1C, HrtA/DegP2/Q/S	*Geobacter metallireducens GS-15*	2.00E-04	25
88	189	no significant hit			
89	336	no significant hit			
90	180	no significant hit			
91	159	no significant hit			
92	504	PilT domain-containing protein	*Thermosediminibacter oceani DSM 16646*	2.00E-16	30
93	1197	DNA-binding protein, putative	*Rhodospirillum centenum SW*	3.00E-67	38
94	867	unclassified family transposase	*Acidobacterium sp. MP5ACTX9*	1.00E-83	54
95	387	two-component hybrid sensor and regulator	*Arthrospira platensis NIES-39*	3.00E-17	40
96	1284	3-phosphoshikimate 1-carboxyvinyltransferase	*Candidatus Koribacter versatilis Ellin345*	6.00E-100	53
97	543	shikimate kinase	*Rhodobacter capsulatus SB 1003*	3.00E-17	42
98	3237	Cna B domain-containing protein	*Terriglobus saanensis SP1PR4*	0	45
99	837	Tyrosine 3-monooxygenase	*Thermobispora bispora DSM 43833*	3.00E-70	50
100	966	DNA polymerase LigD, polymerase domain-containing protein	*Dyadobacter fermentans DSM 18053*	2.00E-67	44
101	582	YceI family protein	*Candidatus Solibacter usitatus Ellin6076*	5.00E-11	30
102	891	UDP-glucose 4-epimerase	*Halothermothrix orenii H 168*	7.00E-95	62
103	660	CmR			

Y. P31G24



ORF	Length (bp)	Top Hit (function)	Top Hit (Microbe)	E value	% Similarity
1	237	gp29			
2	159	no significant hit			
3	1842	p68			
4	222	hypothetical protein			
5	780	apramycin acetyl transferase			
6	144	int			
7	225	hypothetical protein EfaeDRAFT_1157			
8	972	plasmid-partitioning protein			
9	1167	protoporphyrinogen oxidase			
10	756	replication protein			
11	294	resolvase			
12	1002	Homoserine dehydrogenase	*Ktedonobacter racemifer DSM 44963*	8.00E-96	54
13	849	thymidylate synthase	*Acinetobacter lwoffii SH145*	5.00E-89	58
14	471	dihydrofolate reductase	*Leuconostoc gasicomitatum LMG 18811*	6.00E-25	40
15	1404	DNA repair protein RadA	*Geobacter uraniireducens Rf4*	2.00E-142	54
16	300	ArsC arsenate reductase	*uncultured organism*	3.00E-19	54

17	1269	Glu/Leu/Phe/Val dehydrogenase	*Thermobaculum terrenum ATCC BAA-798*	3.00E-127	57
18	801	TonB family protein	*Candidatus Solibacter usitatus Ellin6076*	2.00E-11	45
19	705	hypothetical protein sce2190	*Sorangium cellulosum*	4.00E-38	42
20	903	hypothetical protein Haur_1275	*Herpetosiphon aurantiacus ATCC 23779*	2.00E-39	46
21	1158	alanine racemase	*Geobacter lovleyi SZ*	9.00E-84	45
22	2751	TPR repeat-containing protein	*bacterium Ellin514*	1.00E-60	31
23	1404	replicative DNA helicase	*Clostridium papyrosolvens DSM 2782*	7.00E-121	50
24	534	ribosomal protein L9	*Acidobacterium sp. MP5ACTX8*	2.00E-36	52
25	318	30S ribosomal protein S18	*Pelotomaculum thermopropionicum SI*	5.00E-20	67
26	414	30S ribosomal protein S6	*Terriglobus saanensis SP1PR4*	8.00E-16	46
27	612	peptidyl-tRNA hydrolase	*Candidatus Koribacter versatilis Ellin345*	9.00E-47	49
28	666	ribosomal 5S rRNA E-loop binding protein Ctc/L25/TL5	*Acidobacterium sp. MP5ACTX9*	2.00E-29	42
29	861	ribose-phosphate pyrophosphokinase	*Acidobacterium sp. MP5ACTX9*	3.00E-101	63
30	225	hypothetical protein SBO_0532	*Shigella boydii Sb227*	3.00E-04	75
31	927	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase	*Candidatus Solibacter usitatus Ellin6076*	5.00E-33	42
32	534	copper amine oxidase domain protein	*Clostridium thermocellum DSM 2360*	3.00E-18	40
33	483	no significant hit			
34	1533	outer membrane assembly lipoprotein YfiO	*Terriglobus saanensis SP1PR4*	5.00E-37	34

35	678	ribulose-phosphate 3-epimerase	*Sphaerobacter thermophilus DSM 20745*	1.00E-57	59
36	651	pasta domain protein	*Prevotella bryantii B14*	1.00E-05	32
37	1338	sun protein	*Syntrophobacter fumaroxidans MPOB*	9.00E-68	39
38	915	methionyl-tRNA formyltransferase	*Calditerrivibrio nitroreducens DSM 19672*	2.00E-67	50
39	1593	1-pyrroline-5-carboxylate dehydrogenase	*Geobacillus kaustophilus HTA426*	0	61
40	516	peptide deformylase	*Acidobacterium capsulatum ATCC 51196*	2.00E-48	54
41	1023	chorismate synthase	*Acidobacterium sp. MP5ACTX8*	2.00E-99	54
42	390	hypothetical protein Psta_1778	*Pirellula staleyi DSM 6068*	2.00E-34	61
43	1434	D-lactate dehydrogenase	*Chloroherpeton thalassium ATCC 35110*	2.00E-111	46
44	666	hypothetical protein RSc3377	*Ralstonia solanacearum GMI1000*	4.00E-19	34
45	180	no significant hit			
46	1986	excinuclease ABC subunit B	*Candidatus Koribacter versatilis Ellin345*	0	69
47	609	PBS lyase HEAT-like repeat	*Beggiatoa sp. PS*	5.00E-05	29
48	1419	proteophosphoglycan 5	*Leishmania major strain Friedlin*	1.00E-07	29
49	897	hypothetical protein GSU2641	*Geobacter sulfurreducens PCA*	2.00E-06	26
50	1125	TPR repeat-containing protein	*Geobacter sp. M18*	8.00E-10	40
51	684	response regulator DrrA	*Salinibacter ruber DSM 13855*	6.00E-53	49
52	474	IG hypothetical 18565	*Planctomyces maris DSM 8797*	7.00E-20	33

53	897	phosphoribosylaminoimidazole-succinocarboxamide synthase	*bacterium Ellin514*	4.00E-89	58
54	258	no significant hit			
55	336	chaperonin, 10 kDa	*Acidobacterium capsulatum ATCC 51196*	7.00E-36	76
56	1653	chaperonin GroEL	*Candidatus Solibacter usitatus Ellin6076*	0	79
57	240	no significant hit			
58	501	putative membrane-associated metalloprotease	*Clostridium ljungdahlii DSM 13528*	1.00E-07	37
59	222	no significant hit			
60	981	sigma-54 dependent response regulator	*Candidatus Nitrospira defluvii*	6.00E-77	50
61	3714	putative Pentapeptide repeats (8 copies)	*uncultured marine crenarchaeote HF4000_ANIW137 N18]*	3.00E-49	33
62	1422	protein of unknown function DUF1501	*Acidobacterium sp. MP5ACTX9*	3.00E-89	41
63	2238	conserved hypothetical protein	*Verrucomicrobiae bacterium DG1235*	5.00E-83	36
64	1140	hypothetical protein GM18_3297	*Geobacter sp. M18*	1.00E-55	34
65	1161	hypothetical protein Acid_1655	*Candidatus Solibacter usitatus Ellin6076*	1.00E-77	43
66	1254	RND family efflux transporter MFP subunit	*Candidatus Solibacter usitatus Ellin6076*	3.00E-75	45
67	711	ABC transporter-like protein	*Candidatus Solibacter usitatus Ellin6076*	3.00E-76	63
68	897	PEGA domain-containing protein	*Terriglobus saanensis SP1PR4*	6.00E-11	28
69	579	dTDP-D-Fucp3N acetylase	*Sulfurihydrogenibium yellowstonense SS-5*	7.00E-49	52
70	1440	glycosyltransferase	*Streptomyces pristinaespiralis ATCC 25486*	1.00E-96	43

71	2226	dolichyl-phosphate-mannose-protein mannosyltransferase	*delta proteobacterium NaphS2*	3.00E-28	23
72	2370	no significant hit			
73	1437	hypothetical protein Acid345_3128	*Candidatus Koribacter versatilis Ellin345*	3.00E-55	34
74	1128	hypothetical protein CfE428DRAFT_2436	*Chthoniobacter flavus Ellin428*	3.00E-19	30
75	1458	hypothetical protein ANT_26880	*Anaerolinea thermophila UNI-1*	1.00E-49	33
76	612	no significant hit			
77	2226	hypothetical protein Cyan7425_2988	*Cyanothece sp. PCC 7425*	2.00E-127	42
78	174	no significant hit			
79	189	no significant hit			
80	165	no significant hit			
81	528	peptidase A24A prepilin type IV	*Thermincola sp. JR*	2.00E-11	28
82	921	Flp pilus assembly CpaB	*Polaromonas sp. JS666*	6.00E-49	45
83	1542	type II and III secretion system protein	*Candidatus Solibacter usitatus Ellin6076*	9.00E-66	38
84	1317	hypothetical protein VFA_000095	*Vibrio furnissii CIP 102972*	3.00E-14	28
85	522	TadE family protein	*Ralstonia pickettii 12J*	1.00E-08	35
86	519	TadE family protein	*Nitrosococcus halophilus Nc4*	5.00E-16	34
87	1155	response regulator receiver protein	*Thermincola sp. JR*	1.00E-41	32
88	1335	type II secretion system protein E	*Geobacter sp. M18*	4.00E-160	65
89	204	no significant hit			
90	438	CmR			

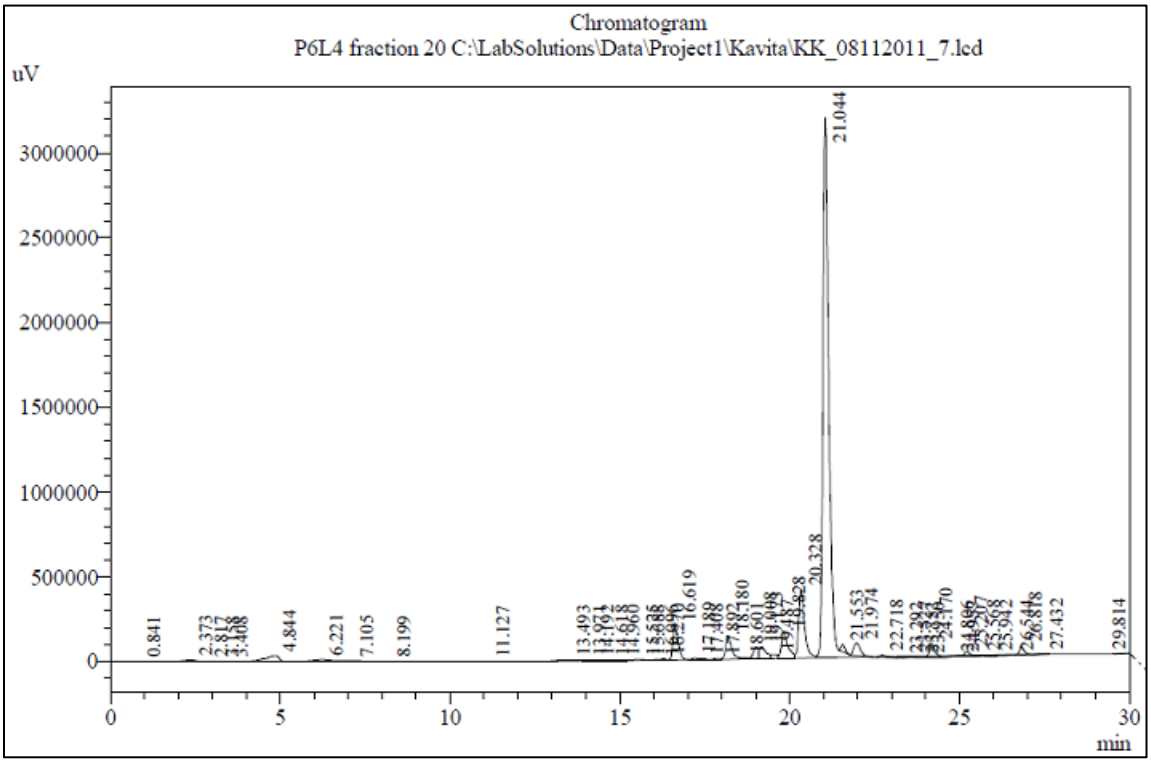
Table 4.2. Summary of anti-MRSA BAC clone annotations.

Clone ID#	Sequenced Insert Size	Fold Coverage (454 FLX)	PROPERTIES
P18N22	82.9 kb	236x	Many of the predicted gene products are transport proteins, or of unknown or hypothetical function.
P20G1	132.9 kb	285x	Contains prophage, gene products predicted to be associated with amino acid/protein synthesis and degradation, penicillin binding protein and penicillin amidase.
P22C4	112.6 kb	108x	Aldo/keto reductase (54% identity) and an isoprenoid biosynthesis gene (35% identity).
P22E10	135.1 kb	74x	Numerous predicted gene products involved in biosynthesis, however no obvious antimicrobial synthesis pathways.
P23K15	132.3 kb	79x	Polyketide enterocin synthesis - 31% identity. Entire prophage from *Acidobacteria* genome. 43% of predicted ORFs with no significant hit.

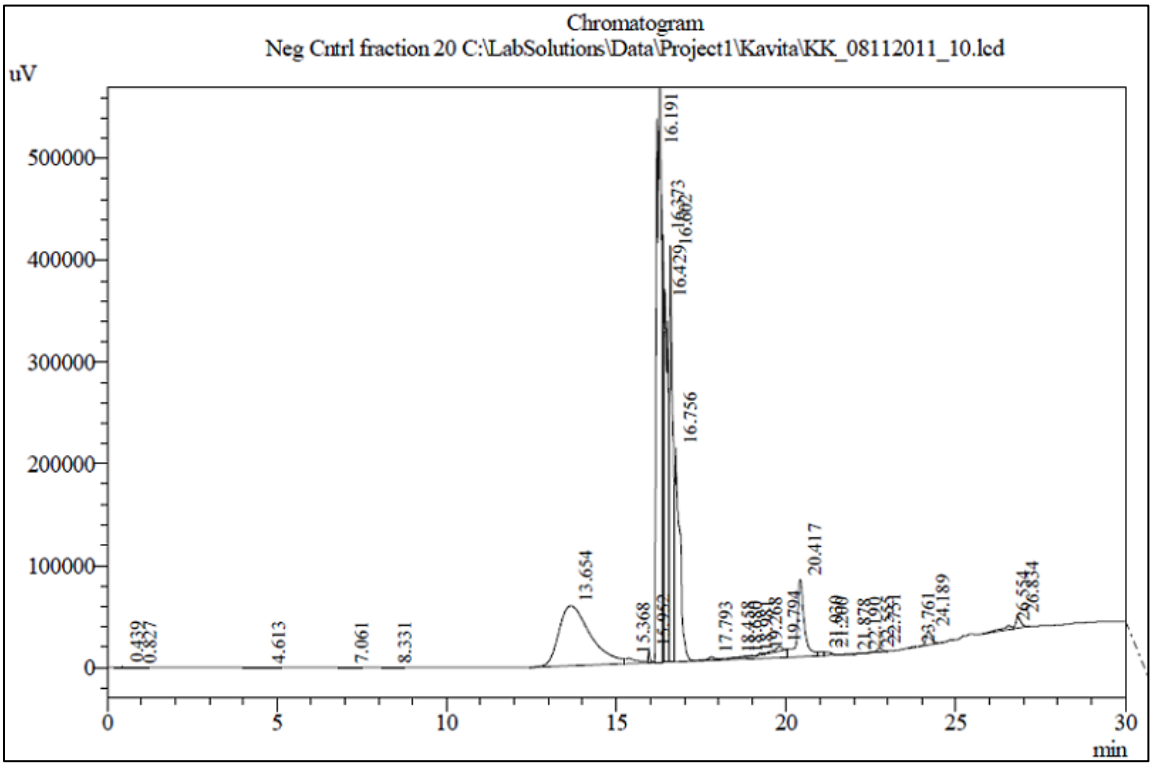
P27K16	108.2 kb	46x	23% of predicted ORFs with no significant hit. Several antibiotic resistance determinants. No clear phylogenetic origin.
P27M10	93.4 kb	161x	Numerous genes for biosynthesis & chemical modification: mandelate racemase (precursor compound), polyketide cyclase, phenazine biosynthesis (known antibiotic). Phylum *Acidobacteria*.
P28H1	105.1 kb	20x	Several unique PKS genes, low % identity (<0.001). No clear phylogenetic origin.
P28I7	136.7 kb	223x	Radical SAM domain protein, beta-ketoacyl synthase, 6-methysalicylic acid synthase. Phylum *Acidobacteria*.
P6L4	119.1 kb	280x	Polyketide cyclase/dehydrase, a hypothetical PKS. No clear phylogentic origin.
P28L21	104.5 kb	189x	33% of predicted ORFs with no significant hit. No clear phylogenetic origin.
P31G24	96.6 kb	129x	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase. Phylum *Acidobacteria*.

Figure 4.4 A. HPLC analysis of the concentrated ethyl acetate extracts from cell free supernatants. Chromatograms for P6L4 (i) and Negative control (ii) and Cm reference standard at 1.25 mg/ml (iii) are depicted below.

4.4 A (i)



4.4 A (ii)



4.4 A (iii)

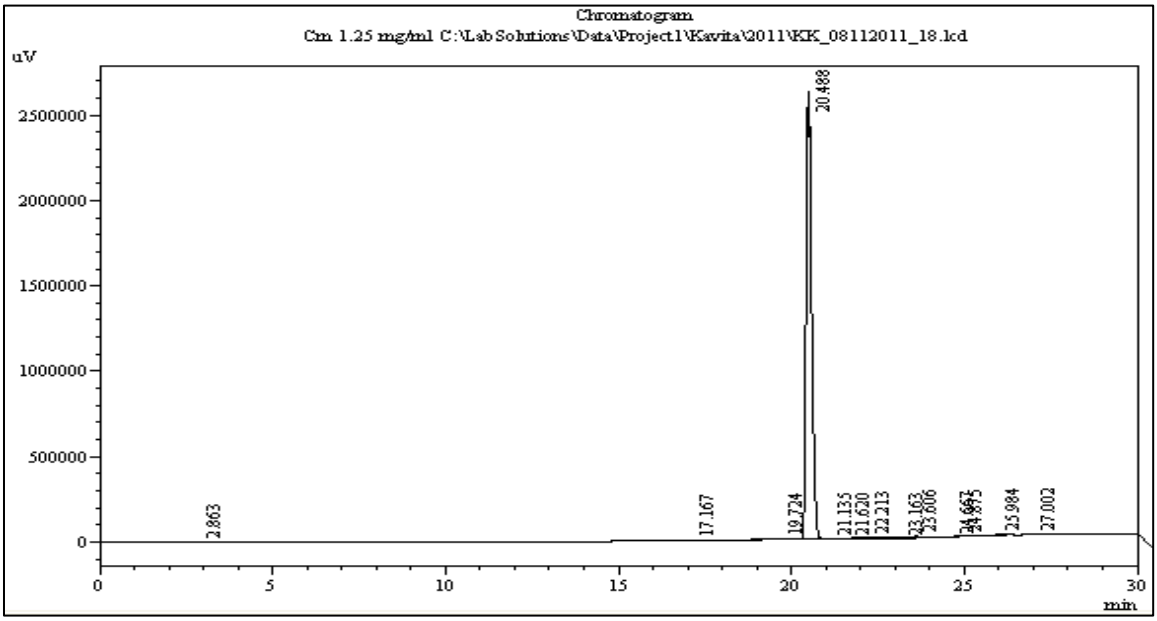
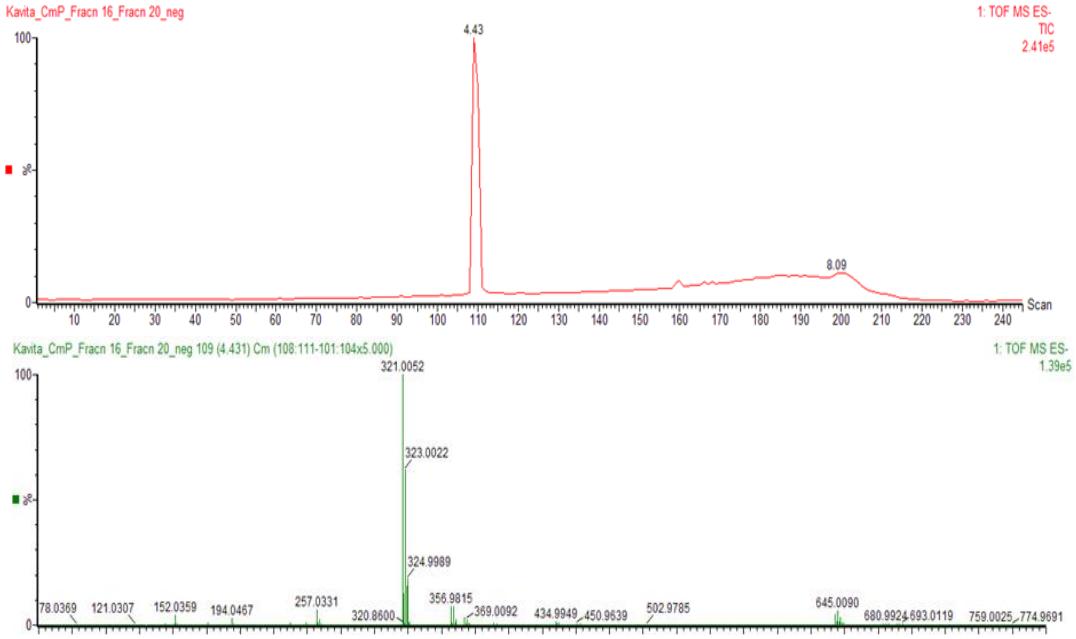


Figure 4.4 B. LC-MS analysis of active HPLC fractions.

Comparison of the Cm control (i) & clone P6L4 (ii) is depicted here.

4.4 B (i)



4.4 B (ii)

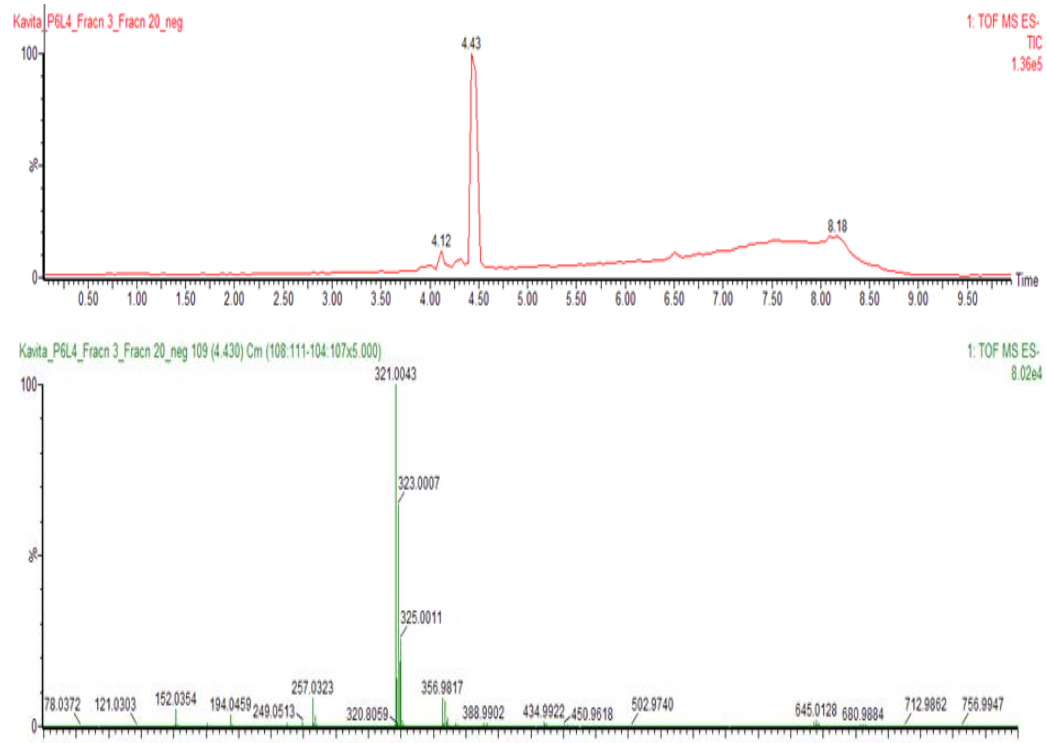


Figure 4.5. **Comparison of negative control & clone P6L4 culture extract with BCAM as a substrate.**

Aliquots withdrawn every 12 hours were processed for extraction with Ethyl acetate and analyzed by TLC.

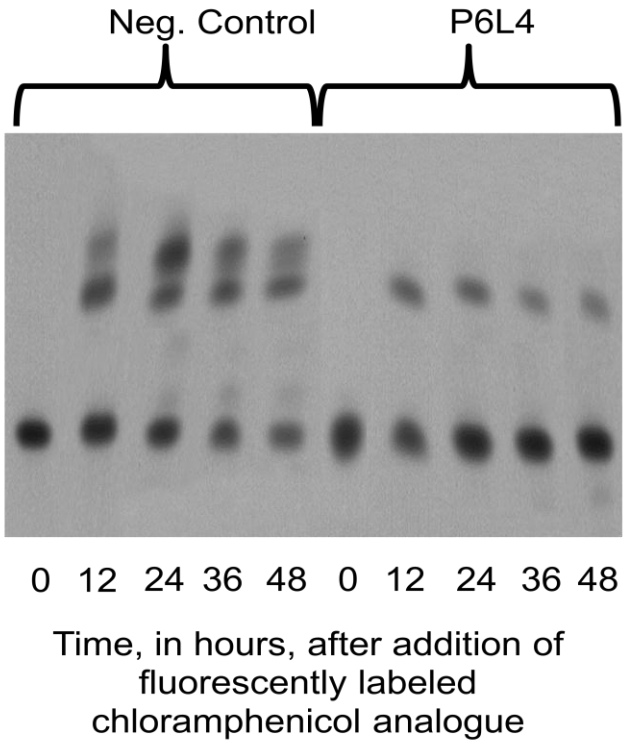


Figure 4.6. Amplification, cloning and induced expression of esterase genes from clone P6L4 using the Expresso Rhamnose SUMO system.

Agarose gel electrophoresis of PCR amplified DNA template from clone P6L4 (A), from subclones pRham-e and pRham-ce (B) and SDS-PAGE image of proteins E and Ce from uninduced (Un) and induced (In) cultures (C).

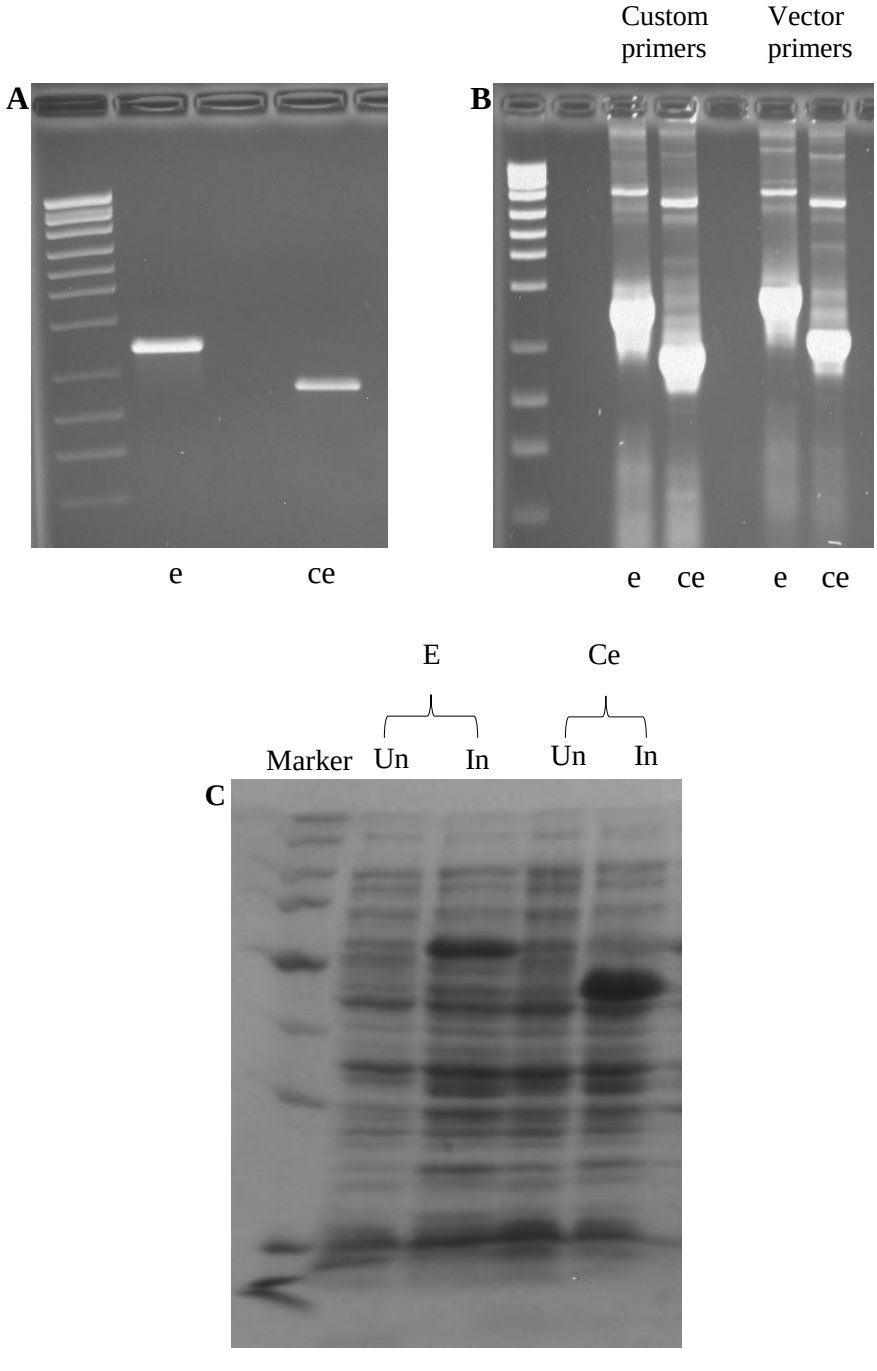


Figure 4.7. Comparison of anti-MRSA activity of the P6L4 subclones pRham-e and pRham-ce in the presence and absence of rhamnose-induced expression.

The graph represents the % growth inhibition (Y axis) of MRSA strain EAMC 30 by the subclones (X axis) relative to the empty vector negative control, considered to have no inhibitory effect and calculated by measuring the fluorescence of reduced resazurin.

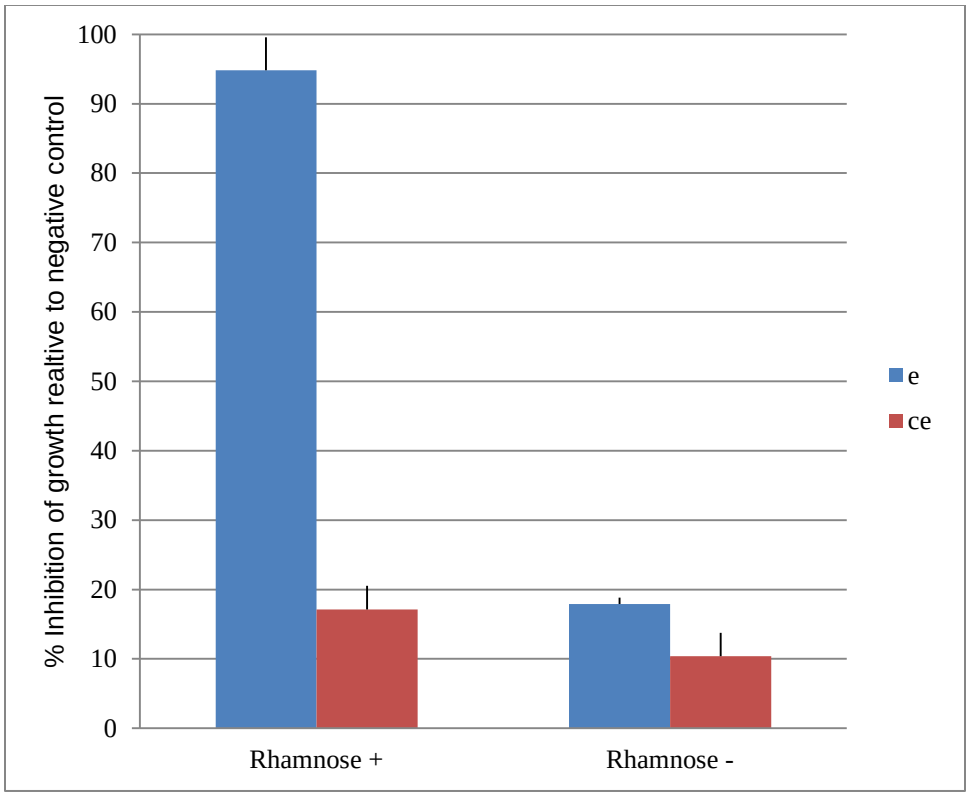
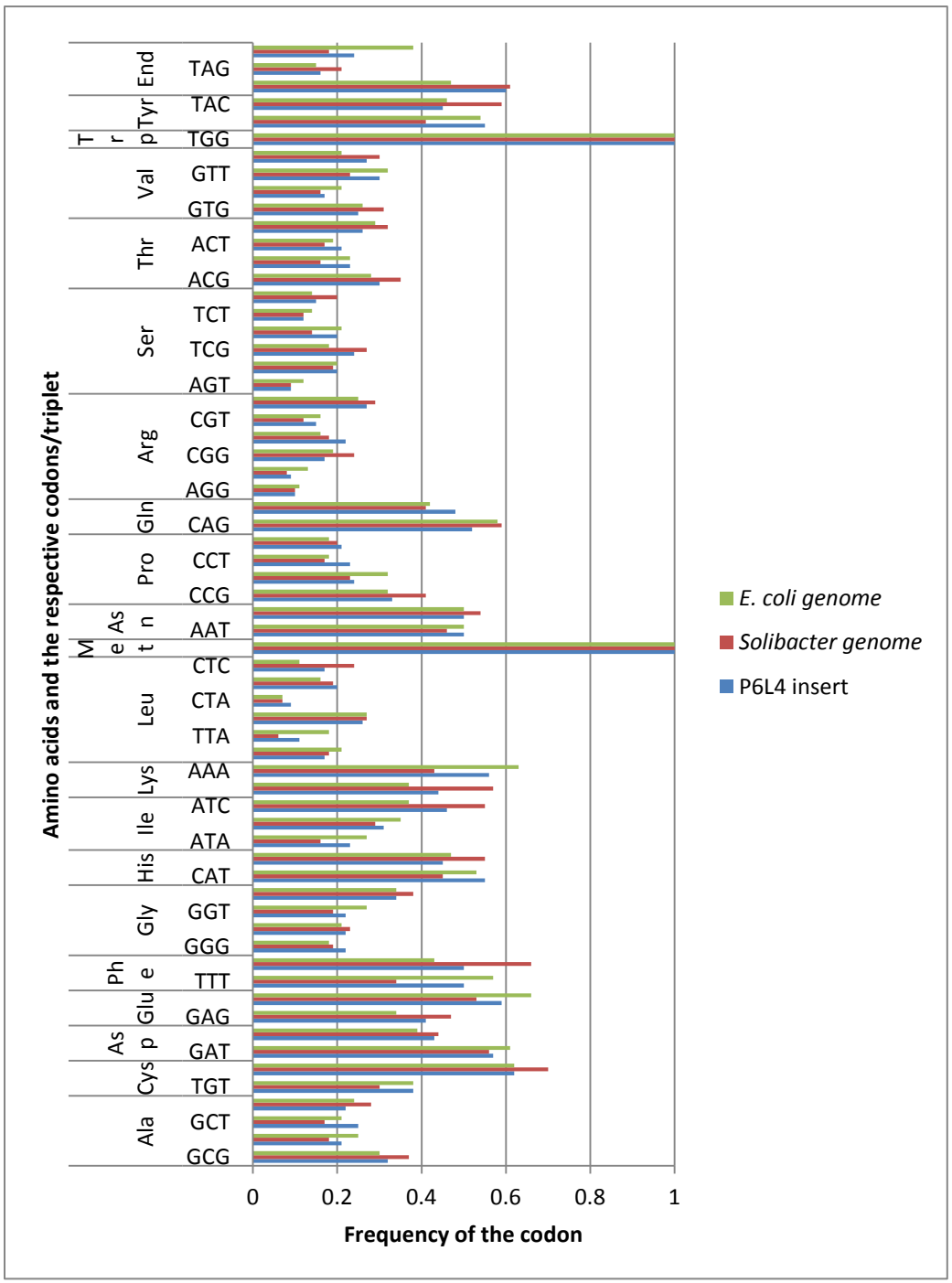


Figure 4.8. Comparison of codon usage.

Comparative codon usage frequencies of the P6L4 complete insert sequence and whole genome sequence of *Candidatus Solibacter usitatus Ellin6076* and *Escherichia coli strain K12 substrain DH10B.*



COMPREHENSIVE BIBLIOGRAPHY

Aakvik, T., K.F. Degnes, R. Dahlsrud, F. Schmidt, R. Dam, L. Yu, U. Volker, T.E. Ellingsen, S. Valla. 2009. A plasmid RK2-based broad-host-range cloning vector useful for transfer of metagenomic libraries to a variety of bacterial species. FEMS Microbiology Letters 296, 149-58.

Altshuler, M.L. 2006. PCR troubleshooting, the essential guide. 1st ed. Caister Academic Press, Wymondham, Norfolk (UK).

Bakken, L.R., V. Lindahl. 1995. Recovery of bacterial cells from soil, p. 13-27, *In* J. T. Trevors and J. D. Van Elsas, eds. Nucleic Acids in the Environment. Springer-Verlag, Berlin.

Beja, O., M.T. Suzuki, E.V. Koonin, L. Aravind, A. Hadd, L.P. Nguyen, R. Villacorta, M. Amjadi, C. Garrigues, S.B. Jovanovich, R.A. Feldman, E.F. DeLong. 2000a. Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. Environmental Microbiology 2, 516-29.

Beja, O., L. Aravind, E.V. Koonin, M.T. Suzuki, A. Hadd, L.P. Nguyen, S.B. Jovanovich, C.M. Gates, R.A. Feldman, J.L. Spudich, E.N. Spudich, E.F. DeLong. 2000b. Bacterial rhodopsin, evidence for a new type of phototrophy in the sea. Science 289, 1902-6.

Berry, A.E., C. Chiocchini, T. Selby, M. Sosio, E.M. Wellington. 2003. Isolation of high molecular weight DNA from soil for cloning into BAC vectors. FEMS Microbiology Letters 223, 15-20.

Bertrand, H., F. Poly, V.T. Van, N. Lombard, R. Nalin, T.M. Vogel, P. Simonet. 2005. High molecular weight DNA recovery from soils prerequisite for biotechnological metagenomic library construction. Journal of Microbiological Methods 62, 1-11.

Blumberg, P.M., J.L. Strominger. 1974. Interaction of penicillin with the bacterial cell: penicillin-binding proteins and penicillin-sensitive enzymes. Bacteriol. Rev. 38:291-335.

Brady, S.F., C.J. Chao, J. Clardy. 2004. Long-chain N-acyltyrosine synthases from environmental DNA. Applied and Environmental Microbiology 70, 6865-70.

Breitbart, M., I. Hewson, B. Felts, J.M. Mahaffy, J. Nulton, P. Salamon, F. Rohwer. 2003. Metagenomic analyses of an uncultured viral community from human feces. Journal of Bacteriology 185, 6220-3.

Breitbart, M., P. Salamon, B. Andresen, J.M. Mahaffy, A.M. Segall, D. Mead, F. Azam, F. Rohwer. 2002. Genomic analysis of uncultured marine viral communities. Proceedings of the National Academy of Sciences of the United States of America 99, 14250-5.

Brennan, Y., W.N. Callen, L. Christoffersen, P. Dupree, F. Goubet, S. Healey, M. Hernandez, M. Keller, K. Li, N. Palackal, A. Sittenfeld, G. Tamayo, S. Wells, G.P. Hazlewood, E.J. Mathur, J.M. Short, D.E. Robertson, B.A. Steer. 2004. Unusual microbial xylanases from insect guts. Applied and Environmental Microbiology 70, 3609-17.

Brulc, J.M., D.A. Antonopoulos, M.E. Miller, M.K. Wilson, A.C. Yannarell, E.A. Dinsdale, R.E. Edwards, E.D. Frank, J.B. Emerson, P. Wacklin, P.M. Coutinho, B. Henrissat, K.E. Nelson, B.A. White. 2009. Gene-centric metagenomics of the

fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. Proceedings of the National Academy of Sciences of the United States of America 106, 1948-53.

Burgess, J.G., E.M. Jordan, M. Bregu, A. Mearns-Spragg, K.G. Boyd. 1999. Microbial antagonism, a neglected avenue of natural products research. Journal of Biotechnology 70, 27-32.

Bycroft B.W., R.E. Shute. 1985. The Molecular Basis for the Mode of Action of Beta-Lactam Antibiotics and Mechanisms of Resistance Pharmaceutical Research Volume 2, Number 1, 3-14, DOI: 10.1023/A:1016305704057

Cabanas, M.J., D. Vazquez, J. Modolell. 1978. Inhibition of ribosomal translocation by aminoglycoside antibiotics. Biochem. Biophys. Res. Commun. 83, 991–997.

Chandler, D.P., Brockman, F.J., Fredrickson, J.K. 1997. Use of 16S rDNA clone libraries to study changes in a microbial community resulting from *ex situ* perturbation of a subsurface sediment. FEMS Microbiology Reviews 20, 217–230.

Chen, I.C., W.D. Lin, S.K. Hsu, V. Thiruvengadam, W.H. Hsu. 2009. Isolation and characterization of a novel lysine racemase from a soil metagenomic library. Applied and Environmental Microbiology 75, 5161-6.

Chen, Y., M.G. Dumont, J.D. Neufeld, L. Bodrossy, N. Stralis-Pavese, N.P. McNamara, N. Ostle, M.J. Briones, J.C. Murrell. 2008. Revealing the uncultivated majority, combining DNA stable-isotope probing, multiple displacement amplification and metagenomic analyses of uncultivated *Methylocystis* in acidic peatlands. Environmental Microbiology 10, 2609-22.

Chu, X., H. He, C. Guo, B. Sun. 2008. Identification of two novel esterases from a marine metagenomic library derived from South China Sea. Applied Microbiology and Biotechnology 80, 615-25.

Chung, E.J., H.K. Lim, J.C. Kim, G.J. Choi, E.J. Park, M.H. Lee, Y.R. Chung, S.W. Lee. 2008. Forest soil metagenome gene cluster involved in antifungal activity expression in *Escherichia coli*. Applied and Environmental Microbiology 74, 723-30.

Cieslinski, H., A. Dlugolecka, J. Kur, M. Turkiewicz. 2009. An MTA phosphorylase gene discovered in the metagenomic library derived from Antarctic top soil during screening for lipolytic active clones confers strong pink fluorescence in the presence of rhodamine B. FEMS Microbiology Letters 299, 232-40.

Clardy, J., C. Walsh. 2004. Lessons from natural molecules. Nature 432, 829-37.

Collins, J., B. Hohn. 1978. Cosmids, a type of plasmid gene-cloning vector that is packageable *in vitro* in bacteriophage lambda heads. Proceedings of the National Academy of Sciences of the United States of America 75, 4242-6.

Cosgrove, L., P.L. McGeechan, P.S. Handley, G.D. Robson. 2010. Effect of biostimulation and bioaugmentation on degradation of polyurethane buried in soil. Applied and Environmental Microbiology 76, 810-19.

Courtois, S., C.M. Cappellano, M. Ball, F.X. Francou, P. Normand, G. Helynck, A. Martinez, S.J. Kolvek, J. Hopke, M.S. Osburne, P.R. August, R. Nalin, M. Guerineau, P. Jeannin, P. Simonet, J.L. Pernodet. 2003. Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. Applied and Environmental Microbiology 69, 49-55.

Cowan, D., Q. Meyer, W. Stafford, S. Muyanga, R. Cameron, P. Wittwer. 2005. Metagenomic gene discovery, past, present and future. Trends in Biotechnology 23, 321-9.

Craig, J.W., F.Y. Chang, S.F. Brady. 2009. Natural products from environmental DNA hosted in *Ralstonia metallidurans*. ACS Chemical Biology 4, 23-8.

Craig, J.W., F.Y. Chang, J.H. Kim, S.C. Obiajulu, S.F. Brady. 2010. Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria. Applied and Environmental Microbiology 76, 1633-41.

Crumplin, G.C., J.T. Smith. 1976. Nalidixic acid and bacterial chromosome replication. Nature 260, 643–5.

Curtis, T.P., W.T. Sloan. 2005. Microbiology. Exploring microbial diversity--a vast below. Science 309, 1331-3.

Davies, J., L. Gorini, B.D. Davies. 1965. Misreading of RNA codewords induced by aminoglycoside antibiotics. Mol. Pharmacol. 1, 93–106.

Davies, J., B.D. Davis. 1968. Misreading of ribonucleic acid code words induced by aminoglycoside antibiotics. The effect of drug concentration. J Biol Chem. Jun 25;243(12):3312-6.

de Lorenzo, V. 2005. Problems with metagenomic screening. Nature Biotechnology 23, author reply 1045–1046.

DeSantis, T.Z., E.L. Brodie, J.P. Moberg, I.X. Zubieta, Y.M. Piceno, G.L. Andersen. 2007. High-density universal 16S rRNA microarray analysis reveals broader

diversity than typical clone library when sampling the environment. Microbial Ecology 53, 371-83.

Dinsdale, E.A., R.A. Edwards, D. Hall, F. Angly, M. Breitbart, J.M. Brulc, M. Furlan, C. Desnues, M. Haynes, L.L. Li, L. McDaniel, M.A. Moran, K.E. Nelson, C. Nilsson, R. Olson, J. Paul, B.R. Brito, Y.J. Ruan, B.K. Swan, R. Stevens, D.L. Valentine, R.V. Thurber, L. Wegley, B.A. White, F. Rohwer. 2008. Functional metagenomic profiling of nine biomes (vol 452, pg 629, 2008). Nature 455, 830-830.

Dumont, M.G., S.M. Radajewski, C.B. Miguez, I.R. McDonald,.C. Murrell. 2006. Identification of a complete methane monooxygenase operon from soil by combining stable isotope probing and metagenomic analysis. Environmental Microbiology 8, 1240-50.

Elend, C., C. Schmeisser, C. Leggewie, P. Babiak, J.D. Carballeira, H.L. Steele, J.L. Reymond, K.E. Jaeger, W.R. Streit. 2006. Isolation and biochemical characterization of two novel metagenome-derived esterases. Applied and Environmantal Microbiology 72, 3637-45.

Entcheva, P., W. Liebl, A. Johann, T. Hartsch, W.R. Streit. 2001. Direct cloning from enrichment cultures, a reliable strategy for isolation of complete operons and genes from microbial consortia. Applied and Environmental Microbiology 67, 89-99.

Faegri, A., V. Torsvik, J. Goksoyr. 1977. Bacterial and fungal activities in soil, separation of bacteria and fungi by a rapid fractionated centrifugation technique. Soil Biology and Biochemistry 9, 105-112.

Feinstein, L.M., W.J. Sul, C.B. Blackwood. 2009. Assessment of bias associated with incomplete extraction of microbial DNA from soil. Applied and Environmental Microbiology 75, 5428-33.

Ferrer, M., O.V. Golyshina, T.N. Chernikova, A.N. Khachane, V.A. Martins Dos Santos, M.M. Yakimov, K.N. Timmis, P.N. Golyshin. 2005. Microbial enzymes mined from the Urania deep-sea hypersaline anoxic basin. Chemistry and Biology 12, 895-904.

Fierer, N., M. Breitbart, J. Nulton, P. Salamon, C. Lozupone, R. Jones, M. Robeson, R. A. Edwards, B. Felts, S. Rayhawk, R. Knight, F. Rohwer, and R. B. Jackson. 2007. Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. Applied and Environmental Microbiology 73,7059-7066.

Frostegard, A., S. Courtois, V. Ramisse, S. Clerc, D. Bernillon, F. Le Gall, P. Jeannin, X. Nesme, P. Simonet. 1999. Quantification of bias related to the extraction of DNA directly from soils. Applied and Environmental Microbiology 65, 5409-20.

Gabor, E.M., E.J. Vries, D.B. Janssen. 2003. Efficient recovery of environmental DNA for expression cloning by indirect extraction methods. FEMS Microbiology Ecology 44, 153-63.

Gale, E.F., J.P. Folkes. 1953. The assimilation of amino-acids by bacteria. 15. Actions of antibiotics on nucleic acid and protein synthesis in *Staphylococcus aureus.* Biochem. J. (London), 53, 493-498.

Garbeva, P., W. de Boer. 2009. Inter-specific interactions between carbon-limited soil bacteria affect behavior and gene expression. Microbial Ecology 58, 36-46.

Gellert, M., K. Mizuuchi, M.H. O'Dea, H.A. Nash. 1976. DNA gyrase: an enzyme that introduces superhelical turns into DNA. Proceedings of the National Academy of Sciences USA 73, 3872–6.

George, I.F., M.R. Liles, M. Hartmann, W. Ludwig, R.M. Goodman, S.N. Agathos. 2009. Changes in soil *Acidobacteria* communities after 2,4,6-trinitrotoluene contamination. FEMS Microbiology Letters 296, 159-66.

Gillespie, D.E., S.F. Brady, A.D. Bettermann, N.P. Cianciotto, M.R. Liles, M.R. Rondon, J. Clardy, R.M. Goodman, J. Handelsman. 2002. Isolation of antibiotics turbomycin a and B from a metagenomic library of soil microbial DNA. Applied and Environmental Microbiology 68, 4301-6.

Ginolhac, A., C. Jarrin, B. Gillet, P. Robe, P. Pujic, K. Tuphile, H. Bertrand, T.M. Vogel, G. Perriere, P. Simonet, R. Nalin. 2004. Phylogenetic analysis of polyketide synthase I domains from soil metagenomic libraries allows selection of promising clones. Applied and Environmental Microbiology 70, 5522-7.

Hain, T., S. Otten, U. von Both, S.S. Chatterjee, U. Technow, A. Billion, R. Ghai, W. Mohamed, E. Domann, T. Chakraborty. 2008. Novel bacterial artificial chromosome vector pUvBBAC for use in studies of the functional genomics of *Listeria* spp. Applied and Environmental Microbiology 74, 1892-901.

Handelsman, J. 2004. Metagenomics, application of genomics to uncultured microorganisms. Microbiology and Molecular Biology Reviews 68, 669-85.

Handelsman, J., M.R. Rondon, S.F. Brady, J. Clardy, R.M. Goodman. 1998. Molecular biological access to the chemistry of unknown soil microbes, a new frontier for natural products. Chemistry and Biology 5, R245-9.

Hao, Y., Winans, S.C., Glick, B.R., Charles, T.C. 2010. Identification and characterization of new LuxR/LuxI-type quorum sensing systems from metagenomic libraries. Environmental Microbiology 12,105-117.

Harry, M., B. Gambier, Y. Bourezgui, E. Garnier-Sillam. 1999. Evaluation of purification procedures for DNA extracted from organic rich samples, interference with humic substances. Analysis 27, 439-442.

Hassink, J., L. A. Bouman, K. B. Zwart, J. Bloem, and L. Brussaard. 1993. Relationships between soil texture, physical protection of organic matter, soil biota, and C and N mineralization in grassland soils. Geoderma 57, 105-128.

Healy, F.G., R.M. Ray, H.C. Aldrich, A.C. Wilkie, L.O. Ingram, K.T. Shanmugam. 1995. Direct isolation of functional genes encoding cellulases from the microbial consortia in a thermophilic, anaerobic digester maintained on lignocellulose. Applied Microbiology Biotechnology 43, 667-74.

Heath, C., X.P. Hu, S.C. Cary, D. Cowan. 2009. Identification of a novel alkaliphilic esterase active at low temperatures by screening a metagenomic library from antarctic desert soil. Applied and Environmental Microbiology 75, 4657-9.

Henne, A., R. Daniel, R.A. Schmitz, G. Gottschalk. 1999. Construction of environmental DNA libraries in Escherichia coli and screening for the presence of genes conferring utilization of 4-hydroxybutyrate. Applied and Environmental Microbiology 65, 3901-7.

Henne, A., R. A. Schmitz, M. Bomeke, G. Gottschalk, and R. Daniel. 2000. Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity of *Escherichia coli*. Appl. Environ. Microbiol. 66:3113-3116.

Holben, W.E., J.K. Jansson, B.K. Chelm, J.M. Tiedje. 1988. DNA Probe Method for the Detection of Specific Microorganisms in the Soil Bacterial Community. Applied and Environmental Microbiology 54, 703-711.

Hopkins, D.W., S.J. Macnaughton, A.G. O'Donnell. 1991. A dispersion and differential centrifugation technique for representatively sampling microorganisms from soil. Soil Biology and Biochemistry 23, 217-225.

Hugenholtz, P., G.W. Tyson. 2008. Microbiology, metagenomics. Nature 455, 481-3.

Hugenholtz, P., B.M. Goebel, N.R. Pace. 1998. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. Journal of Bacteriology 180, 4765-74.

Huse, S.M., J.A. Huber, H.G. Morrison, M.L. Sogin, D.M. Welch. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biology 8, R143.

Huson, D.H., A.F. Auch, J. Qi, S.C. Schuster. 2007. MEGAN analysis of metagenomic data. Genome Research 17, 377-86.

Izaki, K., M. Matsuhashi, J.L. Strominger. 1968. Biosynthesis of peptioglycan in cell walls. J. Biol. Chem. 243:3180-3192.

Jacobsen, C.S., O.F. Rasmussen. 1992. Development and application of a new method to extract bacterial DNA from soil based on separation of bacteria from soil with cation-exchange resin. Applied and Environmental Microbiology 58, 2458-2462.

Jeon, J.H., J.T. Kim, S.G. Kang, J.H. Lee, S.J. Kim. 2009. Characterization and its potential application of two esterases derived from the arctic sediment metagenome. Marine Biotechnology (New York, N.Y.) 11, 307-16.

Jiang, C., G. Ma, S. Li, T. Hu, Z. Che, P. Shen, B. Yan, B. Wu. 2009. Characterization of a novel beta-glucosidase-like activity from a soil metagenome. Journal of Microbiology 47, 542-8.

Jordan, D. C., P. E. Reynolds. 1967. Vancomycin, p.102-116. In D. Gottlieb and P. D. Shaw (ed.), Antibiotics: mechanism of action, vol. 1. Springer-Verlag, Heidelberg.

Kalyuzhnaya, M.G., A. Lapidus, N. Ivanova, A.C. Copeland, A.C. McHardy, E. Szeto, A. Salamov, I.V. Grigoriev, D. Suciu, S.R. Levine, V.M. Markowitz, I. Rigoutsos, S.G. Tringe, D.C. Bruce, P.M. Richardson, M.E. Lidstrom, L. Chistoserdova. 2008. High-resolution metagenomics targets specific functional types in complex microbial communities. Nature Biotechnology 26, 1029-34.

Kim, B.S., S.Y. Kim, J. Park, W. Park, K.Y. Hwang, Y.J. Yoon, W.K. Oh, B.Y. Kim, J.S. Ahn. 2007. Sequence-based screening for self-sufficient P450 monooxygenase from a metagenome library. Journal of Applied Microbiology 102, 1392-400.

Kim, U.J., H. Shizuya, P.J. de Jong, B. Birren, M.I. Simon. 1992. Stable propagation of cosmid sized human DNA inserts in an F factor based vector. Nucleic Acids Research 20, 1083-5.

King, R.W., J.D. Bauer, S.F. Brady. 2009. An environmental DNA-derived type II polyketide biosynthetic pathway encodes the biosynthesis of the pentacyclic polyketide erdacin. Angewandte Chemie (International ed. in English) 48, 6257-61.

Knaebel, D.B., R.L. Crawford. 1995. Extraction and purification of microbial DNA from petroleum-contaminated soils and detection of low numbers of toluene, octane

and pesticide degraders by multiplex polymerase chain reaction and Southern analysis. Molecular Ecology 4, 579-91.

Knietsch, A., S. Bowien, G. Whited, G. Gottschalk, R. Daniel. 2003. Identification and characterization of coenzyme B12-dependent glycerol dehydratase- and diol dehydratase-encoding genes from metagenomic DNA libraries derived from enrichment cultures. Applied and Environmental Microbiology 69, 3048-60.

Kreader, C.A. 1996. Relief of amplification inhibition in PCR with bovine serum albumin or T4 gene 32 protein. Applied and Environmental Microbiology 62, 1102-6.

Kunin, V., A. Copeland, A. Lapidus, K. Mavromatis, P. Hugenholtz. 2008. A bioinformatician's guide to metagenomics. Microbiology and molecular biology reviews 72, 557-78, Table of Contents.

Lammle, K., H. Zipper, M. Breuer, B. Hauer, C. Buta, H. Brunner, S. Rupp. 2007. Identification of novel enzymes with different hydrolytic activities by metagenome expression cloning. Journal of Biotechnology 127, 575-92.

Lee, S.W., K. Won, H.K. Lim, J.C. Kim, G.J. Choi, K.Y. Cho. 2004. Screening for novel lipolytic enzymes from uncultured soil microorganisms. Applied microbiology and biotechnology 65, 720-6.

Li, X., L. Qin. 2005. Metagenomics-based drug discovery and marine microbial diversity. Trends in Biotechnology 23, 539-43.

Liles, M.R., L.L. Williamson, R.M. Goodman, J. Handelsman. 2004. Isolation of high molecular weight genomic DNA from soil bacteria for genomic library construction, p. 839-852, *In* G. A. Kowalchuk, et al., eds. Molecular microbial

ecology manual, 2nd ed. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Liles, M.R., B.F. Manske, S.B. Bintrim, J. Handelsman, R.M. Goodman. 2003. A census of rRNA genes and linked genomic sequences within a soil metagenomic library. Applied and Environmental Microbiology 69, 2684-91.

Liles, M.R., L.L. Williamson, J. Rodbumrer, V. Torsvik, R.M. Goodman, J. Handelsman. 2008. Recovery, purification, and cloning of high-molecular-weight DNA from soil microorganisms. Applied and Environmental Microbiology 74, 3302-5.

Liles, M.R., Turkmen, O., Manske, B.F., Zhang, M., Rouillard, J.M., George, I.F., Balser, T., Billor, N., Goodman, R.M. . 2010. A phylogenetic microarray targeting 16S rRNA genes from the bacterial division *Acidobacteria* reveals a lineage-specific distribution in a soil clay fraction. Soil Biology and Biochemistry 42, 739-747.

Lowy, F.D. 2003. "Antimicrobial resistance: the example of *Staphylococcus aureus*". The Journal of Clinical Investigation 111: 1265–1273.

Macdonald, R.M. 1986. Sampling soil microflora-dispersion of soil by ion-exvhange and extraction of specific microorganisms from suspension by elutriation. Soil Biology and Biochemistry 18, 399-406.

Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V. Gomes, B.C. Godwin, W. He, S. Helgesen, C.H. Ho, G.P. Irzyk, S.C. Jando, M.L. Alenquer, T.P. Jarvie, K.B. Jirage, J.B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant,

B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, J.M. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376-80.

Martinez, A., S.J. Kolvek, J. Hopke, C.L. Yip, M.S. Osburne. 2005. Environmental DNA fragment conferring early and increased sporulation and antibiotic production in *Streptomyces* species. Applied and Environmental Microbiology 71, 1638-41.

Martinez, A., S.J. Kolvek, C.L. Yip, J. Hopke, K.A. Brown, I.A. MacNeil, M.S. Osburne. 2004. Genetically modified bacterial strains and novel bacterial artificial chromosome shuttle vectors for constructing environmental libraries and detecting heterologous natural products in multiple expression hosts. Applied and Environmental Microbiology 70, 2452-63.

Mathur, E., Toledo, G., Green, B.D., Podar, M., Richardson, T.H., Kulwiec, M., Chang, H.C. 2005. A biodiversity-based approach to development of performance enzymes, Applied metagenomics and directed evolution. Industrial Biotechnology 1,283-287.

Metzker, M.L. 2005. Emerging technologies in DNA sequencing. Genome Res 15, 1767-76.

Meyer, F., D. Paarmann , M. D'Souza , R. Olson , E. M. Glass , M. Kubal , T. Paczian , A. Rodriguez , R. Stevens, A. Wilke , J. Wilkening and R. A. Edwards. 2008. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9:386.

Miller, R.V., M.J. Day. 2004. Microbial evolution, gene establishment, survival, and exchange. ASM Press, Washington, D.C.

Mirete, S., C.G. de Figueras, and J.E. Gonzalez-Pastor. 2007. Novel nickel resistance genes from the rhizosphere metagenome of plants adapted to acid mine drainage. Applied and Environmental Microbiology 73, 6001-11.

Misumi, M., T. Nishimura, T. Komai, N. Tanaka. 1978. Interaction of kanamycin and related antibiotics with the large subunit of ribosomes and the inhibition of translocation. Biochem. Biophys. Res. Commun. 84, 358–365.

Moffitt, M.C., B.A. Neilan. 2003. Evolutionary affiliations within the superfamily of ketosynthases reflect complex pathway associations. Journal of Molecular Evolution 56, 446-57.

Muyzer, G., E.C. de Waal, A.G. Uitterlinden. 1993. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. Applied and Environmental Microbiology 59, 695-700.

Nieto, M., H.R. Perkins. 1971. Modifications of the acyl-D-alanyl-D-alanine terminus affecting complexformation with vancomycin. Biochem. J. 123:789-803.

Ogram, A., G.S. Sayler, T.J. Barkay. 1987. DNA extraction and purification from sediments. Journal of Microbiological Methods 7, 57-66.

Ono, A., R. Miyazaki, M. Sota, Y. Ohtsubo, Y. Nagata, M. Tsuda. 2007. Isolation and characterization of naphthalene-catabolic genes and plasmids from oil-contaminated soil by using two cultivation-independent approaches. Applied Microbiology and Biotechnology 74, 501-10.

Osborn, A.M., C.J. Smith. 2005. Molecular microbial ecology Taylor & Francis, New York ; Abingdon [England].

Osoegawa, K., P.Y. Woon, B. Zhao, E. Frengen, M. Tateno, J.J. Catanese, P.J. de Jong. 1998. An improved approach for construction of bacterial artificial chromosome libraries. Genomics 52, 1-8.

Overbeek, R., T. Begley, R.M. Butler, J.V. Choudhuri, H.Y. Chuang, M. Cohoon, V. de Crecy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E.D. Frank, S. Gerdes, E.M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A.C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G.D. Pusch, D.A. Rodionov, C. Ruckert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko, V. Vonstein. 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Research 33, 5691-702.

Palackal, N., C.S. Lyon, S. Zaidi, P. Luginbuhl, P. Dupree, F. Goubet, J.L. Macomber, J.M. Short, G.P. Hazlewood, D.E. Robertson, B.A. Steer. 2007. A multifunctional hybrid glycosyl hydrolase discovered in an uncultured microbial consortium from ruminant gut. Applied Microbiology and Biotechnology 74, 113-24.

Park, S.J., C.H. Kang, J.C. Chae, S.K. Rhee. 2008. Metagenome microarray for screening of fosmid clones containing specific genes. FEMS Microbiol Lett 284, 28-34.

Parsley, L.C., E.J. Consuegra, K.S. Kakirde, A.M. Land, W.F. Harper Jr., M.R. Liles. (2010) Identification of diverse antimicrobial resistance determinants carried on bacterial, plasmid, or viral metagenomes from an activated sludge microbial assemblage. Applied and Environmental Microbiology 76, 3753-3757.

Pel J., D. Broemeling, L. Mai, H.L. Poon, G. Tropini, R.L. Warren, R.A. Holt, A. Marziali. 2009. Nonlinear electrophoretic response yields a unique parameter for separation of biomolecules. Proceedings of the National Academy of Sciences of the United States of America 106 (35), 14796-801.

Perkins, H.R., M. Nieto. 1972. The molecular basis for the antibiotic action of vancomycin, ristocetin and related drugs, p. 363-387. In E. Munoz, F. Garcia-Ferrandiz, and D. Vazquez (ed.), Molecular mechanisms of antibiotic action on protein biosynthesis and membranes, proceedings of a symposium, Granada, June 1971. Elsevier Scientific Publishing Company, Amsterdam.

Perkins, H.R., M. Nieto. 1973. The significance of D-alanyl-D-alanine termini in the biosynthesis of bacterial cell walls and the action of penicillin, vancomycin, and ristocetin. Pure Appl. Chem. 35:371-381.

Perkins, H.R., M. Nieto. 1974. The chemical basis for the action of the vancomycin group of antibiotics. Ann. N.Y. Acad. Sci. 235:348-363.

Pfeifer, B.A., C. Khosla. 2001. Biosynthesis of polyketides in heterologous hosts. Microbiology and Molecular Biology Reviews 65, 106-18.

Pignatelli, M., G. Aparicio, I. Blanquer, V. Hernandez, A. Moya, J. Tamames. 2008. Metagenomics reveals our incomplete knowledge of global diversity. Bioinformatics 24, 2124-5.

Quaiser, A., T. Ochsenreiter, H.P. Klenk, A. Kletzin, A.H. Treusch, G. Meurer, J. Eck, C.W. Sensen, C. Schleper. 2002. First insight into the genome of an uncultivated crenarchaeote from soil. Environmental Microbiology 4, 603-11.

Radajewski, S., P. Ineson, N.R. Parekh, J.C. Murrell. 2000. Stable-isotope probing as a tool in microbial ecology. Nature 403, 646-9.

Riaz, K., C. Elmerich, D. Moreira, A. Raffoux, Y. Dessaux, and D. Faure. 2008. A metagenomic analysis of soil bacteria extends the diversity of quorum-quenching lactonases. Environmental Microbiology 10, 560–570.

Rice,P., I. Longden, and A. Bleasby. 2000 EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16, (6) pp276—277.

Richardson, T.H., X. Tan, G. Frey, W. Callen, M. Cabell, D. Lam, J. Macomber, J.M. Short, D.E. Robertson, C. Miller. 2002. A novel, high performance enzyme for starch liquefaction. Discovery and optimization of a low pH, thermostable alpha-amylase. The Journal of Biological Chemistry 277, 26501-7.

Riesenfeld, C.S., R.M. Goodman, J. Handelsman. 2004a. Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. Environmental Microbiology 6, 981-989.

Riesenfeld, C.S., P.D. Schloss, J. Handelsman. 2004b. Metagenomics, genomic analysis of microbial communities. Annual Review of Genetics 38, 525-52.

Robertson, D.E., J.A. Chaplin, G. DeSantis, M. Podar, M. Madden, E. Chi, T. Richardson, A. Milan, M. Miller, D.P. Weiner, K. Wong, J. McQuaid, B. Farwell, L.A. Preston, X. Tan, M.A. Snead, M. Keller, E. Mathur, P.L. Kretz, M.J. Burk, J.M. Short. 2004. Exploring nitrilase sequence space for enantioselective catalysis. Applied and Environmental Microbiology 70, 2429-36.

Roesch, L.F.W., R.R. Fulthorpe, A. Riva, G. Casella, A.K.M. Hadwin, A.D. Kent, S.H. Daroub, F.A.O. Camargo, W.G. Farmerie, E.W. Triplett. 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. ISME Journal 1, 283-290.

Ronaghi, M. 2001. Pyrosequencing sheds light on DNA sequencing. Genome Research 11, 3-11.

Rondon, M.R., P.R. August, A.D. Bettermann, S.F. Brady, T.H. Grossman, M.R. Liles, K.A. Loiacono, B.A. Lynch, I.A. MacNeil, C. Minor, C.L. Tiong, M. Gilman, M.S. Osburne, J. Clardy, J. Handelsman, R.M. Goodman. 2000. Cloning the soil metagenome, a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Applied and Environmental Microbiology 66, 2541-7.

Roose-Amsaleg, C.L., E. Garnier-Sillam, M. Harry. 2001. Extraction and purification of microbial DNA from soil and sediment samples. Applied Soil Ecology 18, 47-60.

Rothberg, J.M., J.H. Leamon. 2008. The development and impact of 454 sequencing, Nature Biotechnology 26, 1117–1124.

Saida, F., M. Uzan, B. Odaert, F. Bontems. 2006. Expression of highly toxic genes in *E. coli*, special strategies and genetic tools. Current Protein and Peptide Science 7, 47-56.

Schatz, A., S. A. Waksman. 1944. Effect of streptomycin and other antibiotic substances upon *Mycobacterium tuberculosis* and related organisms. Proceedings of the Society for Experimental Biology and Medicine. Society for Experimental Biology and Medicine (New York, N.Y.) 57, 244-248.

Schipper, C., Hornung, C., Bijtenhoorn, P., Quitschau, M., Grond, S., Streit, W. R. 2009. Metagenome-Derived Clones Encoding Two Novel Lactonase Family Proteins Involved in Biofilm Inhibition in *Pseudomonas aeruginosa*. Applied and Environmental Microbiology 75, 224-233

Schirmer, A., R. Gadkari, C.D. Reeves, F. Ibrahim, E.F. DeLong, C.R. Hutchinson. 2005. Metagenomic analysis reveals diverse polyketide synthase gene clusters in microorganisms associated with the marine sponge *Discodermia dissoluta*. Applied and Environmental Microbiology 71, 4840-9.

Schloss, P.D., J. Handelsman. 2003. Biotechnological prospects from metagenomics. Current Opinion in Biotechnology 14, 303-10.

Schreiber, F., P. Gumrich, R. Daniel, P. Meinicke. Treephyler, fast taxonomic profiling of metagenomes. Bioinformatics 26, 960-1.

Schwartz, I. 2000. Microbial genomics, from sequence to function. Emerging Infectious Diseases 6, 493-5.

Sebat, J.L., F.S. Colwell, R.L. Crawford. 2003. Metagenomic profiling, microarray analysis of an environmental genomic library. Applied and Environmental Microbiology 69, 4927-34.

Selenska, S., W. Klingmuller. 1991. Direct detection of nif-gene sequences of *Enterobacter agglomerans* in soil. FEMS Microbiology Letters 80, 243-246.

Sessitsch, A., A. Weilharter, M.H. Gerzabek, H. Kirchmann, E. Kandeler. 2001. Microbial population structures in soil particle size fractions of a long-term fertilizer field experiment. Applied and Environmental Microbiology 67, 4215-24.

Shizuya, H., B. Birren, U.J. Kim, V. Mancino, T. Slepak, Y. Tachiiri, M. Simon. 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. Proceedings of the National Academy of Sciences of the United States of America 89, 8794-7.

Simon, C., R. Daniel. 2009. Achievements and new knowledge unraveled by metagenomic approaches. Applied Microbiology and Biotechnology 85(2), 265-76

Simon, C., J. Herath, S. Rockstroh, R. Daniel. 2009. Rapid identification of genes encoding DNA polymerases by function-based screening of metagenomic libraries derived from glacial ice. Applied and Environmental Microbiology 75, 2964-8.

Sleator, R.D., C. Shortall, C. Hill. 2008. Metagenomics. Letters in Applied Microbiology 47, 361-6.

Smarr, L. 2006. The ocean of life, Creating a community cyberinfrastructure for advanced marine microbial ecology research and analysis (a.k.a. CAMERA). Friday Harbor (Washington), Strategic News Service.

Smith, A.E., K. Hristova, I. Wood, D. M. Mackay, E. Lory, D. Lorenzana, K. M. Scow. 2005. Comparison of biostimulation versus bioaugmentation with bacterial strain PM1 for treatment of groundwater contaminated with methyl tertiary butyl ether (MTBE). Environmental Health Perspectives 113, 317-332.

Sogin, M.L., H.G. Morrison, J.A. Huber, D. Mark Welch, S.M. Huse, P.R. Neal, J.M. Arrieta, G.J. Herndl. 2006. Microbial diversity in the deep sea and the

underexplored "rare biosphere". Proceedings of the National Academy of Sciences of the United States of America103, 12115-20.

Solbak, A.I., T.H. Richardson, R.T. McCann, K.A. Kline, F. Bartnek, G. Tomlinson, X. Tan, L. Parra-Gessert, G.J. Frey, M. Podar, P. Luginbuhl, K.A. Gray, E.J. Mathur, D.E. Robertson, M.J. Burk, G.P. Hazlewood, J.M. Short, J. Kerovuo. 2005. Discovery of pectin-degrading enzymes and directed evolution of a novel pectate lyase for processing cotton fabric. The Journal of Biological Chemistry 280, 9431-8.

Sosio, M., F. Giusino, C. Cappellano, E. Bossi, A.M. Puglia, S. Donadio. 2000. Artificial chromosomes for antibiotic-producing actinomycetes. Nature Biotechnology 18, 343-5.

Staley, J.T., A. Konopka. 1985. Measurement of *in situ* activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annual Review of Microbiology 39, 321-346.

Staunton, J. & K.J. Weissman. 2001. Polyketide biosynthesis: a millennium review. Nat Prod Rep 18: 380–416.

Steffan, R.J., J. Goksoyr, A.K. Bej, R.M. Atlas. 1988. Recovery of DNA from soils and sediments. Applied and Environmental Microbiology 54, 2908-15.

Stein, J.L., T.L. Marsh, K.Y. Wu, H. Shizuya, E.F. DeLong. 1996. Characterization of uncultivated prokaryotes, isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. Journal of Bacteriology 178, 591-9.

Stinear, T.P., A. Mve-Obiang, P.L.C Small PLC et al. 2004. Giant plasmid-encoded polyketide synthases produce the macrolide toxin of Mycobacterium ulcerans. P Natl Acad Sci USA 101:1345–1349.

Suenaga, H., T. Ohnuki, K. Miyazaki. 2007. Functional screening of a metagenomic library for genes involved in microbial degradation of aromatic compounds. Environmental Microbiology 9, 2289-97.

Sukchawalit, R., P. Vattanaviboon, R. Sallabhan, S. Mongkolsuk. 1999. Construction and characterization of regulated L-arabinose-inducible broad host range expression vectors in *Xanthomonas*. FEMS Microbiology Letters 181, 217-23.

Sul, W.J., J. Park, J.F. Quensen, 3rd, J.L. Rodrigues, L. Seliger, T.V. Tsoi, G.J. Zylstra, J.M. Tiedje. 2009. DNA-stable isotope probing integrated with metagenomics for retrieval of biphenyl dioxygenase genes from polychlorinated biphenyl-contaminated river sediment. Applied and Environmental Microbiology 75, 5501-6.

Sylvia, D.M. 2005. Principles and applications of soil microbiology. 2nd ed. Pearson Prentice Hall, Upper Saddle River, N.J.

Tebbe, C.C., W. Vahjen. 1993. Interference of humic acids and DNA extracted directly from soil in detection and transformation of recombinant DNA from bacteria and a yeast. Applied and Environmental Microbiology 59, 2657-65.

Tien, C.C., C.C. Chao, W.L. Chao. 1999. Methods for DNA extraction from various soils, a comparison. Journal of Applied Microbiology 86, 937-943.

Torsvik, V., L. Ovreas. 2002. Microbial diversity and function in soil, from genes to ecosystems. Current Opinion Microbiology 5, 240-5.

Tringe, S.G., C. von Mering, A. Kobayashi, A.A. Salamov, K. Chen, H.W. Chang, M. Podar, J.M. Short, E.J. Mathur, J.C. Detter, P. Bork, P. Hugenholtz, E.M. Rubin. 2005. Comparative metagenomics of microbial communities. Science 308, 554-7.

Tsai, Y.L., B.H. Olson. 1991. Rapid method for direct extraction of DNA from soil and sediments. Applied and Environmental Microbiology 57, 1070-4.

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

Uchiyama, T., T. Abe, T. Ikemura, K. Watanabe. 2005. Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. Nature Biotechnology 23, 88-93.

Van Elsas, J.D., V. Mantynen, A.C. Wolters. 1997. Soil DNA extraction and assessment of the fate of *Mycobacterium chlorophenicolum* strain PCP-1 in different soils by 16S ribosomal RNA gene sequence based most-probable-number PCR and immunofluorescence. Biology and Fertility of Soils 24, 188-195.

Veluci, R.M., D.A. Neher, T.R. Weicht. 2006. Nitrogen fixation and leaching of biological soil crust communities in mesic temperate soils. Microbial Ecology 51, 189-96.

Venter, J.C., K. Remington, J.F. Heidelberg, A.L. Halpern, D. Rusch, J.A. Eisen, D. Wu, I. Paulsen, K.E. Nelson, W. Nelson, D.E. Fouts, S. Levy, A.H. Knap, M.W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-

Tillson, C. Pfannkoch, Y.H. Rogers, H.O. Smith. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304, 66-74.

Vogel, T.M. 1996. Bioaugmentation as a soil bioremediation approach. Current Opinion in Biotechnology 7, 311-6.

Voget, S., C. Leggewie, A. Uesbeck, C. Raasch, K.E. Jaeger, W.R. Streit. 2003. Prospecting for novel biocatalysts in a soil metagenome. Applied and Environmental Microbiology 69,6235-42.

Wang, C., D.J. Meek, P. Panchal, N. Boruvka, F.S. Archibald, B.T. Driscoll, T.C. Charles. 2006. Isolation of poly-3-hydroxybutyrate metabolism genes from complex microbial communities by phenotypic complementation of bacterial mutants. Applied and Environmental Microbiology 72, 384-91.

Wang, G.Y., E. Graziani, B. Waters, W. Pan, X. Li, J. McDermott, G. Meurer, G. Saxena, R.J. Andersen, J. Davies. 2000. Novel natural products from soil DNA libraries in a *Streptomycete* host. Organic Letters 2, 2401-4.

Ward, D.M., R. Weller, M.M. Bateson. 1990. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. Nature 345, 63-5.

Ward, N.L., J.F. Challacombe, P.H. Janssen et al. 2009. Three genomes from the phylum Acidobacteria provide insight into the lifestyles of these microorganisms in soils. Appl Environ Microb 75: 2046–2056.

Warnecke, F., P. Luginbuhl, N. Ivanova, M. Ghassemian, T.H. Richardson, J.T. Stege, M. Cayouette, A.C. McHardy, G. Djordjevic, N. Aboushadi, R. Sorek, S.G. Tringe, M. Podar, H.G. Martin, V. Kunin, D. Dalevi, J. Madejska, E. Kirton, D. Platt, E. Szeto, A. Salamov, K. Barry, N. Mikhailova, N.C. Kyrpides, E.G.

Matson, E.A. Ottesen, X. Zhang, M. Hernandez, C. Murillo, L.G. Acosta, I. Rigoutsos, G. Tamayo, B.D. Green, C. Chang, E.M. Rubin, E.J. Mathur, D.E. Robertson, P. Hugenholtz, J.R. Leadbetter. 2007. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature 450, 560-565.

Wawrik, B., L. Kerkhof, G.J. Zylstra, J.J. Kukor. 2005. Identification of unique type II polyketide synthase genes in soil. Applied and Environmental Microbiology 71, 2232-8.

Waxman D.J., J.L. Strominger.1983. Penicillin-binding proteins and the mechanism of action of beta-lactam antibiotics. Annu Rev Biochem. 52:825-69.

Wellington, E.M., A. Berry, M. Krsek. 2003. Resolving functional diversity in relation to microbial community structure in soil, exploiting genomics and stable isotope probing. Current Opinion Microbiology 6, 295-301.

Wild, J., Z. Hradecna, W. Szybalski. 2002. Conditionally amplifiable BACs, switching from single-copy to high-copy vectors and genomic clones. Genome Research 12, 1434-44.

Williamson, L.L., Borlee, B.R., Schloss, P.D., Guan, C., Allen, H.K., and Handelsman, J. 2005. Intracellular screen to identify metagenomic clones that induce or inhibit a quorum-sensing biosensor. Applied and Environmental Microbiology 71,6335–6344.

Woese, C.R. 1987. Bacterial evolution. Microbiological Reviews 51, 221-71.

Wommack, K.E., J. Bhavsar, J. Ravel. 2008. Metagenomics, read length matters. Applied and Environmental Microbiology 74, 1453-63.

Xia, X., J. Bollinger, A. Ogram. 1995. Molecular genetic analysis of the response of three soil microbial communities to the application of 2,4-D. Molecular Ecology 4, 17-28.

Zhou, J., M.A. Bruns, J.M. Tiedje. 1996. DNA recovery from soils of diverse composition. Applied and Environmental Microbiology 62, 316-22.

Aakvik, T., Degnes, K.F., Dahlsrud, R., Schmidt, F., Dam, R., Yu, L., Volker, U., Ellingsen, T.E., Valla, S., 2009. A plasmid RK2-based broad-host-range cloning vector useful for transfer of metagenomic libraries to a variety of bacterial species. FEMS Microbiol. Lett. 296, 149-58.

Alexeyev, M. F., Shokolenko, I. N., Croughan, T. P., 1995. Improved antibiotic-resistance gene cassettes and omega elements for Escherichia coli vector construction and in vitro deletion/insertion mutagenesis. Gene 160(1), 63-67.

Craig, J.W., Chang, F.Y., Kim, J.H., Obiajulu, S.C., Brady, S.F., 2010. Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria. Appl. Environ. Microbiol. 76, 1633-41.

Gay, P., Le Coq D., Steinmetz M., Berkelman T., Kado C.I., 1985. Positive selection procedure for entrapment of insertion sequence elements in Gram-negative bacteria. J. Bacteriol. 164(2), 918–921.

Handelsman, J., Rondon, M. R., Brady, S., Clardy, J., Goodman, R. M., 1998. Molecular biology provides access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem. Biol. 5, R245-R249.

Herrero, M., de Lorenzo, V., Timmis, K.N., 1990. Transposon vectors containing non-antibiotic resistance selection markers for cloning and stable chromosomal insertion of foreign genes in Gram-negative bacteria. J. Bacteriol. 172, 6557–6567.

Kim, U.-J., Birren, B. W., Slepak, T., Mancino, V., Boysen, C., Kang, H.-L., Simon, M. I., Shizuya, H., 1996. Construction and characterization of a human bacterial artificial chromosome library. Genomics 34, 213-218.

Liles, M.R., Manske, B.F., Bintrim, S.B., Handelsman, J., Goodman, R.M., 2003. A census of rRNA genes and linked genomic sequences within a soil metagenomic library. Appl. Environ. Microbiol. 69 (5), 2684-2691.

Liles, M. R., Williamson, L. L., Rodbumrer, J., Torsvik, V., Goodman, R. M., Handelsman, J., 2008. Recovery, purification, and cloning of high molecular weight genomic DNA from soil microorganisms. Appl. Environ. Microbiol. 74, 3302-3305.

Martinez, A., Kolvek, S.J., Yip, C.L.T., Hopke, J., Brown, K.A., MacNeil, I.A., Osburne, M.S., 2004. Genetically modified bacterial strains and novel bacterial artificial chromosome shuttle vectors for constructing environmental libraries and detecting heterologous natural products in multiple expression hosts. Appl. Environ. Microbiol. 70, 2452–2463.

Perri, S., Helinski, D. R., Toukdarian, A., 1991. Interactions of plasmid-encoded replication initiation proteins with the origin of DNA replication in the broad host-range plasmid RK2. J. Biol. Chem. 266, 12536– 12543.

Rine, J., Hansen, W., Hardeman, E., Davis, R.W., 1983. Targeted selection of recombinant clones through gene dosage effects. Proc. Natl. Acad. Sci. USA 80, 6750–6754.

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., Osburne, M. S., Clardy, J., Handelsman, J., Goodman, R. M., 2000. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl. Environ. Microbiol. 66, 2541-2547.

Simon, R., U. Priefer, Puhler, A., 1983. A broad host range mobilization system for in vivo genetic engineering: transposon mutagenesis in Gram negative bacteria. Bio/Technology 1, 784–791.

Thomas, C.M., Stalker, D.M., Helinski, D.R., 1981. Replication and incompatibility properties of segments of the origin region of replication of the broad host range plasmid RK2. Mol. Gen. Genet. 181(1), 1–7.

Wang, G.Y., Graziani, E., Waters, B., Pan, W., Li, X., McDermott, J., Meurer, G., Saxena, G., Andersen, R. J., Davies, J., 2000. Novel natural products from soil DNA libraries in a streptomycete host. Org. Lett. 2, 2401–2404.

Wild, J., Hradecna, Z., Szybalski, W., 2002. Conditionally Amplifiable BACs: Switching From Single-Copy to High-Copy Vectors and Genomic Clones. Genome Res. 2002 12, 1434-1444.

Wild, J., Szybalski, W., 2004a. Copy-control pBAC/oriV vector for genomic cloning, in: Balbas P., Lorence A. (Eds.), Methods in Molecular Biology, Recombinant Gene Expression. Reviews and Protocols. J.M. Walker, Series Ed., Vol. 267, Chap. 10. Humana Press Inc., Totowa NJ, pp. 145-154.

Wild, J., Szybalski, W, 2004b. Copy-control tightly regulated expression vectors based on pBAC/oriV, in: Balbas P., Lorence A. (Eds.), Methods in Molecular Biology, Recombinant Gene Expression. Reviews and Protocols. J.M. Walker, Series Ed., Vol. 267, Chap. 11. Humana Press Inc., Totowa NJ, pp. 155-167.

Wisseman Jr, C.L., F.E. Hahn, H.E. Hopps, J.E. Smadel. 1953 Chloramphenicol inhibition of protein synthesis. Federation Proc. 12, 466.

Wisseman Jr, C.L., J.E. Smadel, F.E.Hahn HAHN, H.E. Hopps. 1954. Mode of action of chloramphenicol. I. Action of chloramphenicol on assimilation of ammonia and on synthesis of proteins and nucleic acids in Escherichia coli. J. Bacteriol., 67, 662-673.