EXPLORING THE FUSION OF METAGENOMIC LIBRARY AND DNA MICROARRAY TECHNOLOGIES

Dan Spiegelman
Department of Natural Resource Sciences
McGill University
Macdonald Campus
Submitted Feb 20, 2006

A thesis submitted to McGill University in partial fulfilment of the requirements of the degree of Master of Science

Your file Votre référence
ISBN: 978-0-494-24806-5
Our file Notre référence
ISBN: 978-0-494-24806-5

Short Title

Exploring the Fusion of Metagenomic Library and DNA Microarray Technologies

Abstract

We explored the combination of metagenomic library and DNA microarray technologies into a single platform as a novel way to rapidly screen metagenomic libraries for genetic targets. In the "metagenomic microarray" system, metagenomic library clone DNA is printed on a microarray surface, and clones of interest are detected by hybridization to single-gene probes. This study represents the initial steps in the development of this technology. We constructed two 5,000-clone large-insert metagenomic libraries from two diesel-contaminated Arctic soil samples. We developed and optimized an automated fosmid purification protocol to rapidly extract clone DNA in a high-throughput 96-well format. We then created a series of small prototype arrays to optimize various parameters of microarray printing and hybridization, to identify and resolve technical challenges, and to provide proof-of-principle of this novel application. Our results suggest that this method shows promise, but more experimentation must be done to establish the feasibility of this approach.

Resumé

Nous avons exploré la possibilité de combiner les technologies de banques métagénomiques et de biopuces d'ADN en tant que moyen rapide de trouver une cible génétique dans une banque métagénomique. Dans ce système de "biopuces métagénomiques", l'ADN des clones est imprimé sur la surface d'une biopuce, et les clones d'intérêt sont détectés par hybridation avec une sonde d'ADN marquée. Ce projet représente les étapes initiales du développement de cette technologie. Deux banques métagénomiques de 5,000 clones furent construites à partir d'échantillons de sols arctiques contaminés avec du diesel. Une méthode automatisée fut développée et optimisée pour rapidement purifier l'ADN des clones dans un format 96-puits. Une série de prototypes de biopuces servirent à optimiser plusieurs paramètres d'impression et d'hybridation des puces, à identifier et résoudre les problèmes techniques, et à valider cette nouvelle application. Bien que les résultats soient prometteurs, de l'expérimentation additionnelle sera requise pour établir définitivement la praticabilité de cette technologie.

Acknowledgments

I would first like to thank Dr. Charles Greer for giving me the opportunity to work in his lab at the Biotechnology Research Institute, and the unique access to expertise and scientific resources which this allowed me. I could not even have conceived of a project like this one in any other environment. I would also like to thank Dr. Greer for his support and encouragement throughout this process, and for his efforts to secure funding so that I could continue to eat during the last few years. I would also like to thank my present and erstwhile co-supervisors at McGill, Drs. Lyle Whyte and Brian Driscoll, for their advice and support over the course of this challenging endeavour.

I would like to thank the entire Environmental Microbiology Group at BRI, for providing their help freely and enthusiastically on a thousand separate occasions, and for creating the most friendly, helpful, positive working environment I could ever have wished for. While this is true of every single member of the group, I would particularly like to thank Nathalie Fortin, Dr. David Juck, Diane Labbé and Sylvie Sanschagrin for the insight and help they have provided on every possible occasion, and for the patience with which they have handled my numerous questions. I would also like to thank Jean-Sebastien Deneault, Mélanie Arbour, François Benoit and Tracy Rigby from the Microarray Group of Dr. André Nantel, for all their considerable help with the robotics and microarray portions of this project.

Finally, I would like to thank the friends I have made over the course of my tenure at BRI. Gavin Whissell, Nancy Perreault, Punita Mehta, Ana Viquez, Sophie Gonin, Nicolas Lopes, Lori Phillips, Chris Newcombe, Nathalie Guibord, Céline Lacroix and Genevieve Bush. Quite simply, I couldn't have made it through this process without you. You have been there to laugh with on good days, to cheer me up on bad days, and to bounce ideas off when I wasn't sure. I am deeply thankful for the time we have shared together, and hope that our time at BRI is but the first chapter in an ongoing story.

Funding for this research was provided by a grant from NRC Canada PERD.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Hydrocarbon contamination, biodegradation and bioremediation in low-temperature environments

The last few centuries of human development have produced events unprecedented in the history of life on Earth. Not since the "great oxidation event" more than 2 billion years ago (66), in which cyanobacterial production of oxygen gas is believed to have transformed the atmosphere of the early Earth, has any one species had such a pervasive impact on the planetary environment. The collective industrial and agricultural activities of the human race have produced massive deforestation and soil erosion, depletion of the protective ozone layer of the atmosphere, the extinction of species, the disruption of entire ecosystems, and the potential destabilization of the planetary climate due to the accumulation of so-called 'greenhouse gasses'.

In addition, human industrial activity has generated a plethora of environmental pollutants, from heavy metals and radioactive materials to a wide range of organic solvents and wastes. One of the largest sources of environmental pollutants is the use of fossil fuels to generate energy. Fossil fuels, in the form of coal, natural gas and petroleum, are used for heating, electrical power generation, and transportation of all kinds. They represent by far the largest source of energy for the human race, accounting for 86% of all energy consumption worldwide in 2003 (42).

Accordingly, the storage, transportation and combustion of fossil fuels releases massive amounts of contaminating hydrocarbons wherever humans engage in industrial and commercial activities. Virtually no environment on Earth is unaffected by hydrocarbon contamination: soils and sediments, groundwater, freshwater and oceans are all subjected to various degrees of contamination. The polar regions, being largely devoid of human habitation, are less affected. However, human activity occurs in these areas in the form of scientific research, military activity, resource exploitation and tourism. Pollution occurs in particular wherever permanent bases are established, or where they existed in the past.

These bases use hydrocarbons for power generation, heating and vehicle operation; consequently, the most common source of pollution in polar regions is from accidental fuel spillage during storage and distribution from storage tanks or pipelines (4).

Diesel is a type of petroleum fuel commonly used in electrical generators and vehicle engines. Like other petroleum fuels, diesel is a complex mixture of many different hydrocarbons. The largest proportion, between 60-90% by volume, consists of normal, branched and cyclic alkanes (of chain length between C_9 and C_{30}). Aromatic compounds, especially alkylbenzenes, constitute 5-40%, while alkenes make up 0-10% by volume. The content of the highly toxic polycyclic aromatic hydrocarbons (PAHs) can exceed 10% by volume, although usually this number is less than 5% (164). However, combustion of diesel fuel increases PAH concentrations, especially the heavier and more toxic multiple-ring compounds (91).

Hydrocarbon fuel spills in polar soils can affect the physical, chemical and biological properties of these environments. Where spills darken the surface, they can affect soil temperature by decreasing albedo, which in some cases can increase daily maximum temperatures by up to 10°C (4). They can also lead to significant increases in the organic carbon content of soils; in instances where this increases microbial growth, the result can be a depletion of nitrates and a decrease in soil pH, possibly from the accumulation of hydrocarbon-derived acidic metabolites (3).

The harmful effects of hydrocarbon spills on marine avians, mammals, fish and invertebrates have been extensively popularized, especially in the wake of such dramatic events as the 1989 Exxon-Valdez oil spill. However, the effects on microbial communities are more complex and not as clearly negative, owing to the broad spectrum of microbial metabolism. While petroleum spills can be toxic to the growth and activity of some microorganisms, the elevated levels of organic carbon can also serve as substrates for growth for other microorganisms (17). Indeed, it has long been documented that hydrocarbon degraders can be isolated from a great number of different environments (180). In studies of petroleum-contaminated polar and alpine soils, most probable numbers (MPN) of hydrocarbon-degrading microorganisms have been measured at up to 10^5-10^7 per gram of soil (3, 4, 99). In other environments, hydrocarbon degraders have even been found in some isolated cases to represent up to 100% of the viable population after a contamination event (9).

By some measures, hydrocarbon contamination can appear beneficial to microbial communities; several studies have demonstrated a 10- to 100-fold increase in the number of cultivatable heterotrophic microorganisms in contaminated versus pristine samples, in Antarctic soils (3), alpine soils (99), Arctic soils (165), and Arctic sea-ice (55). However, population size is only one measure of a microbial community, and other metrics reveal a different picture. It has been observed in the above cases and in other studies that contamination negatively impacts microbial species diversity compared to uncontaminated controls (3, 47, 55, 75, 128). However, there is evidence that these results may be sample-dependent (75), and some studies have observed neutral or contradictory results (75, 90, 146).

When petroleum products are spilled or leaked into polar or alpine soils, they are affected by a number of physical and chemical processes that alter their distribution and composition. Once spilled, hydrocarbons tend to migrate downwards through the soil; lighter molecules of lower viscosity volatilize more readily and migrate more quickly, while heavier and more viscous materials are less mobile and less volatile. This downward movement stops at the permafrost boundary, upon contact with an ice-saturated layer only permeable through small pores and fissures. Along the way, some hydrocarbons are lost to chemical dissolution, while others adsorb to colloids and humus particles (101). One study of diesel-contaminated alpine soils measured the loss of contaminants by such abiotic processes at 30% (103). The same study observed that after a finite period of time, contaminant losses due to abiotic processes ceased; similarly, the same group in another study noted that no such abiotic decontamination occurs in chronically-contaminated soils (104). Thus, abiotic processes play a significant, but transient role in the removal of hydrocarbon contaminants.

A far greater role in decontamination is played by the microbial communities indigenous to contaminated sites. The process by which this occurs is known as biodegradation, which has been defined as the metabolic ability of microorganisms to transform or mineralize organic contaminants into less harmful substances, which are then integrated into natural biogeochemical cycles (100). It has been noted that microorganisms can attack almost all hydrocarbons, from methane to the heaviest paraffins (143). Several experiments have shown that microorganisms from polar soils possess the ability to degrade the most common components of hydrocarbon contamination: n-alkanes (11), monoaromatics (2), and PAHs (46). Although these classes cover hundreds of individual compounds, the general mechanism is similar: a series of key oxidation steps releases carbon from complex hydrocarbons, producing intermediate compounds that can be integrated into the central metabolic pathways of the cell (9, 170). An important secondary mechanism is co-metabolism, where complex hydrocarbons are partially degraded, but no metabolic energy is derived from the process (68).

The most extensively-characterized catabolic pathways for hydrocarbon degradation are the *alk*, *xyl* and *ndo* pathways, responsible for the degradation of $>C_5$ n-alkanes, aromatic hydrocarbons, and PAHs respectively (100). These pathways are most often encoded on catabolic plasmids: the OCT, TOL and NAH plasmids, which have been isolated from a number of uncontaminated environments and laboratory enrichment studies (130). Despite the wide range of hosts associated with these catabolic plasmids, studies of hydrocarbon biodegradation in Arctic and Antarctic soils suggest that hydrocarbon-degrading populations in these environments are dominated by a few key microbial species, especially of the *Rhodococcus*, *Sphingomonas* and *Pseudomonas* genera (4, 46, 168). The results of Whyte et al. (168) in particular suggest that *Rhodococcus* species are the principal degraders of the n-alkanes which are predominant among environmental hydrocarbons. Although natural sources of hydrocarbons exist in polar soils, such as the long-chain n-alkanes and n-alkenes likely produced by cyanobacteria and green algae (4), these catabolic genes are most often isolated from sites associated with anthropogenic hydrocarbon contamination (148, 165, 166).

Many factors affect the rate of microbial biodegradation of hydrocarbons in soil and other environments. One important factor is the availability of an electron donor: many studies have established the importance of oxygen in aerobic biodegradation (reviewed in (170)). In anaerobic habitats, denitrification may play a similar role in the oxidation of organic compounds, including pollutants such as the n-alkane hexadecane (25). Another factor is the nature of the compounds being degraded. Among the alkanes, nC_{10} to nC_{18} compounds are most readily attacked by degradative enzymes, $<nC_{10}$ molecules are often toxic due to membrane damage, while $>nC_{18}$ alkanes are only degraded slowly due to their highly hydrophobic nature. Alkanes larger than nC_{22} have very low solubility, making them even harder to degrade (143). In general, the viscosity and solubility of the offending compounds affect their ease of biodegradation; the low temperatures characteristic of polar environments decrease the solubility and bioavailability of aliphatic hydrocarbons and PAHs, and decrease the volatility of toxic short-chain alkanes (100, 101). Nutrient availability is another important consideration. Several studies investigating the effect of nutrients on biodegradation have found this to be a key limiting factor, especially nitrogen and phosphorous (18, 25, 143).

If biodegradation is the microbially-catalyzed breakdown of environmental pollutants, then bioremediation is defined as any technology that uses biodegradation to remove pollutants from the environment. In recent years, bioremediation has become one of the most rapidly developing areas in the field of environmental restoration (40). Several different bioremediation strategies have been employed, with differing degrees of success: natural attenuation, biostimulation, bioaugmentation and bioengineering (15, 40, 77, 102, 151, 155, 165). Natural attenuation simply refers to the natural biodegradation that occurs in response to contamination. The sudden input of carbon in the form of hydrocarbon contamination often leads to a depletion of the nutrients available to the community, as a result of increased microbial growth (110). If the addition of nutrients or augmentation of other degradation-limiting environmental variables such as aeration are used to accelerate microbial degradative activity, this process is referred to as biostimulation. Bioaugmentation, by contrast, involves not the modification of environmental conditions, but the modification of the microbial community itself, by adding specific microbes with degradative abilities. Bioengineering strategies for bioremediation consist of genetically modifying microorganisms for enhanced biodegradative ability. Engineering options include enzyme-tailoring and DNA-shuffling to optimize enzymatic activities, combining multiple degradation pathways in a single organism, increasing resistance to bioremediation-hampering factors such as short-chain alkane toxicity, and adding genes for biosurfactant production, to increase substrate availability (40).

Several studies of hydrocarbon contamination in low-temperature soil environments have addressed the question of which bioremediation strategy is the most effective. Although a recent review has claimed that "assisted bioremediation" (biostimulation) is likely to be the bioremediation method of choice in polar environments (4), a cursory review of the experimental literature produces a more complex picture. Several studies comparing natural attenuation with biostimulation cover the range of results from a clear preference for biostimulation (102) to a small transient preference for biostimulation (77) to a clear preference for natural attenuation (15). Meanwhile, some studies of bioaugmentation and biostimulation conclude that the addition of a bioaugmentation treatment does little to improve the overall results of a biostimulation treatment alone (155), while others have found a combination of these treatments to be optimal (151). A more nuanced conclusion from a similar study is that bioaugmentation decreases the lag time in community response to contamination, which in cold sites with short summer seasons could be a particular asset (165). Perhaps the most salient conclusion to be drawn from the contradictory results of these comparative studies is that site-specific characteristics are of great importance in determining which bioremediation strategy is most effective (15). Thus any framework to decide upon which strategy to pursue must take into account the composition and biochemistry of the pollutants in questions (i.e. how many different compounds and how amenable they are to biodegradation), the bioavailability of the contaminants, and how much potential exists to optimize the microbiological activity at the site (40).

1.2 The Eureka Bioremediation Project

In the spring of 1947, American and Canadian personnel established the Eureka Joint Arctic Weather Station in Eureka, on Ellesmere Island in the Northwest Territories (now Nunavut), as part of a joint project to build and operate five such stations in the Canadian High Arctic. For 50 years, the Eureka station has been providing vital climate information. More recently, the renamed and Canadian-run Eureka High Arctic Weather Station has also become the site of the Arctic Stratospheric Ozone Observatory (ASTRO), and has been host to teams of scientific researchers and to dozens of tourists yearly (44).

In 1990, the Eureka High Arctic Weather Station also suffered a serious hydrocarbon contamination event, when leaking pipelines spilled some 37,000 litres of diesel fuel, contaminating approximately 3200m^3 of soil. A feasibility study was undertaken by the Biotechnology Research Institute of the National Research Council of Canada (NRC-BRI) to determine the potential for bioremediation of the site (167). The conclusions of that study were that there was considerable bioremediation potential during the short Arctic summer season of 4-6 weeks. The recommendations of the Phase 1 study formed the basis for the bioremediation project that has been conducted at the site since the summer of 2000.

The bioremediation strategy undertaken at Eureka was one of biostimulation. The treatment consisted of the addition *in situ* of a commercial fertiliser (C:N:P ratio of 20:20:20), coupled with a tilling regime that increased aeration up to a foot below the surface. Control areas were those which had been contaminated, but received no treatment. To accompany the bioremediation project, NRC-BRI has been conducting a bioremediation monitoring program of the site. This monitoring has three main components. First, the indigenous microbial communities from various sampling sites have been enumerated in terms of total viable aerobic bacteria (culturable by the spread-plate technique on MSM-YTS medium), in terms of diesel-degradative populations (culturable by most probable number (MPN) analysis at 5°C), and in terms of the levels of culturable bacteria possessing alkane- and PAH-degradative genotypes (*P. putida alkB and Rhodococcus alkB2,* and *ndoB*, respectively, monitored by colony hybridizations). Second, the hydrocarbon mineralization potential has been monitored, in soil microcosm studies using ^{14}C-hexadecane (for C_{16} alkanes) and ^{14}C-naphthalene (for PAHs), at 5°C. Third, sampling sites have been monitored for total petroleum hydrocarbons (TPH). Samples at each site were taken from the active (upper 80-100cm) and permafrost (>100cm) layers of the soil (85).

As of February 2004, the results of the bioremediation study were as follows: both total viable aerobic and diesel-degrading microbial populations were generally higher in treated soils than in untreated, although there were some discrepancies from year to year. By the 2003 sampling, hexadecane mineralization activity was stronger in the treated samples, as was naphthalene mineralization, with a few exceptions. Soil TPH levels in the treated and untreated soils had been significantly reduced over the study period (85).

The samples used in the current study were taken from the Eureka bioremediation study, from the 2003 sampling. These samples were a treated active layer sample BRI-1 (designated 1A3) and an untreated active layer sample BRI-6 (designated 6A3). Specific properties of these samples, and the rationale for their selection, will be discussed later in this report.

1.3 Microbial communities, DNA libraries and metagenomics

1.3.1 Community characterisation and DNA libraries

In microbiology, studying an environmental sample is synonymous with studying a microbial community. Microbial communities exist in every environment on Earth, underpin the food webs of many ecosystems, and play a crucial role in the biogeochemical cycles of many key elements (34, 36, 107). A wide variety of methods exist for the characterisation of microbial communities (reviewed in (150)). Traditional microbiological approaches rely on the cultured growth of a small subset of the microbial community. However, since some 99% of environmental microbes are not culturable by standard methods, these approaches do not provide a comprehensive view (7). Another set of methods characterises microbial communities by analysing the biochemical properties and molecular composition of key cellular biomolecules such as membrane lipids and respiratory quinones, usually producing profiles characteristic to a single community (150). The most powerful and most commonly-used methods are based on the analysis and differentiation of microbial DNA.

Most DNA-based methods for community characterisation rely on the polymerase chain reaction (PCR) to amplify specific genes from DNA extracted directly from an environmental sample. A large majority of such studies focus on taxonomically-differentiated genes – highly conserved genes that mutate at a slow but constant rate over time, such as the gene for the 16S ribosomal RNA. These genes can be used as a "molecular chronometer" to measure taxonomic distances between species, based on variations in their DNA sequence (171). Such DNA polymorphism-based methods are used to produce community profiles, and to provide species (or other taxon) identification of community members by DNA sequencing of PCR fragments, based on their homology with other sequences stored in bioinformatic databases. Another subset of PCR-based methods looks at other functional genes, often genes that code for a catabolic function of interest in the environmental sample. These assays can serve as an indirect presence-or-absence test for a specific catabolic function, or they can be used for taxonomic identification of a subset of the microbial population, based on sequence variation in a shared catabolic gene (150).

For all the analytical power of PCR-based methods of community characterisation, these methods all share a set of biases and limitations imposed by the use of PCR. In general, the more manipulation a sample undergoes prior to analysis, the more opportunities exist to introduce bias into that analysis. At the heart of the PCR process is a cyclical series of manipulations: double-stranded DNA denaturation, primer binding, enzymatic primer extension (DNA replication), then denaturation again. Each of these steps has the possibility of introducing bias. The %GC content of the template DNA affects the kinetics of denaturation; fragments with a higher %GC denature less efficiently, causing differential amplification. Once denatured, single-stranded DNA can form higher-order structures such as hairpin loops, which interfere with DNA extension by *Taq* polymerase. This can particularly be a problem when amplifying rRNA genes, which depend on such higher-order structures in their transcribed RNA for proper biological function. Proper primer binding depends on an exact match between primer and template sequences, which can be difficult to achieve when amplifying many different forms of the same gene in an environmental sample; improper specificity between template and primer can lead to the under-representation or loss of those template sequences. Heterologous hybridization between highly-similar but non-identical template sequences in an environmental sample has been shown to interfere with primer binding as well. Even when this is not the case, heterologous binding can lead to the formation of chimeric molecules, artefacts that can form at frequencies of several percent of the total amplified DNA (150). During primer extension, replication errors can be introduced by *Taq* polymerase, which lacks a proofreading function. Finally, errors of manipulation such as tube-to-tube contamination or reagent contamination can occur. Though all these errors can be small, the exponential nature of PCR amplification can greatly magnify even the most minute errors (150).

DNA clone libraries represent a technology for amplifying DNA that is free of the biases and limitations of PCR. In this method, genomic or environmental community DNA is extracted and ligated into a cloning vector, a mobile genetic element which is placed inside a host cell and amplified using the host's DNA replication machinery. Ligation of the insert DNA into the vector requires the generation of compatible ends of the respective DNA molecules; this can be achieved either by digestion of the insert and vector DNA with restriction endonucleases, or by end-repair of sheared insert DNA for blunt-end cloning (45). Although some instances of cloning bias have been reported in studies of (PCR-based) single-gene amplicon clone libraries, the explanations suggested were concerned with restriction sites existing within the single gene being cloned (97), which is not a factor in PCR-independent cloning of environmental DNA. Other suggested (but unsubstantiated) sources of cloning bias are random error due to undersampling of community diversity, toxicity of vector inserts, choice of cloning kit, and differences due to cloning strategy (blunt- vs. sticky-end cloning) (97, 118, 153). Another possible source of error not discussed is the effect of methylation, which may block restriction endonuclease action. What is clear is that there is much less evidence for bias arising from the manipulations of cloning than from the manipulations of PCR, other than the common biases associated with DNA extraction (150).

Besides relative freedom from certain biases, libraries can match and surpass the information provided by PCR-based methods of community characterisation. PCR can still be performed on the library sample to find any number of target genes, and the amplicons produced are very amenable to sequencing. But with libraries, since the clones bearing the target genes can be identified, it is possible to obtain sequence information from outside the fragment amplified by PCR. In recent years, with the advent of large-insert vectors suitable for gene expression (142), the breadth of library-based analyses of environmental DNA has increased still further.

1.3.2 Metagenomics and metagenomic libraries

The Human Genome Project spawned and spurred a host of new technologies and technological innovations. Critical to this endeavour was the ability to clone and sequence very large fragments of DNA, a key requirement to generate physical maps over multi-megabase genetic distances. This was originally done using Yeast Artificial Chromosomes (YACs), which can routinely hold insert sequences up to 500kb, and even above 1Mb in some cases (22, 87). However, YACs are prone to several important disadvantages: low cloning efficiency, high incidence of chimeric clones, instability of insert DNA and difficulties purifying YAC DNA from host cells all complicated the use of these vectors (8). The development of Bacterial Artificial Chromosomes (BACs) represented a major advance over existing technology. Based on the *E. coli* F factor, this vector maintains tight replication control (at 1-2 copies per cell), produces a much lower proportion of chimeric clones than YACs, and can stably maintain inserts larger than 300kb (142), in some cases as large as 600kb (179). A similar innovation produced by the same group of researchers was the fosmid cloning vector, a hybrid of traditional cosmid vectors (based on bacteriophage λ) and the *E. coli* F factor, which allowed the cloning of cosmid-sized inserts (30-50kb) with BAC-like stability, replication control, and low incidence of chimerism (81).

With these advances in DNA library technology in hand, researchers in microbiology quickly realised that new avenues of research were now open. For those who wished to characterise the physiological and metabolic potential of the huge majority of uncultivated microorganisms, the new generation of high-capacity cloning vectors were the key. These vectors allowed for the direct cloning of the collective genomes of the microbial species present in an environmental sample, termed the 'metagenome' of the sample (62).

From the start it was clear that this approach was a very powerful tool for discovery. The first study to use this method found a single marine archaeal fosmid clone bearing a suite of genes never before characterised in this phylum (152). Since then, metagenomic studies have been conducted in a wide variety of environments, including soil, fresh- and saltwater, human feces, the human oral cavity, and the 'hospital metagenome'; the scales of metagenomic analysis have ranged from single genes and pathways to whole organisms and communities (reviewed in (121)). These studies have generated a broad array of exciting new discoveries: discoveries of pure science include a novel form of marine bacterial phototrophy (12), recovery of an entire rRNA operon from a ubiquitous uncharacterised crenarchaeote (117), and sequencing of 3 Mb of the even-more ubiquitous uncultured *Acidobacterium* division of bacteria (92, 134). Discoveries of industrial importance from metagenomics are numerous, and include several classes of novel antibiotics (19, 56, 163), antibiotic resistance genes (38), chitinases (32), a 4-hydroxybutyrate-metabolizing enzyme (64) and an entire biosynthetic pathway for biotin (43).

One of the truly novel benefits of metagenomic library technology is that it allows researchers to find links between function and phylogeny in the uncultured fraction of the microbial community. This has primarily been accomplished in two ways. First, in phylogenetic studies that identify clones bearing rDNA genes: once a large-insert clone bearing a rDNA gene has been identified, the rest of that clone is sequenced and the genes are compared to database sequences to infer their function. Despite the haphazard "fishing expedition" nature of this approach, it has nevertheless led to discoveries of an array of functional genes from novel organisms, including archaeal DNA polymerase (117), archaeal RNA helicase (152), and the first-ever documented rhodopsin of bacterial origin (12). The second means of linking phylogeny to function is in a sense the inverse of the first method: studies that identify functional genes in library clones can sequence genes flanking the target gene, in the hope that both the target genes and the flanking genes can provide clues to phylogeny by homology with database sequences from identified species (43, 82, 115). In cases where phylogenetic identifications are made from rDNA sequences in metagenomic clones, the sequences of flanking genes can similarly be used to strengthen phylogenetic association (121).

When constructing a metagenomic library, it is very important to consider the desired insert size, because this decision has several important ramifications. The larger the insert, the more genes can be included on a single fragment. This is important for metagenomic expression studies, which often rely on the presence of entire biochemical pathways in a single clone to produce the desired metabolic activity. Conversely, small inserts are more suited to sequence-based inquiries and bulk sequencing efforts (121). Desired insert size also affects the choice of DNA extraction methods and cloning strategies; harsh extraction methods such as bead-beating are more likely to lyse a greater range of cell types (84), but much gentler extraction methods (with their associated extraction biases) are needed to obtain the very large inserts used in BAC and similar large-insert vectors. Moreover, since restriction digestion can greatly reduce the size of extracted DNA fragments, the preparation of large fragments for cloning requires the use of partial digestions (8, 13, 92, 123), or blunt-end (end-repair) cloning strategies (117).

Insert size also directly impacts library size and the extent of genomic coverage. The number of clones needed (N) to represent a specific DNA sequence with a certain degree of probability (P) can be calculated according to the following equation (26):

$$(1) \qquad N = \frac{\ln(1-P)}{\ln(1-L/G)}$$

where L is the average length of the insert DNA in base pairs, and G is the haploid [meta]genome size in base pairs. Two important conclusions can be drawn from this formulation: first, large-insert libraries require fewer clones to achieve the same level of metagenome coverage as small-insert libraries. Second, for complex environments such as soil, where estimates for number of species present range from 1,000 to 10,000 per gram of soil (156, 157), the number of clones needed for extensive representation is enormous. Assuming an average diversity of 4,500 species per gram soil, and an average prokaryotic genome size of 3.1Mb (53), to obtain 99% coverage of the metagenome would require ~12.8 million 5kb inserts (small plasmids), ~1.6 million 40kb inserts (cosmids/fosmids), or ~428,000 150kb inserts (BACs). According to a recent survey (121), most metagenomic studies use libraries that are several orders of magnitude too small to capture the full microbial diversity present in the environment under study. The abundance of discoveries made using metagenomic libraries is thus even more astounding considering this limitation, and speaks volumes to the deep reservoir of information hidden within natural microbial communities around the world, that we have only recently begun to tap.

1.3.3. Metagenomic library screening methods

To obtain useful information from a metagenomic library requires some means of ordering and sorting the huge volumes of information and focusing in on specific clones of interest, a process known as screening. Library screening methods generally fall into two broad categories: sequence-based and function-based screening. In addition, sub-populations of metagenomic DNA can be screened prior to library construction, allowing for the creation of smaller, targeted libraries.

Sequence-based screening methods seek to identify clones bearing specific sequences, based on their homology to other known sequences. There are three basic approaches to this end. One method is to use successive rounds of PCR to detect the presence of a desired gene in increasingly small subsets of the library, until an individual clone(s) can be identified. The standard PCR screening method identifies first a gene-positive superpool of multiple microtiter plates, then a single microtiter plate, and finally the row and column co-ordinates of the gene-positive clone (although there is some variation in the precise details of pooling) (23). One of the most efficient variants of this method can identify a positive clone from a full library in two PCR reactions; one to identify the tens- and ones-column of the positive plate number, and one to identify the row and column co-ordinates of the positive clone (8). It should be noted that this particular use of PCR is less affected by the inherent biases of this technique. The formation of chimeric molecules and other replication errors still occur, but since the amplified DNA is only used to signal the presence of the target sequence among the clones being screened, artificial sequence changes are irrelevant. PCR biases that lead to the non-amplification of certain sequences, of course, remain a potential problem.

Hybridization screening represents another sequence-based approach. In this method, metagenomic library DNA is fixed in gridded sets on a series of filters and incubated in the presence of labelled gene-specific nucleic acid probes. These probes hybridize selectively to individual clones bearing the gene of interest, and are detected by chemiluminescence or autoradiography. The advantage of this method over PCR-based screening is its ability to screen large numbers of individual clones in parallel in a single assay. However, the relatively lower level of sensitivity and higher levels of background non-specific interactions can yield large numbers of false positives and false negatives (23). In addition, to screen a large library in this manner requires a very large number of filters, which can be cumbersome and costly. An alternative is a multistep screening process analogous to the PCR-based processes described above: superpools of clone DNA representing individual high-density filters are first screened by southern hybridization to a labelled probe; thus only the high-density filters identified as positive for the gene of interest need be subject to probe hybridization (8).

Thanks to cost-reducing advances in DNA sequencing technology, bulk sequencing of entire metagenomic libraries has recently emerged as a third method of sequence-based metagenomic library screening. Given the size requirements for comprehensive library coverage discussed above, such bulk sequencing represents an enormous undertaking. Not surprisingly, the infrequent uses of this approach have been confined to environments of relatively lower biological complexity (such as seawater or biofilm communities) (158, 161), or else to a relatively small number of clones representing a small fraction of potential metagenomic diversity (135). In the latter scenario, the task of analysis is limited to performing database searches of the clone sequences, in order to identify as many genes as possible. However, sequencing a more representative metagenomic library involves a more challenging step: the re-assembly of the disparate clone sequences into (ideally) complete organism genomes. To do this, sequences are aligned into multi-clone scaffolds, which are sorted into tentative "organism bins" based on %GC content, read depth (frequency of sequence re-occurrence in the library), and similarity to sequence from reference species in existing databases (158, 161).

The power of this approach is self-evident, and this method presents specific advantages over other sequence-based methods: researchers can look *in silico* for as many genes as desired, at no extra cost in screening materials, and without the possibility of false positive and false negative bias inherent to molecular biological screening methods (122). In addition, bulk sequencing and database searching allows researchers to uncover new homologues of functional genes that would escape detection using PCR- and hybridization-based screening, which require a high degree of homology between the target genes and the primer/probes used for screening (134). Although the biggest limitation associated with the bulk sequencing approach is cost, several other challenges and limitations must be considered. During sequence assembly well-conserved regions such as rRNA genes may assemble across species, which breaks scaffolds; similarly, closely-related species may also cross-assemble. Low-abundance organisms allow for fewer mate-links between contigs, and thus offer less statistical support for scaffolds built from those contigs (161). Another limitation, alluded to earlier, is that for the time being, bulk sequencing efforts cannot be applied to complex communities like soil, due to prohibitive costs in time and resources. Furthermore, the input of massive amounts of sequence from bulk sequencing efforts can skew sequence databases. For example, immediately after the publication of the Venter group's findings (161), their Sargasso Sea sequences alone accounted for some 5% of the GenBank database (121). Finally, massive sequencing projects are still limited by the available databases for gene identification. For example, only 35% of the genes identified among the 1.6 billion bases sequenced by Venter et al. produced significant hits during database searches (121, 161).

The second major category of metagenomic library screening methods is known as expression-based screening or functional screening. In this strategy, clones are screened for expression of a desired trait, by use of a broad range of functional assays. Functional screening is made possible by two factors: first, large-insert vectors are capable of cloning a large number of genes in a single insert. For example, one study isolated two cosmid clones that contained 22 open reading frames (ORFs) each, in an average insert of 34.3kb (135). Second is the fact that genes for natural-product biosynthetic pathways (such as antibiotics) tend to cluster together in prokaryotes (62). Among metagenomic studies, the expressed-based approach has been responsible for the vast majority of discoveries of industrial importance (such as novel antibiotics, biocatalysts and biosynthetic pathways) (19, 32, 38, 43, 56, 60, 64, 65, 123, 139, 163). Though many of these screening studies usually involve a rapid, simple plate assay for the desired function, other approaches are possible. Metagenomic libraries can be used to complement specific activity-deficient mutants of a heterologous host (64, 98). Alternately, the use of liquid cultures for screening allows researchers to identify genes involved in "mosaic pathways", where the desired function is accomplished by the activity of multiple species acting in concert (134).

The probability of finding a desired gene in a metagenomic library is a function of its abundance in the environmental sample, the average insert size, the length of the gene, and the presence of functional expression signals recognised by the host. More specifically, heterologous expression requires the cloning of a transcriptional promoter and a ribosomal binding site (in the appropriate position relative to the start codon), both of which must be acceptable to the host cell (53). In the majority of metagenomic expression studies, *E. coli* is the host of choice; it has relaxed requirements for promoter recognition and transcription initiation compared to other hosts, and has been known to express genes from very divergent clades such as *Thermus*, *Bacillus cereus* and *Corynebacterium* (62, 95). In an *in silico* study of 32 complete prokaryotic genome sequences, researchers searched for expression signals known to function in *E. coli*. They concluded that about 40% of enzymatic activities from heterologous bacteria could be recovered using *E. coli* as a host, although this number varied from 7-73% depending on the species of origin (53).

The above finding reflects one of the major limitations of expression-based screening: the dependence on successful heterologous expression of cloned genes. One possible solution that has been employed is to use different hosts. To this end, both *Streptomyces lividnas* and *Pseudomonas putida* have been used (33, 105, 163). Another limitation of functional screening is that an assay for the desired function must exist, and it must be amenable to large-scale use in an efficient and high-throughput manner, due to the low frequency of hits typically found in large metagenomic libraries (134). Furthermore, unless the desired function is the product of a single gene, this approach requires that the necessary genes are clustered together in an area small enough to fit on one clone (175).

However, expression-based screening has one major advantage over sequence-based screening. Indeed, the two method classes complement each other in terms of their advantages and limitations. Sequence-based methods are limited by the need for homology between primers/probes and clone sequences, and rely on databases for gene identification; but most functionally important genes are too divergent to identify new homologues by PCR or hybridization, which require conserved regions in well-studied genes (134). Expression-based methods on the other hand, because they rely exclusively on function, can detect genes of entirely novel sequence (175). Conversely, because sequence-based screens do not require expression of gene products, these methods can recover genes which would go undetected in functional screens due to undetectable expression in incompatible hosts (95).

Recently, a third screening method has appeared in the literature, a modification of the standard expression-based methods, termed substrate-induced gene expression or SIGEX (159, 175). The impetus for the development of this method is that traditional expression screening is labour-intensive, time-consuming and inefficient, screening a huge number of clones for a very small number of hits. The conceptual basis of this approach is the premise that expression of catabolic genes is generally induced by substrates or metabolites of the enzymes in question, often under the control of proximate genetic elements. In this method, metagenomic DNA is cloned into a GFP-fusion expression vector. Thus, expression of catabolic genes in the presence of substrate is measured by the expression of GFP (175). The power of this method comes from the use of fluorescence-activated cell sorting (FACS), a high-throughput way to isolate those clones expressing GFP and presumably, clones expressing the desired catabolic activity. This method is ideally suited for attempts to isolate a catabolic gene whose enzymatic activity is not easily analysed by traditional expression screens, or for which no traditional assay is available (159).

There are several limitations associated with this method. The substrate or its metabolites must be able to reach the clone DNA in order to activate transcription, thus substrates which cannot migrate into the cytoplasm are excluded (175). Also, this method is not appropriate for very large-insert libraries: it has been calculated that using inserts longer than 15kb is not useful for transcriptional fusions, due to the abundance of transcriptional terminators (53). More important, this method shares the main limitation of all expression-based screening methods, namely its reliance on heterologous expression. However in this case the limitation can be mitigated somewhat by reducing the stringency of the FACS, allowing lower levels of GFP expression to be selected as a positive hit (with the natural consequence of increasing the incidence of false positives) (159). The main advantages of this system are its use of high-throughput screening by FACS, and its ability to greatly reduce the size of the primary metagenomic library. In this way, SIGEX is analogous to methods such as stable-isotope probing (41) and bromodeoxyuridine enrichment (160), which can isolate actively-metabolising sub-populations based on incorporation of modified nucleic acids. Although to date these methods have not been used to pre-segregate metagenomic sub-populations for library construction, they offer great potential for substrate-based pre-screening of metagenomic libraries.

1.4 DNA Microarrays

While BACs and other high-capacity cloning vectors are a technology that has revolutionised modern microbiology, another recent technological advance, represents no less of a revolution: DNA microarrays. Indeed, the widespread use of these tools is transforming biology as a whole, opening up entire new avenues of research. A DNA microarray is an extremely high-density matrix of thousands of individual DNA sequences arranged in a well-defined grid that has been immobilised on a 2-dimensional surface. Data is produced when fluorescently-labelled DNA or RNA hybridises to immobilised sequences on the array. In essence, a microarray hybridization is a massively parallel search by each labelled molecule for a matching partner on the array (94).

Microarray technology emerged out of the genomics revolution of biology, to become one of the principal tools of genomic analysis. The ever-expanding numbers of fully- or partially-sequenced genomes have generated enormous amounts of sequence data. However, sequence alone offers very few clues to the function of that which is sequenced. Where the traditional approach to elucidate the function of genes depends on step-by-step analysis of individual genes, in a whole-genome context a global approach is more appropriate, or at least an essential first step to unravel such massive complexity (72).

If the demands of the genomics revolution provided the impetus to develop microarray technology, then recent advances in biochemistry and robotics have provided the occasion. Certainly the concept of nucleic acid hybridization on a fixed surface is not new; dot-blot hybridization has been an established technique in molecular biology for more than a quarter of a century (78). Arrays of nucleic acids have been used for years for a variety of purposes, including measuring mRNA expression levels (89), analysis of RNA secondary structures (149), and DNA sequencing (80). However, it was the development of techniques to imprint biomolecules on glass surfaces (50, 106) and the use of robotics for large-scale oligonucleotide synthesis which by 1994 had opened the door for high-density, miniaturised arrays (51, 114).

There are two major types of DNA microarrays that are widely used today: oligonucleotide-based arrays and amplicon-based arrays. The former are spotted with oligonucleotides typically ranging from 25 to 70 nucleotides in length, which are often synthesised *in situ* robotically and printed directly on the slide (54). The latter rely on PCR to produce amplicons from DNA or mRNA template, which are then purified, plated and printed (174). The two classes of microarray are generally applied to different types of study, though there is some overlap. Oligonucleotide arrays are better suited to genome-wide diversity analyses, sequencing by hybridization, investigation of intergenic sequences, single-nucleotide polymorphism investigations and other mutational analyses, while the main applications for amplicon arrays are in cDNA- (mRNA amplicon)-based expression studies (54, 93, 144). The two classes cover a wide range of hybridization specificities: amplicon-based arrays can produce reliable hybridization signals with up to 20% sequence difference between probe and target, while small oligonucleotides are used to detect single-base mismatches

(59). For long, oligonucleotide arrays were burdened by high costs of production; however oligonucleotide synthesis prices have been falling steadily for years, to a point where large oligonucleotide arrays are approaching the cost of amplicon arrays [André Nantel, personal communication].

DNA microarrays represent a significant advance over earlier fixed-surface hybridization methods. The most obvious advantage is in sheer numbers. As of 2005, it is possible to obtain commercially manufactured arrays that contain more than 1.3 million unique features on a single chip (1). Microarrays also possess several advantages due to their very small size: compared to other hybridization platforms microarrays use minimal reaction volumes, which has the effect of reducing reagent consumption, increasing sample concentrations and accelerating reaction kinetics (132). Another key advantage of microarrays is the ability to hybridize multiple samples on the same array, by use of multiple fluorophores for labelling multiple samples (133). Such comparative analyses using older hybridization methods would require using multiple membranes, or else a damaging process of membrane stripping and re-probing. In contrast, the multiplexing ability of microarrays minimises variables inherent to comparing samples in separate experiments: membrane-to-membrane variation and variations in experimental conditions (132).

Nevertheless, the range of manipulations involved in a microarray experiment, from producing the chip to scanning a hybridization image, offer many opportunities to introduce error or bias, including cross-contamination of samples used for array printing, generation of PCR artefacts (for amplicon arrays), irregularities in spot printing, informatic errors leading to misidentification of array spots, inefficient sample labelling, uneven hybridization, and errors in image acquisition and processing (113, 174).

Thus, proper study design and methods of data normalisation are critical to minimise these effects. Key issues of study design are sample size and the proper use of replicates. Commonly used replicate types include printing duplicate/triplicate spots, technical replicates such as dye swaps, and biological replicates. There is general agreement that some form of power analysis is necessary and that larger samples with more replicates provide more power, but there is currently no consensus on methods to determine sample size (5).

Data normalisation is the process by which microarray spot intensities are adjusted to take into account the variability across different experiments and platforms (5). Two important normalisations are the use of labelling standards and internal printing controls. Standardisation of labelling is used to account for differences in labelling efficiency and probe amounts, by reference to the hybridization signal from a standard amount of control DNA added to every labelling reaction. Internal standards for printing generally rely on comparing the signal from a standard amount of control DNA added to each spot to be printed, and correcting for measured differences. Because microarray experiments generate such a vast amount of data, statistical analysis of results is an essential step to derive quantitative or even qualitative conclusions. However, statistical analysis of microarrays is still very much a developing field, and there is still a lack of consensus on important questions of study design, data pre-processing, statistical inference, and classification and validation of results (5, 113).

Although the first use of microarrays for research was to study gene expression in plants (133), the most common uses of this technology since then has been in biomedical research and clinical diagnostics (10). Applications to this end include single-nucleotide polymorphism mapping and annotation of the human genome (131, 162), target validation, biomarker identification and toxicology in drug discovery (70, 129, 137), and analysis of gene expression in human disease, from simple diseases like β-thalassemia to complex neurological disorders and different types of cancer (27, 39, 57, 58, 109).

In recent years, DNA microarrays have also been applied to microbiology in a wide range of studies. One important area of application is the study of human pathogens, where the use of microarrays to characterise the human cell response to infection has allowed researchers to develop an understanding of the pathogenic process without the need to culture dangerous or difficult pathogens (10, 14, 30, 71). Microarrays have also been developed into tools to detect pathogens in clinical samples (29). Comparative genomics studies of expression and genome sequence in pathogenic and non-pathogenic strains have provided insight into the genetic determinants of virulence (79, 136). Indeed, comparative genomics studies between closely-related strains and species are equally used to suggest function-specific genetic clusters in non-pathogenic organisms (111), or core clusters which define the taxonomic group (125). Microarrays can also be used to define a regulon – the set of genes controlled by a single regulator – by hybridization to RNA from cells where the regulator is mutated or over-expressed (174, 178). Another innovative use of microarrays is for *in vitro* protein-binding studies, where protein-binding specificities are determined by protein-binding to specific sequences, detected by protein-specific antibodies and labelled secondary antibodies (21).

More recently, microarrays have also been applied with increasing frequency and success to environmental studies of complex microbial communities. The most common application is the characterisation of microbial communities based on the simultaneous detection of large numbers of microbial genes (59). These applications subdivide into studies which look for specific functional genes (173) and studies which seek taxonomic affiliations of community members using rDNA-based microarrays (96). The other main use of microarrays in environmental studies is expression profiling. In this case, functional gene arrays are still used, but instead of labelling total extracted or PCR-amplified community DNA (120, 145), researchers label total community mRNA in order to generate an expression profile for the entire microbial community (37). In all such studies, where the sequences of the genes of interest in the community can differ substantially from those printed on the array, the unique properties of amplicon-based arrays can be a great asset: under low-stringency hybridization conditions, target sequences with as little as 60-70% identity to the immobilised probes can still be detected. However, these relaxed requirements for sequence specificity can cut both ways, creating problems when it comes to generating quantitative data: because of the great sequence variation in the environment, it is difficult to distinguish differences in signal intensity due to population abundance with differences due to sequence divergence (177).

Although most applications of microarrays in microbial systems resemble those described above, some groups have devised novel applications which warrant a brief discussion. In one method, termed 'Library-on-a-slide', the entire genomes of 72 different strains of *E. coli* were printed individually into single array spots, to be hybridised to single genes or gene alleles; the purpose of this method being to explore genetic differences between closely-related strains (176). The key innovation of this method is the printing of very large DNA, where normally only small amplicons or oligonucleotides are used. In another novel application, devised as an alternative for expression studies where no annotated genomes are available, PCR-amplified genomic library inserts of a single species were arrayed on a slide, to be hybridised to total RNA from the same species under various biological conditions (63). Microarrays have also been applied to screening metagenomic libraries: in a test project, the end sequences of a large number of metagenomic cosmid clones were amplified by PCR, producing 1kb fragments that were printed on an array. This array was hybridised to the labelled genomic DNA from single isolates, pooled isolates, and total community DNA isolated from the experimental sample; the goal being to isolate those library clones representing uncultured microbes (138).

1.5 The current study: Metagenomic Microarrays

1.5.1 Metagenomic microarrays

The current study represents another novel application of DNA microarrays: they are used as a platform to rapidly screen a metagenomic library for the presence of a single gene, and to identify specific clones bearing the gene of interest. To do this requires three conceptual steps: first, construction of a metagenomic library in high-capacity cloning vectors. Second, purification of plasmid DNA from each clone, and printing of purified clone DNA from the entire library on a microarray slide. We term this construct a 'metagenomic microarray'. Third, hybridization experiments using a single labelled gene as a probe, to identify specific clones bearing the target gene. Following this, the identified clones can be subject to further analysis, including sequencing of the target gene and its surrounding genetic information. Thus in theory, starting from a printed metagenomic microarray, one could screen a metagenomic library for the presence of any number of specific genes in a single overnight assay, at the cost of a few slides per gene screened (depending on the size of the library) and a few micrograms of labelled probe.

In context of the earlier discussion in section 1.3.3 of this review, metagenomic microarrays are a sequence-based method of library screening, specifically a method of hybridization screening. As a method of hybridization screening, this technology represents a significant advance over existing methods. Since clones can be spotted at a much greater density on microarray slides than on traditional membranes, a complete metagenomic library that would require a large number of membranes can be compressed onto a single microarray slide, or a small number thereof, greatly reducing the complexity of the experiment. Similarly, because microarrays require such small reaction volumes (on the order of 20-150μl), array-based screening represents a great reduction in the amount of expensive labeling reagents consumed. Finally, it is easy to produce a large number of identical microarrays from a very small amount of starting materials, due to the incredibly small volumes of material required for printing (less than 1nl per spot), and the extensive use of automation. Thus the metagenomic library can easily be screened for a large number of genes in parallel array experiments, with much greater ease and rapidity than can be achieved by standard membrane hybridization screening.

The greatest technical challenges of this method lie in printing large-insert library clones on a microarray slide, in quantities and of quality sufficient to obtain a reliable hybridization signal from clones bearing specific genes of interest. The technical complications are twofold in this respect. First, the use of large DNA fragments presents several obstacles to signal recovery: large fragments can only be printed at low molar amounts, since higher amounts increase printing solution viscosity to unacceptable levels. Also, if the desired target is a single gene, this represents only a small fraction of the DNA contained in a single large-insert clone. Second, the plasmid DNA to be printed must be of a sufficient purity that contaminants do not interfere with printing or hybridization.

There have been no reports in the literature of printing large-insert library clones directly onto microarrays, although a recent review suggests that this method should be explored (10). Researchers have reported printing PCR fragments amplified from large-insert libraries, providing either partial (138) or full representation (73, 154) of the clones used as template. However, considering the large costs associated with the thousands of PCR reaction required for such an application, it would be of great value to develop an alternative that does not require this added expense and extra manipulation. Metagenomic microarrays, on which unmodified clones are printed directly, present one such alternative.

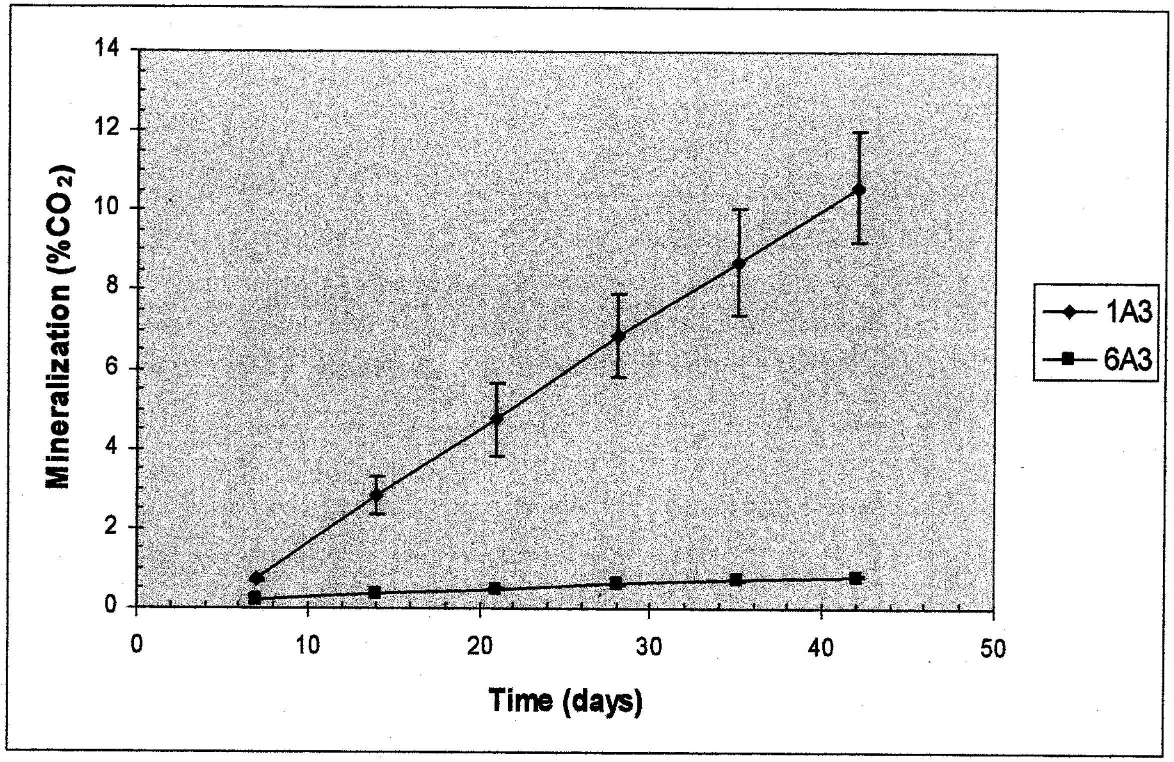
1.5.2 Research goals

In the broadest sense, the goal of the current project was to take the first steps in the development of the novel technology of metagenomic microarrays. The construction of prototypes thus figured prominently among the subordinate goals of this broader endeavour. However, the construction of prototypes even of limited scope required a series of intermediate steps which must be counted as important goals in their own rights. Therefore, the first goal of this project was to successfully develop metagenomic libraries from two similar but varied ecosystems; the microbial communities from the diesel-contaminated and bioremediation-treated BRI-1A3 soil, and from the contaminated but untreated BRI-6A3 soil. These libraries would supply the clones to be used for construction of prototype metagenomic microarrays. The second goal of this project was to develop an automated process of clone preparation for microarray printing, one which could provide purified clone DNA from a large number of library clones in a short amount of time, while minimizing the costs associated with such production. Only once these two goals were accomplished could we begin the process of developing and testing this novel technology.

Originally, the research goals of this project concerning metagenomic microarrays were twofold. First, to design and build two prototype metagenomic microarrays, from 5,000-clone metagenomic libraries of the 1A3 and 6A3 microbial communities. Second, to test the effectiveness of these prototype metagenomic microarrays by using them to track changes in a soil microbial community as a result of bioremediation treatment. Although these goals were later revised to be less expansive, the original goals of the microarray portion of this project bear some discussion here because they were germane to the choice of metagenome samples and gene probes used to screen the libraries. The updated research goals of the microarray portion of this study are presented at the end of the following discussion.

At the time of sampling, the samples chosen for this study had undergone three years of differential treatment. Both had been contaminated in the original fuel spill, but BRI-1A3 had undergone three years of bioremediation treatment (nutrient amendment and tilling), while BRI-6A3 had received none. The 2004 report of the Eureka Bioremediation project (85) had shown that BRI-1A3 could readily mineralize ^{14}C-hexadecane in microcosms at 5°C, while BRI-6A3 showed virtually no such activity [figure 1]. This suggests that BRI-1A3 has an active alkane-degrading population, while BRI-6A3 does not.

In order to use metagenomic microarrays to track differences between the two soil samples required genetic targets that reflected the differences between the samples. Based on the results of the prior study presented above, one target chosen for the current study was *alkB*, encoding alkane monooxygenase. This enzyme catalyzes the initial terminal oxidation of the alkane degradative pathway, which includes hexadecane (168). Thus, *alkB* was to be used as a probe to hybridize with the metagenomic microarrays from the 1A3 and 6A3 samples. The measurement of differences between these two samples was to be based on the differential prevalence (frequency of occurrence in the library) and variability (number of different forms, determined by sequencing) of this gene in both libraries. Similarly, to reflect the differential fertilizer treatment that the two samples had undergone for three years, nitrite reductase was chosen as a target, the key enzyme in the dissimilatory denitrification process (20). Specifically, we chose two genes: *nirS* – encoding a cytochrome cd_1-containing nitrite reductase, and *nirK* – encoding a copper-containing nitrite reductase, as targets to reflect the differences between the two samples.



Adapted from Labbé et al. (2004)

Figure 1. Mineralization of hexadecane at 5°C in microcosms of Eureka samples BRI-1A3 and BRI-6A3. Microcosm experiments were conducted in triplicate. $^{14}CO_2$ evolution was monitored by liquid scintillation spectrometry. Experiments were performed by the staff of the BRI Environmental Microbiology lab, as part of the Eureka Bioremediation Project. This data was adapted from Labbé et al. 2004.

However, the original microarray-based research goals of this project were overly expansive. As the magnitude of the undertaking became clearer with time and experimentation, we chose to adopt research goals that were more limited, more practical, and more realistically attainable in the framework of a Master's-level research project. In particular, a series of persistent technical challenges to basic clone hybridization signal detection and experimental control forced us to abandon the prospect of constructing full 5,000-clone metagenomic microarrays, thus preventing any thorough comparison of the two soil communities using these tools. Instead, the microarray portion of this study was confined to producing a series of smaller test arrays containing no more than 400 clones. These test arrays were used to optimize various parameters of the experimental format, to analyze the technical problems of this novel technology, and resolve them where possible.

Consequently, the microarray-based research goals were reformulated to reflect the practical realities of this project: First, to develop and optimize the process of preparing and printing large-insert plasmid DNA on microarrays. Second, to assess the feasibility of this technology as a means of rapidly screening a collection of metagenomic library clones for a specific gene. Third, to use small prototype arrays (test arrays) to optimize various parameters of metagenomic microarray experiment design, to identify the technical challenges of this experimental format, and to resolve these challenges to the best of our ability.

1.5.3 Study design

This study is aimed at developing and testing a new technology. The questions it seeks to answer are technical, and the key analytical steps consist of method optimization and troubleshooting. Thus, the design of this study centers around the production of a series of deliverables. Excluding a brief but necessary characterization of the community samples used, the bulk of the current study can be subdivided into three sections, representing major steps in the construction of a metagenomic microarray. The first section consists of the construction and characterization of two metagenomic libraries from samples BRI-1A3 and BRI-6A3. The second section of this study is concerned with producing purified metagenomic clone DNA, for array printing. This section is characterized by the development and optimization of an automated, high-throughput and cost-effective method to purify cloned metagenomic library DNA from its amplification host. The third section of this study comprises the construction of a series of test arrays, designed to determine standard conditions and parameters of metagenomic microarray experiment design, to define the technical obstacles to the construction of full metagenomic microarrays, to resolve these obstacles where possible, and to assess the feasibility of this approach as a means of screening a metagenomic library.

2 Materials and Methods

2.1 DNA extraction and sample characterization

2.1.1 Sample collection from Arctic contaminated soils

Soil samples BRI-1A3 and BRI-6A3 were collected, using aseptic techniques, from the active layer of the soil (30-100cm below the surface) as part of the NRC-BRI Eureka bioremediation project (85). Several hundred grams of each sample were placed in sterile bags and immediately frozen. Samples were transported frozen from Eureka, Nunavut to Montreal, Quebec, and were stored at –20°C until ready for DNA extraction.

2.1.2 DNA extraction from Arctic contaminated soils

To extract total community metagenomic DNA from the BRI-1A3 and BRI-6A3 soil samples, we adapted the method of Fortin et al., designed to extract microbial DNA from polluted soils (52). Forty grams of soil from each sample were suspended in 80ml of soil buffer 1 (50mM Tris-HCl [pH 8.3], 200mM NaCl, 5mM EDTA [pH 8.0], 0.05% Triton X-100, 1% PVP-10, 1% PVP-40) by vortexing and inversion, then pelleted by centrifugation at 3,110 x g for 3 minutes at 4°C. This was repeated two more times, in 80ml of soil buffer 2 (50mM Tris-HCl [pH 8.3], 200mM NaCl, 5mM EDTA [pH 8.0]) and in 80ml of soil buffer 3 (50mM Tris-HCl [pH 8.3], 0.1mM EDTA [pH 8.0]). Cells were disrupted by incubating with 10mg/ml lysozyme for 30 minutes at 30°C with moderate shaking, followed by a 30 minute incubation at 37°C. Five µl of 20mg/ml proteinase K were added and samples were incubated for 1 hour at 37°C. Lysis occurred upon addition of 50µl of 20% SDS and incubation for 30 minutes 85°C. Samples were centrifuged to remove cellular debris. One-half volume of 7.5M ammonium acetate was added and samples were incubated for 15 minutes on ice. One volume of 2-propanol was then added and samples were precipitated overnight at –20°C. DNA was pelleted by centrifugation at 17,400 x g for 30 minutes at 4°C, rinsed once with 70% ethanol, then again with 100% ethanol. After air drying for 1h, the DNA was resuspended in 100µl 10mM Tris-HCl [pH

8.0] + 0.1mM EDTA [pH 8.0] (TE 10/0.1). RNAse treatment was not included since RNA removal occurred during size selection (section 2.2.1).

2.1.3 16S rRNA gene PCR amplification

To prepare materials for DGGE, PCR of the 16S rRNA gene was performed on both samples after size-selection and gel extraction, using the universal bacterial primers U341 and U758 [Table 1]. Between 300pg and 60ng of metagenomic template DNA was added to 25pmol of each primer, in the presence of 0.2mM dNTPs, 1mM $MgCl_2$, 125µg/ml BSA and 1x *Taq* buffer (Amersham Biosciences). Initial denaturation occurred at 94°C for 5 min, after which 1.2U of *Taq* polymerase was added while holding at 80°C. Thermal cycling was performed following a "touch down" procedure, to minimize amplification of non-specific primer-binding events: denaturation at 94°C for 1 min, annealing at 65°C for 1 min, primer extension at 72°C for 1 min. The annealing temperature was then decreased by 1°C per cycle until it reached 55°C, then remained at 55°C for an additional 20 cycles. Final primer extension was at 72°C for 7 min. Template concentrations were varied to determine optimal conditions for amplification.

2.1.4 DGGE

To characterize the microbial communities from the two soil samples, denaturing gradient gel electrophoresis was performed using a DCode Universal Mutation Detection System (Bio-Rad) for electrophoretic resolution of PCR-amplified 16S rRNA gene fragments on the basis of %GC content. Fragments of the 16S rRNA gene (750ng) amplified using the U341GC and U758 primers [Table 1] from each sample was loaded into 8% acrylamide gels containing denaturant gradients of 40-60%, 60-70% or 40-80%. Gels were run at 80V for 16 hours, stained with Vistra Green (GE Healthcare) for 30 minutes and destained in 1x TAE for 30 minutes. Gels were visualized under UV using a FluorImager System (Molecular Dynamics; Sunnyvale, CA)

In order to determine the relatedness of the two microbial community samples in each of the different denaturant gradients, we applied Sorensen's index of similarity to calculate band sharing between samples (147). First, the total number of different bands was determined by visual inspection of the DGGE profiles for each sample. Bands from different samples were considered common if they migrated the same distance in the gel. We then applied the following equation to generate the coefficient of similarity S_{AB} between the two samples:

(2) $S_{AB} = 2J / (A+B)$

Where A is the number of bands in sample A, B is the number of bands in sample B, and J is the number of bands common to A and B.

Table 1. primers used in the current study

primer name	gene	description	sequence (5'-3')	source
U341	16S rRNA	universal bacterial primers	CCTACGGGAGGCAGCAG	Muyzer et al. 1993
U758			CTACCAGGGTATCTAATCC	Lee et al. 1993
U341GC	16S rRNA	GC clamped for DGGE	CGCCCGCCGCGCGCGGCGGGCGGGGCGGGGGCACGGGGGGCCTACGGGAGGCAGCAG	Muyzer et al. 1993
alkH1F	*alkB*	alkane monooxygenase concensus primers	CIGIICACGAIITIGGICACAAGAAGG	Chénier et al. 2003
alkH3R			IGCITGITGATCIII GTGICGCTGIAG	
nirS F	*nirS*	nitrite reductase (cytochrome cd1) concensus primers	CGGCTACGCGGTGCATATCTCGCGTCTGTC	Ren et al. 2000
nirS R			GATGGACGCCACGCGCGGCTCGGGGTGGTA	
cu-nir F	*nirK*	nitrite reductase (copper) concensus primers	GGGCATGAACGGCGCGCTCATGGTGCTGCC	Ren et al. 2000
cu-nir R			CGGGTTGGCGAACTTGCCGGTGGTCCAGAC	
FOS-cosF	pCC1FOS *cos site*	pCC1FOS-specific amplicon	ACATGAGGTTGCCCCGTATTCAGN	current study
FOS-cosR			ACTTCCATTGTTCATTCCACGGAN	
FOS-CmF	pCC1FOS chloramphenicol resistance gene	pCC1FOS-specific amplicon	AAACGGCATGATGAACCTGAN	current study
FOS-CmR			GATGTGGCGTGTTACGGTGAN	
mmoX1	*mmoX*	soluble methane monooxygenase concensus primers	CGGTCCGCTGTGGAAGGGCATGAAGCGCGT	Miguez et al. 1997
mmoX2			GGCTCGACCTTGAACTTGGAGCCATACTCG	
pmoA-A189	*pmoA*	particulate methane monooxygenase concensus primers	GGNGACTGGGACTTCTGG	Holmes et al. 1995
pmoA-mb661			CCGGMGCAACGTCYTTACC	Costello and Lidstrom 1999
luxAb 941	*luxA*	luciferase enzyme alpha subunit	CCGACTGCCCATCCGGTTCGACAAGC	Cohn et al. 1985
luxAe 1231			CTCCGCGACGACATAAACAGGAGCACCACC	
gfp-F1	*gfp*	Green fluorescent protein	TGTGGTCTCTCTTTTCGTTGGG	Juck et al. 2005
gfp-R1			TGGTGTTCAATGCTTTGCGAG	
uidA 858	*uidA*	*E. coli*-specific beta-glucuronidase	ATCACCGTGGTGACGCATGTCGC	Juck et al. 1996
uidA 1343			CACCACGATGCCATGTTCATCTGCC	

2.1.5 PCR of target catabolic genes from Arctic soil DNA extracts

PCR of the *alkB*, *nirS* and *nirK* catabolic genes was performed on both samples to determine their presence or absence in the samples. Reaction conditions were similar to those described in section 2.1.3, with the following modifications: no "touch down" procedure was followed. Instead, annealing temperatures were 57°C for *alkB*, 55°C for *nirS* and 65°C for *nirK*. The primers used were alkH1F and alkH3R for *alkB*; nirSF and nirSR for *nirS*; cu-nirF and cu-nirR for *nirK* [Table 1]. Thirty pg of plasmid (pDrive) containing each of the three catabolic genes (cloned by Sylvie Sanschagrin at BRI) were used as positive controls for the reaction.

The above primers were chosen because they are concensus primers, targeted to conserved regions in their respective genes and thus able to amplify these genes from a broad range of organisms, such as might be found in an environmental sample. The *alkB* primers anneal to bases 495-521 (forward) and 1018-1044 (reverse), and were designed from conserved regions from histidine boxes 1 (forward) and 3 (reverse) from all database entries for *alkB* as of February 2003, except for *Acinetobacter*. The *nirS* primers anneal to bases 852-881 (forward) and 1138-1167 (reverse), and were designed from conserved regions in *Paracoccus denitrificans* Pd1222, *Pseudomonas aeruginosa* PAO1, *Pseudomonas stutzeri* ATCC 14405, and *Ralstonia eutropha*. The *nirK* primers anneal to bases 560-589 (forward) and 906-935 (reverse), and were designed from conserved regions in *Achromobacter cycloclastes*, *Pseudomonas aeruginosa* G179, and *Pseudomonas aureofaciens* ATCC 13985

2.2 Metagenomic library construction

2.2.1 Size-selection and gel extraction of total community DNA

To purify 30-50kb fragments of metagenomic DNA for subsequent cloning, extracted total community DNA was run overnight at 20V on a 0.8% SeaPlaque GTG low-melting TAE-agarose gel (Cambrex; East Rutherford, NJ), alongside a λ Mono Cut Mix ladder (New England Biolabs; Ipswitch, MA). The 30-50kb region of unstained sample was excised from the gel under long-wave UV using EtBr-stained ladder and a small amount of stained sample as a guide. As RNA migrates quickly through the gel, this process also serves to remove RNA. Samples were removed from the gel using the Gelase agarose gel-digesting preparation (Epicentre; Madison, WI), following the manufacturer's protocol with the following modifications: samples were digested for 1-2h to ensure complete digestion, and were centrifuged 3x10 minutes at room temperature to remove undigested gel. Purified DNA was resuspended in TE (10/0.1)

2.2.2 End-repair

To generate blunt-end fragments for cloning, size-selected metagenomic DNA was incubated for 45 minutes at room temperature with End-Repair Enzyme Mix (Epicentre; Madison, WI) following the manufacturer's protocol. After inactivation at 70°C, remaining enzyme was removed from the DNA solution by phenol:chloroform (1:1) extraction and ethanol precipitation, and the DNA was resuspended in 20μl sterile water. At this stage, DNA was quantified by PicoGreen (Molecular Probes; Eugene, OR). Briefly, dsDNA-binding reagent was added to a small amount of sample DNA, and quantified by comparison to reference amounts of λ DNA.

2.2.3 Ligation

Blunt-end ligation to dephosphorylated pCC1FOS Fosmid vector was performed with a 10:1 vector:insert molar ratio. Ligation was performed using the CopyControl Fosmid Library Production Kit (Epicentre; Madison, WI) following the manufacturer's protocols, with the following modifications: 325ng size-selected, blunt-ended metagenomic DNA and 650ng fosmid were added to a 13μl ligation reaction, which proceeded overnight at 16°C. One 25ng-insert sample was removed prior to ligation, and two 25ng-insert samples were removed after ligation to verify the success of the reaction, leaving 10.84μl ligated DNA representing 250ng insert and 500ng vector. A control reaction was similarly performed using 325ng of 'Fosmid control DNA' from the CopyControl Fosmid kit. To verify the success of the reaction, a 25ng-insert sample was digested with

1 unit of *Not*1 restriction endonuclease in NEBuffer 3 (New England Biolabs; Ipswitch, MA) supplemented with 100μg/ml BSA for 1 hour at 37°C, to cut the vector on either side of the multicloning site. This was run alongside a 25ng-insert sample taken prior to ligation, and a 25ng-insert sample of uncut, ligated DNA on a 0.3% TBE gel run overnight at 20V. The absence of a linearized vector band at 8.1kb in the ligated, uncut sample confirmed successful ligation.

2.2.4 Phage packaging

In vitro packaging of ligated, concatomerized DNA into λ phage heads for transfection of the *E. coli* cloning host was performed using MaxPlax Packaging Extracts (Epicentre; Madison, WI). A 10.84μl ligation reaction containing 250ng of insert DNA was incubated at 30°C for 90 minutes with 25μl of packaging extract, followed by an additional 90 minutes at 30°C with an additional 25μl of packaging extract. Packaged phage was diluted for storage in a final volume of 1ml Phage Dilution Buffer (10mM Tris-HCl [pH 8.3], 100mM NaCl, 10mM $MgCl_2$), supplemented with 2.5% chloroform.

2.2.5 Determining library titer

In order to plate the proper density of colonies for efficient recovery, it is imperative to determine the titer of the metagenomic library (the number of clone colonies formed per ml of packaged phage). To this end, packaged phage was diluted serially from 1:10 to $1:10^4$ in Phage Dilution Buffer. Ten μl of each above dilution was incubated for 20 minutes at 37°C with 100μl of *E. coli* EPI300-T1 cells (Epicentre; Madison, WI) grown from an overnight culture to an OD_{600} of 0.8-1.0 in LB supplemented with 10mM $MgSO_4$. Packaged control DNA from the CopyControl Fosmid kit (Epicentre; Madison, WI) was similarly used to transfect EPI300-T1 cells, at dilutions of $1:10^2$ to $1:10^6$. Transfected cells were plated on LB plates containing 12.5μg/ml chloramphenicol and incubated overnight at 37°C to select for fosmid-bearing clones. Library titer (in cfu/ml) was determined by counting colonies and applying the following formula:

$$\text{library titer} = \frac{(\text{\# of colonies})(\text{dilution factor})(1000\mu l/ml)}{(\text{volume of packaged phage plated})}$$

Titer was averaged over all dilution plates. Plates with fewer than 30 colonies were not included in averaging calculations.

2.2.6 Plating the metagenomic libraries

To generate individual colonies bearing cloned metagenomic DNA, *E. coli* EPI300-T1 cells were transfected with packaged phage at a 10:1 (v:v) cell:phage particle ratio, incubated for 20 minutes at 37°C. Plating volumes were approximately 1ml of transfected cell suspension per 22.5cm x 22.5cm plate (LB + 12.5μg/ml chloramphenicol), so the library was diluted accordingly prior to transfection. Plates were incubated overnight at 37°C. Final colony densities were on the order of 4,000-5,000 per plate, which was the recommended density for optimum automated colony picking. Three such plates from each library were provided for colony picking.

2.2.7 Automated colony picking and library plating

Automated colony picking was employed to circumvent the intensive labour required to pick 5,000 clones from each metagenomic library. Automated colony picking was a paid service provided by the Centre for Structural and Functional Genomics at Concordia University in Montreal. Colonies bearing metagenomic library DNA were picked and sorted into 96-well microtiter plates using a VersArray Colony Picker and Arrayer (BioRad). Picked colonies were transferred to a 25% glycerol + 50%LB/chloramphenicol solution for long-term storage at –80°C.

2.2.8 Characterization of library clones by Restriction Fragment Length Polymorphism

Clones (48) from each library were characterized by RFLP analysis to verify the successful cloning of multiple non-repetitive fragments of metagenomic DNA. One fosmid preparation's worth (~700-1500ng) of cloned DNA from the 1A3 library plate 49, rows A-D and the 6A3 library plate 11, rows E-H were subjected to restriction digestion using 3U of *Not*1 (New England Biolabs) overnight at 37°C, followed by incubation for 20 minutes at 65°C to stop the reaction. Five μl of each reaction was run on 0.5% TAE-agarose at 60V for 15 minutes, then at

100V for 2.5-3 hours, until the bromophenol blue dye band reached the bottom of the gel. Fragment sizes were estimated by reference to DNA fragments from the GeneRuler 1kb DNA Ladder (Fermentas) and λ Mono Cut Mix ladder (New Englad Biolabs).

2.3 Plasmid purification and microarray preparation

2.3.1 Automated fosmid preparation by alkaline lysis

Final optimized protocol

We developed the following protocol to purify fosmid DNA from metagenomic library glycerol stock plates. All steps were performed on a Biomek FX Laboratory Automation Workstation (Beckman Coulter; Fullerton, CA), except where specified. Cells from glycerol stock plates were transferred to deep-well (2ml) Costar 96-well plates (Corning; Corning, NY) containing 1ml of LB + 12.5μg/ml chloramphenicol using a manual Library Copier (V&P Scientific; San Diego, CA). Cultures were grown to saturation overnight at 37°C shaking at 300rpm, in 96-well blocks sealed with Bioseal breathable tape (CLP; San Diego, CA). Overnight culture (200μl) was transferred to new deep-well plates containing 1ml (per well) of 2xYT, containing 12.5μg/ml chloramphenicol, and 1x CopyControl Induction Solution (Epicentre; Madison, WI), a patented formulation for induction of high-copy replication of the pCC1FOS vector. Cultures were grown to an OD_{600} of 1.4, representing a near-saturation level of cellular growth in our chosen 96-well format, then centrifuged for 5 minutes at 1,811 x g to pellet the cells. After decanting the supernatant, cells were resuspended by vortexing in 300μl of chilled STE solution (10mM Tris-HCl [pH 8.0], 100mM NaCl, 1mM EDTA [pH 8.0]), then centrifuged and pelleted as before. Cells were resuspended by vortexing in 200μl of chilled GTE solution (50mM glucose, 25mM Tris-HCl [pH 8.0], 10mM EDTA [pH 8.0]). Six μl of 20mg/ml proteinase K was added using a multichannel pipet, and 96-well blocks were incubated 30 minutes at 37°C, with moderate shaking. Four hundred μl of alkaline lysis solution (0.2N NaOH, 4% SDS) were added and mixed by pipeting. Three hundred μl of chilled 7.5M ammonium acetate were added to precipitate the lysate, and culture blocks were incubated 15 minutes at 4°C. Samples were centifuged at 1,811 x g for 30 minutes at room temperature to pellet cellular debris (all subsequent centrifugations were at room temperature as well). All supernatant from more than 2mm above the bottom of the wells was transferred to a new 96-well block, then centrifuged and transferred as before to a third 96-well block. Nucleic acids were precipitated by adding 800μl of room-temperature 2-propanol, mixing by pipetting, then centrifuging at 1,811 x g for 30 minutes. After decanting the supernatant, the pellet was air-dried for 30 minutes, then resuspended in 200μl of 100μg/ml RNAse in water and incubated for 30 minutes at 37°C with moderate shaking. RNAse was removed by adding 100μl of chilled 7.5M ammonium acetate, incubation on ice for 15 minutes, then centrifugation at 1,811 x g for 30 minutes. All supernatant from more than 2mm above the bottom of the wells was transferred to a new 96-well block, and 600μl of 100% ethanol were added and mixed by pipetting. DNA was precipitated by centrifugation at 1,811 x g for 30 minutes, and supernatant was decanted. The DNA pellet was washed with 1ml of 70% ethanol and centrifuged at 1,811 x g for 10 minutes. After decanting the supernatant, the pellet was air-dried at room temperature for 30 minutes and resuspended in 30μl of sterile deionized water. All clones printed on test arrays from TA 5.1 onward were purified using the final optimized protocol.

Protocol optimizations

Optimization of precipitant solutions was performed by comparing 4 different precipitants. Immediately following alkaline lysis in 400μl of 0.2N NaOH + 4% SDS as described above, we allowed 0.1 volumes of 3M sodium acetate, or ½ volume of 7.5M ammonium acetate, or ½ volume of potassium acetate to precipitate cellular debris from a single 96-well block of clones, followed by centrifugation and transfer twice and 2-propanol precipitation as described above. For CTAB precipitation, a working concentration of 0.2% CTAB (w/v) was used in accordance with Lander et al. (86), who suggest that this concentration of CTAB selectively precipitates plasmid DNA, leaving protein, RNA and lipopolysaccharides in solution. Following centrifugation, the pellet containing the plasmid DNA was resuspended in 0.7M NaCl and precipitated again by adding 1 volume of 2-propanol. From this stage onward, all optimization sample plates were treated identically. DNA from the same 12 clones from each plate was quantified by PicoGreen (Molecular Probes; Eugene, OR), and relative protein contamination was assessed by comparing the spectrophotometric absorbance of each sample at 260nm (for DNA) and 280nm (for protein) using a ND-1000 Spectrophotometer (Nanodrop Technologies; Wilmington, DE)

For optimization of centrifugation parameters, the method used was identical to the final optimized protocol described above, with the following exceptions. All manipulations subsequent to clone growth were performed manually in 1.5ml microcentrifuge tubes. Centrifugations were carried out at 15,000 x g or 1,811 x g, at 4°C or at room temperature, in a Biofuge Fresco table-top microcentrifuge (Heraeus Instruments; Hanau, Germany). All 1,811 x g centrifugations were for 30 minutes, while 15,000 x g centrifugations were for 15 minutes.

For initial optimization of induction culture time, clones were grown for 3h or 5h, then fosmid DNA was purified as described in the final optimized protocol, except that the proteinase K digestion step was omitted (referred to in section 3 as the "basic method"). All fosmid clones printed on Test Arrays 1 through 4.2 were grown for 5 hours and purified by the basic method. For later optimization of induction culture time, cultures were grown to OD_{600} values of 0.6, 0.8, 1.1 or 1.4, measured by absorbance at 600nm in a Turner Model 340 Spectrophotometer (Testwave LLC; Sparks, NV)

For optimization of additional purification steps, 6 replicates of each of 4 clones grown to each of 4 OD_{600} values underwent fosmid purification by the basic method, with the following modifications for each purification treatment. Proteinase K treatment was as described in the final optimized method. Size-exclusion filtration was performed using a 0.2μm polyvinylidene fluoride (PVDF) filter (Corning; Corning, NY). Vacuum filtration on the Biomek FX replaced the two rounds of centrifugation and supernatant recovery described in the final optimized protocol. Glass fiber DNA-binding was accomplished using a Multiscreen-FB Glass Fiber Type B 1μm filter (Millipore; Billerica, MA). Immediately after RNAse digestion, 1 volume of "binding buffer" (7M Guanidine-HCl in 200mM MES (2-[N-morpholino]ethane-sulfonic acid) buffer [pH 5.6]) was added to samples and mixed by pipeting, then liquid was passed through the filter by vacuum on the Biomek FX. Filters were washed in 1ml of 80% Ethanol, and DNA was resuspended in 50μl sterile water and recovered by centrifugation for 3 minutes at 1,811 x g. Fosmid purification by commercial kits was not performed in the 96-well format. Instead, 100ml of each clone were grown in Erlenmeyer flasks. Twenty ml samples were removed at each target OD_{600} and were used for either maxi-preparation (OD 0.8, 1.1 and 1.4) or midi-preparation (OD 0.6), using Plasmid Midi and Maxi Kits (Qiagen; Hilden, Germany).

2.3.2 Origin of internal standards for fosmid clone DNA printed on microarrays

We chose two candidates to serve as internal standards for the fosmid clone DNA printed on microarrays: *FOS-cos* and *FOS-Cm*. Each corresponded to a unique feature of the pCC1FOS vector, not expected to appear in the cloned metagenomic DNA. The *FOS-cos* probe was produced by amplifying a 399bp segment of pCC1FOS corresponding to the entire *cos* site of the vector [Figure 2]. The *cos* site is the locus of DNA binding and cutting by the lambda terminase enzyme critical to λ phage packaging (169), and as such would not likely be found in the metagenome of a contaminated Arctic soil microbial community. *FOS-Cm* corresponds to a 219bp segment within the vector's chloramphenicol resistance gene, a feature much more likely found in an enteric environment or in contaminated aquaculture than in the communities sampled in this study (16, 127).

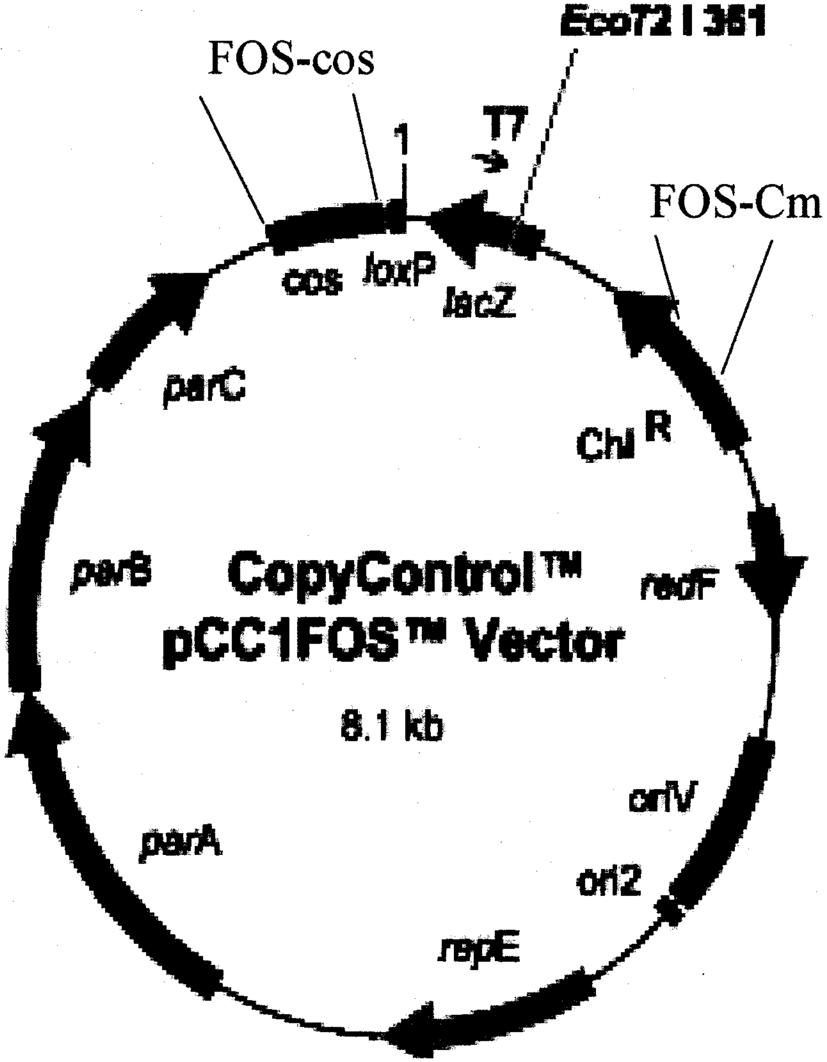


Figure 2. Origin of internal standard probe candidates. A schematic vector map of the pCC1FOS cloning vector used in this study, with major genetic elements labeled. The *FOS-cos* (bases 7684-8083) and *FOS-Cm* (bases 851-1070) amplified regions are indicated. Other notable features of this vector are the *Eco*72 I 361 blunt-end cloning site, and the extra origin of replication *ori2*, used for the induction of high-copy replication in the presence of proprietary induction solution.

Figure modified from the original Epicentre product information page for pCC1FOS, on the web at: http://www.epibio.com/item.asp?ID=385&CatID=125&SubCatID=60

2.3.3 PCR for probe positive controls

Positive controls for labelled probes in microarray hybridizations were produced by PCR as described in section 2.1.5, with the following exceptions: $MgCl_2$ was omitted from *mmoX* and *pmoA* amplifications. Following PCR, samples were purified using the QiaQuick PCR Purification kit (Qiagen) following the manufacturer's protocol, except that final elution of samples was in sterile deionized water. DNA was quantified by PicoGreen (Molecular Probes; Eugene, OR). The primers used were *FOS-cosF* and *FOS-cosR* for *FOS-cos*; *FOS-CmF* and *FOS-CmR* for *FOS-Cm*; *mmoX1* and *mmoX2* for *mmoX*; *pmoA-A189* and *pmoA-mb661* for *pmoA*; *luxAb* and *luxAe* for *luxA*; *gfpF* and *gfpR* for *gfp*; *uidA 858* and *uidA 1343* for *uidA*. Annealing temperatures were 52°C for *pmoA and FOS-Cm*, 57°C for *gfp*, 58°C for *luxA*, 60°C for *FOS-cos* and *uidA*, 65°C for *mmoX* [Table 1]. Thirty pg of plasmid (pDrive) containing each catabolic gene (cloned by Sylvie Sanschagrin at BRI) was used as positive control for all genes with the following exceptions: 30pg of pCC1FOS was used as template for both *FOS-cos* and *FOS-Cm*, while 5ng of *E. coli* genomic DNA was used as template for *uidA*. Conditions and primers for *alkB*, *nirS* and *nirK* were as described in section 2.1.5.

2.3.4 E. coli genomic DNA purification for microarray positive controls

Purification of genomic DNA from *E. coli* EPI300-T1 cells was done using the Genomic-tip System (Qiagen; Hilden, Germany) with 100/G tips for 10ml culture volumes, following the manufacturer's protocol, except that final elution was in sterile, deionized water. Purified DNA was quantified by PicoGreen (Molecular Probes; Eugene, OR).

2.3.5 Microarray sample preparation and printing

All DNA samples to be printed were loaded into 384-well microtiter plates. If sample volumes were above 80µl, the samples were first dried in a SpeedVac dessicator until their volume was less than 80µl. The samples were dessicated, then resuspended in 5µl of sterile, distilled water overnight at room temperature. Five µl of 100%DMSO was added to each sample and incubated at room temperature for 24 to 48 hours. Microarray printing was performed on GAPS II Amino-Silane-coated slides (Corning; Corning, NY) using a Virtek arrayer (Bio-Rad; Hercules, CA) with Stealth Micro Spotting pins (TeleChem International; Sunnyvale, CA). Quality control hybridizations were performed by the staff of the BRI Microarray Lab, using the Paragon DNA Microarray Quality Control Stain Kit (Molecular Probes; Eugene, OR) or the Spot QC Kit (Integrated DNA Technologies; Coralville, IA), total DNA hybridization probes based on labelled random oligonucleotides.

2.3.6 Washing and sterilizing 96-well blocks for re-use

In order to reduce equipment costs, we developed a protocol to clean and sterilize the deep-well 96-well blocks used for plasmid purification (section 2.3.1) to allow their re-use. First, 96-well blocks were autoclaved on liquid cycle (20' steam sterilizing) filled with either tap water or unused bacterial culture in an industrial autoclave (Alfa Medical; Hempstead, NY), to facilitate disposal of culture wastes by rendering them biologically inactive. Autoclaved blocks were then emptied, rinsed with water, and immersed in a diluted solution of industrial bleach for no less than 1 day, to kill any remaining microbes. Blocks were then rinsed again with tap water, then washed in an industrial labware washer (Hoplab; Beauport, Quebec). Finally, dried blocks were sterilized by autoclaving on gravity cycle (20' sterlizing, 20' drying) and left wrapped in aluminum foil until ready for re-use.

2.4 Microarray hybridization and proof-of-principle experiments

2.4.1 Probe labelling with DIG and membrane hybridization

As a preliminary test of a subset of the gene probes to be used for subsequent microarray hybridizations, PCR amplicons were labelled with dioxygenin (DIG) and hybridized to unlabelled PCR amplicons of the same genes, spotted on membranes. The labelling reaction was performed using the PCR DIG Probe Synthesis Kit (Roche; Basel, Switzerland) following the manufacturer's protocol. *FOS-Cm* and *FOS-cos* were amplified from the pCC1FOS vector, while *gfp*, *mmoX* and the *E. coli 16S rRNA* gene were amplified from genomic DNA

(*Pseudomonas Cam-1*, *Methylosinus trichosporium* and *E. coli,* respectively). Thirty pg of plasmid DNA or 5-10ng of genomic DNA was used as template for DIG-labelling. Proper labelling was verified by running 5µl of labelled product alongside unlabelled controls on a 2% agarose-TAE gel, and confirmed by the slower migration of the labelled DNA. Nylon membranes (Roche; Basel, Switzerland) were spotted with six serial dilutions of control DNA from 100fg to 10ng, using a Minifold 1 Dot-Blot System (Schleicher and Schuell; Dassel, Germany) and cross-linked to the membrane using a UV Stratalinker (Stratagene; La Jolla, CA) at 305nm for 3 minutes. Pre-hybridization was performed in 25ml of hybridization solution (5x SSC-0.1% N-lauroyl sarcosine-0.02% SDS-1% Blocking Reagent (Roche; Basel, Switzerland)) for 2 hours. Prior to hybridization, 15µl of probe was added to 25ml of hybridization solution and denatured at 100°C for 10 minutes, then placed on ice. Hybridization was performed in 25ml of hybridization solution + probe at 65°C for 1 hour. Membranes were washed twice for 15 minutes in 250ml of 2x SSC-0.1% SDS at room temperature, then twice for 15 minutes in 250ml of 1x SSC-0.1% SDS at the hybridization temperature, then once for 15 minutes in 250ml of 0.5x SSC-0.1% SDS at the hybridization temperature, with moderare agitation during all washes. Chemiluminescent detection of probes was performed as follows: membranes were stabilized for 5 minutes with moderate agitation in 150ml of detection solution 1 (0.1M maleic acid-0.15M NaCl-0.3% Tween-20 [pH 7.5]), then incubated for 90 minutes in detection solution 2 (detection solution 1 supplemented with 1% Blocking Reagent (Roche; Basel, Switzerland)) at room temperature. Antibody binding was performed in 100ml of the detection solution 2 supplemented with 0.75U of Anti-Dioxygenin-alkaline phosphatase Fab fragment, for 30 minutes at room temperature with low agitation. Membranes were washed twice in 100ml of detection solution 1 at room temperature with medium agitation, then stabilized in detection solution 3 (100mM Tris-HCl [pH 9.5]-100mM NaCl) for 2 minutes at room temperature with medium agitation. For colour detection, 25µl of CDP-Star chemiluminescent substrate (Roche; Basel, Switzerland) was added to 5ml of detection solution 3 and incubated with the membranes for 30 seconds at room temperature with low agitation, then placed in a hybridization bag. Chemiluminescence was detected by exposure to X-Omat AR film in an X-Omatic cassette (Kodak; Rochester, NY) for 5 minutes. Probe yield was assessed by spotting serial dilutions of probe from 10^{-1} to 10^{-5} on a nylon membrane and UV crosslinking as described above. The membrane was washed in detection solution 1 for 1 minute, then incubated in 100ml detection solution 2 for 30 minutes at room temperature with agitation. The membrane was then incubated in 20ml of detection solution 2 containing 3U of anti-Dioxygenin-AP for 30 minutes at room temperature, washed twice in 100ml of detection solution 1 for 15 minutes, then stabilized in 20ml of detection solution 3 for 2 minutes. For colour detection, the membrane was incubated overnight in 20ml of 100mM Tris-HCl [pH 9.5]-0.1M NaCl containing 90μl of NBT and 70μl of X-phosphate from the DIG labelling kit. Colour development was stopped by washing the membrane in 50ml of sterile water for 5 minutes.

2.4.2 Probe labelling with Cy fluorophores

In order to produce probes for microarray hybridization, 100ng of PCR-amplified or metagenomic DNA was used as template for labelling with the Cy3 or Cy5 fluorophores. Fifty pg of a *luxA* PCR amplicon (1:2000 of the total template DNA) were added to each labelling reaction to be used as a labelling control for hybridization. In each 50μl reaction, 20μl of 2.5x random octamer primer solution from the BioPrime Labelling Kit (Invitrogen; Carlsbad, CA) was added to the template DNA, and incubated for 5 minutes at 95°C, followed by 5 minutes on ice. On ice, 5μl of dNTPs (1.2mM dA/G/TTP, 0.6mM dCTP) and 2μl of Cy3- or Cy5-dCTP (0.6mM) were added. Forty units of Klenow polymerase from the BioPrime kit was added to bring the final reaction volume to 50μl. The reaction proceeded for 3 hours at 37°C, and stopped upon addition of 5μl of 0.5M EDTA [pH 8.0]. Labelled samples were purified using the QiaQuick PCR Purification Kit (Qiagen; Hilden, Germany), following the manufacturer's protocol, with the following modifications: Prior to the first buffer step (buffer PB), 2.5μl of 3M sodium acetate [pH 5.2] were added to the labelled samples. Columns were washed four times with PE washing solution, instead of once. Labelled DNA was eluted twice with 30µl of EB buffer, heated to 50°C. Samples were quantified using a ND-1000 Spectrophotometer (Nanodrop Technologies; Wilmington, DE): DNA was quantified by its absorbance at 260nm, Cy3 by its absorbance at 633nm, and Cy5 at 543nm.

2.4.3 Microarray hybridization

First, a brief word on hybridization terminology: "probe" is used here to refer to the DNA fragment of known sequence, hybridized to the "target" of generally unknown sequence. In all test array experiments described in this report, the probe was a single gene amplicon, labelled with a fluorescent dye; the target was a DNA sequence fixed on a microarray surface, either a metagenomic clone of unknown sequence (hopefully bearing a "target" gene complementary to the labelled probe), or a PCR amplicon of known sequence used as a hybridization control. These designations are in a sense the opposite of those encountered in functional gene array hybridizations of environmental samples, where the "probes" are bound to the microarray surface, while the "targets" are labelled total community DNA. However, the conceptual differentiation of unknown sequences (targets) and known sequences (probes) allows the use of consistent terminology between different microarray applications.

Microarray slides were prehybridized with 125µl of 5x SSC-0.1% SDS-1% BSA for 1 hour at 42°C, washed three times in 0.1x SSC and once in 2-propanol, and dried in a Spectrafuge Mini (Labnet; Edison, NJ) microcentrifuge for slides. Prior to hybridization, labelled probes (amplicons) or targets (metagenomic DNA) were concentrated to 2-3µl in a SpeedVac dessicator then resuspended in 20-30µl (depending on microarray coverslip dimensions) of DIG Easy Hyb hybridization reagent (Roche), supplemented with 5µg tRNA and 5µg salmon sperm DNA. Just before loading, labelled materials were denatured for 2 minutes at 95°C. Hybridization proceeded for 16 hours at 42°C. Following hybridization, slides were washed 3 times at 42°C for 10 minutes in 0.1x SSC-0.1% SDS, then rinsed 3 times in 0.1x SSC, and finally once in 2-propanol. Slides were dried by centrifugation in a Spectrafuge Mini microcentrifuge and stored in dry containers in the dark to prevent photobleaching of fluorescent dyes.

2.4.4 Microarray image analysis

Microarray hybridizations were scanned using a ScanArray Lite Microarray Analysis System (Packard BioChip Technologies; Billerica, MA) and ScanArray Express software (Perkin Elmer; Wellesley, MA) for image production, spot finding and quantification. Scans were performed at wavelengths of 633nm (Cy3) and 543nm (Cy5). Spots were quantified using the adaptive circle method to define signal and background pixels.

2.4.5 Microarray data normalization and hybridization-positive designation

Microarray data was normalized by two different methods, depending on the source of the data to be normalized. The first method, Normalization Technique A, was applied to all quantitative data generated by the ScanArray Express software. Median pixel intensity, signal-to-noise ratios and average background signal for each microarray spot were calculated by the software. The average of all backround values of each hybridization was subtracted from the median signal for each spot, to provide a hybridization signal intensity value corrected for background noise. Signal intensity values standardized by this technique are reported in the results with the designation "(corrected)".

The second standardization method, Normalization Technique B, was applied only in cases where quantitative data from multiple hybridizations were pooled for analysis. To correct for differences in probe amounts and labelling intensity, data from disparate hybridizations were normalized on the basis of *luxA* labelling control microarray spot intensity. The average intensity of all *luxA* control spots in all hybridizations to be pooled was divided by the average intensity of all *luxA* spots in a single hybridization, generating a relative correction factor for each hybridization. All signal intensity values (corrected) were multiplied by this correction factor prior to data pooling. Normalization Technique B was only used in conjunction with Normalization Technique A. Thus, all signal intensity values derived by this method are reported in the results with the designation "(corrected and normalized)".

Designation of a microarray spot as hybridization-positive was based on the signal-to-noise value for that spot. A signal-to-noise ratio greater than or equal to 3 was considered to constitute a positive signal. When calculating average hybridization signal intensity for various figures, only hybridization-positive spots were considered.

2.4.6 DNA sequence analysis

Pairwise sequence alignments of *FOS-cos, mmoX, GFP, pmoA* and the pCC1FOS vector were performed using the MacVector program (Accelrys; San Diego, CA) to identify any areas of sequence similarity. Nucleotide-nucleotide alignments were also performed using Basic Local Alignment Search Tool (BLASTn) provided by the National Center for Biotechnology Information (6) (http://www.ncbi.nlm.nih.gov/BLAST).

The *alkB* PCR amplicon produced from the *alkB*-positive clone 1A3-18 F10 was isolated for sequencing by excision from EtBr-stained 1.4% TAE-agarose under UV illumination, and purified using the Ultrafree-DA Centrifugal Filter Unit (Millipore; Billerica, MA), following the manufacturer's instructions. DNA sequencing services were provided by the McGill University and Génome Québec Innovation Centre. To identify the sequenced fragment, single-stranded DNA sequence was converted to amino acid sequence using the DNA-protein sequence conversion tool provided online by the ExPASy proteomics server of the Swiss Institute of Bioinformatics (http://ca.expasy.org). Candidate protein sequences were compared with all entries in the GenBank protein sequence database using the BLASTp search program.

2.4.7 PCR screening of metagenomic library

The 1A3 metagenomic library was screened for the presence of *alkB*, *nirS* and *nirK* by PCR. DNA was extracted from cell culture by boiling lysis: briefly, culture was incubated for 5 minutes in a boiling water bath or in a themal cycler at 99°C, then centrifuged at ~1,800 x g for 5 minutes to pellet cellular debris. 1 μl of lysate was used as template for PCR. First, PCR was performed on each of 53 pools of clones, representing each 96-well microtiter plate in the library. Clones from every plate identified as bearing the desired gene were then pooled into column and row pools (a total of 20 for each plate) and lysed, then were subject to another round of PCR. The results of this second PCR identified the gene-positive clones by providing row and column coordinates. PCR conditions were as described in section 2.1.5.

3 Results and Discussion

3.1 DNA sample characterization

After extracting total community DNA from samples BRI-1A3 and BRI-6A3, dilutions of 1:1, 3:10, 1:10 and 3:100 of the extracts were run on a TAE-agarose gel to confirm successful extraction, and to quantify the extracted DNA using the 10kb band of the High DNA Mass Ladder (Invitrogen) as a reference [Figure 3a]. The results indicate that community DNA was successfully extracted. Before undertaking the construction of metagenomic libraries from 1A3 and 6A3, these samples were subjected to a short series of tests aimed at differentiating the two samples and confirming their suitability for subsequent manipulation and analysis.

Using size-selected, gel purified DNA as a template, PCR amplification of the 16S rRNA gene was performed to confirm that DNA was successfully extracted and purified from the soil samples. Universal bacterial primers were used to amplify the 16S rRNA gene from several dilutions of 1:1, 1:3, 1:10, 1:30 and 1:100 of each sample (corresponding to 30ng to 300pg for 1A3, and 60ng to 600pg for 6A3) [Figure 3b]. In all samples the correct 16S rRNA fragment was successfully amplified, confirming successful extraction of community DNA of sufficient purity for enzymatic manipulation.

The above test was also used to determine the optimal dilution of extracted DNA to be used for subsequent amplification of the 16S rRNA gene for DGGE analysis. For both samples, 1:30 was chosen because this dilution produced a minimum of multiple banding, visible in Figure 3b as a single thick band, as visualized on TAE-agarose. DGGE was performed first on a denaturant gradient of 40-80% [Figure 4a], and the resulting profiles were expanded by performing additional DGGE on denaturant gradients of 40-60% and 60-70% [Figure 4b, 4c]. DGGE is often used in environmental studies to differentiate samples based on taxonomic (sequence) differences, in band intensity or for the presence or absence of specific bands (150). As can be seen in Figure 4, the DGGE profiles from the two samples exhibited a strong degree of similarity, but also possessed a number of unique or enriched bands. Sorensen's coefficients of similarity (S_{AB}) between samples were 0.41 for the 40-80% gradient, 0.71 for the 40-60% gradient, and 0.91 for the 60-70% gradient. This suggests that the two samples are different but related, a likely conclusion given the origin of the samples.

The DNA samples were also tested for the presence of the catabolic genes *alkB*, *nirS* and *nirK*. Since the metagenomic libraries from samples 1A3 and 6A3 were to be screened for the presence of these genes, it was necessary to first establish that they were present in the extracted total community DNA. PCR amplification of each of these genes was performed on serial dilutions of both size-selected, gel-purified samples. All three genes could be amplified from both samples; both *alkB* and *nirS* could be amplified from 1:10 sample dilutions, while *nirK* could only be amplified faintly from the undiluted samples (data not shown).

3.2 Metagenomic library production

There are numerous options available in the literature for cloning high molecular weight DNA extracted from environmental samples; the choice of DNA extraction methods, cloning vectors and cloning strategies each present several alternatives with various advantages and disadvantages. We briefly explored the possibilities of extracting very high molecular weight DNA from agarose plugs (24, 117, 142) and agarose microbeads (83, 172) and cloning in BACs (142). However, we abandoned these efforts in the face of a number of technical obstacles and a lack of available expertise with these methods. In the end, we chose to use the CopyControl Fosmid

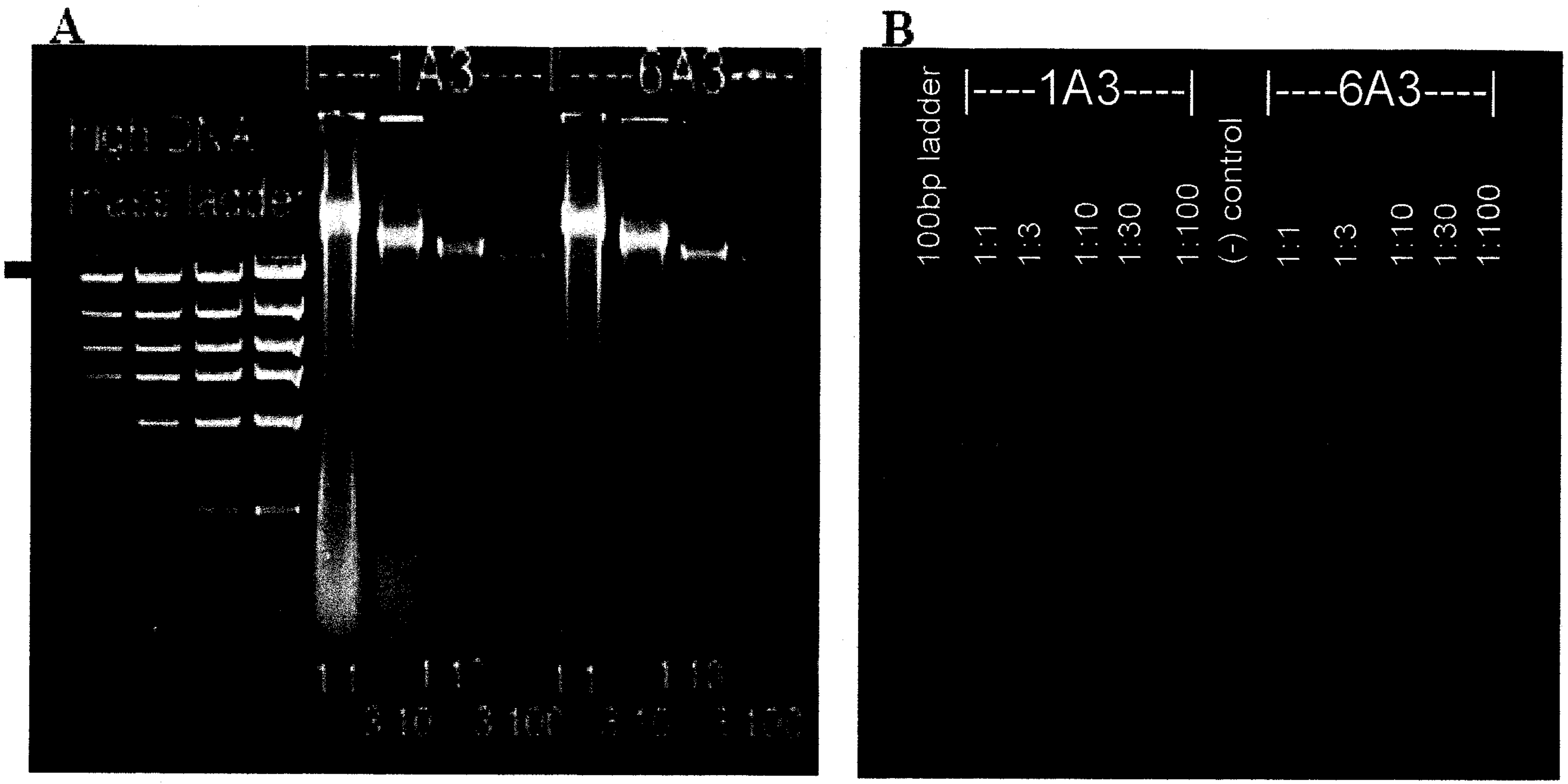


Figure 3. Initial DNA sample characterization. (A) Community DNA extracts of samples 1A3 and 6A3, resolved on 1% agarose-TAE stained with ethidium bromide. Arrow indicates 10kb band in High DNA Mass Ladder. (B) 417bp 16S-PCR amplicons of both samples, resolved on 1.4% agarose-TAE stained with ethidium bromide. PCR negative control (lane 7): no DNA.

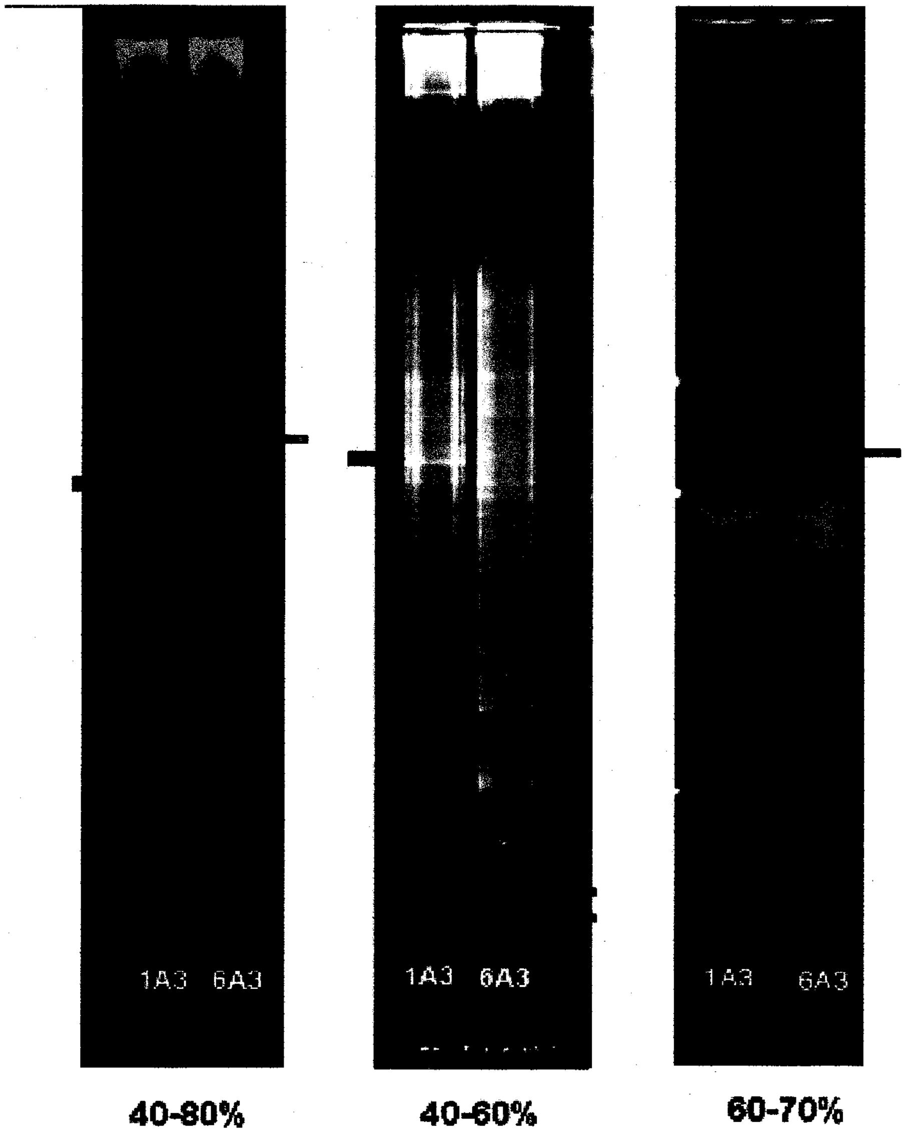


Figure 4. DGGE profiling of samples 1A3 and 6A3. Banding patterns of the two samples were compared on three different denaturant gradients: 40-80% (left), 40-60% (middle) and 60-70% (right). Patterns were mostly similar, but unique or enriched bands were visible at all three denaturant gradients, most prominently at 40-60% denaturant. Arrows indicate examples of unique or enriched bands.

Library Production Kit (Epicentre; Madison, WI) because of the many advantages associated with this method, discussed below.

One of the great advantages of any λ phage-based cloning system is that phage particles will only properly package DNA from a very specific size range, between 38kb and 51kb of total vector-plus-insert (48, 49). Using the 8139bp pCC1FOS vector, this translates into insert sizes of approximately 30-43kb. When performing DNA extraction by the method described in section 2.1.1, we noticed that DNA often fell into this size range with no additional manipulation required. Thus, cloning in the CopyControl Fosmid system allowed us to make use of a DNA extraction method already developed by the Environmental Microbiology group at BRI. The main modifications we made to this method, namely the omission of steps designed to remove RNA and co-extracted organic acids, came as a natural consequence of the size-selection process since RNA and organic acids (such as humic and fulvic acids) migrate much farther down the gel than the 30-50kb fragments that were excised for cloning. Indeed, this RNA is clearly visible at the bottom of the DNA extract lanes in Figure 3, while the organic acids were visible in the gel as an orange/brown stain that roughly co-migrated with the RNA (not visible in Figure 3).

Other advantages of the CopyControl Fosmid system similarly simplified the task of metagenomic library production. The blunt-end cloning strategy avoided the need for restriction endonuclease digestion of extracted DNA to generate compatible ends for cloning. Instead, end-repair of the DNA fragments sheared during the extraction process and cloning into the blunt *Eco*72 I cloning site maximized the recovery of extracted DNA fragments that fell into the critical size range. At the colony-picking stage, selection of transformants was guaranteed simply by the growth of colonies in the presence of chloramphenicol: since phage packaging would not occur unless 38-51kb of total DNA was present, and since the vector bearing the chloramphenicol resistance determinant was only ~8kb, this guaranteed at least 30kb of metagenomic DNA in every colony (concatomerization of vector alone into 38-51kb units was impossible since all vector was dephosphorylated).

One of the main landmarks of this study was the successful production of two 5,000-clone metagenomic libraries from each of the soil samples BRI-1A3 and BRI-6A3. Once these libraries were constructed, 48 clones from each library were chosen to be characterized by RFLP. *Not*1 was chosen for this purpose, a rare-cutting restriction enzyme with an eight-nucleotide recognition site. The pCC1FOS vector has 2 *Not*1 sites flanking the cloning site at nucleotide positions 1-8 and 642-649; thus digestion of clone DNA with *Not*1 produces a vector band of 7,490bp and a banding pattern unique to the insert DNA of each clone [Figure 5]. Average clone sizes were 34.8kb ± 1.15kb (standard error) in the 1A3 clones and 35.7kb ± 859bp in the 6A3 clones, which corresponds well to the expected range of 30-43kb. Visual inspection of the banding patterns of 48 clones from each library reveals only one possible occurrence of multiple identical clones in the 6A3 subset (indicated by vertical arrows), and none in the 1A3 subset, indicating that there is very little overlap between library clones. This is hardly surprising given the degree of metagenomic coverage represented by 5,000 clones, which can be illustrated by restating equation (1) in terms of P (the probability that a specific sequence is represented, a measure of library coverage):

(3) $P = 1 - (1 - L/G)^N$

If we retain our earlier assumptions for the value of G, set N at 5,000 clones, and assume an average insert size of 35.25kb (the average of the 1A3 and 6A3 libraries), we can see that 5,000 clones only provides a 1.26% probability of locating a specific target sequence. Despite this very low degree of coverage, we decided that this was an appropriate number of clones because the goal of this project was to develop methods and construct small-scale prototypes; 5,000 clones per library was large enough to warrant the use of high-throughput robotics, yet small enough to be logistically manageable. As well, the numbers derived above are somewhat misleading: the figure of 1.26% is a base probablility, representing the odds of locating a specific sequence *that occurs only once in the community metagenome*. We chose target genes that were known beforehand (in the case of the alkane degradation gene *alkB*, (85)) or assumed (in the case of the denitrification genes *nirS* and *nirK*) to be present in greater proportion, owing to the biological properties of the soil community, and thus were more likely to be found even among a small metagenomic sample.

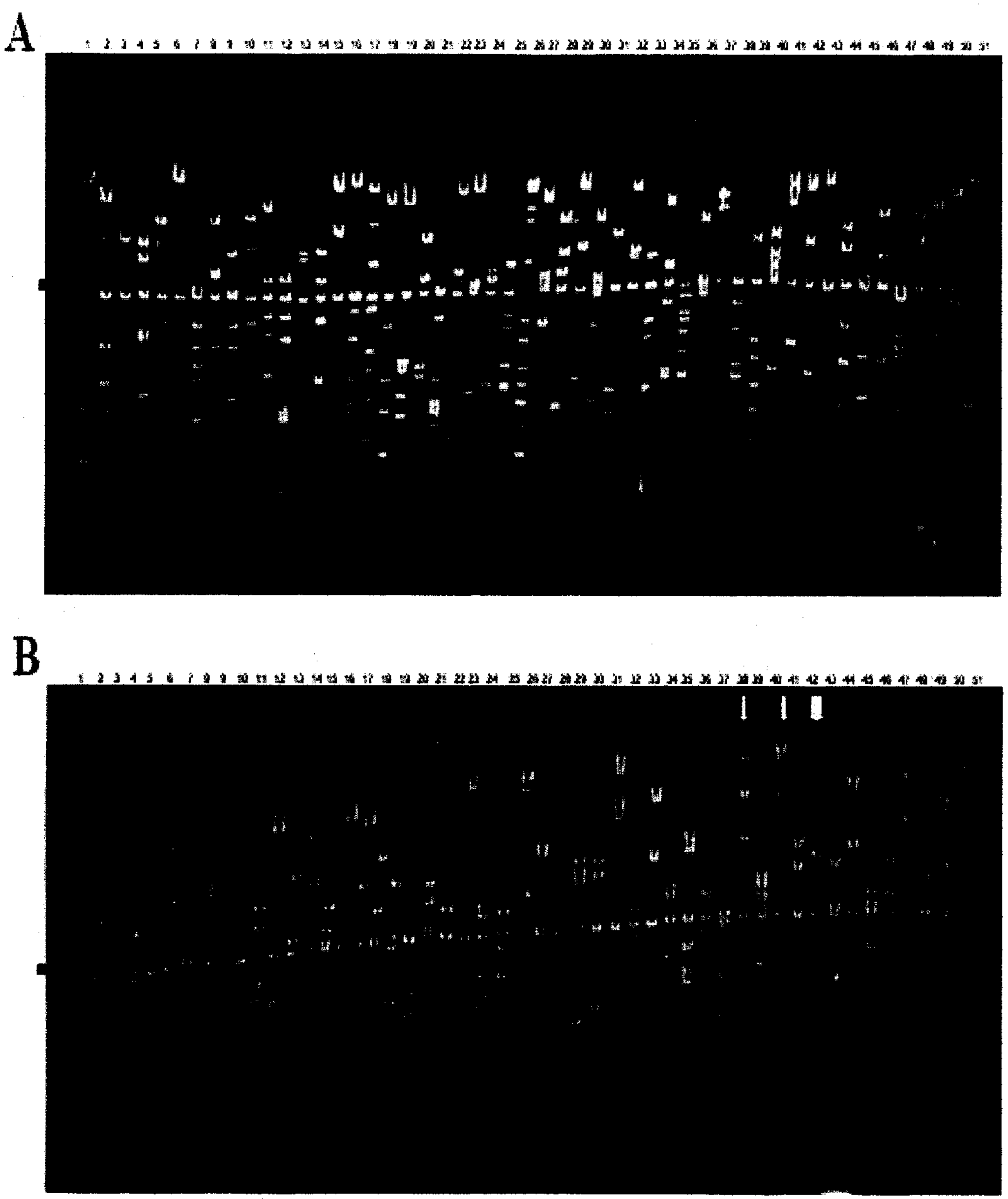


Figure 5. Restriction fragment length polymorphism (RFLP) analysis of 48 sample clones from 1A3 and 6A3 libraries digested with *Not*I. Size markers, a combination of 1kb DNA ladder (1-10kb) and Lambda Mono Cut Mix (10-48kb), are indicated with an "L" (lanes 1, 26 and 51). Red arrows indicate the 7,499bp vector band. (A) 1A3 library plate 49, clones A1-D12. (B) 6A3 library plate 11, clones E1-H12.Yellow arrows indicate possible clone duplications.

3.3 Development of an automated fosmid purification protocol

One of the major goals of this project was to develop a protocol to purify cloned fosmid DNA from its bacterial host. The primary consideration in the development of this protocol was that it should be in a high-throughput format that would allow rapid purification of an entire 5,000-clone library. To this end, the protocol was designed for use on an automated liquid handler, the Biomek FX Laboratory Automation Workstation (Beckman Coulter; Fullerton, CA), in a 96-well format. There were several secondary considerations guiding protocol design as well. First, the purified fosmid DNA had to be of sufficient quality to allow for printing on microarrays. Second, the protocol had to incur a minimum of cost, in particular avoiding the use of commercial kits. Third, the protocol had to be appropriate to the equipment available for use at the BRI. A great deal of optimization was required for the development of this protocol, due in no small part to the fact that these various considerations were often at odds with one another. The protocol as it appears in section 2.3.1 represents the finalized version. The modifications and optimizations that were performed to produce this final protocol are discussed below.

The basis for the fosmid purification protocol developed in this study are the protocols for purification of BAC and plasmid DNA by alkaline lysis, from the third edition of *Molecular Cloning: A Laboratory Handbook* (126). These protocols were modified according to the dictates of the chosen cloning system, the need for a high-throughput format, and the desired end-use of the purified fosmid DNA (printing on microarrays).

3.3.1 Modifications without optimization

Some modifications were made to the template protocols immediately, without optimization. The first was the addition of a second culturing step: the standard overnight clone culture was used to inoculate a day culture grown in the presence of CopyControl Induction Solution. This is an essential step in the

CopyControl cloning system, where this patented solution is used to activate replication of the normally low-copy number vector at a level of up to 50 copies per cell. Another modification was the addition of an RNAse digestion step, followed by the removal of RNAse enzyme by precipitation with ammonium acetate. Since the purified nucleic acids were to be printed on microarrays, it was necessary to remove the RNA as it would dilute the fosmid DNA when printed on the array. Similarly, the RNAse needed to be removed since residual enzyme would interfere with the DNA printing process. Phenol:chloroform extraction to remove RNAse was not feasible in the 96-well format, therefore we chose to incorporate precipitation with ½ volume of 7.5M ammonium acetate based on reports of the success of this method in the scientific literature (35).

In order to accommodate the 96-well format, a number of other modifications were introduced to the standard alkaline lysis protocol as well. Although the 96-well plates used in these experiments are commercially listed as having a 2ml capacity, we found it impractical to work with volumes greater than 1.8ml during the automated process, for fear of cross-contamination between wells. For this reason we chose to perform the post-lysis alcohol precipitation in 2-propanol rather than in ethanol, since the former requires the addition of only 1 volume for precipitation, as opposed to 2-2.5 volumes of the latter. This in turn allowed us to maximize reaction volumes at the alkaline lysis stage, ensuring a more complete lysis of the harvested cells. During the cultured growth stages of the fosmid purification protocol, maximum volumes were restricted still further to no more than 1.2ml, to ensure that no cross-contamination would occur as a result of shaking at 300rpm (to increase culture aeration). Another limitation imposed by the 96-well format was the inability to transfer supernatants post-centrifugation by decanting. Our solution was to program the Biomek FX to gently pipet the supernatant from a height of 2mm above the bottom of the wells, to avoid disrupting the pellet. For the precipitation and supernatant transfer immediately following alkaline lysis, this strategy resulted in the transfer of visible amounts of cellular debris with the supernatant, therefore we added a second supernatant transfer step to remove all visible traces of the unwanted cell debris. Although 96-

well PVDF filters (Corning; Corning, NY) are a commercially-available alternative to centrifugation and supernatant transfer, we avoided their use because this would represent a significant increase in cost if used on a library-wide scale.

3.3.2 Optimization of precipitant solution

Other modifications to the standard alkaline lysis protocols required optimization. To precipitate cellular debris from the lysate, the standard protocols call for the use of potassium acetate. However, the staff of the BRI Microarray Lab advised us to avoid introducing potassium ions into our purified DNA, as even trace amounts of these ions can cause large irregularities in microarray spot morphology and printing efficiency. Consequently, we tested three additional precipitants for their ability to recover purified fosmid DNA from the lysate solution: ammonium acetate, sodium acetate and the cationic detergent cetyltrimethylammonium bromide (CTAB) were each compared to potassium acetate in terms of the yield of DNA recovered from the fosmid purification procedure, and in terms of sample purity as measured by the ratio of spectrophotometric absorbance at 260nm (DNA) and 280nm (protein). Each precipitant was used to prepare the same 96-well plate of fosmid library clones in the manner described in section 2.3.1. As can be seen in Table 2, ammonium acetate is superior to CTAB and sodium acetate both in DNA recovery and sample purity, and comparable to potassium acetate in both parameters. Thus, ammonium acetate replaced potassium acetate as the precipitant of choice for the automated fosmid purification protocol.

3.3.3 Optimization of centrifugation conditions

Another set of optimizations was concerned with DNA precipitation and centrifugation. One of the limitations of working in the 96-well format is that the available 96-well blocks could only be spun at a maximum speed of 1,811 x g, or 3150rpm in a Beckman Allegra-6 swinging-bucket centrifuge (Beckman Coulter

Inc; Fullerton, CA). However, the protocols used as template for our fosmid purification protocol repeatedly called for centrifugation to pellet DNA at speeds well in excess of 10,000 x g (at maximum speed in a microfuge). We set out to test if the speed constraints imposed by the 96-well format would cause unacceptable reductions in DNA recovery. In addition, there was disagreement about centrifugation temperature among the template plasmid purification protocols and various DNA precipitation protocols, whether samples should be centrifuged at 4°C or at room temperature. This question was highlighted by the fact that the 96-well plate centrifuges available were not equipped for 4°C centrifugation. Thus, we sought to test what effect the twin limitations of our centrifugation equipment might have on DNA recovery by our fosmid purification method.

In order to centrifuge samples faster than 1,811 x g, these tests were performed in microcentrifuge tubes, which were not bound by the same speed constraints as 96-well plates. The two centrifugation parameters of speed and temperature were varied in a binary fashion: centrifugations

Table 2. Optimization of automated fosmid purification protocol: choice of precipitant

precipitant	[DNA] (ng/ul)	OD 260/280*
CTAB	4.0 ± 1.6	1.16
NaOAc	105.8 ± 23.8	1.09
KOAc	353.5 ± 79.8	1.34
NH4OAc	394.0 ± 57.3	1.39

* values are averaged over 12 replicates

proceded at 15,000 x g or 1,811 x g, and were performed at 4°C or at room temperature. Thus, replicate samples of a single fosmid clone were subjected to a set of 4 different centrifugation treatments. The results indicated that the slower centrifugation speed of 1,811 x g resulted in far greater DNA recovery then centrifugation at the higher speed [Figure 6a]. Since this parameter had such a large effect on DNA recovery, data from the different speeds are presented separately in the analysis of centrifugation temperature [Figure 6b]. This latter parameter was found to have a small but significant effect on DNA recovery at both centrifugation speeds, with slightly higher DNA recoveries for room temperature centrifugations.

This surprising result that the slower centrifugations recovered more DNA may well have been due to an uncontrolled variable in this experiment. In order to minimize the expected DNA losses at the slower centrifugation speed, all 1,811 x g centrifugations were extended to 30 minutes, while the high-speed centrifugations lasted for 5 minutes as recommended by the template protocols. Thus, the higher DNA yields may have been as much a factor of centrifugation time as of centrifugation speed. In effect, these experiments were not so much an outright optimization as a comparison between the conditions recommended by the template protocols and the conditions imposed by the limitations of the centrifugation equipment. Fortunately, as the results indicated, our constraint-imposed modifications were not only comparable but superior to the conditions recommended by our template protocols.

It should be mentioned that in all the protocol optimizations discussed thus far, a determination of superior total DNA yield was assumed to signify a superior yield of clone DNA as well. It is possible that superior yields may have been produced as a result of increased recovery of host genomic DNA alone. However, had this possibility been explored and found to be true, the final optimized conditions would most likely not have changed. Certainly, no acceptable alternatives existed for centrifugation conditions, due to equipment limitations. As for precipitant solution optimization (section 3.3.2), the only viable alternative to ammonium acetate was sodium acetate, since potassium ions could not be used at all, and use of CTAB resulted in near-total DNA loss (Table 2). With a total DNA yield using ammonium acetate nearly four times that obtained using sodium acetate, it strains credibility to assume that so large a difference could be accounted for solely by increased recovery of host genomic DNA. In subsequent optimization experiments conducted using clones printed on microarrays, this question of host genomic DNA contamination was addressed directly, using the *E. coli*-specific *uidA* gene probe.

3.3.4 Optimization of induction culture time/OD$_{600}$

The duration of the induction culture step of the fosmid purification protocol, during which fosmid clones were induced to high-copy replication, was another key parameter optimized. More extensive growth of the cloning host clearly produces a greater amount of fosmid DNA. However, there was some concern that if cultures were grown for too long, past the exponential phase and into the stationary phase, an excess of cellular debris and exopolysaccharides might reduce fosmid DNA purity and interfere with the microarray printing process. This question was addressed twice over the course of this project. In the first instance, it was approached in terms of total time of growth: on the first test array (Test Array 1 or TA 1) two 96-well plates of clones were printed in triplicate spots, one for which the induction culture lasted for 3 hours, and one which had been grown for 5 hours. The 3 hour time was selected based on prior quantification of fosmid preparations from single clone cultures grown at intervals from 2 hours to 6 hours, because after 3 hours DNA yield was approximately 75% of its maximal value (data not shown); the 5 hour time was selected because this was the time suggested by the manufacturer of the cloning system. When 1μg of a labelled probe specific to the fosmid vector DNA (*FOS-cos,* see section 3.4.1) was hybridized to Test Array 1, a simple visual inspection of the resulting hybridization profile was enough to confirm that the signal from the 3-hour clones was inferior to that from the 5-hour clones, and indeed was insufficient for informatic signal detection [Figure 7]. Informatic signal detection from a series of similar hybridizations with amounts of *FOS-cos* probe ranging from 100ng to 4μg added a quantitative dimension to this result: only 1.2% of all 3-hour clone spots were detected, compared to 40.9% of all 5-hour clone spots (data not shown). (Specifics of test array probes, controls and experiments will be discussed in greater detail in section 3.4). Based on the results of this experiment, the induction culture stage of the fosmid purification protocol was set at 5 hours, and fosmid clones for all test arrays prior to Test Array 5.1 were prepared in this manner.

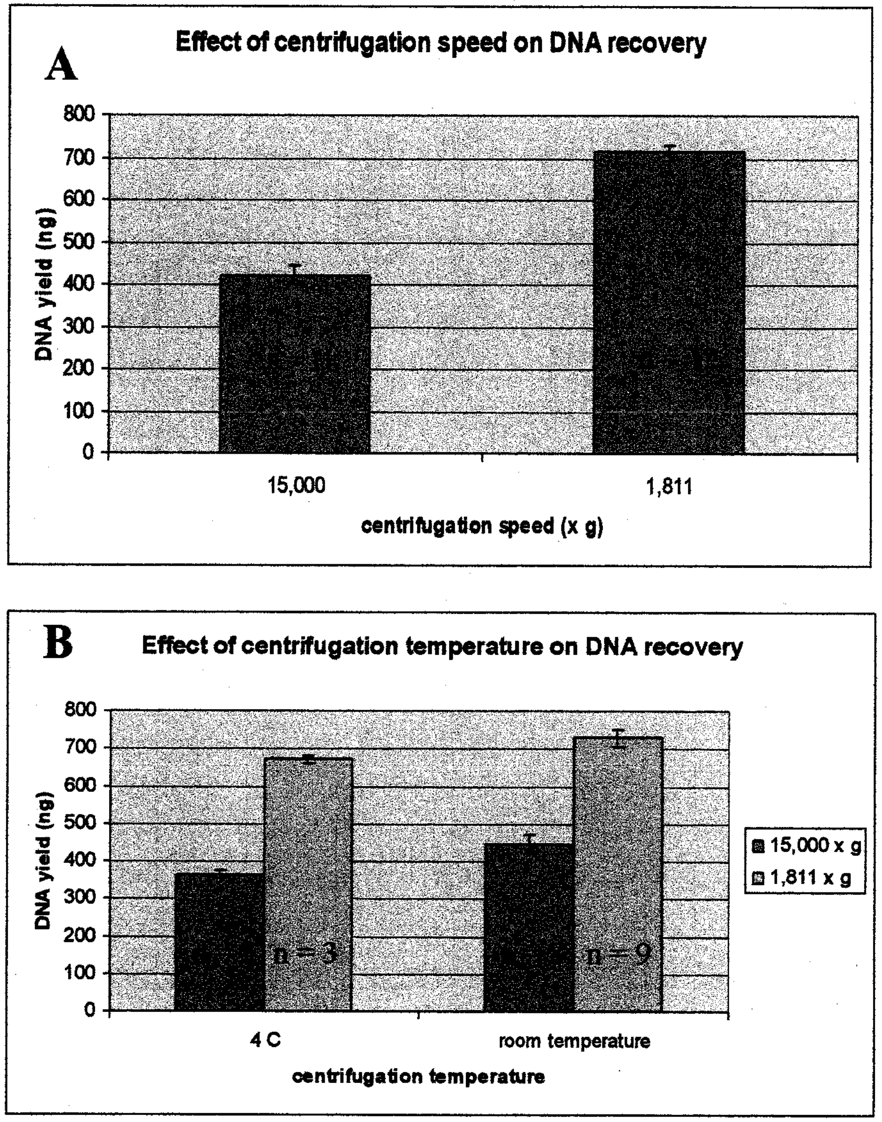


Figure 6. Optimization of automated fosmid preparation: centrifugation conditions. Fosmid purification from samples subjected to different centrifugation speeds/times and temperatures were compared to determine the losses in DNA yield, if any, due to limitations of the available centrifugation equipment. Sample sizes for each data point are indicated. (A) Comparison of DNA yield from samples centrifuged at 1,811 x g for 30 minutes and samples centrifuged at 15,000 x g for 5 minutes. (B) Comparison of DNA yield from samples centrifuged at 4°C and at room temperature. Data from the two centrifugation speed variables are presented separately to better visualize the effect of centrifugation temperature.

Note: error bars in all figures in this report denote standard error of the mean.

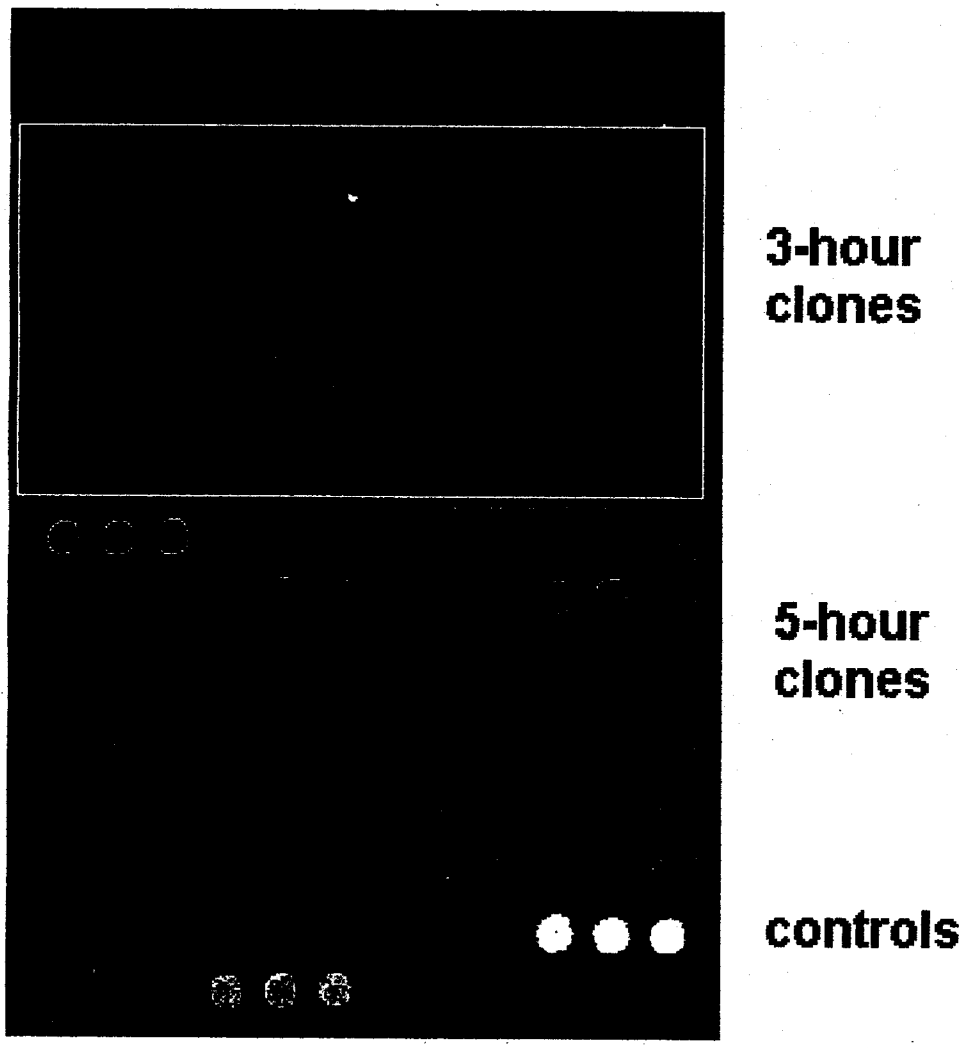


Figure 7. Clone-specific hybridization to 3-hour and 5-hour clones. The image represents a single subarray from Test Array 1, hybridized to 1µg of the fosmid internal standard probe *FOS-cos* (Cy3). These results are representative of the other subarrays of this hybridization, and the other hybridizations of this probe on this test array series. Specific probes, controls and test arrays are discussed in greater detail in section 3.4.

Signal intensity key: Blue < Green < Yellow < Red <White (saturated)

Optimization of induction culture incubation time was more thoroughly explored in an experiment on Test Array 5.1. In this experiment, 4 different clones were grown to an OD_{600} (spectrophotometric absorbance at 600nm) of 0.6, 0.8, 1.1 or 1.4. No higher OD_{600} was explored, as cultures grown in our chosen 96-well format often tended to reach saturation between OD 1.2 and OD 1.6 (data not shown). Fosmid DNA was then purified, and DNA from each fosmid preparation was printed on a microarray in triplicate spots at a uniform concentration of 200ng/µl (TA 5.1).[*] OD_{600} was chosen for this experiment as a more standardized measure of growth than incubation time, as we had noticed a great deal of variation in OD_{600} between cultures that had been grown for a set time in previous experiments. Microarray hybridizations were performed with 1, 2, 3, 4, and 5µg of vector-specific *FOS-cos* (Cy3) probe (see section 3.4.1) and data from all 5 hybridizations were normalized based on the *luxA* labelling control and pooled. Hybridization intensities for the 4 different OD_{600} values were compared, and are presented in Figure 8.

DNA from cultures grown to an OD_{600} of 1.4 produced the greatest vector-specific hybridization signal [Figure 8a]. In addition, clones grown to an OD_{600} of 1.4 were detected at a rate approximately 7.2x higher than clones grown to an OD_{600} of 0.6, and approximately 1.5x higher than clones grown to an OD_{600} of 0.8. Since all clones were printed at a total DNA concentration of 200ng/µl, there are a few possible explanations for these results. One possibility is that clones grown to a lower OD_{600} had less time to induce high-copy fosmid replication, resulting in a higher proportion of host genomic DNA in the final fosmid purification product. Since PicoGreen quantification allows no distinction between fosmid DNA and host genomic DNA, there was no way to determine,

[*] Another experiment on Test Array 5.1 compared the effects of 5 different fosmid purification regimes on hybridization signal. Clones from each of the 5 treatments were grown to each of the 4 OD values. The data used to calculate average values for each of the 4 ODs were averaged over all 5 treatments. In theory, this produced 20 sets of triplicate data points for each OD value per hybridization. This number actually ranged higher or lower, because some clones were never printed due to insufficient materials, and because additional clone series from some OD groups were printed for other experiments contained on TA 5.1. The actual number of triplicate spot sets for each OD value in each hybridization was 12 for OD 0.6, 15 for OD 0.8, 18 for OD 1.1 and 44 for OD 1.4.

prior to printing, how much of the 200ng/µl fosmid prep sample was in fact fosmid DNA. However, hybridization of Test Array 5.1 with 500ng of the *E. coli*-specific probe *uidA* (see section 3.4.1) suggested that this is not the case, or at least not the entire explanation [Figure 8b]. At the lowest OD_{600}, there was no detectable signal from *uidA* whatsoever. The *uidA* signal did increase significantly up to OD 1.1, but was lower at OD 1.4 than at OD 0.8 or OD 1.1. Interestingly, the proportion of clones that hybridized to the *uidA* probe varied less than 10% between the three highest OD_{600} values despite significant differences in signal intensity.

The fact that equal microgram amounts of similar proportions of *uidA*-positive clones produced less *E. coli*-specific hybridization signal when grown to the highest OD_{600} suggests a selective enrichment of fosmid DNA at this OD_{600}, a finding supported by the vector-specific hybridization results [Figure 8a]. However, the coincidence of lower vector-specific and *E. coli*-specific hybridization signal from the two lowest OD_{600} data sets suggested that another factor was lowering overall hybridization signal despite the equal amounts of total DNA printed for each sample. Thus, another possibility is that DNA preparations from lower-OD_{600} cultures contained a larger amount of contaminating materials, since they required the concentration of much larger volumes of fosmid purification product to attain the 200ng/µl printing concentration. These contaminating materials may have interfered with DNA printing on the arrays and consequently reduced any hybridization signal that could be retrieved from these spots.

Based on the results presented in Figure 8, we selected a culture OD_{600} of 1.4 as optimal for our fosmid purification protocol. This represents a growth time longer than that recommended by

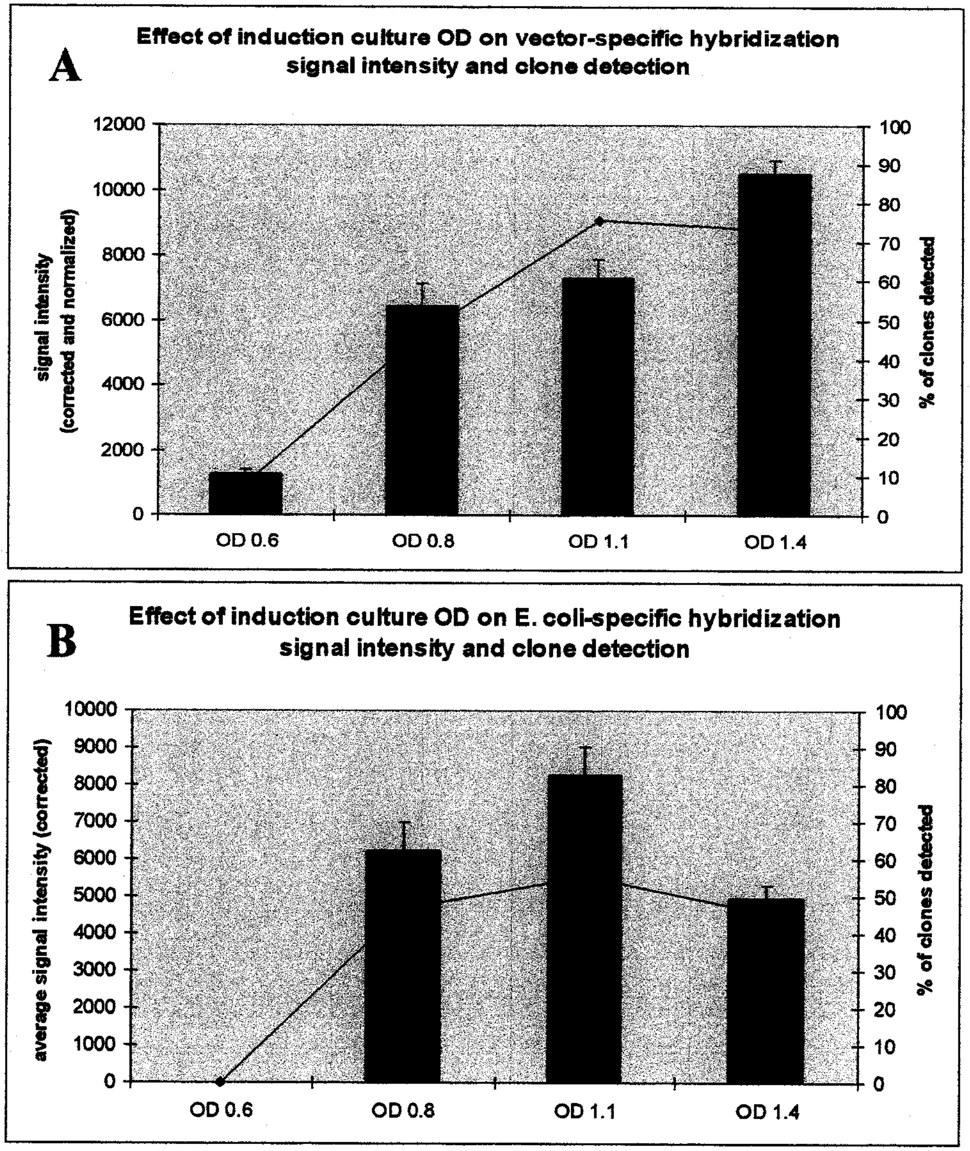


Figure 8. Optimization of automated fosmid preparation: culture OD_{600}. Different culture OD_{600} values were compared on the basis of average clone hybridization signal intensity (bars) and clone detection rate (lines) from hybridizations to Test Array 5.1. Clone detection rate was defined as the number of triplicate clone spots informatically detected as hybridization-positive, as a percentage of the total triplicate clone spots printed on the array. The signal intensity value designations of "corrected" and "corrected and normalized" are defined in section 2.4.5. (A) Composite quantitative data from hybridizations of 1-5μg of vector-specific *FOS-cos* (Cy3) probe (B) Quantitative data from hybridization of 500ng of *E. coli*-specific *uidA* (Cy5) probe

the manufacturer of the cloning system, who suggests that clones be harvested while cultures are still in the exponential phase of growth. The decision to harvest cells during the stationary phase was made despite the possibility that the results may have been skewed to favour the highest OD_{600} by virtue of the disproportionate amount of contaminants in the other OD_{600} samples on Test Array 5.1. This decision can be justified for two reasons. First, the average volume of automated fosmid preparation material used to prepare the printed OD_{600} 1.4 samples was 56μl, approximately equal to the 60μl volume (2 sets of fosmid preparations) of material chosen as the optimal microarray printing quantity (see section 3.4.3). Thus, results of this experiment reflect the amount of DNA and contaminants selected as optimal for microarray printing in other experiments. The second reason is one of practical feasibility: to attain the fosmid DNA concentrations necessary for proper microarray printing, from cultures grown to lower OD_{600} values, would require an impractically large number of fosmid preparations for each plate of clones (at least 3), and thus an unacceptable cost in time and materials.

3.3.5 Optimization of additional purification steps

The final optimization of the automated fosmid purification protocol aimed at reducing the amount of contaminating non-DNA material in the final fosmid preparation. Other automated alkaline lysis methods commercially available or reported in the literature incorporate an extra element of purification, either by filtration of the cellular lysate (116, 124) or by affinity-binding and washing of plasmid DNA (74). In this series of optimization experiments, we added three separate purification elements to our automated protocol, singly and in combination: proteinase K treatment to remove proteins, size-exclusion filtration to clarify the cellular lysate, and glass fiber DNA binding to remove impurities from the final fosmid DNA product. These modifications were compared both to the standard automated protocol and to a commercial plasmid preparation representing the best possible level of purificaiton. Altogether, a total of 9 different treatments were initially compared. As part of the OD_{600} optimizations discussed in section 3.3.4, each of the treatments was performed on 4 different clones grown to the 4 experimental OD_{600} values, for a total of 16 samples per treatment.

Proteinase K treatment consisted of an initial incubation in 600μg/ml proteinase K immediately prior to alkaline lysis. During alkaline lysis, the concentration of proteinase K was reduced to 200μg/ml by addition of lysis solution; it was at this stage that most protein digestion was meant to occur, once cells had been lysed and intracellular proteins were exposed. Size-exclusion filtration was performed using a 0.2μm polyvinylidene fluoride (PVDF) filter (Corning; Corning, NY). Glass fiber DNA-binding was accomplished using a Multiscreen-FB Glass Fiber Type B 1μm filter (Millipore; Billerica, MA). The commercial kit fosmid purification was done using Plamid Midi and Maxi Kits (Qiagen; Hilden, Germany).

The initial basis for comparison of the different treatments was DNA yield. On the basis of these results, presented in Table 3, we eliminated further analysis of all treatments that used glass fiber DNA-binding, as virtually all DNA was lost in these samples. Of all the purification treatments except for the commercial kit, the proteinase K treatment gave the highest average yield, almost exactly double that of the basic method. The PVDF treatment yielded approximately as much DNA as the basic method, but that yield increased nearly to the level of the proteinase K treatment when the two purification methods were combined, which further suggests that proteinase K alone is responsible for greatly increasing DNA recovery. It is possible that the incubation with proteinase K prior to alkaline lysis increased the efficiency of cell lysis by

Table 3. Optimization of automated fosmid purification protocol: purification treatment

treatment	OD	average [DNA], all ODs (ng/ul)	treatment average [DNA] (ng/ul)
proteinase K	0.6	15.45	26.69
	0.8	23.37	
	1.1	21.08	
	1.4	46.87	
PVDF filter	0.6	1.00	12.25
	0.8	6.93	
	1.1	11.25	
	1.4	29.82	
glass fiber	0.6	0.03	0.54
	0.8	0.29	
	1.1	0.26	
	1.4	1.58	
proteinase K + PVDF	0.6	10.22	21.41
	0.8	17.73	
	1.1	19.75	
	1.4	37.93	
protienase K + glass fiber	0.6	0.10	0.57
	0.8	0.21	
	1.1	0.27	
	1.4	1.68	
PVDF + glass fiber	0.6	0.02	1.22
	0.8	0.26	
	1.1	0.47	
	1.4	4.11	
protienase K + PVDF + glass fiber	0.6	0.28	1.37
	0.8	0.62	
	1.1	0.78	
	1.4	3.79	
basic method*	0.6	0.45	13.30
	0.8	5.98	
	1.1	7.82	
	1.4	38.94	
commercial kit**	0.6	6.18	197.43
	0.8	35.81	
	1.1	159.29	
	1.4	588.42	

* the basic method is the same as described in section 2.3.1, minus the proteinase K step

** culture volumes were 20x larger, and final volumes were 16.6x larger than other treatments digesting membrane proteins and permeabilizing cell membranes, resulting in the release of more DNA. The five non-glass-fiber-based treatments were then selected for further analysis on Test Array 5.1.

In a manner similar to the OD_{600} comparisons, Test Array 5.1 was used to compare the effectiveness of the different purification treatments on the basis of hybridization intensities of vector-specific and *E. coli*-specific probes. This comparison was made using only clones grown to an OD of 1.4, as these were most likely to produce detectable hybridization signals. Triplicate spots of the 4 different clones from each treatment were arranged into a single subarray for visual comparison, and OD 1.4 spots from the 5 other subarrays (each subarray representing a single purification treatment at 4 different ODs) were also used to generate data for this experiment; in all, data was obtained from a total of 8 triplicate sets of spots per hybridization for each treatment, except the commercial kit which derived data from 12 triplicate sets of spots.

The main measure of the effectiveness of purification was hybridization signal intensity from 1, 2, 3, 4, and 5μg of vector-specific *FOS-cos* probe, pooled and normalized to the *luxA* labelling control as in the previous experiment. As can be seen in Figure 9a, both the proteinase K and PVDF treatments result in hybridization signals significantly higher than the basic method, from a far greater proportion of clones (though still far inferior to the signal from commercially-purified clones). Despite the high DNA yield of the combined proteinase K + PVDF treatment, the hybridization signal intensity from this treatment was not significantly different than the basic method, and the clone detection rate was even lower. *E. coli* genomic contamination of the fosmid samples, as measured by the intensity of hybridization to 500ng of the *E. coli*-specific *uidA* probe, was lowest in the combined proteinase K + PVDF treatment and the commercial kit treatment [Figure 9b]. There was no statistical difference between the basic method and the individual proteinase K and PVDF treatments in this respect.

Based on the results presented in Table 3 and Figure 9, we chose to incorporate a proteinase K treatment into the basic automated fosmid purification protocol developed for this project. In all experiments prior to this result, including the construction of Test Arrays 1 through 4.2, fosmid DNA was purified by the basic method (identical to the finalized method, but lacking the proteinase K treatment). Although there was no statistical difference in the average hybridization signal intensities of vector-specific and host DNA-specific probes between the proteinase K and PVDF treatments, and the clone/spot detection rates were very close, we nevertheless concluded that proteinase K was a superior purification option for a few reasons. First, since the average DNA yield from the proteinase K treatment was more than double that of the PVDF treatment [Table 3], it would be more cost-effective to use the former to produce the necessary amount of DNA for printing. Second, the use of PVDF filters on a library-wide scale represents a large increase in cost over the basic method, and many times more expensive than the ~11.5mg of proteinase K required per 96-well plate of library clones (600mg per 5,000-clone library).

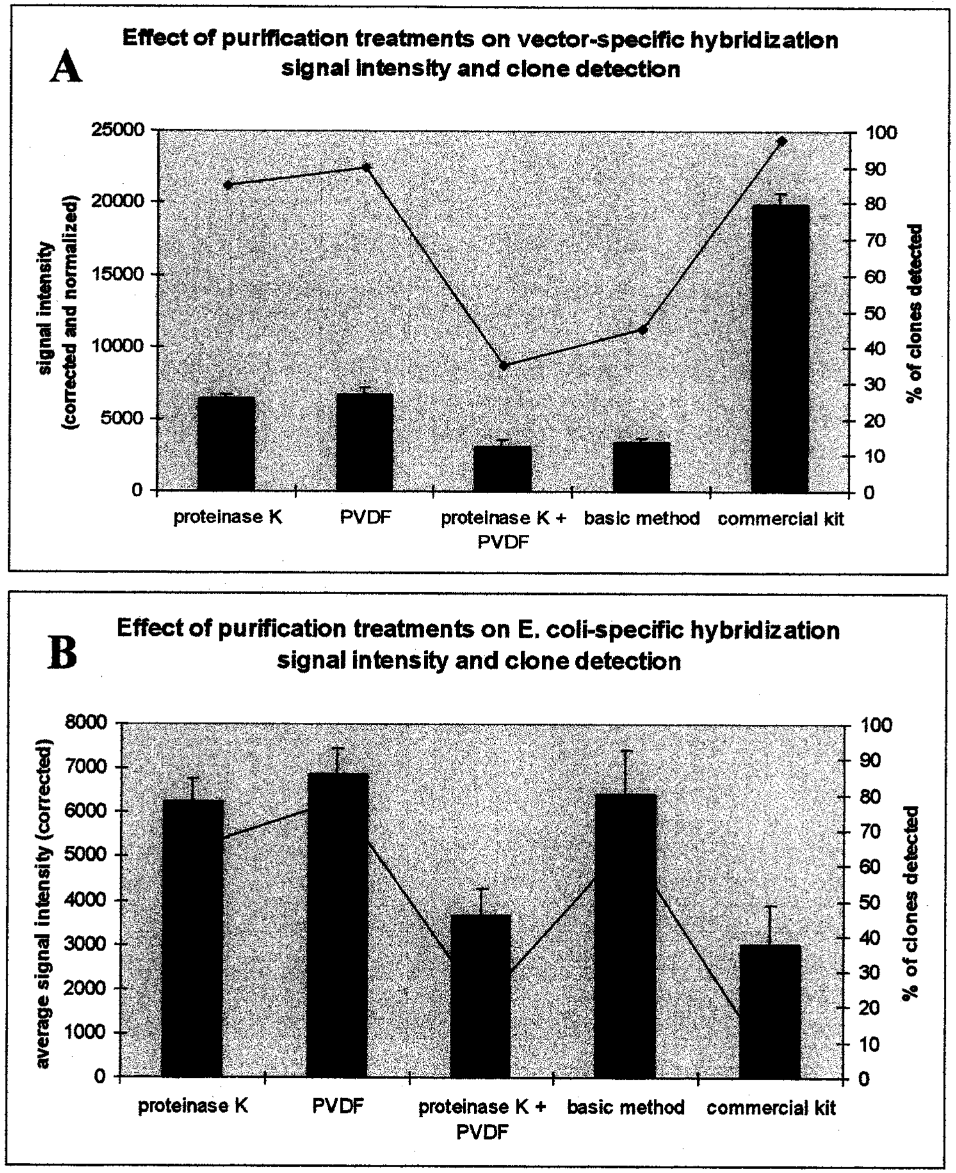


Figure 9. Optimization of automated fosmid preparation: effect of purification treatment. Two purification treatments (proteinase K treatment and PVDF filtration) were added to the basic fosmid purification protocol and clones were compared, alone and in combination, to the basic method and to clones purified using a commercial kit. The different purification treatments were compared on the basis of average clone hybridization signal intensity (bars) and clone detection rate (lines) from hybridizations to Test Array 5.1. (A) Composite quantitative data from hybridizations of 1-5μg of vector-specific *FOS-cos* (Cy3) probe (B) Quantitative data from hybridization of 500ng of *E. coli*-specific *uidA* (Cy5) probe

3.4 Microarray experiments: design, production and results

Once the metagenomic libraries had been produced and the automated fosmid purification protocol had been established, the remaining research goals of this project centered around the production of a series of small-scale metagenomic microarrays. These small test arrays were used to experimentally refine parameters of sample printing and hybridization, in order to determine the necessary conditions for eventually printing and using full-scale metagenomic microarrays. The test arrays were also used in proof-of-principle experiments designed to assess the feasibility of using metagenomic microarrays to rapidly screen a metagenomic library for clones bearing a desired target gene. These experiments are discussed below, following an explanation of the probes and controls employed. The major technical obstacles encountered and the steps taken to resolve them will also be discussed.

3.4.1 Probes and controls used in microarray experiments

3.4.1.1 Design of internal standards for printed fosmid DNA

Internal standards are used in microarray experiments to control for the amount of DNA printed in each spot, allowing the intensity of the experimental hybridization signal to be corrected for differences in printing efficiency between spots. The standard approach for microarray experiments in the Environmental Microbiology group at NRC-BRI is to include a small fixed amount of λ phage DNA in every sample to be printed. However this approach to internal standardization is inappropriate for the unique application of microarrays in the current study. This is because the most meaningful quantity to standardize in this case is not total DNA printed but total *fosmid* DNA printed, since any target genes would be located exclusively in the cloned DNA. Ideally, every clone to be printed would be set to a standard concentration of fosmid DNA before printing. For the large amounts of clones involved in creating a full metagenomic microarray, this would require an inordinate amount of time and effort to quantify every single clone and to adjust concentrations for each clone accordingly. But even if this were done, the amount of *fosmid* DNA in each preparation could still differ, owing to differential levels of induction and purification of fosmid DNA between clones. Thus, the simplest and most instructive approach would be to create a probe specific to the fosmid DNA found in every metagenomic clone, which could be used to define the relative differences in the amount of fosmid DNA printed in every spot.

We initially chose two candidates to serve as internal standards for the fosmid DNA: *FOS-cos* and *FOS-Cm*. To assess if the vector-specific probes would hybridize to fosmid clone DNA, we performed membrane hybridizations with DIG-labelled *FOS-cos* and *FOS-Cm* probes. Both candidates for internal standard hybridized successfully and specifically hybridized to the clone DNA, and to their own positive controls (data not shown). The two candidates for internal standards were then compared as part of the first set of test array experiments; the results of that comparison are presented as part of the discussion of Test Array 1, in section 3.4.2.

In typical microarray applications, the use of internal standards is most useful when applied to questions with a quantitative signal component. A general example of environmental relevance would be to monitor changes in the proportions of a particular functional gene in a microbial community in response to changing environmental conditions, using an array of catabolic genes (59). In this example, chip-to-chip variation in the printing of the desired gene probe must be taken into account to properly measure any changes in signal intensity due to differences between experimental samples. However, for application of full metagenomic microarrays envisioned by the current study, such quantitative comparisons are irrelevant, since screening a library for the presence of a specific gene is a simple binary test of presence or absence; this is a measure defined informatically by the signal-to-noise ratio of a particular spot, which would be unaffected by any mathematical internal standard-based correction of hybridization signal. Hybridization of the metagenomic microarray to the vector-specific internal standard is primarily useful for identifying exceptionally strong or weak vector signals, to alert researchers to potential sources of false positive or false negative results. In the current study of small prototype arrays however, even this application of fosmid internal standards was secondary. Instead, the use of a vector-specific probe was most valuable for its own sake, to produce a vector-specific signal used to optimize various parameters of fosmid purification and clone printing.

3.4.1.2 Experimental and control probes

In addition to the two internal standards described above, a variety of different genes were used as experimental and control probes throughout the microarray testing phase of the current study. The *uidA* gene was used as an *E. coli* genomic DNA-specific probe, as primers for this gene had previously been developed by researchers as a means to detect low levels of *E. coli* in potable water (76). The gene most often used as a negative control was *GFP*, encoding a green fluorescent protein first isolated in protein extracts from the luminescent hydrozoan jellyfish *Aequorea* (141). This was considered a negative control because of the extremely low likelihood that this gene would appear in the sample microbial communities.

The three main experimental genes used to screen 1A3 and 6A3 metagenomic library clones for the various test array experiments were *nirS, nirK* and *alkB* (described in section 1.5.2). In addition, a clone-specific *alkB* dubbed *alkBc*, produced from the *alkB*-PCR amplicon of 1A3 plate 18 clone F10, was also used in Test Array 6.2 experiments (section 3.4.5). Three additional genes appear in the test array experiments as part of a parallel project conducted by McGill Master's student Gavin Whissell to develop metagenomic microarray technology using different soil samples: two of them, *mmoX* and *pmoA*, respectively code for subunits of the soluble methane monooxygenase and particulate methane monooxygenase enzymes (69). The third additional gene, *pmoAc*, was a clone-specific *pmoA* PCR amplicon analogous to *alkBc*, derived from Mr. Whissell's libraries in a manner similar to *alkBc* in the current study (section 3.4.5).

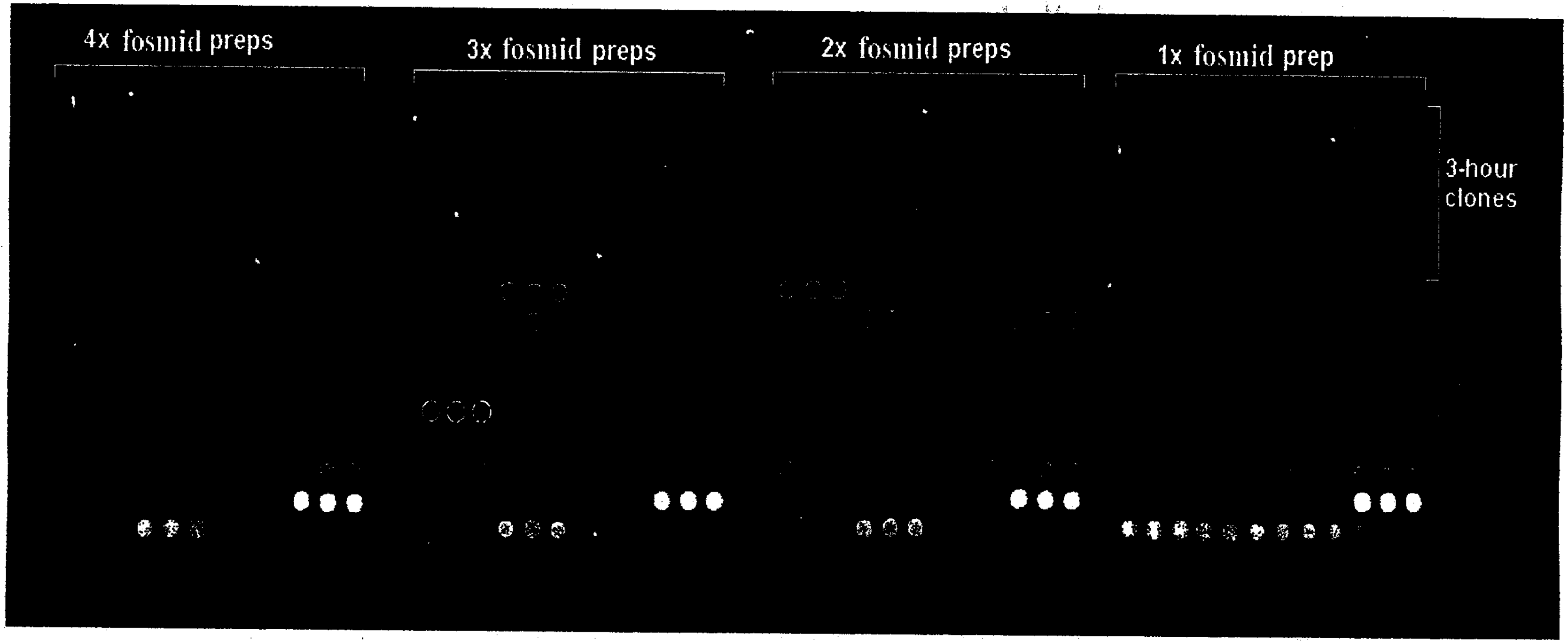
In general, if a gene probe was to be used in hybridization experiments on a given test array, then a PCR amplicon of that gene was printed on the test array as a positive control for hybridization. This was not true of the *luxA* amplicon, which was printed on all test arrays but never used as a probe. The *luxA* gene encodes the α-subunit of the light-emitting luciferase protein from bacterial *Vibrio* symbionts of luminescent marine invertebrates (28). This gene was used throughout the course of the current study as a labelling control: each probe labelling reaction was spiked with a constant amount of *luxA* amplicon. When data from more than one hybridization was pooled, the intensities of the *luxA* control spots were used to normalize the two sets of data (using Normalization Technique B, section 2.4.5).

3.4.2 Initial test array experiments

The first round of test array experiments was designed to answer the most basic questions about this experimental format. First and foremost: could fosmid clones purified by the automated fosmid purification method and printed on microarrays be detected by hybridization? Also addressed in these initial experiments was the question of which internal standard to use for subsequent experiments. Concerning the printing of fosmid DNA on microarrays, another important question was how many preparations were necessary to produce enough fosmid DNA for reliable hybridization signal detection (discussion of this question is deferred to section 3.4.3). Another experiment, discussed in section 3.3.4, was the choice of fosmid induction culture incubation time.

To answer the first question, Test Array 1 was hybridized with 100ng, 500ng, 1μg, 2μg and 4μg of *FOS-cos* (Cy3) and *FOS-Cm* (Cy5) probes. In every case, clone spots could be detected, although at a very low level of signal. Figure 10 shows a sample hybridization of an entire array with 1μg of *FOS-cos* probe. The dark blue colour of the clone spots represents the lowest level of signal detectable: We concluded from this result that fosmid clones printed on an array could be detected, albeit only faintly. This result represents the first reported instance of large-insert metagenomic DNA clones printed on a microarray being detected by hybridization, and provided a starting point for further development of the metagenomic microarray platform.

In order to choose which of *FOS-cos* and *FOS-Cm* would serve as internal standard for clone DNA, we originally decided to compare clone detection rates, i.e. the fraction of clones which could be informatically detected. Both probes were hybridized to Test Array 1 in the 5 different amounts described above and the data for each probe were pooled. Positive clone detections were measured informatically as described in section 2.4.5, based on clone signal from the 5-hour clones only (due to the low detection rates of 3-hour clones). Many of the clones (66.3%) could be detected using the *FOS-Cm* probe, while only 40.9% clones could be detected with *FOS-cos*. This would suggest *FOS-Cm* as the internal standard of choice. However, *FOS-Cm* hybridizations were consistently



Signal intensity key: Blue < Green < Yellow < Red <White (saturated)

Figure 10. Sample Test Array 1 hybridization. This is the full image of a single hybridization of 1μg *FOS-cos* (Cy3) probe to Test Array 1. This figure illustrates the multiple-experiment design of this test array: testing at once the effects of clone growth time (3h vs. 5h) and amounts of printed material (1-4 fosmid preps). Figure 6 was derived from the "2x fosmid prep" subarray in this picture. Control spots are identified in Figure 11.

plagued by a serious problem: this probe cross-hybridized with all other controls on Test Array 1, except for *FOS-cos*, at saturation or near-saturation levels of signal. In fact, *FOS-cos* also cross-hybridized to other TA 1 controls, but at much lower levels of signal [Figure 11]. This suggested that using *FOS-cos* as an internal standard might be less problematic, as its specificity was more assured.

Furthermore, we later discovered that the informatic clone detection results cited above were misleading: in these experiments, the *FOS-cos* probe was always labelled with Cy3, while *FOS-Cm* was always labelled with Cy5. These two dyes possess different fluorometric properties, in particular, Cy3 produces a higher background signal than Cy5 (140), which directly impacts (negatively) upon the informatic determination of positive hybridization signal. To correct for this difference between dyes, we re-scanned the 4μg *FOS-cos* hybridization, this time raising the scanner PMT (photon multiplier tube) setting from the standard 85% setting to 100%, in order to boost threshold signals above the higher background noise of the Cy3 dye. Under these conditions, a similar informatic comparison of Cy3-*FOS-cos* (PMT 100%) to Cy5-*FOS-Cm* (PMT 85%) showed that clone detection rates were almost equal, with 71.9% of clones detected by *FOS-Cm* and 73.2% detected by *FOS-cos*. With the apparent clone detection superiority of *FOS-Cm* no longer clear, we decided upon *FOS-cos* as the internal standard of choice due to its lesser propensity to cross-hybridize with other microarray controls.

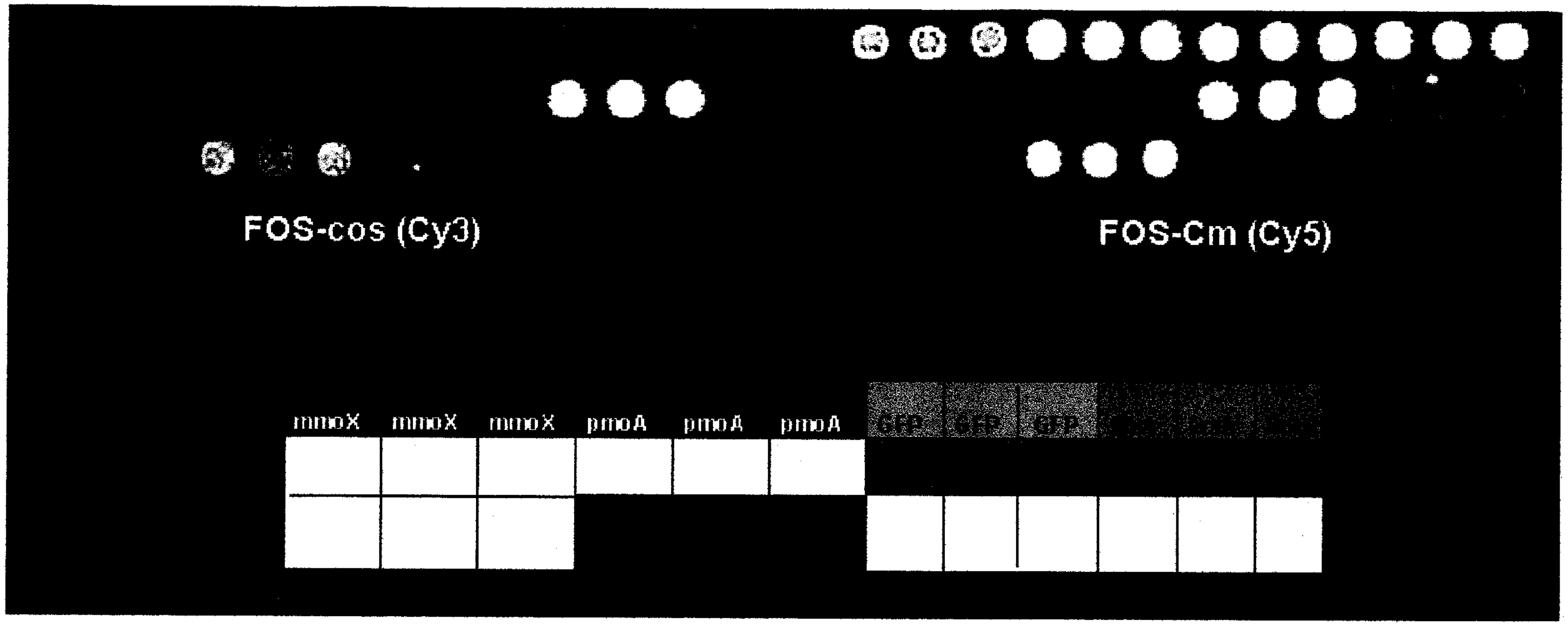
3.4.3 Optimization of printing and hybridization parameters

Over the course of the test array experiments, two principal parameters of printing and hybridization were subject to optimization: the amount of clone DNA printed on the arrays, and the amounts of probes used for hybridization. The former quantity was optimized in the first Test Array 1 experiments, as it was essential to the design of subsequent test arrays. The latter quantity was optimized as part of the experiments on Test Array 5.1.

3.4.3.1 Optimization of fosmid clone quantity

It became apparent before and during the preparation of materials for the first test array that the amount of DNA recovered by the automated fosmid purification protocol could be highly variable from plate to plate, clone to clone and even between replicate purifications of the same clone. For instance, among the TA1 clones that were quantified prior to printing, total purified DNA was 424ng ± 235ng (standard deviation) for 3-hour clones, and 831ng ± 73ng for 5-hour clones. Meanwhile, preparation of clones for TA 4.2 by the same method (with 5-hour induction cultures) produced an average DNA recovery of 1438ng ± 596ng. Instead of quantifying every single clone and standardizing concentrations prior to printing (an unmanageable task were these prototype arrays to be scaled up to contain full clone libraries), we chose instead to quantify printed DNA in terms of how many fosmid preparations were pooled in the sample wells. In the Test Array 1 experiment, we compared DNA pooled from 1 to 4 replicate fosmid preparation plates [Figure 10]. The different clone amounts were compared on the basis of average hybridization signal intensity and clone detection rate, as in other experiments.

Data for this experiment were taken from hybridizations of 100ng, 500ng, 1μg, 2μg and 4μg of *FOS-cos*. Only the 5-hour clones were used in this analysis because so few 3-hour clones were detectable. *FOS-Cm* data was not incorporated into this analysis because signals from the *luxA* labelling control spots were at saturation levels in virtually every hybridization, which



Signal intensity key: Blue < Green < Yellow < Red <White (saturated)

Figure 11. Choosing internal standards: cross-hybridization of controls. This figure demonstrates the problem of control cross-hybridization encountered throughout this study. The images in this figure were taken from a 1μg *FOS-cos* (Cy3) hybridization and a 1μg *FOS-Cm* (Cy5) Test Array 1 hybridization, performed separately. The control regions of a single subarray from each these hybridizations were combined to produce this figure. Cross-hybridizations are here defined as any visible hybridization signal originating from a probe-target combination that is not self-self (e.g. *FOS-Cm* probe and *gfp* printed controls).

Note: hybridization between either probe and the cloning vector (pCC1FOS) or the labeling control (*luxA*) was not considered cross-hybridization; in both cases some signal was expected.

prohibited pooling and normalization of the data from disparate hybridizations. The results show a significant increase in average hybridization signal from 1x to 2x to 3x fosmid preparations, while the signal from 4x preparations dropped to the level of 1x [Figure 12]. Clone detection rates present a slightly different picture, with a greater detectable proportion of 2x clones than of 3x clones, despite the much greater average signal intensity of the 3x clones.

The 4x fosmid preparation clones were detected at the lowest frequency, with less than 19% of clones positively identified by hybridization to *FOS-cos*, and the average signal from these clones was indistinguishable from the average 1x clone signal [Figure 12]. There are two factors that likely contributed to this result. First, as was mentioned in the discussion of the OD_{600} optimization experiments (section 3.3.4), the concentration of large volumes of fosmid preparation likely concentrated contaminants as well, hindering proper printing of these clones on the array. As well, the concentration of DNA in the 4x clone preparations well exceeded the 200ng/μl maximum spotting concentration recommended by the staff of the BRI Microarray Lab, averaging 332ng/μl $\pm$ 29.2ng/μl (derived from the quantification numbers presented above). As a result, the DNA solution may have been too viscous for proper printing, resulting in a loss of material from the printed spots. These factors may also account for the decrease in clone detection from 2x to 3x clone preparations, despite the higher average signals from the 3x clones. Based on these results, and based on considerations of resource usage, we chose 2x fosmid preparations as the optimal printing amount for subsequent test arrays.

3.4.3.2 Optimization of probe quantity

In all hybridization experiments prior to Test Array 5.1, no standard amount of probe had been formally established, although most hybridizations used 500ng or 1μg of probe. It was apparent (and intuitive) that higher amounts of probe produced more intense hybridization signals. However, in context of trying to resolve the problem of low clone signals and large numbers of clones evading detection (section 3.4.4.2), an important question to resolve was what impact *FOS-cos* probe amounts had on clone detection. Earlier comparisons of different probe amounts on TA 1 had been compromised by cross-contamination of probe solutions between adjacent experiments on the same chip, essentially eliminating the possibility of obtaining reliable data on this question from TA 1 hybridizations. With this human error eliminated as of the initial TA 5.1 experiments, we set out to answer this question by comparing clone detection rates on TA 5.1 using 1µg, 2µg, 3µg, 4µg and 5µg of *FOS-cos* probe.

The results of this experiment are presented in Figure 13a. Data are presented in terms of average hybridization signal intensity and clone detection rate. Average clone hybridization signal increased over the range of probe amounts, but the largest increase in signal occurred between 3µg and 4µg of probe, representing more than a 3.5-fold increase in signal. The results also show a steady increase in the clone detection rate over the range of probe quantities tested, although the greatest fold-increase occurred again between 3µg and 4µg of probe (1.17-fold). These results did not suggest that a detection plateau had been reached, but do present the possibility of diminishing returns in terms of hybridization signal and clone detection at *FOS-cos* probe amounts greater than 4µg. Based on these results, we chose 4µg as the optimal amount of internal standard probe for subsequent experiments. However, it must be noted that at this amount of probe, the intensity of cross-hybridization with various array controls became rather severe [Table 4]. Thus we were faced with a trade-off where amelioration of one technical problem (low clone signal) aggravated another (cross-hybridization).

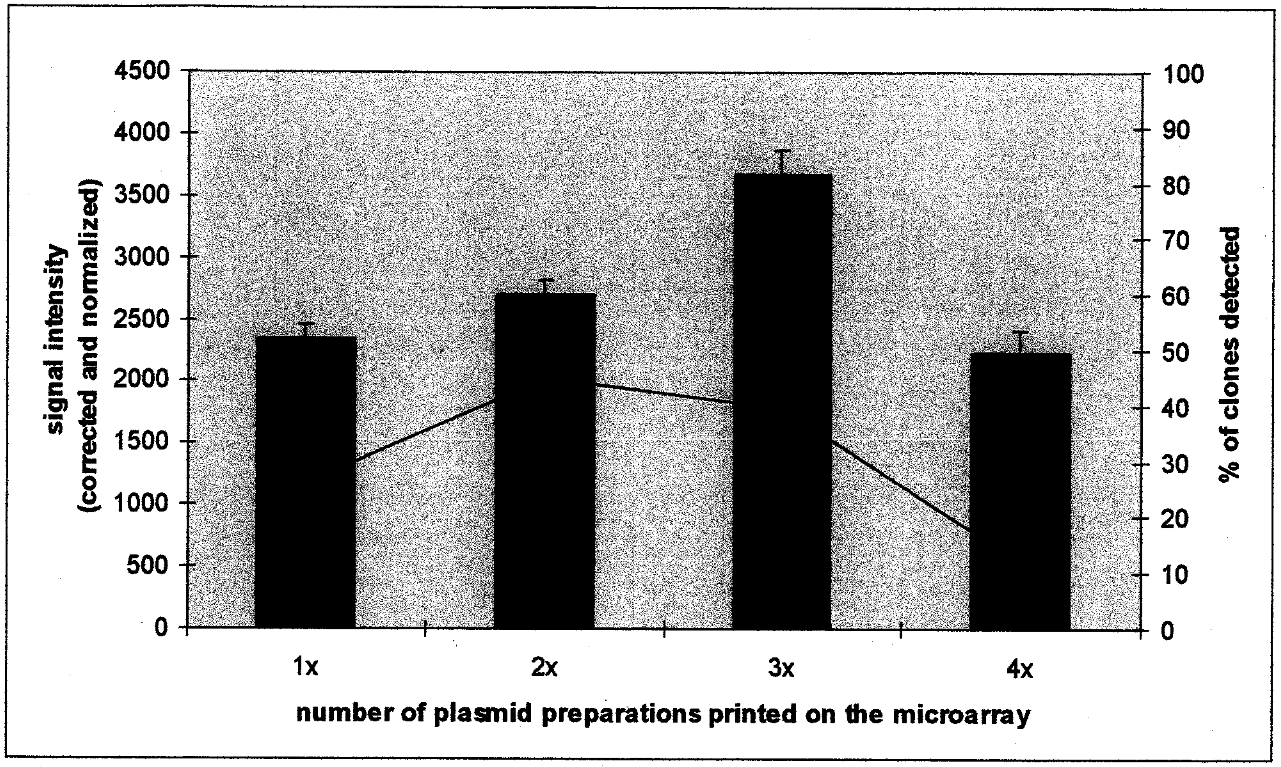


Figure 12. Optimization of microarray clone amounts printed: effect of number of plasmid preparations on average hybridization signal intensity and clone detection Clone amounts were defined by the number of fosmid preparations used to prepare DNA (1-4 preps). 100ng, 500ng, 1µg, 2µg and 4µg of *FOS-cos* (Cy3) probe were hybridized to Test Array 1. Composite quantitative data from all five hybridizations were used to compare average clone hybridization signal intensity (bars) and clone detection rate (lines) for the different amounts of clone DNA printed. Only hybridizations to 5-hour clones on Test Array 1 were used for this figure, as very few 3-hour clones were detectable in any hybridization (see Figure 10).

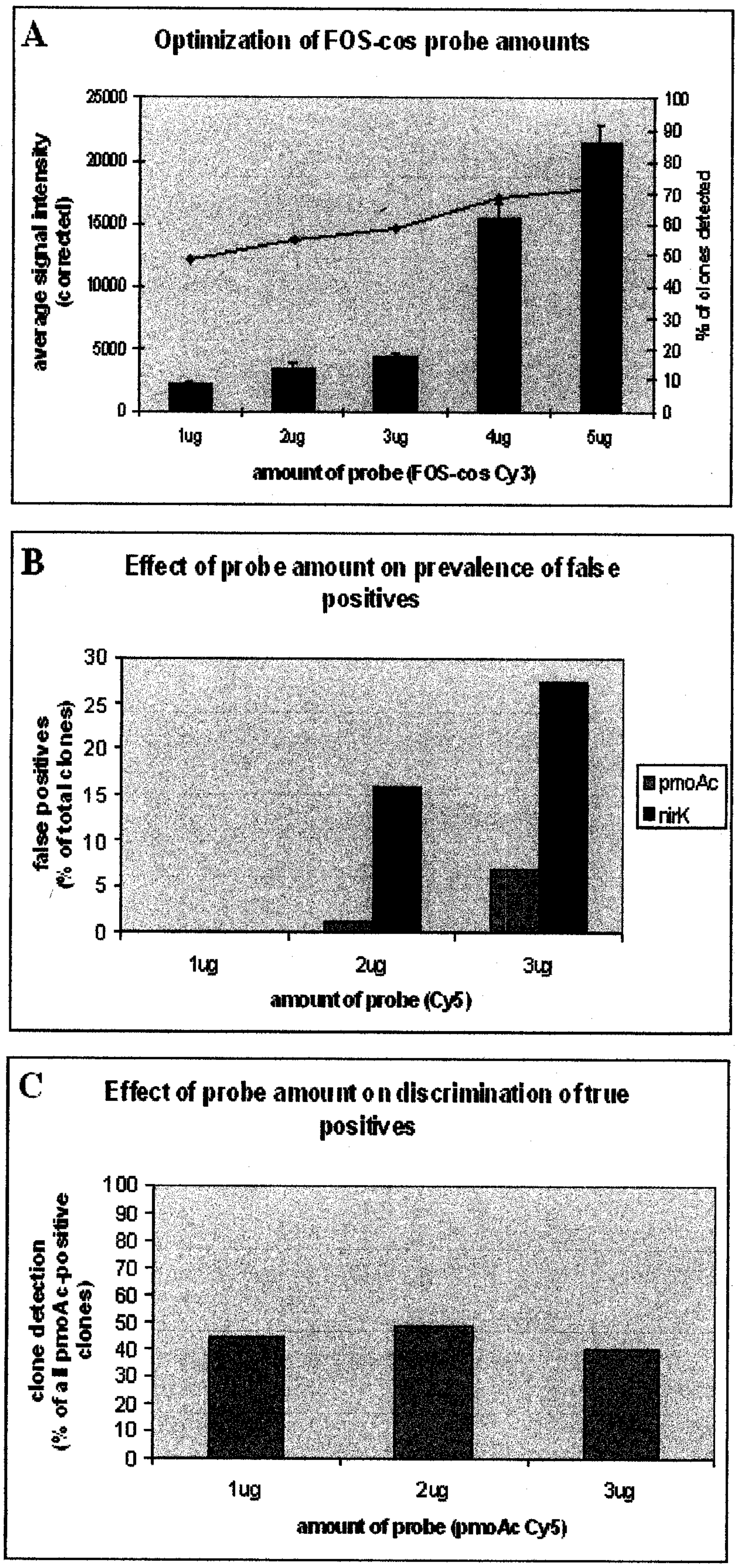


Figure 13. Optimization of probe amounts.

Hybridization experiments on Test Array 5.1 were performed to optimize amounts of internal standard and experimental probes used for subsequent hybridization.

(A) Comparison of 1-5μg of *FOS-cos* (Cy3) in terms of average clone hybridization signal intensity (bars) and clone detection rate (lines). (B) Comparison of 1-3μg of *pmoAc* and *nirK* (Cy5) probes in terms of rate of false-positive identification.

(C) Comparison of 1-3μg of *pmoAc* (Cy5) probe in terms of rate of true-positive identification. *pmoAc* is specific to one of the 4 clones used to create TA 5.1. See section 3.4.3.2 for definitions of false-positive and true-positive hybridizations.

Note: no error bars are included in (B) or (C) because each hybridization produced only a single value, and experiments were not replicated.

Table 4. Optimization of probe amounts: average *FOS-cos* (Cy3) hybridization signal from microarray amplicon controls

		amount of probe			
target	1ug	2ug	3ug	4ug	5ug
FOS-cos	39160	50453	52654	65201	65133
alkB	0	0	0	0	0
GFP	2047	4619	7486	44229	60741
mmoX	670	1834	3413	1547	2639
nirK	0	0	0	0	0
nirS	644	1339	1968	3683	7227
pmoA	3040	7069	11180	15842	27328
uidA	0	0	0	0	1080

The optimized amount of 4µg of internal standard (*FOS-cos*) probe represented a nearly 10^4-fold molar excess compared to the total amount of clone targets printed on TA 5.1 (282 spots, 6amol per spot; see section 3.4.4.2). At this level of excess, it was possible that false-positive hybridization could present a serious problem, not with the internal standard probe, but with an experimental probe used to find a few copies of a target gene of interest from among a large number of library clones, as might be the case in a typical application of a full metagenomic microarray. To explore this possibility, we took advantage of the fact that one of the 4 clones used to create Test Array 5.1 had been previously found by Mr. Whissell to be PCR-positive for *pmoA* (data not shown). A probe created from the *pmoA* PCR amplicon of this clone (dubbed *pmoAc* for "clone-specific *pmoA*") would thus be specific to a quarter of the clones printed on TA 5.1, providing a vehicle to explore the effect of probe amounts on both false positive and true positive identification of clones by hybridization.

To optimize the quantity of experimental probe (i.e. non-internal standard) to be used in subsequent hybridization experiments, we compared the incidence of false positive detection of clones by 1µg, 2µg and 3µg of two different probes, *pmoAc* and *nirK*. The former, as mentioned above, was specific to only one of the four clones printed on TA 5.1. Thus, positive detection of any of the other three clones by *pmoAc* was considered as a false positive. None of the four clones had been found to be PCR-positive for *nirK* (data not shown), so this gene probe was used as an additional source of data on the assumption that positive detection of any clone on the array was a false positive. Data for detection of full triplicate sets of clone spots are presented in Figure 13b. No error bars are presented because the experiments were not replicated, due to a limited supply of microarray slides and the high cost of fluorescent dyes needed to produce large quantities of labelled probe. The data show an increasing incidence of false positives from 1µg to 2µg to 3µg of both probes. No higher amount was tested due to the unacceptably high proportion of false positives at 3µg probe, approaching 30% with *nirK*. Notably, no full triplicate sets of false positives appear at 1 µg of probe, suggesting that this is the optimal amount of experimental probe. It also reinforces the notion that only full triplicate sets of spots be considered when identifying positive hybridization events.

However, the above results begged the question: would using only 1μg of experimental probe negatively affect the identification of true positives? In the context of screening libraries by hybridization to metagenomic microarrays, the identification of false positives is less problematic than the appearance of false negatives; false positives can always be weeded out by subsequent analysis of the identified clones, but false negatives represent clones of interest that escape detection. To explore this question, we compared the ability of 1μg, 2μg and 3μg of *pmoAc* probe to detect true positive clones, the 25% of TA 5.1 clones known to contain the *pmoAc* sequence. The data are presented in Figure 13c, again without error bars, for reasons previously described. The results indicate that hybridization with 1μg of probe produces more false positives (is able to discriminate fewer true positives) than 2μg of probe, but paradoxically produces fewer false positives than 3μg of probe. Without replicate experiments to provide an estimate of measurement error, we can only conclude that there is no reason to assume more experimental probe will result in fewer false positives. In other words, the data suggest that the quantity of experimental probe has no bearing on the ability to detect true positives. Based on the results of Figures 13b and 13c, we decided upon an optimal amount of experimental probe (i.e. non-internal standard) of 1μg for all subsequent experiments.

3.4.4 Major technical challenges encountered

Initial experiments on Test Array 1 identified two major problems that have to be resolved before any attempt at constructing a full-scale metagenomic microarray can be undertaken. The first problem was the extensive cross-hybridization of microarray controls. The second problem was the chronically low hybridization signal from fosmid clones on all test arrays. Many experiments were conducted on a number of different test arrays to address these problems, and both have been alleviated to a degree in this study, in particular, the problem of low clone signal. However, further experimentation in future studies will be required for development of full metagenomic microarrays. Both problems are discussed below, along with the steps taken to resolve them and their results.

3.4.4.1 Cross-hybridization of microarray controls

The cross-hybridizations observed in Test Array 1 experiments came as a surprise, because preliminary membrane hybridizations of DIG-labelled *FOS-cos*, *FOS-Cm*, *GFP* and *mmoX* probes revealed no non-specific interactions (data not shown). Although the *FOS-Cm* probe produced by far the highest levels of cross-hybridization on TA 1, from this first test array onward this was a problem to a certain degree for every probe used. Test Arrays 2 and 3 were designed exclusively with the goal of solving this problem, and Test Arrays from the 4.x and 5.x series were at least partly dedicated to this problem as well. A chance observation during a Test Array 5.1 experiment uncovered the source of a large part of this problem. All microarrays from TA 1 to TA 5.1 had been designed in such a way that three full arrays were printed on each slide, allowing up to 6 hybridizations at a single time, under three separate coverslips. The revealing observation was that liquid from each coverslip was coming into contact with adjacent coverslips, creating a channel through which probes from adjacent experiments were crossing freely. Upon observing this we repeated all Test Array 5.1 experiments that had been performed to date, leaving a large space free between coverslips by using only the top and bottom arrays on each chip, so that no liquid could come into contact between arrays. Table 5 summarizes the patterns of cross-hybridiation that remained after correcting this human error. It is clear from this table that a noticeable degree of cross-hybridization existed that could clearly not be attributed to the effect of human error.

All of the microarray data presented thus far in this report have been carefully selected from experiments in which the aforementioned human error could not have affected the results. All data presented from TA 5.1 (Table 5) was obtained from experiments performed after the discovery of this problem.

Table 5. Cross-hybridization between microarray controls: summary of TA 5.1 experiments

	Target (immobilized)								
Probe (labeled)	pmoA	uidA	GFP	mmoX	nirS	luxA	alkB	nirK	Fos-cos
pmoA	+++		++	+		+			+
uidA		+++							
GFP			+++			+			+
mmoX	+		++	+++		+			+
nirS					+++				++
luxA	++		+			+++			+
alkB							+++		
nirK								+++	

Legend

+= medium (blue)   ++= high (aqua/green)   +++= saturated (red/white)

▨ = labelling control

In all reported experiments from TA 1, *FOS-cos* was the only Cy3-labelled probe present on the chip, while *FOS-Cm* was the only Cy5-labelled probe, thus all cross-hybridization observed in those experiments was not artifactual. Similarly, the following discussion of cross-hybridization troubleshooting in Test Arrays 2, 3 and 4 will refer only to cross-hybridizations that cannot be explained by cross-contamination of probes on adjacent arrays (i.e. only cross-hybridization from probes not loaded on the same chip or labelled with the same dye will be considered).

A number of potential causes for control cross-hybridization were explored, but none offered a solution to the problem. At best, some steps resulted in a slight reduction of the intensity or range of cross-hybridization. Table 6 outlines the various approaches taken to address this problem, and the effects of these troubleshooting steps. These are discussed below briefly to illustrate the intractability of the problem, but no data is presented. To investigate if excessively high concentrations of control DNA were promoting non-specific interactions, we printed dilution series of all TA 1 controls, from 1:10 to 1:10^4 of their original 200ng/µl spotting concentration. To investigate possible contamination during control production, all controls were re-amplified from completely new reagents and templates and spotted alongside the original TA 1 controls. Neither of these approaches produced more than sporadic improvement in individual cross-hybridizations; no systematic improvement was achieved. To investigate whether the amplicon-amplicon hybridization format (unique to this novel microarray application) itself was somehow responsible, we hybridized our amplicon probes to an array of catabolic and taxonomic gene amplicons produced by the Environmental Microbiology group at BRI. Upon finding instances of extensive cross-hybridization, we then printed the same host of catabolic and taxonomic genes, cloned into the pDrive vector (generously provided by Sylvie Sanschagrin of BRI), to see if avoiding the amplicon-amplicon hybridization format could alleviate the problem, but to no effect. Sequence alignments of *FOS-cos, mmoX, GFP, pmoA* and the pCC1FOS cloning vector revealed no significant DNA sequence similarity between any of the sequences. Among all the troubleshooting approaches attempted, only increasing hybridization stringency by raising hybridization temperatures from 42°C to 50°C had any marked effect, in particular removing the cross-hybridizations that occurred only faintly at 42°C. However as a solution to be incorporated permanently into the metagenomic microarray hybridization protocol, this was unacceptable, as the 50°C hybridization temperature reduced clone hybridization intensity and the detection rate nearly to zero (data not shown).

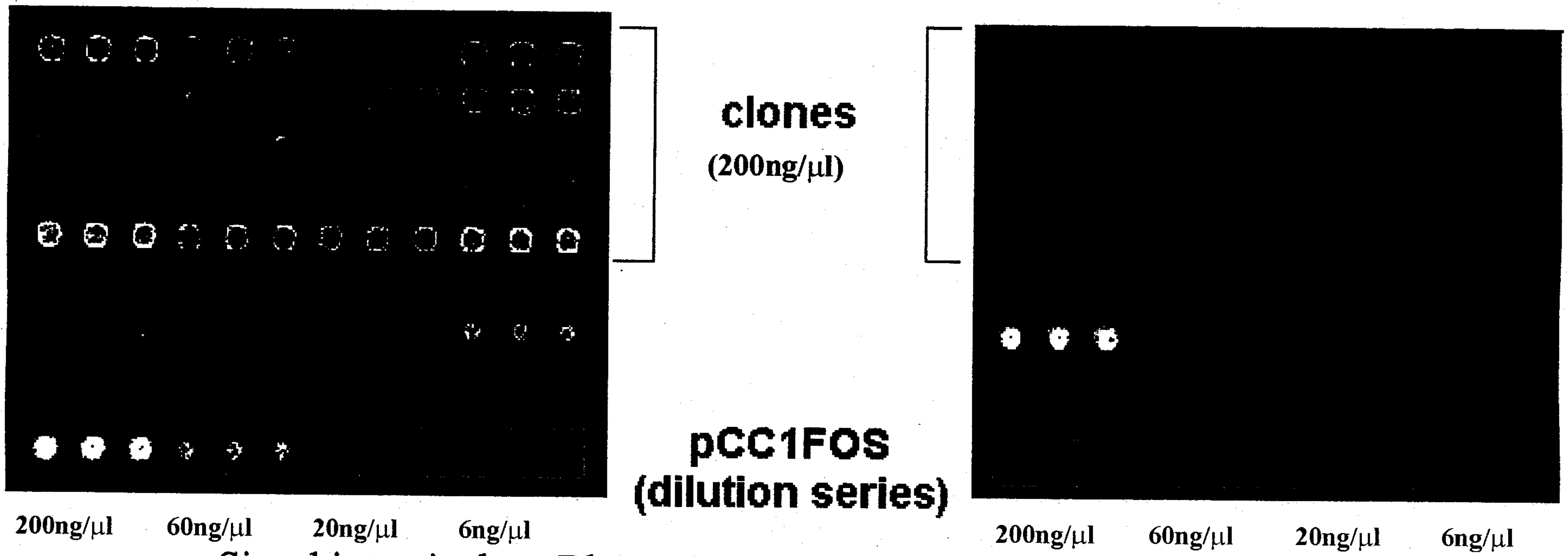
By the end of the current study, the problem of control cross-hybridization has still not been solved completely. However as a result of troubleshooting investigations, *GFP* has been eliminated as a negative control, as this gene proved one of the most problematic; Test Array 6.2 instead uses *nirS* as a negative control, and it is our recommendation that *GFP* not be used in any subsequent development of metagenomic microarrays. As well, *FOS-cos* remains prone to a broad spectrum of cross-hybridization, though admittedly at low levels of signal [Table 5]. Future work should explore other possibilities for fosmid-specific internal standards, compared in particular on the basis of their relative freedom from cross-hybridization to other array controls.

The problem of cross-hybridization was most noticeable among the amplicon controls printed on the arrays, but it was not restricted to this area alone. At times, some probes would hybridize to clones known to be PCR-negative for the probe gene, creating false positives (see section 3.4.3.2). Figure 14 demonstrates this effect among a small subset of clones and controls on Test Array 5.1. The left panel shows a 4μg *FOS-cos* hybridization, and the right panel shows a 3μg *nirK* hybridization of the same array spots. In these case as in other instances of false-positive identification, the offending probe also cross-

Table 6. Summary of experimental steps attempted to eliminate cross-hybridization

Troubleshooting approach	Result
dilution of printed microarray controls	localized improvements and deteriorations; no systematic improvement
Increase hybridization temperature	overall reductions in all signals, elimination of low-level cross-hybridizations
Re-production of all controls from fresh reagents and templates	no effect
hybridization to catabolic gene array	extensive cross-hybridization of some probes, no cross-hybridization of others
printing vector-borne controls	low-level cross-hybridization to *all* printed controls, some stronger cross-hybridization
Informatic sequence alignment of controls	no significant alignment detected

* Provided by Sylvie Sanschagrin, BRI Environmental Microbiology group



Signal intensity key: Blue < Green < Yellow < Red <White (saturated)

Figure 14. Cross-hybridization to clone DNA: effect of amounts of printed fosmid DNA. A small subset of the clone and control spots on Test Array 5.1 are presented to illustrate the cross-hybridization of clone and vector spots by experimental probes (here represented by *nirK*). *Left*: hybridization of 4µg *FOS-cos* (Cy3). The most intense signals correspond to the highest concentrations of fosmid clones or vector. *Right*: hybridization of 3µg nirK (Cy5). The visible clone and vector spots in this hybridization correspond to the highest *FOS-cos* signals in the left pane.

Note: the cross-hybridization visible from 3µg *nirK* (right pane) is not representative of the level of clone cross-hybridization arising from from lower amounts of probe. The high amount of probe was chosen to more dramatically illustrate the problem.

hybridized to the vector DNA printed as a control (a dilution series in the red box), suggesting that the vector was the source of cross-hybridization between probe and clones. Furthermore, the clones exhibiting cross-hybridization were the clones with the greatest amounts of fosmid DNA, as detected by *FOS-cos* hybridization. Not all probes exhibited this behaviour; for example *GFP*, despite its consistent and intense cross-hybridization to other controls, never cross-hybridized to vector control or clone spots. Further clouding this issue, no sequence similarity was found between the cloning vector and any of the probes which cross-hybridized with the vector, using the MacVector or BLASTn sequencing software (data not shown). As in the case of cross-hybridization between amplicon controls, this problem remained ultimately unsolved by the end of the current study.

3.4.4.2 Low clone signal

The other main technical problem encountered in this study was apparent from the first experiments on Test Array 1: the vector-specific hybridization signal from metagenomic clones was faint, producing dark blue spots near to or below the limit of informatic detection. This is due in good part to the nature of the metagenomic microarray format, which places a very low upper limit on the amount of clone DNA that can be printed. When we initially began to produce test arrays, the staff of the BRI Microarray Lab advised us not to exceed a printing concentration of 200ng/µl, the standard concentration used for oligonucletide and amplicon arrays. Microarray printing sample concentrations must fall within a narrow range, bounded below by questions of detectability, and bounded above by questions of sample viscosity. Too low a concentration, and hybridization signals will be undetectable. Too high, and the viscosity of the DNA solution will interfere with proper printing.

For large-insert clones, this range of acceptable concentrations is particularly narrow. The recommended 200ng/µl of large-insert fosmid DNA represents a much smaller number of molecules than does 200ng/µl of PCR amplicons or oligonucleotides. A 200ng/μl fosmid preparation of a 50kb clone contains 1/100th of the molecules contained in a 200ng/μl solution of a 500bp amplicon, and 1/1,000th of the number of 50-base oligonucleotides. In concrete terms, the average 1A3 clone printed at 200ng/μl in 0.7nl array spots contains only 6amol of clone molecules (6 x 10^{-18} moles), assuming the average clone size reported for 1A3 in section 3.2. As this number is 2-3 orders of magnitude lower than the amount of DNA in standard microarray applications, the low clone signals reported in this study are hardly surprising. Indeed, the fact that any signal at all was detected was never a foregone conclusion, and speaks to the sensitivity of the metagenomic microarray system.

Despite the innate factors contributing to the problem of low clone signal, we sought experimental means to boost these signals. The optimization of fosmid growth and purification conditions in Test Array 5.1 experiments had a dual purpose, optimizing the fosmid purification protocol while at the same time boosting clone detection. As reported in section 3.3.5, the addition of a proteinase K purification step to the basic method increased average clone detection by 1-5μg of *FOS-cos* probe from 53% to 88% of printed clones, and nearly doubled average signal intensity [Figure 9a]. The optimization of *FOS-cos* probe quantity, presented in section 3.4.3.2, also did much to improve average clone signal and clone detection rate, producing a 4.9x and 1.2x improvement in these respective quantities over the full range of probe amounts compared in this experiment [Figure 13a].

Another contributor to the problem of low clone signals can be described as the problem of "stacking elephants". In essence this is a problem of the physical space occupied on the microarray surface by the printed DNA: when the printed materials consist of small oligonucleotides or amplicons, a very large number can be compressed into the small physical area of an array spot. However individual large, bulky fosmid clones occupy a much greater 3-dimensional space. As a result of steric hindrance between clone molecules, the available space on the slide surface may be occupied by a relatively small number of clone molecules, leaving the rest of the printed material attached in loose aggregates to the clone molecules more firmly bound to the slide surface (like stacked elephants). This may become a problem during post-hybridization washes, where the robust treatment of microarray slides could remove some of these loosely-bound clone molecules, and the probe molecules bound to them, resulting in decreased hybridization signal. To investigate this possibility, we added a vigorous 0.1x SSC washing step before microarray pre-hybridization in order to remove any poorly-bound clone molecules *before* any probe was hybridized, and compared this to the standard hybridization protocol in terms of average clone hybridization signal and clone detection rate from a 1μg *FOS-cos* hybridization. The results, presented in Figure 15, indicate a nearly 9% increase in clone detection, and a small but significant increase in average signal intensity. As a result, we incorporated a 0.1x SSC pre-pre-hybridization washing step into all subsequent array experiments.

Altogether, the experiments discussed above provided significant improvement to the chronic problem of low clone signal in the test array experiments. However, when compared to the hybridization signal from clone spots produced from a commercial kit (representing the "best possible method" of purification), it is clear that there is still much room for improvement. This was illustrated in an experiment on Test Array 6.2: all optimizations established in this study were incorporated into the preparation of two 96-well plates of clones printed on this array. In addition, four individual clones were purified using a commercial kit and were printed on this array in a dilution series from 800ng/ul to 2ng/μl. When this array was hybridized to 4μg of *FOS-cos* (Cy3) probe (not shown), we compared the average signal intensity from the plate-purified clones to those from the various dilutions of kit-purified clones. Not only were the kit-purified clones informatically detectable right down to their lowest dilution (representing about 1/100th of the total DNA printed in the plate-purified clone spots), but the average hybridization signal intensity from the lowest dilution of kit-purified clones was superior to that of the average plate-purified clone spots. In quantitative terms, the 2ng/μl dilution of kit-purified clones produced an average hybridization signal intensity of 8500 ± 1391 units (standard error, N =

12), compared to an average intensity of 5606 ± 600 units (N = 444) for the plate-purified clones, spotted at a concentration of 150-200ng/µl. Only those plate-purified clones that were informatically detected in triplicate (148 out of 192) were considered for these calculations.

The problem of low clone signal may be solved more fully in future studies. Two approaches in particular were not attempted in this study, that were suggested by the researchers who pioneered the "Library-on-a-slide" format, where entire genomes were printed in single array spots (176). They found that sonication of samples prior to printing and use of a secondary labelling system such as Tyramide Signal Amplification (TSA) (PerkinElmer; Wellesley, MA) both resulted in significant increases in signal intensity. Sonication can decrease the viscosity of a high-concentration solution of large DNA, allowing higher spotting concentrations to be printed with fewer viscosity-induced printing anomalies. The TSA system uses a system of unlabelled antibodies that recognize labelling molecules (such as biotin) incorporated into DNA probes, and secondary antibodies to recognize the primary antibodies and catalyze the deposition large amounts of Cy3- or Cy5-labelled reagent (176). Considering the successes scored by these two approaches in the cited study, these possibilities should be explored in any further development of metagenomic microarrays. In addition, future work should more explicitly explore the question of how much clone DNA can be printed on a microarray before further increases become counterproductive due to printing problems. That question was addressed only semi-quantitatively in the current study, when optimizing the number of fosmid preparations should be printed on metagenomic microarrays (section 3.4.3.1). Establishing a firm relationship between fosmid clone spotting concentrations (in ng/µl) and hybridization signal and clone detection would be an asset, especially for exploring of the usefulness of sample sonication.

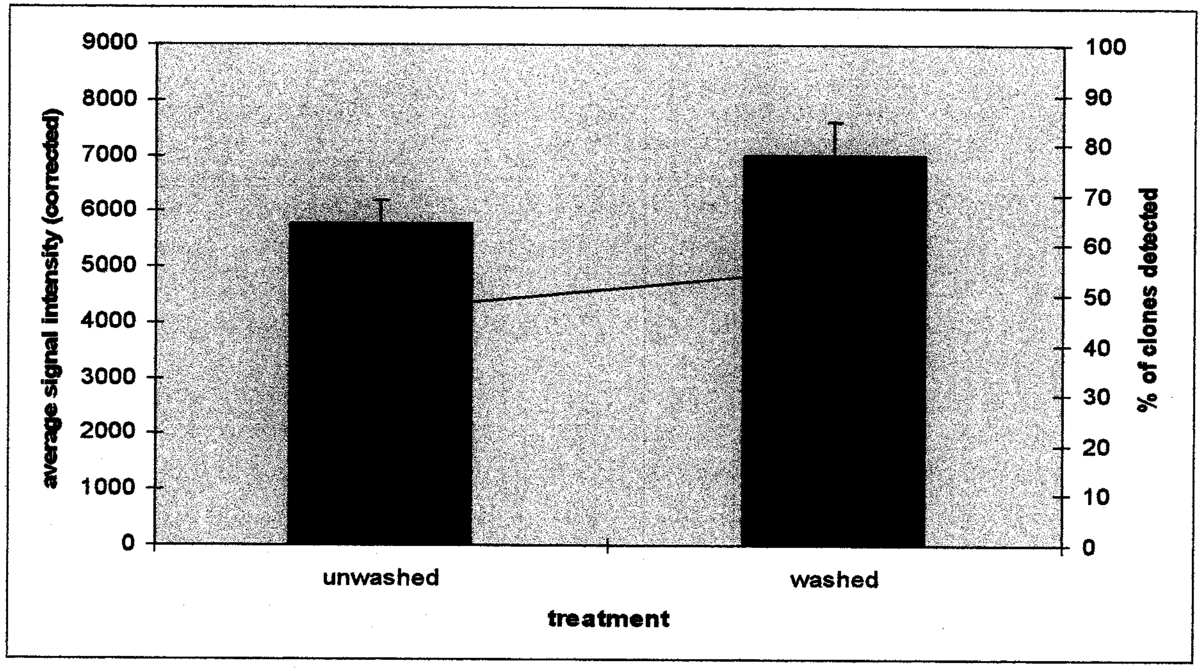


Figure 15. Low clone signal: effect of pre-pre-hybridization washing on hybridization signal intensity and clone detection. In order to mitigate the signal loss due to improperly printed clone spots arising from the presumed "stacked elephant" effect (described in section 3.4.4.2), we tested the effect of incorporating a microarray washing step prior to pre-hybridization. Both hybridizations were performed with 1μg of *FOS*-cos (Cy3) probe. Washed samples differ from *unwashed* samples only in the inclusion of this pre-pre-hybridization washing step. The effect was measured in terms of average clone hybridization signal intensity (bars) and clone detection rate (lines).

3.4.5 Proof-of-principle experiments

Since final resolution of major technical challenges must await further study, justification of ongoing research in this area depends on proof that the experimental concept of metagenomic microarrays is sound, that we can indeed print metagenomic library DNA on a microarray and use this tool to successfully screen the library for a target gene. We sought to provide that proof in a final set of experiments on Test Array 6.2. Briefly, an *alkB*-bearing clone was identified by PCR screening of the 1A3 library, and the entire plate of 96 clones bearing this clone was printed on the test array. The array was hybridized to an *alkB* probe to see if this could selectively identify the *alkB*-positive clone.

The 1A3 library was first screened by PCR to identify clones bearing the three experimental targets *alkB*, *nirS* and *nirK*. Multiple clones were found to be PCR-positive for *alkB* and *nirS*. Figure 16 shows the results from the *alkB* screening that identified plate 18 clone F10 (clone 1A3-18 F10) as *alk-B* positive, results representative of experiments on other plates and other genes. Two fosmid preparations' worth of DNA from all clones from 1A3 plate 18 (PCR *alkB*-positive) and plate 49 (PCR *alkB*-negative) were then purified by the optimized method and printed on Test Array 6.2. Large amounts fosmid DNA from four clones were additionally purified using the Qiagen Plasmid Maxi Prep kit (Qiagen; Hilden, Germany): 1A3-18 F10 (the *alkB*-positive clone) and 1A3-18 A2, 1A3-49 B11 and 1A3-49 H11 (random *alkB*-negatvie clones), and printed as controls as a dilution series from 2ng/ul to 800ng/ul spotting concentration (in addition a suite of amplicon clones was also printed).

Test Array 6.2 was initially hybridized with the standard *alkB* used in other experiments, which had originally been amplified from *Rhodococcus sp.* strain Q15 using concensus *alkB* primers. An image scan of this hybridization is presented in Figure 17a. Neither visual nor quantitative analysis of this hybridization could identify the *alkB*-positive clone from among the 96 clones purified by the automated method (all spot locations are visible in the SpotQC hybridization, presented as a visual reference in Figure 17c; clone 1A3-18 F10 spots are highlighted in all three figures). Among the commercially purified clones, only clone F10 (*alkB*-positive) was informatically detected at all, at spotting concentrations of 400, 600 and 800ng/µl (data not shown).

To investigate whether a more clone-specific probe could positively identify the *alkB*-positive clone, we purified the clone F10 *alkB*-PCR amplicon. We used this amplicon as template to produce the *alkBc* probe, which was thus identical to the sequence found in clone 1A3-18 F10. Before performing the hybridization to Test Array 6.2, we sequenced the PCR amplicon from this clone to ensure that it came from an *alkB*. BLAST protein sequence analysis of the translated protein sequence from a single strand of DNA identified this amplicon as having a 91% protein sequence identity and an 89% DNA sequence identity with the alkane monooxygenase from *Nocardioides sp.* Strain CF8 (61), strongly suggesting that this was indeed an *alkB* gene (data not shown). The hybridizaion to *alkBc* was then performed, and the image scan of that hybridization is presented in Figure 17b. The spots in the red box represent clone 1A3-18 F10. Although these spots visibly stand out from all surrounding clone spots, they were not detected as positive hybridizations by standard analysis of the quantification data. However, when we reduced the signal-to-noise ratio cutoff for positive hybridizations (section 2.4.5) from 3 to 1.5, all three replicates of this clone spot were successfully and exclusively detected from the two plates of clones. Among the dilution series of commercially purified clones, the lower signal-to-noise criteria permitted the false detection of only the highest concentration of 1A3-49 clones B11 and H11 (data not shown).

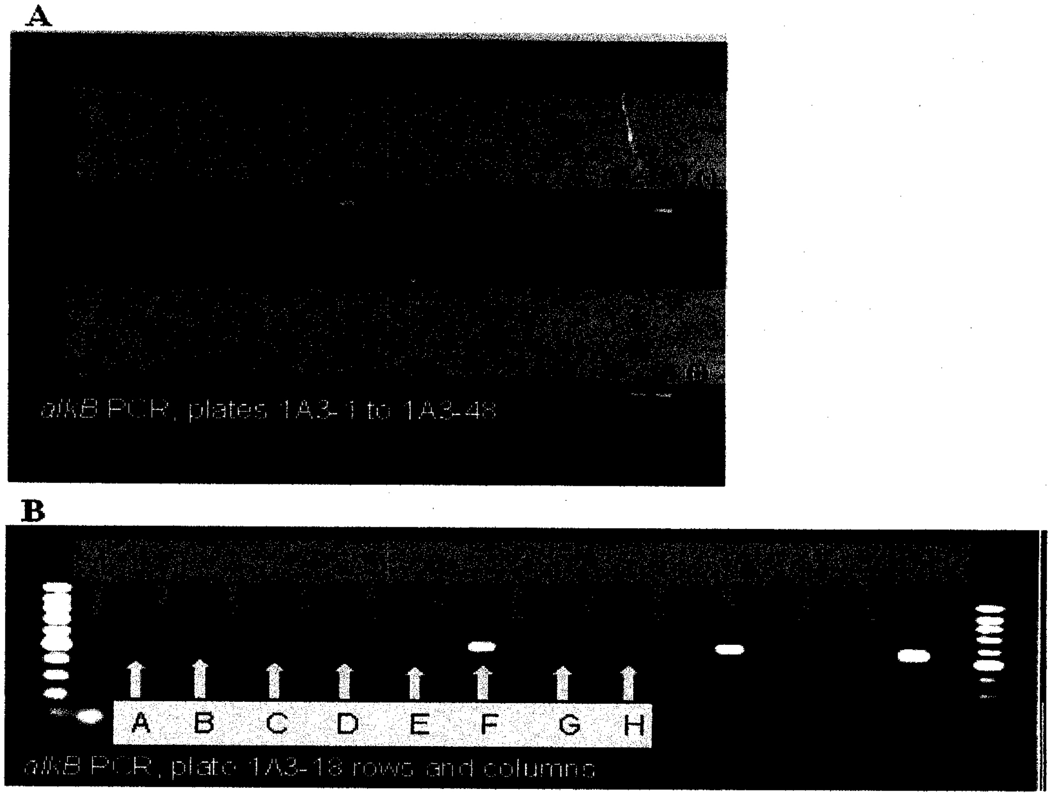

A
alkB PCR, plates 1A3-1 to 1A3-48
B
A B C D E F G H
alkB PCR, plate 1A3-18 rows and columns

Figure 16. Isolation of an *alkB*-positive clone by PCR screening. PCR positive control was 30pg of pDrive containing *Rhodococcus sp.* Q15 *alkB*; negative control was no DNA. (A) PCR was performed first on DNA pooled from each plate (96 clones) to identify plates bearing *alkB*-positive clones. Arrows indicate positive results. Marker: λ Mono Cut Mix. (B) For every *alkB*-positive plate, a second PCR was performed on clones pooled from each row (12 clones) and column (8 clones) on the plate, providing row and column coordinates for *alkB*-positive clones. This figure portrays one such experiment, which identified clone 1A3-18 F10 as *alkB*-positive. Marker: 1kb DNA Ladder.

Note: PCR results from plates 1A3-49 through 1A3-53 are not shown in (A) for simplicity, as they were run on a different gel.

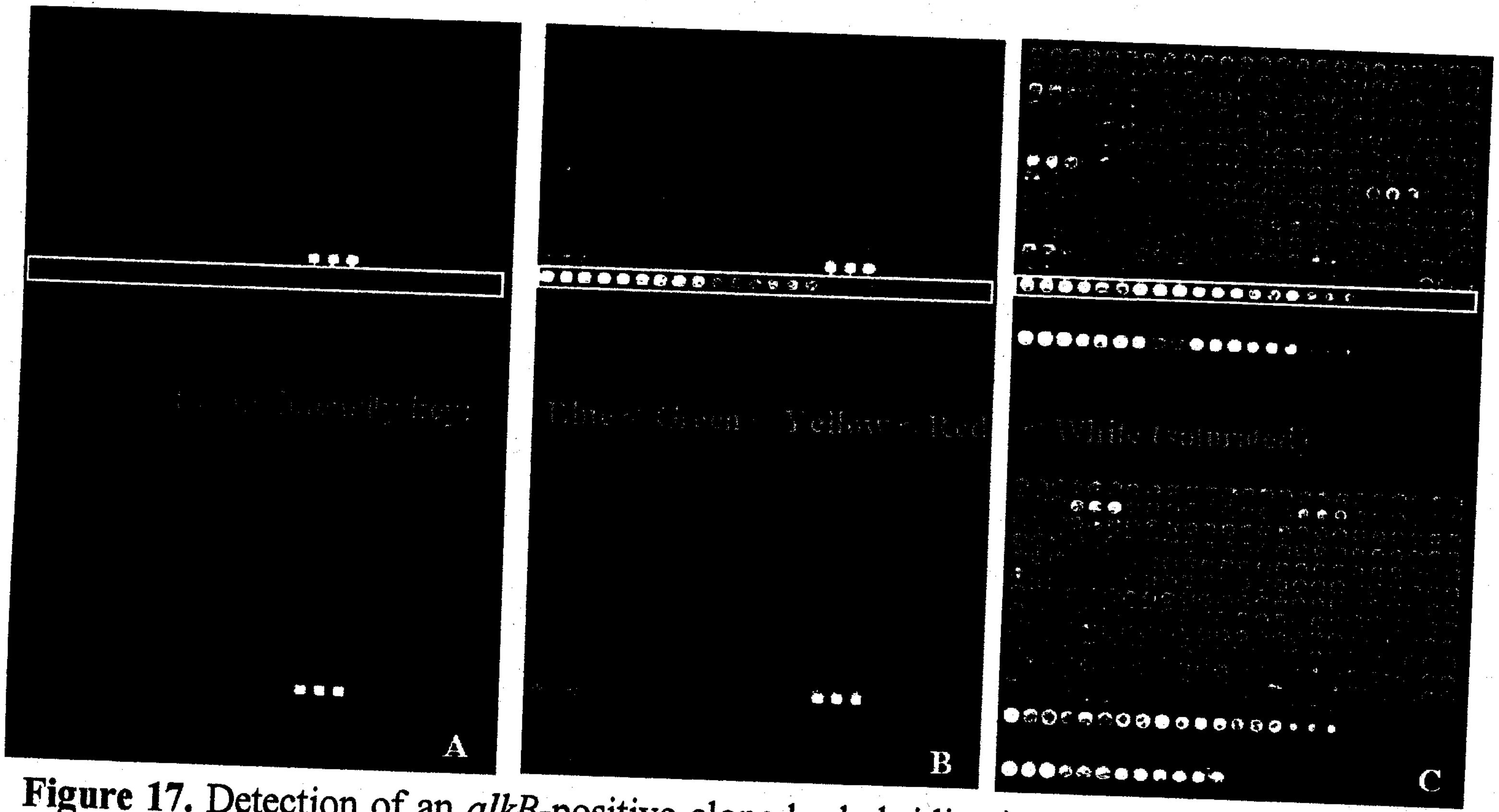


Figure 17. Detection of an *alkB*-positive clone by hybridization. A known *alkB*-positive clone is identified from among a larger number of clones on Test Array 6.2 by hybridization to *alkB* probes. Red box: clone 1A3-18 F10 (*alkB*-PCR-positive), purified by automated fosmid purification protocol. Yellow box: same clone purified by a commercial kit. *(*A) Hybridization of 1µg *alkB* (Cy5), produced from *Rhodococcus sp.* Q15 *alkB*, cloned into pDrive (standard *alkB*). (B) Hybridization of 1µg *alkBc* (Cy5), produced from *alkB* PCR amplicon of clone 1A3-18 F10 (clone-specific *alkB*). (C) SpotQC hybridization for total DNA (random nonamer probes)

The failure to detect the *alkB*-positive robot-purified clone by standard quantitative analysis of the *alkBc* hybridization can be explained by the results of the internal standard control hybridization performed alongside the experimental *alkBc* hybridization: informatic detection of *FOS-cos* positve spots also failed to detect the clone of interest, suggesting that not enough DNA from this clone had been purified during automated alkaline lysis. This was an unfortunate coincidence, as the *alkB*-positive clone was among the only 10 clones that could not be detected even under the less stringent requirements of a 1.5:1 signal-to-noise ratio (data not shown).

Unfortunately, the data reported here do not constitute proof-of-principle for the metagenomic microarray format. However, these results do not fully disprove the principle either.

Although informatic analysis of the *alkBc* hybridization under the standard conditions applied throughout this study failed to detect the *alkB*-positive automated alkaline lysis-prepared clone, reducing the stringency of the detection criteria resulted in selective identification of this clone, despite the fact that *FOS-cos* could not detect the clone even using less stringent detection criteria. When using the standard *alkB* probe (*Rhodococcus sp.* strain Q15 *alkB3*), the commercially-purified *alkB*-positive clone was detected at spotting concentrations above 400ng/µl (data not shown), suggesting that while it is theoretically possible to detect gene target-positive clones by hybridization to a similar, but non-identical gene probe, the current metagenomic microarray experiment design is too stringent to do so reliably.

Several options can still be explored in future studies to validate the potential of the metagenomic microarray. Instead of screening with a probe amplified from a reference organism (such as *Rhodococcus sp.* strain Q15 *alkB3* in this study) which may lack sufficient homology for detectable hybridization, or on the other extreme screening with a gene probe specific for a single library clone (which defeats the purpose of metagenomic microarray library screening), a third alternative should be explored. A probe could be constructed from the target gene amplified from the community DNA sample, which presumably would include all different forms of the target gene found in the library, although it is possible that PCR bias could lead to the exclusion of some particular genes from the probe set. This preparation may contain enough probe specific to each gene-positive clone (PCR bias notwithstanding) to allow for informatic detection. To aid in detection, signal-to-noise detection criteria can be relaxed, though this may increase incidence of false positive results. Alternately, clone DNA can be printed in higher amounts, although this may cause problems of sample loss and would raise the costs of metagenomic microarray production. As another possibility, clone DNA could be purified using commercial products, though this would increase production costs still further.

4 Conclusions

4.1 Future perspectives

It is unfortunate that the results of the proof-of-principle experiments could not unambiguously demonstrate the feasibility of the metagenomic microarray method developed over the course of the current project. It is particularly unfortunate that the results of the proof-of-principle experiments were overshadowed by the fact that the all-important *alkB*-positive automated protocol-purified clone was one of the very few which printed badly on Test Array 6.2. Consequently, the most optimistic conclusions that can be drawn from this study about the feasibility of the metagenomic microarray approach is that the evidence is inconclusive, but promising.

Two major technical obstacles to further development of metagenomic microarrays were identified in this project. A series of troubleshooting experiments produced a good deal of improvement on one of these problems, the low levels of signal detectable from large clones printed in extremely small amounts on the arrays. Based on the results reported by the group that developed the "Library-on-a-slide" (176) (who faced similar problems from large DNA printed in exceeding small amounts), we are confident that remedies such as sample sonication and secondary probe labelling can further minimize the problem of low clone signal. Such advancements push back further the limits of sensitivity of the microarray format, a process we have begun here with the successful detection of single-digit attomole quantities of fosmid DNA (600,000 to 6,000,000 individual molecules).

The other main technical obstacle, cross-hybridization of one gene probe with microarray controls for other probes, the cloning vector and even a few clones, remained unsolved at the end of this project. However, it is important to note that this problem was not distributed uniformly among all probes used over the course of the test array phase of this project. Some genes, such as *alkB*, exhibited no cross-hybridization that could not be attributed to human error. Others, like *GFP* proved so problematic that they had to be removed from the study. While this would not solve the problem entirely, it may be very useful in future explorations of metagenomic microarrays to pre-screen potential genetic targets for cross-hybridization behaviour prior to engaging in full-scale experiments. If the most problematic candidates are removed prior to the experiment and microarray design, it is quite possible that any remaining cross-hybridization be confined to background or near-background levels of signal. In particular, probes that only cross-hybridize to other probe controls at low levels of signal would not be likely to cross-hybridize detectably to library clones, which must be printed at much lower molar amounts than amplicon controls.

Assuming that the metagenomic microarray principle can eventually be proven to be unambiguously feasible, the future use of metagenomic microarrays in environmental microbiology applications ultimately requires an answer to a broader research question than could be addressed by the current study: is this tool able to successfully discern differences between closely-related microbial community samples? If it can be shown that library screening by metagenomic microarrays is at least as sensitive as other methods of community analysis, then the only obstacles to implementation of this method would be questions of cost and accessibility to the means of production.

However, the cost of this method in equipment, materials and technical expertise are high, and the speed and depth of analysis promised by metagenomic microarrays would likely exceed the practical requirements of most investigations in environmental microbiology (Craig Venter's Sargasso Sea research notwithstanding). This tool is much better suited to exploitation of the genetic wealth of microbial communities, to discover novel drugs, antibiotics, biocatalysts and other enzymatic or chemical products. Truly, the research and development budgets of biotechnology companies may be more appropriate to harness the technical and logistical resources required for this method. The steps taken in the current study to minimize production costs are helpful, but metagenomic microarrays still require a robotic liquid handler and an automated arrayer, as well as an automated colony picker.

Ultimately, metagenomic microarrays could find their best use as part of an integrated genomics-based approach to develop novel microbially-derived products, combining sequence-based screening (using metagenomic microarrays) and expression-based screening. Expression screens could uncover sequences that confer a desirable trait or function. These sequences would then serve as a basis for screening metagenomic microarrays of the same library (or another one) for different forms of the same gene which may prove more efficient, more amenable to large-scale culturing and production, or else to locate clones from the same organism that can help provide taxonomic identification.

In this scheme, glycerol stocks of metagenomic libraries stored at –80°C serve as a long-term record of all the genetic information stored in the microbial community. Once a list of genetic targets is compiled (and candidate probes are pre-screened to eliminate those most prone to cross-hybridization), a set of metagenomic microarrays could then be printed to conduct the desired screening. Since the total time required to go from –80°C stocks to printed microarrays is relatively short (by rough estimate, some 2,000 clones could be purified per day in a well-synchronized process, plus about two weeks for clone growth and microarray preparation and printing), new batches of arrays could be produced for new sets of genetic targets. The more targets for which to screen, the larger the library to be screened, the greater the benefit of this approach, whose main strength lies in the ability to rapidly screen huge amounts of sequence for multiple targets.

Admittedly, such applications are still far away from being realized; the current study represents a first step in a long process of development. It is our belief that the results presented here make the case for further study of the metagenomic microarray method, and it is our hope that others will do so in the future, in order to realize the potential of this powerful tool.

4.2 Summary

The current study was designed to explore if and how the two technologies of metagenomic libraries and DNA microarrays could be combined into a single platform, the metagenomic microarray. To the best of our knowledge, this approach represented a novel application of microarrays, in which DNA from metagenomic library clones was purified and printed on a microarray surface. Once printed, the library clones could then be rapidly screened for the presence of desired target genes by hybridization to single-gene probes. Ultimately, this approach is meant to be an alternative approach to library screening, one which is high-throughput and capable of screening for multiple targets with ease. One of the goals of the current study was to provide an initial exploration of the feasibility of this approach, and of the technical problems that must be surmounted before full-scale metagenomic microarrays can be produced.

The other goals of this project centered around the intermediate steps in producing a series of test arrays, small prototype metagenomic microarrays containing a very limited number of clones. Thus the first goal was the production of the necessary biological materials, in the form of 5,000-clone metagenomic libraries, constructed from two Arctic diesel-contaminated soils. These libraries were successfully produced and are currently stored at –80°C, forming a semi-permanent record of the microbial communities they represent (at least the cloned portion), whose sequence-level information can be accessed at any future time.

Another necessary intermediate goal was the development of a protocol that would purify clone DNA from an entire metagenomic library in a relatively short time. We thus developed a modified alkaline lysis method adapted for use on a robotic liquid handling workstation, which could purify DNA in an automated high-throughput manner appropriate to the isolation of clone DNA from an entire metagenomic library. We optimized this protocol to be suitable for the automated format and to be compatible with the equipment available at our research institution, to minimize associated costs and to produce cloned DNA of sufficient quality to allow hybridization signal detection when printed on microarrays. Although the final optimized method met all these requirements, the DNA purified was still of inferior quality to DNA purified by a commercial plasmid purification kit.

A small subset of the purified clone DNA was printed on a series of prototype test arrays. We used these arrays to address basic questions of method feasibility and sensitivity, to identify and address technical obstacles to reliable signal detection, and to optimize various parameters of future metagenomic microarray experiments such as necessary amounts of printed material, optimal amounts of labelled probe and other hybridization conditions.

Two major technical challenges arose over the course of this study: cross-hybridization between microarray controls, and chronically low hybridization signals from clone spots. We produced a good deal of improvement in the latter problem; some additional problem-solving approaches were identified but not attempted, which we have reason to believe would provide further improvement on this question. On the problem of control cross-hybridization, we provided substantial improvement by eliminating a consistent source of human error, but a lesser degree of cross-hybridization could only slightly be improved by a series of troubleshooting steps.

Experiments aimed at providing proof-of-principle for library screening by metagenomic microarrays provided evidence supporting the feasibility of this approach. We selectively identified a clone by hybridization bearing a desired target gene from a collection of nearly 200 printed clones, though only with a highly clone-specific probe, under detection conditions of relaxed stringency. The ambiguity of the results suggest that more experimentation be done to establish the feasibility of this method.

References

1. **Affymetrix** 2005, posting date. Products and Applications: Array Manufacturing. Affymetrix. [Online.]
2. **Aislabie, J., J. Foght, and D. Saul.** 2000. Aromatic hydrocarbon-degrading bacteria from soil near Scott Base, Antarctica. Polar Biol. **23:**183-188.
3. **Aislabie, J., R. Fraser, S. Duncan, and R. L. Farrell.** 2001. Effects of oil spills on microbial heterotrophs in Antarctic soils. Polar Biol. **24:**308-313.
4. **Aislabie, J. M., M. R. Balks, J. M. Foght, and E. J. Waterhouse.** 2004. Hydrocarbon spills on Antarctic soils: effects and management. Environ. Sci. Technol. **38:**1265-1274.
5. **Allison, D. B., X. Cui, G. P. Page, and M. Sabripour.** 2006. Microarrays data analysis: from disarray to consolidation and consensus. Nature Rev. Genet. **7:**55-65.
6. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403-410.
7. **Amann, R., W. Ludwig, and K.-H. Scleifer.** 1995. Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. Microbiol. Rev. **59:**143-169.
8. **Asakawa, S., I. Abe, Y. Kudoh, N. Kishi, Y. Wang, R. Kubota, J. Kudoh, K. Kawasaki, S. Minoshima, and N. Shimizu.** 1997. Human BAC library: construction and rapid screening. Gene **191:**69-79.
9. **Atlas, R. M.** 1981. Microbial degradation of petroleum hydrocarbons: an environmental perspective. Microbiol. Rev. **45:**108-209.
10. **Ball, K. D., and J. T. Trevors.** 2002. Bacterial genomics: the use of DNA microarrays and bacterial artificial chromosomes. J. Microbiol. Methods **49:**275-284.
11. **Bej, A. K., D. Saul, and J. Aislabie.** 2000. Cold-tolerant alkane-degrading *Rhodococcus* species from Antarctica. Polar Biol. **23:**100-105.
12. **Béjà, O., L. Aravind, E. V. Koonin, M. T. Suzuki, A. Hadd, L. P. Nguyen, S. B. Javanovich, C. M. Gates, R. A. Feldman, J. L. Spudich, E. N. Spudich, and E. F. DeLong.** 2000. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science **289:**1902-1906.
13. **Béjà, 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, and E. F. DeLong.** 2000. Construction and analysis of bactyerial artificial chromosome libraries from a marine microbial assemblage. Environ. Microbiol. **2:**516-529.
14. **Belcher, C. E., J. Drenkow, B. Kehoe, T. R. Gingeras, N. McNamara, H. Lemjabbar, C. Basbaum, and D. A. Relman.** 2000. The transcriptional responses of respiratory epithelial cells to *Bordtella pertussis* reveal host defensive and pathogen counter-defensive strategies. Proc. Nat. Acad. Sci. USA **97:**13847-13852.
15. **Bento, F. M., F. A. Camargo, B. C. Okeke, and W. T. Frankenberger.** 2005. Comparative bioremediation of soils contaminated with diesel oil by natural attenuation, biostimulation and bioaugmentation. Bioresour. Technol. **96:**1049-1055.
16. **Bischoff, K. M., D. G. White, P. F. McDermott, S. Zhao, S. Gaines, J. J. Maurer, and D. J. Nisbet.** 2002. Characterization of chloramphenicol resistance in beta-hemolytic *Escherichia coli* associated with diarrhea in neonatal swine. J. Clin. Microbiol. **40:**389-394.
17. **Bossert, I., and R. Bartha.** 1984. The fate of fuel spills in soil ecosystems, p. 435-473. *In* R. M. Atlas (ed.), Petroleum Microbiology. Macmillan, New York.
18. **Braddock, J. F., M. L. Ruth, J. L. Walworth, and K. A. McCarthy.** 1997. Enhancement and inhibition of microbial activity in hydrocarbon-contaminated arctic soils: implications for nutrient-ammended bioremediation. Environ. Sci. Technol. **31:**2078-2084.
19. **Brady, S. F., and J. Clardy.** 2000. Long-chain N-acyl amino acid antibiotics isolated from heterologously expressed environmental DNA. J. Am. Chem. Soc. **122:**12903-12904.

20. **Braker, G., A. Fesefeldt, and K.-P. Witzel.** 1998. Development of PCR primer systems for amplification of nitrite reductase genes (*nirS* and *nirK*) to detect denitrifying bacteria in environmental samples. Appl. Environ. Microbiol. **64:**3769-3775.

21. **Bulyk, B. L., X. Huang, Y. Choo, and G. M. Church.** 2001. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. Proc. Nat. Acad. Sci. USA **98:**7158-7163.

22. **Burke, D. T., G. F. Carle, and M. V. Olson.** 1987. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science **236:**806-812.

23. **Campbell, T. N., and F. Y. M. Choy.** 2002. Approaches to library screening. J. Mol. Microbiol. Biotechnol. **4:**551-554.

24. **Chen, L., R. Kosslak, and A. G. Atherly.** 1994. Mechanical shear of high molecular weight DNA in agarose plugs. Biotechniques **16:**228-229.

25. **Chénier, M. R., D. Beaumier, R. Roy, B. T. Driscoll, J. R. Lawrence, and C. W. Greer.** 2003. Impact of seasonal variations and nutrient inputs on nitrogen cycling and degradation of hexadecane by replicated reiver biofilms. Appl. Environ. Microbiol. **69:**5170-5177.

26. **Clarke, L., and J. Carbon.** 1976. A colony bank containing synthetic Col El hybrid plasmids representative of the entire *E. coli* genome. Cell **9:**91-99.

27. **Clarke, P. A., R. t. Poele, and P. Workman.** 2004. Gene expression microarray technologies in the development of new therapeutic agents. Eur. J. Cancer **40:**2560-2591.

28. **Cohn, D. H., A. J. Mileham, M. I. Simon, K. H. Nealson, S. K. rausch, D. Bonam, and T. O. Baldwin.** 1985. Nucleotide sequence of the *luxA* gene of *Vibrio harveyi* and the complete amino acid sequence of the [alpha] subunit of bacterial luciferase. J. Biol. Chem. **260:**6139-6146.

29. **Conjero-Goldberg, C., E. Wang, C. Yi, T. E. Goldberg, L. Jones-Brando, F. M. Marincola, M. J. Webster, and E. F. Torrey.** 2005. Infectious pathogen detection arrays: viral detection in cell ilnes and postmortem brain tissue. Biotechniques **39:**741-751.

30. **Coombes, B. K., and J. B. Mahony.** 2001. cDNA array analysis of altered gene expression in human endothelial cells in response to *Chlamydia pneumoniae* infection. Infect. Immunol. **69:**1420-1427.

31. **Costello, A. M., and M. E. Lidstrom.** 1999. Molecular characterization of functional and phylogenetic genes from natural populations of methanotrophs in lake sediments. Appl. Environ. Microbiol. **65:**5066-5074.

32. **Cottrell, M. T., J. A. Moore, and D. L. Kirchman.** 1999. Chitinases from uncultured marine microorganisms. Appl. Environ. Microbiol. **65:**2553-2557.

33. **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, and J. L. Pernodet.** 2003. Recombinant environmental libraries provide access to microbial diveristy for drug discovery from natural products. Appl. Environ. Microbiol. **69:**49-55.

34. **Croal, L. R., J. A. Gralnick, D. Malasarn, and D. K. Newman.** 2004. The Genetics of Geochemistry. Annu. Rev. Genet. **38:**175-202.

35. **Crouse, J., and D. Amorese.** 1989. Ethanol precipitation: ammonium acetate as an alternative to sodium acetate. Focus **9:**3-5.

36. **Delille, D.** 2004. Abundance and function of bacteria in the Southern Ocean. Cell. Mol. Biol. (Noisy-le-grand) **50:**543-551.

37. **Dennis, O., E. A. Edwards, S. N. Liss, and R. Fulthorpe.** 2003. Monitoring gene expression in mixed microbial communities by using DNA microarrays. Appl. Environ. Microbiol. **69:**769-778.

38. **Diaz-Torres, M. L., R. McNab, D. A. Spratt, A. Villedieu, N. Hunt, M. Wilson, and P. Mullany.** 2003. Novel tetracycline resistance determinant from the oral metagenome. Antimicrob. Agents Chemother. **47:**1430-1432.

39. **Drobyshev, A., N. Mologina, V. Shik, D. Pobedimaskaya, G. Yershov, and A. Mirzabekov.** 1997. Sequence analysis by hybridization with oligonucleotide microchip: identification of beta-thalassemia mutations. Gene **188:**45-52.

40. **Dua, M., A. Singh, N. Sethunathan, and A. K. Johri.** 2002. Biotechnology and bioremediation: successes and limitations. Appl. Microbio. Biotechnol. **59:**143-152.

41. **Dumont, M. G., and J. C. Murrell.** 2005. Stable isotope probing - linking microbial identity to function. Nature Rev. Microbiol **3:**499-504.

42. **Energy Information Administration** 2005, posting date. International Energy 2003. Energy Information Administration. [Online.]

43. **Entcheva, P., W. Liebl, A. johann, T. Hartsch, and W. R. Streit.** 2001. Direct cloning from enrighment cultures, a reliable strategy for isolation of complete operons and genes from microbial consortia. Appl. Environ. Microbiol. **67:**89-99.

44. **Environment Canada** 1998-04-07 2005, posting date. High Arctic Weather Stations - 50 Years of Operation. Environment Canada. [Online.]

45. **Epicentre Biotechnologies** 2005, posting date. EpiFOS fosmid library production kit. Epicentre Biotechnologies. [Online.]

46. **Eriksson, M., E. Sodersten, Z. Yu, G. Dalhammar, and W. M. Mohn.** 2003. Degradation of polycyclic aromatic hydrocarbons at low temperature under aerobic and nitrate-reducing conditions in enrichment cultures from northern soils. Appl. Environ. Microbiol. **69:**275-284.

47. **Fahy, A., G. Lethbridge, R. Earle, A. S. Ball, K. N. Timmis, and T. J. McGenity.** 2005. Effects of long-term benzene pollution on bacterial diversity and community structure in groundwater. Environ. Microbiol. **7:**1192-1199.

48. **Feiss, M., R. A. Fisher, M. A. Crayton, and C. Egner.** 1977. Packaging of the bacteriophage lambda chromosome: effect of chromosome length. Virology **77:**281-293.

49. **Feiss, M., and D. A. Siegele.** 1979. Packaging of the bacteriophage lambda chromosome: dependence of *cos* cleavage on chromosome length. Virology **92:**190-200.

50. **Fodor, S. P., J. L. Read, M. C. Pirrung, L. Stryer, A. T. Lu, and D. Solas.** 1991. Light-directed, spatially addressable parallel chemical synthesis. Science **251:**767-773.

51. **Fodor, S. P. A., R. P. Rava, X. C. Huang, A. C. Pease, C. P. Holmes, and C. L. Adams.** 1993. Multiplexed biochemical assays with biological chips. Nature **364:**555-556.

52. **Fortin, N., D. Beaumier, K. Lee, and C. W. Greer.** 2004. Soil washing improves the recovery of total community DNA from polluted and high organic content sediments. J Microbiol Methods **56:**181-91.

53. **Gabor, E. M., W. B. L. Alkema, and D. B. Janssen.** 2004. Quantifying the accessibility of the metagenome by random expression cloning techniques. Environ. Microbiol. **6:**879-886.

54. **Gao, X., E. Gulari, and X. Zhou.** 2004. In situ synthesis of oligonucleotide microarrays. Biopolymers **73:**579-596.

55. **Gerdes, B., R. Brinkmeyer, G. Dieckmann, and E. Helmke.** 2005. Influence of crude oil on changes of bacterial communities in Arctic sea-ice. FEMS Microbio. Ecol. **53:**129-139.

56. **Gillespie, D. E., S. F. Brady, A. D. Bettermann, N. P. Cianciotto, M. R. Liles, M. R. Rondon, J. Clardy, R. M. Goodman, and J. Handelsman.** 2002. Isolation of antibiotics turbomycin A and B from a metagenomic library of soil microbial DNA.

57. **Grant, G. M., A. Fortney, F. Gorreta, M. Estep, L. D. Giacco, A. V. Meter, A. Christensen, L. Appalla, C. Naouar, C. Jamison, A. Al-Timimi, J. Donovan, J. Cooper, C. Garrett, and V. Chandhoke.** 2004. Microarrays in cancer research. Anticancer Res. **24:**441-448.

58. **Greenberg, S. A.** 2001. DNA microarray gene expression analysis technology and its application to neurological disorders. Neurology **57:**755-761.

59. **Greer, C. W., L. G. Whyte, J. R. Lawrence, L. Masson, and R. Brousseau.** 2001. Genomics technologies for environmental science. Environ. Sci. Technol. **35:**360A-366A.

60. **Gupta, R., Q. K. Beg, and P. Lorenz.** 2002. Bacterial alkaline proteases; molecular approaches and industrial applications. Appl. Microbio. Biotechnol. **59:**15-32.

61. **Hamamura, N., C. M. Yeager, and D. J. Arp.** 2001. Two distinct monooxygenases for alkane oxidation in *Nocardioides sp.* Strain CF8. Appl. Environ. Microbiol. **67:**4992-4998.

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

63. **Hayward, R. E., J. L. DeRisi, S. Alfadhli, D. C. Kaslow, P. O. Brown, and P. K. Rathod.** 2000. Shotgun DNA microarrays and stage-specific gene expression in *Plasmodium falciparum* malaria. Mol. Microbiol. **35:**6-14.

64. **Henne, A., R. Daniel, R. A. Schmitz, and G. Gottschalk.** 1999. Construction of environmental DNA libraries in *Escherichia coli* and screening for the presence of genes conferring utilization of 4-hydroxybutyrate. Appl. Environ. Microbiol. **65:**3901-3907.

65. **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 on *Escherichia coli.* Appl. Environ. Microbiol. **66:**3113-3116.

66. **Holland, H. D.** 1999. When did the earth's atmosphere become oxic? A reply. Geochem. News **100:**20-23.

67. **Holmes, A. J., A. Costello, M. E. Lidstrom, and J. C. Murrell.** 1995. Evidence that participate methane monooxygenase and ammonia monooxygenase may be evolutionarily related. FEMS Microbiol. Lett. **132:**203-208.

68. **Horvath, R. S.** 1972. Microbial co-metabolism and the degradatino of organic compounds in nature. Bacteriol. Rev. **36:**146-155.

69. **Horz, H. P., M. T. Yimga, and W. Liesack.** 2001. Detection of methanotroph diversity on roots of submerged rice plants by molecular retrieval of pmoA, mmoX, mxaF and 16S rRNA, including pmoA-based terminal restriction fragment length polymorphism profiling. Appl. Environ. Microbiol. **67:**4177-4185.

70. **Hu, Y. F., J. kaplow, and Y. He.** 2005. From traditional biomarkers to transcriptome analysis in drug development. Curr. Mol. Med. **5:**29-38.

71. **Ichikawa, J. K., A. Norris, M. G. Bangera, G. K. geiss, A. B. v. t. Wout, R. E. Bumgarner, and S. Lory.** 2000. Interaction of *Pseudomonas aeruginosa* with epithelial cells: identification of differentially regulared genes by expression microarray analysis of human cDNAs. Proc. Nat. Acad. Sci. USA **97:**9659-9664.

72. **Imbeaud, S., and C. Auffray.** 2005. *'The 39 steps'* in gene expression profiling: critical issues and proposed best practices for microarray experiments. Drug Discov. Today **10:**1175-1182.

73. **Ishkanian, A., S. Watson, C. Malloff, B. Coe, R. DeLeeuw, M. Krzywinski, M. marra, C. MacAulay, and W. Lam.** 2003. Construction of a DNA microarray with complete coverage of the human genome. Lung Cancer **41:**S60.

74. **Itoh, M., T. Kitsunai, J. Akiyama, K. Shibata, M. Izawa, J. Kawai, Y. Tomaru, P. Carninci, Y. Shibata, Y. Ozawa, M. Muramatsu, Y. Okazaki, and Y. Hayashizaki.** 1999. Automated filtration-based high-thoughput plasmid preparation system. Genome Res. **9:**463-470.

75. **Juck, D., T. Charles, L. G. Whyte, and C. W. Greer.** 2000. Polyphasic microbial community analysis of petroleum hydrocarbon-contaminated soils from two northern Canadian communities. FEMS Microbiol. Ecol. **33:**241-249.

76. **Juck, D. F., J. Ingram, M. Prévost, J. Coallier, and C. W. Greer.** 1996. Nested PCR protocol for the rapid detection of *Escherichia coli* in potable water. Can. J. Microbiol. **42:**862-866.

77. **Ka, J. O., Z. Yu, and W. W. Mohn.** 2001. Monitoring the size and metabolic activity of the bacterial community during biostimulation of fuel-contaminated soil using competitive PCR and RT-PCR. Microbial Ecol. **42:**267-273.

78. **Kafatos, F. C., C. W. Jones, and A. Efstratiadis.** 1979. Determination of nucleic acid sequence homologies and relative concentrations by a dot hybridization procedure. Nucleic Acids Res. **7:**1541-1552.

79. **Kato-Maeda, M., J. T. Rhee, T. R. Gingeras, H. Salamon, J. Drenkow, N. Smittipat, and P. M. Small.** 2001. Comparing genomes within the species *Mycobacterium tuberculosis*. Genome Res. **11:**547-554.

80. **Khrapko, K. R., Y. P. Lysov, A. A. Khorlin, I. B. Ivanov, G. M. Yershov, S. K. Vasilenko, V. L. Florentiev, and A. D. Mirzabekov.** 1991. A method for DNA sequencing by hybridization with oligonucleotide matrix. DNA Seq. **1:**375-388.

81. **Kim, U.-J., H. Shizuya, P. J. d. Jong, B. Birren, and M. I. Simon.** 1992. Stable propagation of cosmid sized human DNA inserts in an F factor based vector. Nucleic Acids Res. **20:**1083-1085.

82. **Knietsch, A., S. Bowien, G. Whited, G. Gottschalk, and R. Daniel.** 2003. Identification and characterization of coenzyme B12-dependent glycerol dehydratase- and diol dehydratase-encoding genes from metagenomic DNA libraries derived from enrighment cultures. Appl. Environ. Microbiol. **69:**3048-3060.

83. **Koob, M., and W. Szybalski.** 1992. Preparing and using agarose microbeads. Methods Enzymol. **216:**13-20.

84. **Kuske, C. R., K. L. Banton, D. L. Adorada, P. C. Stark, K. K. Hill, and P. L. Jackson.** 1998. Small-scale DNA sample preparation method for field PCR detection of microbial cells and spores in soil. Appl. Environ. Microbiol. **64:**2463-2472.

85. **Labbé, D., L. G. Whyte, J. Hawari, and C. W. Greer.** 2004. Eureka Project: Bioremediation treatment of hydrocarbon contaminated soils from Eureka, Nunavut. Phase 3 - Final Report. NRC Biotechnology research Institute.

86. **Lander, R. J., M. A. Winters, F. J. Meacle, B. C. Buckland, and A. L. Lee.** 2002. Fractional precipitation of plasmid DNA from lysate by CTAB. Biotechnol. Bioeng. **79:**776-784.

87. **Larin, Z., A. P. Monaco, and H. Lehrach.** 1991. Yeast artificial chromosome libraries containing large inserts from mouse and human DNA. Proc. Nat. Acad. Sci. USA **88:**4123-4127.

88. **Lee, S., C. Malone, and P. F. Kemp.** 1993. Use of multiple 16S rRNA-targeted fluorescent probes to increase signal strength and measure cellular RNA from natural planktonic bacteria. Mar. Ecol. Prog. Ser. **101:**193-201.

89. **Lennon, G. G., and H. Lehrach.** 1991. Hybridization analyses of arrayed cDNA libraries. Trends Genet. 7:314-317.

90. **Li, H., Y. Zhang, C. G. Zhang, and G. X. Chen.** 2005. Effect of petroleum-containing wastewater irrigation on bacterial diversities and enzymatic activities in a paddy soil irrigation area. J. Environ. Qual. **34:**1073-1080.

91. **Liang, F., M. Lu, T. C. Keezer, Z. Liu, and S.-J. Khang.** 2005. The organic composition of diesel particulate matter, diesel fuel and engine oil of a non-road diesel generator. J. Environ. Monit. **7:**932-988.

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

93. **Lipshutz, R. J., S. P. Fodor, T. R. Gingeras, and D. J. Lockhart.** 1999. High density synthetic oligonucleotide arrays. Nature Genet. **21:**20-24.

94. **Lockhart, D. J., and E. A. Winzeler.** 2000. Genomics, gene expression and DNA arrays. Nature **405:**827-836.

95. **Lorenz, P., K. Liebeton, F. Niehaus, and J. Eck.** 2002. Screening for novel enzymes for biocatalytic processes: accessing the metagenome as a resource of novel functional sequence space. Curr. Opin. Biotechnol. **13:**572-577.

96. **Loy, A., A. Lehner, N. Lee, J. Adamczyk, H. Meier, J. Ernst, K. H. Schleifer, and M. Wagner.** 2002. Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. Appl. Environ. Microbiol. **68:**5064-5081.

97. **Lueders, T., and M. Friedrich.** 2000. Archaeal population dynamics during sequenctial reduction processes in rice field soil. Appl. Environ. Microbiol. **66:**2732-2742.

98. **Majernik, A., G. Gottschalk, and R. Daniel.** 2001. Screening of environmental DNA libraries for the presence of genes conferring Na(+)(Li(+))/H(+) antiporter activity on

Escherichia coli: characterization of the recovered genes and the corresponding gene products. J. Bacteriol. **183:**6645-6653.

99. **Margesin, R., D. Labbé, F. Schinner, C. W. Greer, and L. G. Whyte.** 2003. Characterization of hydrocarbon-degrading microbial populations in contaminated and pristine alpine soils. Appl. Environ. Microbiol. **69:**3085-3092.

100. **Margesin, R., and F. Schinner.** 2001. Biodegradation and bioremediation of hydrocarbons in extreme environments. Appl. Microbio. Biotechnol. **56:**650-663.

101. **Margesin, R., and F. Schinner.** 1999. Biological decontamination of oil spills in cold environments. J. Chem. Technol. Biotechnol. **74:**381-389.

102. **Margesin, R., and F. Schinner.** 2001. Bioremediation (natural attenuation and biostimulation) of diesel-oil-contaminated soil in an alpine glacier skiing area. Appl. Environ. Microbiol. **67:**3127-3133.

103. **Margesin, R., and F. Schinner.** 1997. Bioremediation of diesel-oil-contaminated alpine soils at low temperatures. Appl. Microbiol. Biotechnol. **47:**462-468.

104. **Margesin, R., and F. Schinner.** 1997. Laboratory bioremediation experiments with soil from a diesel-oil contaminated site: Significant role of cold-adapted microorganisms and fertilizers. J. Chem. Technol. Biotechnol. **70:**92-98.

105. **Martinez, A., S. J. kolvek, C. L. T. Yip, J. Hopke, K. A. Brown, I. A. MacNeil, and 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. Appl. Environ. Microbiol. **70:**2452-2463.

106. **Maskos, U., and E. M. Southern.** 1992. Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ. Nucleic Acids Res. **20:**1679-1684.

107. **Meyers, S. P.** 2000. Developments in aquatic microbiology. Int. Microbiol. **3:**203-211.

108. **Miguez, C. B., D. Bourque, J. A. Sealy, C. W. Greer, and D. Groleau.** 1997. Detection and isolation of methanotrophic bacteria possessing soluble methane monooxygenase (sMMO) genes using the polymerase chain reaction. Microbial Ecol. **33:**21-31.

109. **Mocellin, S., M. Provenzano, C. R. Rossi, P. Pilati, D. Nitti, and M. Lise.** 2005. DNA array-based gene profiling: from surgical specimen to the molecular portrait of cancer. Ann. Surg. **241:**16-26.

110. **Morgan, P., and R. J. Watkinson.** 1989. Hydrocarbon degradation in soils and methods for soil biotreatment. CRC Crit. Rev. Biotechnol. **8:**305-333.

111. **Murray, A. E., D. Lies, G. Li, K. Nealson, J. Zhou, and J. M. Tiedje.** 2001. DNA/DNA hybridization to microarrays reveals gene-specific differences between closely related microbial genomes. Proc. Nat. Acad. Sci. USA **98:**9853-9858.

112. **Muyzer, G., E. C. d. Waal, and 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. Appl. Environ. Microbiol. **59:**695-700.

113. **Nadon, R., and J. Schoemaker.** 2002. Statistical issues with microarrays: processing and analysis. Trends Genet. **18:**265-271.

114. **Pease, A. C., D. Solas, E. J. Sullivan, M. T. Cronin, C. P. Holmes, and S. P. A. Fodor.** 1994. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc. Nat. Acad. Sci. USA **91:**5022-5026.

115. **Piel, J.** 2002. A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of Paederus beetles. Proc. Nat. Acad. Sci. USA **99:**14002-14007.

116. **Qiagen** 2003, posting date. R.E.A.L. Prep 96 Biorobot Kit. Qiagen Inc. [Online.]

117. **Quaiser, A., T. Ochsenreiter, H.-P. Klenk, A. Kletzin, A. H. Treusch, G. Meurer, J. Eck, C. W. Sensen, and C. Schleper.** 2002. First insight into the genome of an uncultivated crenarchaeote from soil. Environ. Microbiol. **4:**603-611.

118. **Rainey, F. A., N. Ward, I. Sly, and E. Stackebrandt.** 1994. Dependence on the taxon composition of clone libraries for PCR amplified, naturally occurring 16S rDNA, on the primer pair and the cloning system used. Experientia **50:**796-797.

119. **Ren, T., R. Roy, and R. Knowles.** 2000. Production and consumption of nitric oxide by three methanotrophic bacteria. Appl. Environ. Microbiol. **66:**3891-3897.

120. **Reyes-Lopez, M. A., A. Mendez-Tenorio, R. Maldonado-Rodriguez, M. J. Doktycz, J. T. Flemmin, and K. L. Beattie.** 2003. Fingerprinting of prokaryotic 16S rRNA genes using oligodeoxyribonucleotide microarrays and virtual hybridization. Nucleic Acids Res. **31:**779-789.

121. **Riesenfeld, C. S., P. D. Schloss, and J. Handelsman.** 2004. Metagenomics: genomic analysis of microbial communities. Annu. Rev. Genet. **38:**525-552.

122. **Rodriguez-Valera, F.** 2004. Environmental genomics, the big picture? FEMS Microbiol. Lett. **231:**153-158.

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

124. **Ruppert, A., B. Szalay, D. v. d. Boom, G. Horst, and H. Koster.** 1995. A filtration method for plasmid isolation using microtiter filter plates. Anal. Biochem. **230:**130-134.

125. **Salama, N., K. Guillemin, T. K. McDaniel, G. Sherlock, and L. Tompkins.** 2000. A whole-genome microarray reveals genetic diversity among *Heliobacter pylori* strains. Proc. Nat. Acad. Sci. USA **97:**14668-14673.

126. **Sambrook, J., and D. W. Russell.** 2001. Molecular Cloning: a Laboratory Manual (Third Edition). Cold Spring Harbor Laboratory Press, Woodbury, NY.

127. **Sasmal, D., T. A. Qureshi, and T. J. Abraham.** 2005. Comparison of antibiotic resistance in bacterial flora of shrimp farming systems. Internet J. Microbiol **1.**

128. **Saul, D. J., J. M. Aislabie, C. E. Brown, L.Harris, and J. M. Foght.** 2005. Hydrocarbon contamination changes the bacterial diversity of soil from around Scott Base, Antarctica. FEMS Microbiol. Ecol. **53:**141-155.

129. **Sausville, E. A., and S. L. Holbeck.** 2004. Transcription profiling of gene expression in drug discovery and development: the NCI experience. Eur. J. Cancer **40:**2544-2549.

130. **Sayler, G. S., S. W. Hooper, A. C. Layton, and J. M. H. King.** 1990. Catabolic plasmids of environmental and ecological significance. Microbial Ecol. **19:**1-20.

131. **Schadt, E. E., S. W. Edwards, D. GuhaThakurta, D. Holder, L. Ying, V. Svetnik, A. Leonardson, K. W. Hart, A. Russell, G. Li, G. Cavet, J. Castle, P. McDonagh, Z. Kan, R. Chen, A. Kasarskis, M. Margarint, R. M. Caceres, J. M. Johnson, C. D. Armour, P. W. Garrett-Engele, N. F. Tsinoremas, and D. D. Shoemaker.** 2004. A comprehensive transcript index of the human genome generated using microarrays and computational approaches. Genome Biol. **5:**R73.

132. **Schena, M., R. A. Heller, T. P. Theriault, K. Konrad, E. Lachenmeier, and R. W. Davis.** 1998. Microarrays: biotechnology's discovery platform for functional genomics. Trends Biotechnol. **16:**301-306.

133. **Schena, M., D. Shalon, R. W. Davis, and P. O. Brown.** 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **270:**467-470.

134. **Schloss, P. D., and J. Handelsman.** 2003. Biotechnological prospects from metagenomics. Curr. Opin. Biotechnol. **14:**303-310.

135. **Schmeisser, C., C. Stöckigt, C. Raasch, J. Wingender, K. N. Timmis, D. F. Wenderoth, H.-C. Flemming, H. Liesegang, R. A. Schmitz, K.-E. Jaeger, and W. R. Streit.** 2003. Metagenome survey of biofilms in drinking-water networks. Appl. Environ. Microbiol. **69:**7298-7309.

136. **Schoolnik, G. K.** 2002. Functional and comparative genomics of pathogenic bacteria. Curr. Opin. Microbiol. **5:**20-26.

137. **Searfoss, G. H., T. P. Ryan, and R. A. Jolly.** 2005. The role of transcriptome analysis in pre-clinical toxicology. Curr. Mol. Med. **5:**53-64.

138. **Sebat, J. L., F. S. Colwell, and R. L. Crawford.** 2003. Metagenome profiling: microarray analysis of an environmental genomic library. Appl. Environ. Microbiol. **69:**4927-4934.

139. **Seow, K. T., G. Meurer, M. Gerlitz, E. Wendt-Pienkowski, C. R. hutchison, and J. Davies.** 1997. A study of iterative type II polyketide synthases, using bacterial genes cloned from soil DNA: a means to access and use genes from uncultured microorganisms. J. Bacteriol. **179:**7360-7368.

140. **Shi, L., W. Tong, Z. Su, T. Han, J. Han, R. K. Puri, H. Fang, F. W. Frueh, F. M. Goodsaid, L. Guo, W. S. Branham, J. J. Chen, Z. A. Xu, S. C. Harris, H. Hong, Q. Xie, R. G. Perkins, and J. C. Fuscoe.** 2005. Microarray scanner calibration curves: characteristics and implications. BMC Bioinformatics **6:**S11.

141. **Shimomura, O., F. H. Johnson, and Y. Saiga.** 1962. Extraction, purification and properties of Aequorin, a bioluminescent protein from luminous Hydromedusan, Aequorea. J. Cell Comp. Physiol. **59:**223-239.

142. **Shizuya, H., B. Birren, U.-J. Kim, V. Mancino, T. Slepak, Y. Tachiiri, and M. Simon.** 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. Proc. Nat. Acad. Sci. USA **89:**8794-8797.

143. **Siddiqui, S., and W. A. Adams.** 2002. The fate of diesel hydrocarbons in soils and their effect on the germination of perennial ryegrass. Environ. Toxicol. **17:**49-62.

144. **Slinger, D. W., K. J. Cheung, R. Mei, E. M. Johansson, C. S. Richmond, F. R. Blattner, D. J. Lockhart, and G. M. Church.** 2000. RNA expression analysis using a 30-base pair resolution *Escherichia coli* genome array. Nature Biotechnol. **18:**1262-1268.

145. **Small, J., D. R. call, F. J. Brockman, T. M. Straud, and D. P. Chandler.** 2001. Direct detection of 16S rRNA in soil extracts by using oligonucleotide microarrays. Appl. Environ. Microbiol. **67:**4708-4716.

146. **Song, D., and A. Katayama.** 2005. Monitoring microbial community in a subsurface soil contaminated with hydrocarbons by quinone profile. Chemosphere **59:**305-314.

147. **Sorensen, T.** 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. K. Dan. Vidensk. Selsk. Biol. Skr. **5:**1-34.

148. **Sotsky, J. B., C. M. Greer, and R. M. Atlas.** 1994. Frequency of genes in aromatic and aliphatic hydrocarbon biodegradation pathways within bacterial populations from Alaskan sediments. Can. J. Microbiol. **40:**981-985.

149. **Southern, E. M., S. C. Case-Green, J. K. Elder, M. Johnson, K. U. Mir, L. Wang, and J. C. Williams.** 1994. Arrays of complementary oligonucleotides for analysing the hybridisation behaviour of nucleic acids. Nucleic Acids Res. **22:**1368-1373.

150. **Spiegelman, D., G. Whissell, and C. W. Greer.** 2005. A survey of the methods for the characterization of microbial consortia and communities. Can. J. Microbiol. **51:**355-386.

151. **Stallwood, B., J. Shears, P. A. Williams, and K. A. Hughes.** 2005. Low temperature bioremediation of oil-contaminated soil using biostimulation and bioaugmentation with a *Pseudomonas* sp. from maritime Antarctica. J. Appl. Microbiol. **99:**794-802.

152. **Stein, J. L., T. L. Marsh, K. Y. Wu, H. Shizuya, and E. F. DeLong.** 1996. Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. J. Bacteriol. **178:**591-599.

153. **Suzuki, M., M. S. Rappé, and S. J. Giovannoni.** 1998. Kinetic bias in estimates of coastal picoplankton community structure obtained by measurements of small-subunit rRNA gene PCR amplicon length heterogeneity. Appl. Environ. Microbiol. **64:**4522-4529.

154. **Thomas, R., A. Scott, C. F. Langford, S. P. Fosmire, C. M. Jubala, T. D. Lorentzen, C. Hitte, E. K. Karlsson, E. Kirkness, E. A. Ostrander, F. Galibert, K. Lindblad-Toh, J. F. Modiano, and M. Breen.** 2005. Construction of a 2-Mb resolution BAC microarray for CGH analysis of canine tumors. Genome Res. **15:**1831-1837.

155. **Thomassin-Lacroix, E. J., M. Eriksson, K. J. Reimer, and W. W. Mohn.** 2002. Biostimulation and bioaugmentation for on-site treatment of weathered diesel fuel in arctic soil. Appl. Microbio. Biotechnol. **59:**551-556.

156. **Torsvik, V., J. Goksoyr, and F. L. Daae.** 1990. High diversity in DNA of soil bacteria. Appl. Environ. Microbiol. **56:**782-787.

157. **Torsvik, V., L. Ovreas, and T. F. Thingstad.** 2002. Prokaryotic diversity - magnitude, dynamics and controlling factors. Science **296:**1064-1066.

158. **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, and J. F. Banfield.** 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature **428:**37-43.

159. **Uchiyama, T., T. Abe, T. ikemura, and K. Watanabe.** 2005. Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. Nature Biotechnol. **23:**88-93.

160. **Urbach, E., K. L. Vergin, and S. J. Giovenannoni.** 1999. Immunochemical detection and isolation of DNA from metabolically active bacteria. Appl. Environ. Microbiol. **65:**1207-1213.

161. **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, and H. O. Smith.** 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science **304:**66-74.

162. **Wang, D. G., J.-B. Fan, C.-J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, L. Hsie, T. Topaloglou, E. Hubbell, E. Robinson, M. Mittmann, M. S. Morris, N. Shen, D. Kilburn, K. Rioux, C. Nusbaum, S. Rozen, T. J. Hudson, R. Lipshutz, M. Chee, and E. S. Lander.** 1998. Large-scale identification, mapping and genotyping of single-nucleotide polymorphisms in the human genome. Science **280:**1077-1082.

163. **Wang, G. Y., E. Graziani, B. Walters, W. Pan, X. Li, J. McDermott, G. Meurer, G. Saxena, R. J. Anderson, and J. Davies.** 2000. Novel natural products from soil DNA libraries in a streptomycete host. Org. Lett. **2:**2401-2404.

164. **WHO Drafting Group on Environmental Health Criteria for Diesel Fuel and Exhaust Emissions, and WHO Task Group on Environmental Health Criteria for Diesel Fuel and Exhaust Emissions** 1996, posting date. Environmental Health Criteria for Diesel Fuel and Exhaust Emissions. IPCS INCHEM (International Programme on Chemical Safety). [Online.]

165. **Whyte, L. G., L. Bourbonnière, C. Bellerose, and C. W. Greer.** 1999. Bioremediation assessment of hydrocarbon-contaminated soils form the High Arctic. Bioremediation **3:**69-79.

166. **Whyte, L. G., L. Bourbonnière, and C. W. Greer.** 1997. Biodegradation of petroleum hydrocarbons by psychrotrophic *Pseudomonas* strains possessing both alkane (*alk*) and naphthalene (*nah*) catabolic pathways. Appl. Environ. Microbiol. **63:**3719-3723.

167. **Whyte, L. G., B. Goalen, J. Hawari, D. Labbé, C. W. Greer, and M. Nahir.** 2001. Bioremediation treatability assessment of hydrocarbon-contaminated soils from Eureka, Nunavut. Cold Reg. Sci. Technol. **32:**121-132.

168. **Whyte, L. G., A. Schultz, J. B. v. Beilen, A. P. Luz, V. Pellizari, D. Labbé, and C. W. Greer.** 2002. Prevalence of alkane monooxygenase genes in Arctic and Antarctic hydrocarbon-contaminated and pristine soils. FEMS Microbio. Ecol. **41:**141-150.

169. **Wieczorek, D. J., and M. Feiss.** 2001. Defining cosQ, the site required for temination of bacteriophage lambda DNA packaging. Genetics **158:**495-506.

170. **Wilson, L. P., and E. J. Bouwer.** 1997. Biodegradation of aromatic compounds under mixed oxygen/denitrifying conditions: a review. J. Ind. Microbiol. Biotechnol. **18:**116-130.

171. **Woese, C. R.** 1987. Bacterial evolution. Microbiol. Rev. **51:**221-271.

172. **Woo, S.-S., J. Jiang, B. S. Gill, A. H. Paterson, and R. A. Wing.** 1994. Construction and characterization of a bacterial artificial chromosome library of *Sorghum bicolor*. Nucleic Acids Res. **22:**4922-4931.

173. **Wu, L., D. K. Thompson, G. Li, R. A. Hurt, J. M. Tiedje, and J. Zhou.** 2001. Development and evaluation of functional gene arrays for detection of selected genes in the environment. Appl. Environ. Microbiol. **67:**5780-5790.

174. **Ye, R. W., T. Wang, L. Bedzyk, and K. M. Croker.** 2001. Applications of DNA microarrays in microbial systems. J. Microbiol. Methods **47:**257-272.

175. **Yun, J., and S. Ryu.** 2005. Screening for novel enzymes from metagenome and SIGEX, as a way to improve it. Microb. Cell Fact. **4:**1-5.

176. **Zhang, L., U. Srinivasan, C. F. Marrs, D. Ghosh, J. R. Gilsdorf, and B. Foxman.** 2004. Library on a slide for bacterial comparative genomics. BMC Microbiol. **4:12.**

177. **Zhou, J., and D. K. Tompson.** 2002. Challenges in applying microarrays to environmental studies. Curr. Opin. Biotechnol. **13:**204-207.

178. **Zimmer, D. P., E. Soupene, H. L. Lee, V. F. Wendisch, V. F. Khodursky, B. J. Peter, R. A. Bender, and S. Kustu.** 2000. Nitrogen regulatory protein C-controlled genes of *Escherichia coli*: scavenging as a defense against nitrogen limitation. Proc. Nat. Acad. Sci. USA **97:**14674-14679.

179. **Zimmer, R., and A. M. V. Gibbins.** 1997. Construction and characterization of a large-fragment chicken bacterial artificial chromosome library. Genomics **42:**217-226.

180. **Zobell, C. E.** 1946. Action of microorganisms on hydrocarbons. Bacteriol. Rev. **10:**1-49.