Bioprospecting soil metagenomes for potential new antibiotics

A thesis presented in partial fulfilment of the requirements for the degree of

Master of Science

In

Genetics

At Massey University, Albany, New Zealand

Kelly Hong

2017

0.1 Abstract

Many soil-dwelling microbes have the natural capacity to produce toxic compounds that inhibit growth of competing bacteria; most traditional antibiotics have been derived from small molecules made by such soil-based microorganisms, of which only a small fraction can be grown in the laboratory. Since techniques that require culturing of these microbes in the lab have been the starting point for studying them in the past, our knowledge of the uncultured majority remains limited. Functional metagenomics is a method that circumvents the need for culturing, and thus has the potential to revel a yet untapped reservoir of antibacterial compounds. Here we present a potential application of functional metagenomics using genes isolated from soil microbes that employs high throughput sequencing to identify microbial genes encoding novel compounds that inhibit bacterial growth.

0.2 Acknowledgements

First and foremost, I would like to thank my supervisor, Nikki Freed, for providing me an environment that was always open to new ideas, and encouraging me to take every opportunity to learn new tools that enabled me to grow as a researcher. I would also like to acknowledge my co-supervisor, Olin Silander, for his constant constructive criticisms that challenged me and made every day a bit more interesting. Together, I could not have wished for better support, and I feel incredibly grateful for their mentorship.

To the friends I have made; especially Danielle, Joanna, and Richard – I now consider us a tiny tribe of science warriors.

Thank you to Wayne Patrick of Otago University for providing us with the plasmid.

And last but not least, mum & dad – thank you for your unconditional love and support.

0.3 Table of Contents

0.4 List of Figures

0.6 List of Tables

0.7 Abbreviations

eDNA: environmental DNA

pBAD: pBAD/*Myc*-His B, the plasmid cloning vector used

NGS: next generation sequencing

PCR: polymerase chain reaction

E. coli: *Escherichia coli*

TOPO: PCR8/GW/TOPO cloning vector (Invitrogen)

TSS: transformation and storage solution

LB: lysogeny broth

SOC: super optimal broth with catabolite repression

bp: base pairs

kbp or kb: kilo base pairs

OD600: optical density at 600nm. Indicator of bacterial growth.

ORF: open reading frame

1 Introduction

Microbes have evolved over billions of years [1] to thrive in almost every habitat on Earth [2-6], from boiling hot springs to arctic glaciers [7-9]. Encoded within their genomes are the blueprints of the myriad of molecules they synthesize, including antibiotics [10]. Antibiotics are compounds that kill bacteria or inhibit bacterial growth [11], and enable the microbes that produce them to enhance their own fitness by limiting the growth of competing microbes [12]. Many commercially available antibiotics are derived from these naturally produced antibiotics. Most traditional antibiotics have been discovered by growing microorganisms in the lab to isolate compounds that inhibit the growth of a bacterial culture [13]. However, it is estimated that over 99% of microbial diversity has not been successfully cultured to date [14-16], and consequently, the large number of antibiotics they may be producing remain unknown.

To cope with the increasing demand for new antibiotics, the biosynthetic potential of the remaining 99% of the gene pool of microbes reluctant to laboratory cultivation could be accessed to increase the chances of finding a novel class of antibiotic. One method that circumvents the need for culturing microbes altogether is functional metagenomics [17-20]. Functional metagenomics is an approach where the collective pool of DNA from mixed microbial populations (known as the metagenome) in an environmental sample is extracted, fragmented, and used to construct a vector-based library. This metagenomic library is then inserted into a heterologous (foreign) host organism that grows well in the laboratory, such as *Escherichia coli* [21]. Genetic information from the uncultured microorganisms can then subjected to functional screening by driving protein expression in the host bacterium. By phenotypically

screening the clones of the library under a particular set of conditions, it is possible to detect desired biological activity. In the search for possible new antibiotics, screening would, for example, seek to identify a clone carrying an insert that significantly impedes the growth of its host bacterium when expressed. By tapping into the unprecedented biosynthetic diversity encoded within the genomes of uncultured microbes, functional metagenomics offers a way forward in natural product discovery by overcoming one major hurdle of microbiology: unculturability.

1.1 Bioprospecting soil for novel antibiotics

Bioprospecting is the search for valuable products from natural resources [22]. With the benefit of billions of years of evolution [1], microbes have adapted to survive extreme conditions [23] and produce an immense array of natural products [13, 22, 24-26] that reflect their complexity and functional diversity. Since Alexander Fleming's discovery of penicillin from the fungus *Penicillium rubens* [27], studying a single species in pure culture has been the gold standard of microbiology, with most commercially successful antibiotics discovered predominantly from purified cultures of the actinobacteria phyla from soil. The isolation of streptomycin from actinomycete *Streptomyces griseus* was the first therapy for tuberculosis [28, 29], and subsequently, tetracycline, chloramphenicol, erythromycin, vancomycin, and rifamycin were all discovered from actinobacteria. However, since this era of rapid antibiotic discovery between the 1950s and 1970s, only two new classes of antibiotics have successfully been released in to the market, partly due to high rates of rediscovery, as exhaustive efforts examining the same cultured microbes result in repeatedly finding the same compounds that these few culturable microbes produce [30]. Culture-independent studies, such as microscopy and 16s rRNA sequencing, have revealed that over 99% of microbial diversity remain

reluctant to culturing, and this represents an unexplored microbial gene pool encoding novel compounds that are available as future antibiotics.

Leondardo da Vinci once said: “We know more about the movement of celestial bodies than about the soil underfoot” [31]. The soil remains a largely unexplored reservoir of microbial diversity, containing up to 10^5 unique species per gram [16, 32]. One of the indicators that cultured microorganisms represent a very small fraction of the total microbial diversity was the “great plate count anomaly” [16]: it was observed that less than 1% of the microbes seen under the microscope in the original sample could survive or grow on rich nutrient growth media [33-36]. Another indication of this unculturable diversity comes from sequencing the hypervariable region of the highly-conserved and universal 16S ribosomal RNA (rRNA) gene [37]. As the gene encoding it is highly conserved between different species of bacteria, 16S rRNA gene sequences provide species-specific signature sequences that is used for phylogenetic studies [32, 38, 39]. 16S ribosomal RNA gene sequences of soil bacteria have shown there exists an unprecedented abundance of bacterial diversity both within known phyla and those undefined, as many individuals still remain unaffiliated to any cultivated relatives or defined phylum [14].

Owing to its rich microbial species diversity and abundance, soil was selected as the source of eDNA for this study, and has been used as the source of eDNA for several published metagenomic libraries [40-42]. Additionally, most of the biomolecules recovered from soil-based studies were only very weakly related to known gene products or were entirely novel [26, 43-51], without many reported rediscovery of genes, indicating that the soil still remains a widely unexplored reservoir of natural products, and a promising source of novel antibiotics.

1.2 Functional metagenomics: A cultivation-independent approach

Functional metagenomics describes both a field of research and a set of tools that is designed to identify functional genes from the collective genomes present in an environmental sample. This is achieved by driving gene expression in a foreign host [20], and as such, it is an experimental platform designed to specifically circumvent the need to culture unknown microbes [52]. Functional metagenomics begins with extracting bulk DNA from samples containing mixed microbial populations. For functional screening, a gene-library is first generated in a workhorse host bacterium that grows well under laboratory conditions. Finally, screening is conducted for phenotypes representative of the desired gene product.

One method of functionally screening a metagenomic library for antibacterial or growth-inhibiting gene products involves searching for compounds that appear to impose significant inhibition on bacterial growth [53]. In this way, yet-uncharacterised genes have a chance of being discovered, as no *a priori* knowledge of function or sequence is required for this type of screening. Other screening methods can be used as well, such as homology-based sequence searches for genes similar to known inhibitory products in the library [54].

1.2.1 Soil as a sample for functional metagenomics

Soil is a widely shared starting point in the construction of many functional metagenomic libraries [44, 46, 49, 55-57], partly owing to the abundance and diversity of soil-dwelling microbes. The soil contains up to 10^5 unique species per gram [58], and is estimated to have a more extensive genetic diversity than most other environmental samples, such as sea water, resulting in a higher reported number of discoveries

attributed particularly to soil [59] in the discovery of useful and novel products [60-62], such as new antibiotics [45, 63]. Several laboratories have reported novel antimicrobial compounds. For example, turbomycin A and B were discovered in 2002 [45] – broad spectrum antibiotics effective against both gram-positive and –negative bacteria. Metagenomic inserts expressing antimicrobial activity were found to encode a cyclin-dependent kinase inhibitor [64], and in another study, antimicrobial activity was associated with production of an indirubin compound [46]. By screening for antimicrobial activity of a metagenomic library from an Arizona soil sample using *Bacillus subtilis* as a host, six antimicrobial compounds (two with cell wall-degrading activity, three proteases, and one lypolytic enzyme) were identified [65]. More recently, screening a staphylococcal-derived metagenomic library in *Staphylococcus aureus (S. aureus)* led to the discovery of Lysostaphin in 2014 [66], which has activity against the host bacterium *S. aureus*: a pathogen that is a major cause of human morbidity and mortality today, due to the high prevalence of drug resistant strains [67].

Functional metagenomics is a method that does not require *a priori* knowledge about genes or their hosts, potentiating the discovery of useful gene products from the 99% [68] of the microbial world that remains uncultured. Although in its infancy, this technology has potential to feed new, novel products into the antibiotics discovery pipeline.

1.3 Steps in generating a metagenomic library

1.3.1 Extracting DNA from soil

The isolation and purification of high quantity and quality DNA is a first step in the metagenomic workflow (Figure 1-1). Other important parameters include fragment size

and representiveness of the microbial community DNA [69]. Often, trying to optimize one of these parameters has an antagonistic effect on another. For example, processes that are used for removing impurities inhibiting cloning also shear the DNA [20]. In a study conducted with NZ soil, bead beating extraction procedures caused increased fragmentation of DNA with more vigorous shaking but yielded higher quantities of DNA due to more efficient cell lysis [55]. This is likely due to gentle lysis failing to break open encapsulated bacteria and spores, thus blocking access to their DNA [70]. Highly sheared DNA (low average fragment size) obstructs larger operons and impedes downstream screening, while a low yield does not allow library construction [71]. PCR using phi29 DNA polymerase and random hexamer primers has been shown to amplify DNA from low-yield extracts without biases [72], but this method is still under debate as to if amplified DNA can be truly representative of all the taxa present in any sample [20, 58, 73].

Soil samples also contain organic inhibitors, such as humic acid, from the natural decay of plant and animal materials [74] that co-purify with DNA and interfere with enzymes used in cloning, such as DNA ligase [55, 75]. Removing humic acid, as well as other organic pollutants, metal ions, and chemical impurities is a major bottleneck in metagenomic studies and many extraction protocols have focused in this step especially, to yield PCR inhibitor-free metagenomic DNA. Reported methods include using polyvinylpolypyrrolidone [76], hexadecyltrimethylammonium bromide [77], and cesium chloride gradients [78].

Currently, commercial DNA isolation kits are routinely used by many laboratories working in metagenomics, as commercial kits generally take less time and produce reasonably high quality and quantity DNA from both cultured and uncultured microbial

species [79-81]. However, kits are not optimal, and a 'golden standard' for the process of isolating of metagenomic community DNA remains in need of standardization.

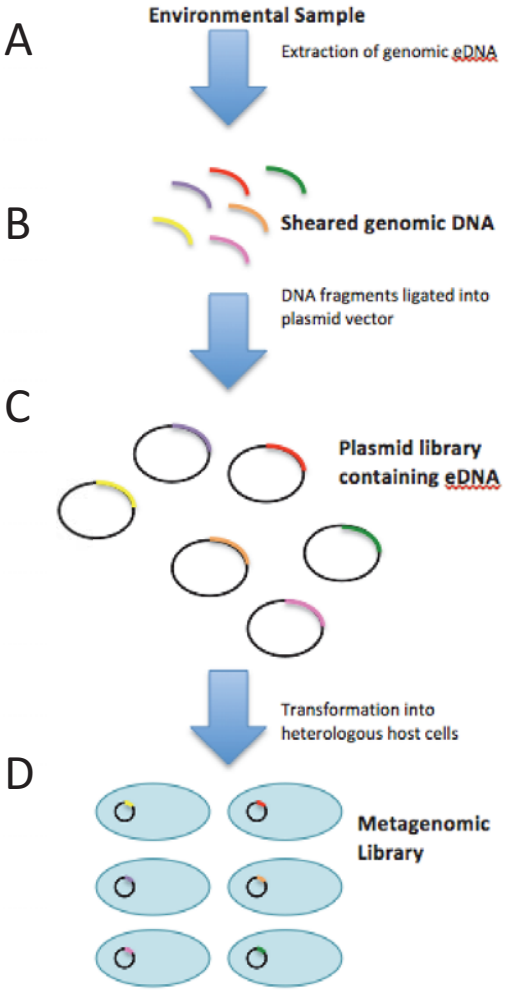

Figure 1-1. Steps in constructing a metagenomic library. (A) The first step is the extraction of environmental DNA (eDNA) directly from a microbial community of an environmental sample. **(B)** Extracted eDNA is then fragmented and **(C)** ligated into an expression vector, then **(D)** cloned into a surrogate host for the generation of a metagenomic library. The host bacterium is induced to express the recombinant DNA, and clones are screened based on desired phenotype.

The PowerSoil DNA Isolation Kit (cat#12888 MoBio Laboratories, Inc.) has been reported in previous studies to effectively extract the soil metagenome with sufficient purity and yields [82-85]. For this reason, this kit was chosen for this study (Materials and Methods). The basic protocol of DNA isolation and purification used this kit is lysis, removal of contaminants, and column purification of DNA. First, cell lysis occurs via mechanical and chemical means. The sample is combined with beads and vigorously mixed to physically break apart cells. An anionic detergent (sodium dodecyl sulfate) and other disruption agents are added and contribute to additional chemical cell lysis. After centrifugation, a reagent is added to the supernatant that precipitates impurities and inhibitors including humic acids. A silica membrane binds DNA. Several wash steps are done to further remove non-binding/non-DNA material before the DNA is eluted from the membrane filter. In this way, much of the soil debris as well as impurities and contaminants that co-purify with DNA are removed to supposedly yield sufficient amounts of high purity DNA in a time-efficient manner.

1.3.2 Fragmenting DNA

DNA is fragmented to narrow the length distribution of the extracted genomic DNA, and prepare them as inserts for subsequent ligation into an expression vector (Figure 1-1A,B). The sizes of fragments (inserts) dictate which types of vectors they can be inserted into (Figure 1-1C). We have used restriction enzymes for fragmentation. However, restriction endonucleases only cut DNA at specific recognition sites, and are thus biased. To increase the coverage of the metagenome and prevent under- or over-representation of certain taxa, fragment sites should be random and sequence-independent.

All current methods, along with restriction enzyme digestion, have limitations [86]: (i) hydrodynamic shearing [87], whereby DNA is sheared randomly by passing them through a narrow tube at high speeds, has clogging issues and is high cost; (ii) nebulization [88], whereby compressed air is used to force the DNA solution through a small hole, resulting in a mist of smaller fragments, has a large size distribution that is difficult to automate; and (iii) sonication [89], whereby physical vibration from sound energy stretches and compresses DNA causing it to rip apart, often results in low cloning efficiency due to DNA damage. An ideal method would fragment DNA without systematic bias, be inexpensive, and integrate minimal DNA loss in the process. As such a method remains to be elucidated, we have used two different restriction enzymes for directional cloning as complimentary sticky ends typically result in a higher ligation efficiency than blunt ends [90] that result from mechanical shearing. Higher ligation efficiencies may lead to a higher diversity and clone number in the final library, which will mean more individuals are available for functional screening.

1.3.3 Selection of insert size and choice of expression vector

The selection of insert size and the vector to be used are based on the desired target functions and uses of the metagenomic library. Small DNA fragments of 10 kilo-base pairs (10kb) or less can be cloned into standard plasmid cloning vectors, such as pUC derivatives or the well-characterized pBAD [91] inducible vector system, and can be used to screen for functions encoded by single or a few genes.

Currently, it is still cheaper and faster to clone small insert clones, rather than large inserts [92], and antibiotic resistance determinants and various novel enzymes have been discovered in metagenomic functional screens of inserts smaller than 10 kb [59]. However, smaller DNA fragments have a low chance at expressing biosynthetic

pathways that involve many genes working together to perform an interlinked function. This includes many secondary metabolite pathways that were found to be expressed in nature, but unable to be replicated as they remained silent under laboratory conditions [93]. Activities and compounds that are encoded by multiple genes are more likely to be captured if larger insert sizes are used. Fosmids have been used for metagenomic libraries composed from inserts 30-40 kb in size, and bacterial artificial chromosomes (BACs) have been used for inserts up to 200kb long [94]. Recently, vectors called pCC1FOS (Genbank accession EU140751; Epicentre Biotechnologies) and pWE15 have been used for cloning large inserts as fosmid and BAC clone inserts have shown problems of phylogenic interference, leading to misrepresentation of the function of the inserts. pCC1FOS vector copy numbers can be modulated by external addition of arabinose, which increases DNA yield. It also carries a chloramphenicol resistance marker, which is better than ampicillin resistance as it does not form satellite colonies, making plating easier for library construction [95]. Successful metagenomic libraries using pCC1FOS have aided discovery of antibiotic resistance determinants, antibiosis, as well as various pigments [96, 97]. Although larger inserts are more likely to capture entire biosynthetic pathways, it is more challenging to achieve larger fragment sizes, and must employ gentle lysis to avoid shearing the DNA. For research purposes, as well as commercial interest, vectors that better facilitate heterologous expression of recombinant pathways of bigger inserts are in development [94].

Additionally, to accurately measure the effect of the expression of a cloned gene, it is often desirable to use an inducible expression system that is tightly regulated. Tight regulation implies that the expression system should be able to achieve synthesis levels sufficient for detection when the inducer is present, and efficiently shut off in the absence of the inducer. The pBAD series of plasmid vectors was developed by Guzman

et al. [91] to be able to be efficiently turned on and off via the tightly regulated ARA promoter. For genes encoding expression products that are toxic to bacterial hosts, it is important that the expression is not leaky, because clones with inserts encoding strong toxins would not be able to be detected in the first place. For this reason, the pBAD expression system was used in this study (see Materials and Methods).

1.3.4 Selecting a heterologous host organism

As the intended end-point of a screening assay is to detect desired phenotypes, a functional metagenomic approach relies on a compatible host that can faithfully express a myriad of foreign genes and gene-clusters [18]. Often, the problem with functional screens are low levels of gene expression in the selected library host [98] or low sensitivity of detection methods [99]. Currently, many functional metagenomic studies have used *Escherichia coli (E. coli*) as a host strain, due to the ability of *E. coli* to express heterologous proteins [100] and the availability of *E. coli* strains that permit greater uptake and maintenance of foreign DNA, such as those with deletions in genes for homologous recombination and restriction systems (*recA* and *mcrA* respectively). Gabor *et al.* [99] showed that *E. coli* can support heterologous expression of around 40% of genes within a subset of 32 taxonomically diverse genomes. *E. coli* hosts are not capable of expressing all the genes of a metagenome, and it has been shown that different host cells express (or lack expression) the same metagenomic library differently [101]. Different hosts have different expression capabilities for the same gene. To maximize the value of a metagenomic library, as well as optimising vectors for the construction of the library, a host that can efficiently express as much of the diversity present in the original DNA extract must be employed (Figure 1-1D). The importance of using a diverse range of different expression hosts has been studied [42,

59], and efforts are being made to express existing metagenomic libraries in an expansive range of different non-*E. coli* hosts, resulting in some reported successes [96, 97, 101]. One metagenomic *E. coli* library from soil DNA that identified novel antibacterial small molecules were taken and expressed in *Streptomyces lividans* and *Pseudomonas putida* as additional non-*E. coli* hosts. Here, different novel activities were detected in different host organisms for the same plasmid library [101]. Under- and over-representation of clones are partially due to the selection of the host. By screening in alternative host systems as well as *E. coli*, it is possible to utilise alternative transcriptional machinery, regulation, and metabolic networks to broaden the scope of gene expression and reduce host-related limitations and bias.

1.4 Screening metagenomic libraries

Metagenomic libraries contain massive amounts of genetic information so screening needs to be sensitive and rapid with adequate detection of the desired activity amongst millions of clones [69]. Several types of screening methods have been developed to accommodate massive libraries with millions of clones. Intracellular screening makes use of a reporter plasmid to detect metabolites within the host cell (metabolite-regulated expression), enabling the detection and isolation of a particular clone with a certain function [102]. For genes that encode compounds sought after by pharmaceutical markets, such as antimicrobial molecules, screening can be as conceptually 'simple' as isolating gene fragments that inhibit growth or have significantly lethal effects on target strains when expressed [60]. As well as expressing in a range of heterologous hosts, efforts have also been made to improve *E. coli* itself as a screening host. Examples include using T7 RNA polymerase to drive transcription [103], as well as introducing sigma factors to guide RNA polymerase to otherwise untranscribed regions [104].

Deleting global negative regulators and/or overexpressing positive regulators have been used to decrease repression of transcription of otherwise-silent gene clusters [105, 106]. These are thought to improve the chances of discovering novel genes, by increasing levels of transcription and translation in attempt to induce expression of a higher proportion of the metagenomic library. A high throughput platform for screening is crucial, as many compounds in a given library are not functional and many successful candidates do not make it through the discovery pipeline [107-109]. There is high demand for improving screening techniques to enable faster and better hit rates. Typically, screens are performed for function [12, 17], then clones with desired phenotypes are isolated, and the candidate gene fragment is characterized by sequencing [110]. This process can be time consuming and labour intensive [95].

To identify novel genes from screens with very low hit rates, metagenomic libraries often consist of millions of clones [110]. Next Generation sequencing (NGS) has allowed a much faster and cheaper means to screen metagenomic libraries which are often comprised of massive amounts of data. Many metagenomics studies have involved sequencing hundreds of gigabytes of DNA, and such scales would not have been possible without NGS platforms such as the 454 pyrosequencer and Illumina [54, 111] that multiply sequence runs from the order of 100 kb per run (Sanger sequencing) by several million-folds by paralleling the sequence reading [112]. NGS is often used in metagenomic studies that aim to identify the members of an environment and elucidate the taxonomic diversity of yet-uncultured microbes [44, 68, 113, 114]. Our approach is different, because here, we propose a way in which NGS can be incorporated to screen clones for their functional phenotypes in a collective pooling way to save time and lessen the intensity of labour.

Identifying putative new antibiotics in a metagenomic library involves identifying clones carrying genes toxic to bacteria using phenotypic screening. This is a rather daunting task if each clone is assayed individually. However, there is strong evidence that many microbial genomes harbour genes that are toxic to *E. coli*. Kimelman and colleagues [53] computationally identified over 15,000 genes that from 393 different microbial genomes that repeatedly failed to be cloned into *E. coli* during Sanger sequencing, indicating that these genes may exhibit some growth inhibition or toxicity when cloned in *E. coli*. Early genome sequencing used a shotgun sanger sequencing approach where multiple copies of the genome being sequenced are fragmented and inserted into plasmids, which are then transformed into *E. coli* and sequenced. After the genome is assembled, it became apparent that there were 'unclonable' regions of many genomes. These cloning gaps were proposed by the Sorek group [115] to contain genes that are toxic to the host and represent novel functions inhibiting bacterial growth. Further studies by Kimelman et al identified an additional 52,330 genes that had significantly reduced coverage from Sanger sequencing. 44 genes were cloned under an inducible promoter and were found experimentally to be highly toxic to *E. coli* host cells when expressed [53]. These studies suggest that there may be many more genes with the potential to inhibit bacterial growth based on the small subset of genomes they studied.

Based on previous work by the Sorek group [53] described above, we hypothesize that a simple method of identifying genes from soil-based microbes encoding inhibitory products could be screening using a negative selection approach. In such a setup, all the clones in the metagenomic library are collectively grown in flasks with and without the inducer present. All clones are expected to grow to similar densities at similar rates, unless the DNA insert being expressed confers toxicity. After growth, the inserts from

the uninduced and induced library are sequenced. Here, multiple copies of each inserted eDNA fragment will be present in the sequence pool. Comparing the two sequence pools will indicate any fragments of DNA that inhibit the growth of their host bacterium as sequences from these fragments will be significantly underrepresented or absent from the sequence pool after expression is induced. Illumina sequencing of the pooled libraries is a realistic means of isolating small molecules that impede bacterial growth, eliminating the need for individual functional assays for each and every clone.

The probability of identifying a functional clone of interest (a positive hit) is relatively low, as the number of genes that are neutral or even beneficial for bacterial growth outnumber those that are toxic [116, 117]. Coupled with the difficulties in expressing cloned genes in heterologous hosts, hit rates for toxic genes have been reported to be lower than 0.01% [59], meaning that a library of at $1x10^6$ unique clones would yield 100 positive "hits" or toxic gene products. Phenotype detection using screening assays that typically have very low hit rates involves screening large numbers of clones, and advances in sequencing technologies enables us to increase the analytical throughput so millions of clones may be screened efficiently.

1.5 Limitations and Challenges

Functional metagenomics is a promising new pipeline for drug discovery, with many researchers making great effort to improve the power of function-based screens that harness biosynthetic diversity into drug discovery pipelines. Improving the methods involved with library construction could result in larger library diversities, which could lead to the successful isolation of novel products from uncultured organisms.

In summary, four key limitations must be addressed to better utilise a functional metagenomic approach for antibiotic discovery. Firstly, improving DNA extraction methods can expand the scope of environmental sampling and capture a greater diversity of novel bioactive compounds. Secondly, enhanced vectors that better facilitate cloning and fine tuning of expression of heterologous pathways (for example, gene clusters that function together) of large inserts must be developed. Thirdly, different host species have different expression capabilities, even for the same gene. To maximize the expressed proportion of any metagenomic library, a range of alternative hosts, other than *E. coli*, should be used for expression to utilise alternative transcription machinery, as well as alternative regulation and metabolic networks in the background to reduce host-related bias so that a better coverage of the metagenome can be captured. Lastly, development of more sensitive and rapid screening techniques that also ensure expression of genes that are normally repressed or silenced so that a greater proportion of the metagenomic library may be expressed.

New classes of antibiotics must be produced to cope with the inevitable ongoing emergence of bacterial resistance [118-120]. The beginning of the pipeline to novel antibiotics begins with developing a robust toolkit to tap into nature's biosynthetic reservoir encoded within the genomes of microbial communities. The development of an efficient screening method would speed up the rate of discovery of novel genes encoding new novel antibiotics, which can then be used to study and feed into the discovery pipeline.

Here we present an experimental framework with potential for high throughput interrogation of soil metagenomes for genes encoding compounds that significantly inhibit the growth of bacteria. We have worked on optimising each step in the

construction of our soil metagenomics library (Figure 1-2) and subjected some clones to phenotypic screening and sanger sequencing. Difficulties arose in creating a large and diverse enough library for this approach. An ideal library would need to contain 10^6 unique clones to theoretically yield 100 "hits" [53]. Our results yielded far fewer unique clones than was theoretically required. Further optimisation of the procedures involved in library construction needs to be done, from extraction to transformation. Once this hurdle is overcome, screening could be improved to be more sensitive and less biased. If successful, such an approach may provide a greater insight into the vast reservoir of natural toxic products that are available as future antibiotics.

1.7 Aims and Objectives of this study

1. Proof-of-principle: testing the vector-host system with a toxin-encoding gene
2. Extract genomic DNA from the soil (Figure 1-2A)
3. Create a plasmid library (Figure 1-2B)
4. Transform plasmid library into a metagenomic library (Figure 1-2C)
5. Arabinose assay: low throughput screening of a subset of clones of the library to identify any genes encoding expression products inhibiting bacterial growth
6. Sanger sequencing to check inserts that appear toxic to *E. coli*
7. Generate a library of at least 10^6 clones
8. Functional screening via NGS (Figure 1-2D)
9. Identify possible antibiotics-encoding genes by comparing the inserts present in grown uninduced and induced libraries (Figure 1-2E)

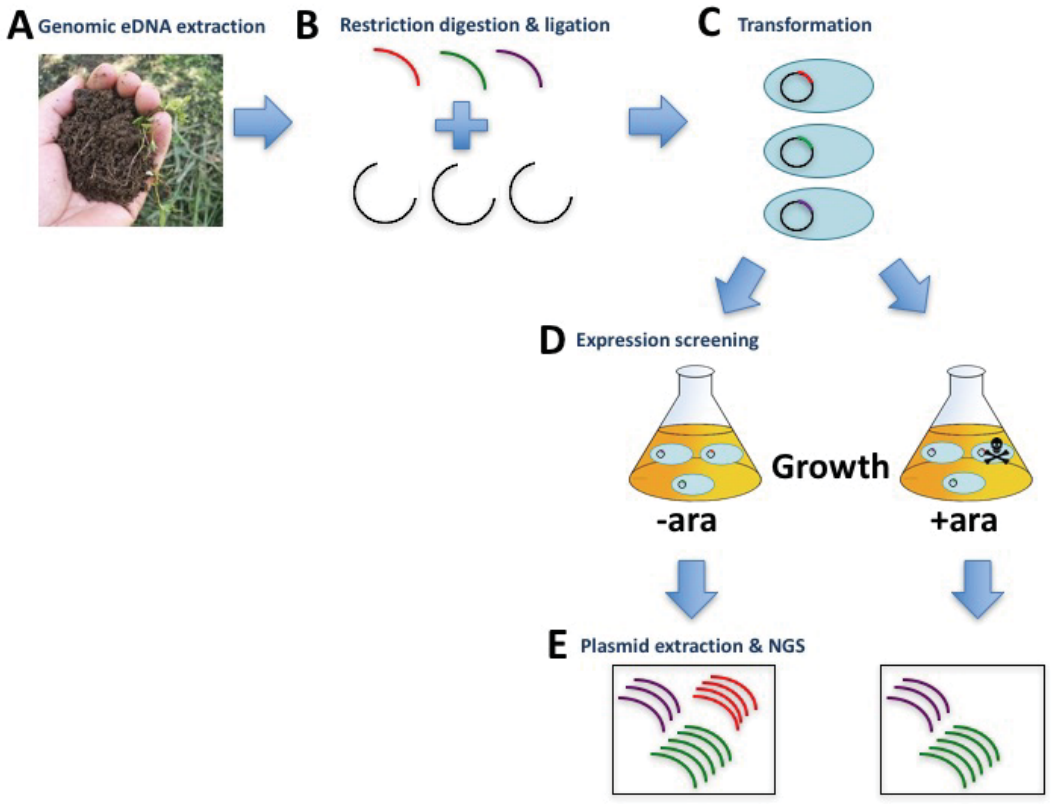

Figure 1-2. Functional metagenomics for identifying toxic genes. A) Environmental DNA (eDNA) is directly extracted from a soil sample. **B)** The extracted eDNA and pBAD/Myc-His B plasmid (pBAD) vector are both doubly digested using two restriction enzymes that have unique recognition sites in the multiple cloning site of the pBAD vector, and the resulting fragmented eDNA is then ligated into this vector to form the plasmid library. **C)** The plasmids carrying eDNA inserts are transformed into Escherichia coli host bacterium to form the metagenomics library. **D)** Since the inserted eDNA fragment is under the araBAD promoter, genes are repressed in the absence of the arabinose inducer (-ara), and expressed in the presence of arabinose (+ara). The entire metagenomics library is grown in the absence of arabinose, and a replicate is grown in the presence of arabinose. Plasmids are extracted from both cultures and sequenced using NGS, such as Illumina, and sequences are compared. **E)** eDNA fragments encoding genes that significantly inhibit the growth of their bacterial host will be absent or significantly underrepresented in the sequence upon expression.

2 Materials and Methods

2.1 Proof-of-principle: testing the vector-host system with a toxin-encoding gene

2.1.1 Using PSTPO as a toxic gene to verify the pBAD system

To ensure the vector (pBAD) and the host (*E. coli)* system was working as expected, a proof-of-principle experiment was conducted. A gene that was reported to be toxic to *E. coli* was used and cloned into the pBAD vector. According to Kimelman et al. [53], (Supplementary table S2), the gene *Psyr* in *Pseudomonas syringae pv. syringae* B728a (NCBI taxon ID: 205918, Replication accession: NC_007005, Locus tag: Psyr_4019) encodes a gene product (a putative transcriptional regulator), which did not perturb host growth when expression was repressed, yet was toxic to *E. coli* host cells when gene expression was induced. An almost-identical orthologous gene (100% identical at the protein level, 96% similarity at the DNA sequence level) from B728a also exists in strain *Pseudomonas syringae pv. syringae* strain DC3000. The gene *PSPTO* (NC_004578) was cloned by using PCR amplification using DNA template from *Pseudomonas syringae* DC3000. *PSPTO* was used instead of *Psyr*, as *pseudomonas DC3000* was immediately available. This gene is was shown [53] to be clonable when expression is repressed and did not affect bacterial growth, but becomes toxic to the host cell upon induction of gene expression.

PCR primers were designed to flank the *PSPTO* gene, with the *Xho*I restriction enzyme recognition sequence linked to the 5' end of the forward primer, and *Kpn*I restriction site linked to the 5' end of the reverse primer (Figure 2-1). This PCR amplified product (insert) is ligated to the pBAD vector between *Xho*I *and Kpn*I, which are unique

restriction sites within the multiple cloning site (MCS) of the pBAD vector, and absent in the insert (checked via Addgene sequence analyser).

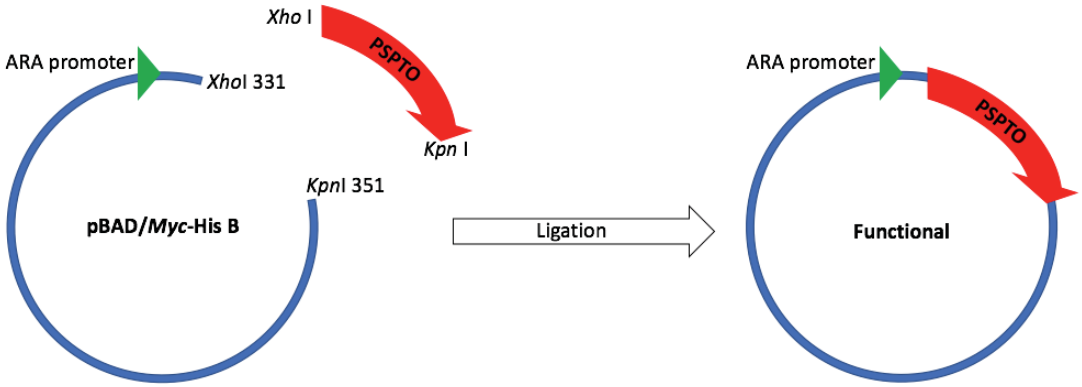

Figure 2-1. Inserting PSPTO in the correct orientation translates to a functional product. *Attaching a XhoI restriction site to the 5' inserted in the correct orientation, with expression regulated by the ARA promoter.*

Primers were also ordered (*Kpn*I site at 5' end of forward primer, and *Xho*I restriction site linked to 5'end of reverse primer) to use as a "backwards" control. This "backwards" control is the same size as the "forward" insert, and contains the same nucleotide sequence, yet will be transcribed in the wrong orientation, and thus yield a nonsense expression product (Figure 2-2).

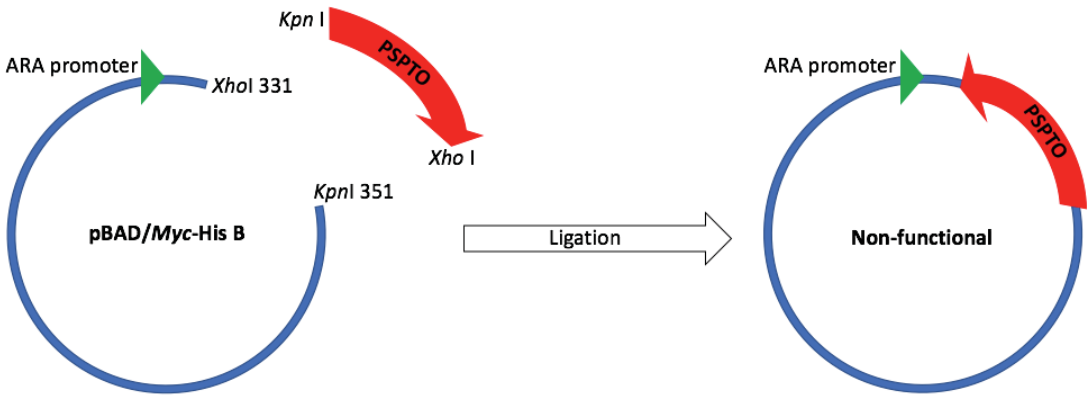

Figure 2-2. Inserting PSPTO in the reverse orientation translates to a non-functional product. By attaching a KpnI restriction site to the 5' end of PSPTO and a XhoI site to the 3' end, the gene is transcribed in the wrong orientation, yielding a nonsense expression product.

PCR was carried out using Taq polymerase (Invitrogen, cat# 10342020) and PCR amplification was done following the Invitrogen Taq DNA Polymerase recombinant kit protocol, with three reactions: the correct "forward" primers, the nonsense "backwards" primers, and a no template (no *DC3000* DNA) control. PCR products (*PSPTO* with restriction site linkers ~380bp) were checked for their size using 1.8% agarose gel electrophoresis, extracted from the gel and purified using the Qiagen QIAquick gel extraction kit (cat# 28704) following the manufacturer's recommendations, except for the final elution step where autoclaved MilliQ water was used instead of the elution buffer.

To generate *PSPTO* inserts with 100% sticky ends, we initially cloned the PCR products (forwards- and backwards- PSPTO) into a TOPO cloning vector using the pCR8/GW/TOPO TA cloning kit (Invitrogen, cat# K250020), as per the manufacturers recommendations. This initial cloning step was done to ensure that when the insert is cut with restriction enzymes, the PSTPO gene would have '100%' sticky ends, as opposed to a digestion of insert directly after PCR, which often yields less than 100% sticky ends.

Chemically competent *E. coli* cells were transformed and transformants grown on LB agar with spectinomycin (spec, 100 μg/mL) as the TOPO vector has a spectinomycin marker. A single colony was used to start larger cultures and plasmids were isolated using the E.Z.N.A. Plasmid Mini Kit I (cat# D6942-00), and run on a 1.8% agarose gel (120 V, 120 A, 1 hour) to validate the TOPO vector contained the *PSPTO* insert, by size (TOPO vector 2817 base pairs + *PSPTO* insert 381 base pairs = 3198 base pairs).

Both the pBAD vector and the TOPO plasmids containing the PSPTO inserts were double digested with *Kpn*I-HF and *Xho*I restriction enzymes from New England Biolabs

at 37°C for 3 hours. pBAD with *Kpn*I-HF and *Xho*I sticky ends, and PSPTO inserts (both forward and reverse) with the same compatible *Kpn*I-HF and *Xho*I sticky ends were identified via size selection on an agarose gel (1% w/v agarose, run at 120 V and 120 A for 90 minutes). These were cut from the gel and purified via a gel extraction kit (Qiagen gel purification kit).

A molar ratio of 1:3 between pBAD (vector) and *PSPTO* insert was used (50ng and 11ng respectively) for ligation, as per manufacturer's protocol (New England Biolabs T4 DNA Ligase cat# M0202). The sticky pBAD vector and sticky PSPTO insert was ligated using T4 DNA ligase. Resulting plasmids were transformed into chemically competent TOP10 *E. coli* cells. Transformants were then plated on LB-ampicillin (100 μg/mL) agar (15% w/v) plates and single colony cultures were prepared.

Positive transformants were induced using a 7-fold range of arabinose concentrations, where expression of the PSPTO gene product of the was expected to result in inhibited growth or death of *E. coli* host cells. These were compared to both the control clones containing the insert in the reverse orientation and clones grown in the absence of arabinose where growth was expected. Growth was measured using optical densities of small culture volumes at regular time intervals for 18 hours in a plate reader at 600 nm.

Growth assays with a 7-fold range of arabinose concentrations were set up using a 96-well plate as follows. Column 1 was set up with 180 μL of LB growth media and 20 μL of sterile water (0% arabinose); column 2 was 180 μL of LB and 20 μL of 20% stock arabinose; columns 3 to 8 were 20 μL of increasing 10-fold dilutions of the 20% stock arabinose with 180 μL LB, i.e. 2%, $2x10^{-1}$%, $2x10^{-2}$%, $2x10^{-3}$%, $2x10^{-4}$%, $2x10^{-5}$%, $2x10^{-6}$% from column 3 to 8 respectively. Rows A to C were overnight cultures of *E. coli* TOP10 transformed with pBAD with the functional *PSPTO* gene inserted in the

correct orientation (see Figure 2-2); row D was a no bacteria negative control to ensure the plate was not contaminated with other bacteria; row E was the *E. coli* TOP10 cells only without any transformation; rows F to H were *E. coli* TOP10 transformed with pBAD with the non-functional *PSPTO* gene inserted in the incorrect, reverse orientation (restriction sites in the backwards direction) (see Figure 2-3).

2.1.2 Using sacB as a toxic gene to test the pBAD system

As *PSPTO* was published as a 'putative transcriptional inhibitor' [53] (supplementary table 2), and no growth inhibition was found when expressed, an alternative, well characterised toxic gene was sought. *SacB* is a well-characterised gene that confers toxicity to *E. coli* [121] and was chosen for further proof of principle experiments.. We designed primers flanking *sacB* in the plasmid vector pkmob18sacB (Genbank accession number FJ437239.1), amplified this by PCR, then extracted *sacB* using agarose gel electrophoresis.

The same procedure was used as with the *PSPTO* gene as describe above, with the following modifications. *SacB* was PCR amplified and inserted into the multiple cloning site of the pBAD vector. Transformants were induced to express the inserted gene via the addition of arabinose. The vector-host system worked as expected, with tight regulation, so we moved onto eDNA.

2.2 Extracting genomic DNA from the soil

DNA was extracted from soil samples taken from three Massey University Albany campus sites (Figure 3-4), using the MoBio PowerSoil kit (cat# 12888), per the manufacturer's protocol, with a change in the final step, where the DNA was eluted in

autoclaved MilliQ water instead of the elution buffer. The extracted genomic DNA from each 0.25 g sample of soil were eluted in 100 μL of autoclaved MilliQ water.

2.3 Preparing the vector for cloning

pBAD/*Myc*-His B plasmid is an arabinose-induced expression vector, with tight regulation and little transcriptional leakage reported [91]. To obtain sufficient copies of the plasmid vector for generating the metagenomic plasmid library, pBAD (4.1 kb) was electroporated into competent *E. coli* TOP10 (Invitrogen) . A positive transformant was transferred to sterile LB and grown overnight at 37°C. Plasmid purifications were done using Qiagen midi and maxi plasmid extraction kits.

2.4 Preparing electrocompetent *E. coli* TOP10

To make electrocompetent cells, OneShot TOP10 chemically competent *E. coli* (Invitrogen cat# C404006) were streaked out on a LB agar plate and incubated at 37°C overnight. A single colony was transferred to 5 mL of LB liquid media and incubated overnight at 37°C. 2 mL of this starter culture was added to 200 mL LB in a 2 L baffled flask, shaking at 250 rpm at 37°C. When the optical density at 600 nm reached between 0.6 – 0.7, cells were quickly chilled, then washed with cold 10% glycerol multiple times by centrifuging and decanting. Treated cells were then finally suspended in 1 mL of 10% glycerol and frozen for storage at -80°C in 70μL aliquots, and used for all electroporations.

2.5 Construction of a mini library

Extracted and purified eDNA was doubly digested with restriction enzymes *Nco*I and *Xba*I and *incubated* at 37°C for 3 hours, then heat inactivated for 20 minutes at 80°C.

Digested eDNA was run on an agarose gel and size selection for fragments between 1000-2000 base pairs was achieved by cutting DNA out of the gel at this size. DNA was then purified and ligated into pBAD. The pBAD vector was prepared by digesting *sacB* out of the pBAD plasmid used for the proof-of-principle screening described above. Once the eDNA was ligated into pBAD, the plasmid library was then transformed into *E. coli* cells. Individual colonies were picked to create a metagenomics library containing 1504 clones. Single clones were arrayed into 96-well microtitre plates and screened for inhibitory effects on growth of the host upon induction and expression of the eDNA insert. The bacterial growth was monitored every 5 minutes for 12 hours by measuring optical density at 600 nm.

2.6 TSS transformation

2x TSS (transformation and storage solution) was made by dissolving 0.8 g bacto-tryptone, 0.5 g yeast extract, 0.5 g NaCl, and 20 g PEG8000 in a total volume of 50 mL with the addition of MilliQ water, autoclaving, then adding 10 mL of sterile 1M $MgSO_4$, 10 mL DMSO, adjusting the pH to 6.5, and making it up to a total volume of 100 mL with autoclaved MilliQ water. To transform *E. coli* with plasmids, a single colony of the target strain (LMG194) was transferred to 5 mL LB and grown at 37°C until the culture was slightly turbid (OD600 between 0.1 and 0.3), then chilled on ice for 10 min. An equal volume of ice-cold 2xTSS was added to the culture and vortexed. After incubating on ice for 10 minutes, 1 mL of this competent-cell mixture was added to 100 ng of the plasmid and incubated on ice for 30 minutes before plating on selective agar plates. Positive pBAD transformants grow on ampicillin plates.

2.7 Expression Screening

Using a range of different concentrations of the arabinose inducer ($2x10^{-9}$% - 10%), effects of expression of inserted genes on growth rate of *E. coli* was investigated by measuring optical densities at 600 nm over 18 hours of growth at 37°C in a plate reader.

2.8 Sanger Sequencing

Universal pBAD primers were used to PCR amplify eDNA fragments inserted into the pBAD vector, using Taq DNA Polymerase (Invitrogen), as per the manufacturer's recommendations. PCR products were then purified using the E.Z.N.A Cycle Pure Kit (D6493-02). Sanger sequencing was conducted by Macrogen Inc. (Seoul) and the resulting sequences were analysed using Geneious R9.1.8.

3 Results and Discussion

3.1 Proof-of-principle experiment with *PSPTO* failed to show cell death upon induction

PSPTO was inserted into the PCR8/GW/TOPO plasmid vector in both the correct and reverse orientations, as described in Materials and Methods. Plasmids were purified from overnight cultures of successful transformants containing *PSPTO* in TOPO, and subjected to restriction enzyme double digests at sites flanking the vector (Figure 3-5) in both forwards (Figure 2-1) and reverse (Figure 2-2) orientations.

Gel purified inserts were then ligated into the pBAD expression vector and subjected to expression screening (Figure 3-1). The expected, idealised results would show that by inducing expression of the toxic *PSPTO* gene, growth of *E. coli* in the presence of increasing concentrations of arabinose would be increasingly inhibited. This would indicate that our vector-host system is working the way it should, allowing downstream screening of toxic genes from soil extracted eDNA. However, we did not observe this result. Strikingly, *PSPTO* did not affect *E. coli* TOP10 growth upon expression. Additionally, our rates of transformation of the correctly-orientated *PSPTO* gene were 100-fold lower than that of the same sized backwards insert. Since *PSPTO* is a toxic gene, we thought at this stage, that expression in TOP10 may be leaky.

Leaky expression implies that transcriptional control of the operon; for pBAD, the araBAD operon; is not efficient and there is always some basal level of transcription. The pBAD vector was specifically selected for its tightly controlled operon, and the level of expression in the absence of the arabinose inducer is expected to be undetectable [91]. The decreased rates of transformation indicate that once *PSPTO* is

transformed into the host, the host then dies and transformants are not recovered. The few clones recovered that did not reduce growth upon induction could have had mutations which somehow decreased the activity of the gene product, or mis-formed vectors. We also contemplated the idea that *E. coli* TOP10 may not be a suitable strain for expression. Although *E. coli* TOP10 is routinely used for high-efficiency cloning and plasmid propagation, re-cloning is often done into other *E. coli* strains that have higher expression levels and additional regulation machineries which make them more suitable for functional evaluation than the TOP10 strain [122-124].

LMG194 is recommended by the developers (Invitrogen) to be used instead of TOP10 for the expression of proteins that are toxic to *E. coli.* LMG194 is adapted to ensure additional repression of the *ara*BAD (P_{BAD}) promoter with the addition of glucose [91] so that uninduced levels can be used as a reference to detect growth-inhibiting effects of inducing any toxic gene. Additionally, both TOP10 and LMG194 are both *E. coli* K-12 strains that are capable of transporting, but unable to metabolise L-arabinose, ensuring levels of L-arabinose inside and outside the cell are constant. This makes both strains suitable for pBAD expression, as neither will break down the L-arabinose inducer (see Table 6-1 in the Appendix for genotypes), thus concentrations will not decrease with time. The pBAD vector containing the *PSPTO* gene insertion was purified from TOP10 and transformed into LMG194 using TSS. However, expression of *PSPTO* did not affect the growth of the LMG194 strain either.

It was not clear at this point whether *PSPTO* was not toxic (functionally different from ortholog *Psyr*), or whether the vector had been incorrectly constructed. Thus, a better characterised *E. coli* toxic gene, *SacB*, was adopted. It was only later that we realised that *PSPTO* had been inserted into the pBAD vector one nucleotide out of frame,

resulting in a frame shift, and thus, the translation of a nonsense product with multiple stop codons. Better care was needed to ensure the sequence of the construct is in frame with the C-terminal (pBAD/*Myc*-His) peptide to ensure the correct protein is translated.

New primers were designed to ensure the reading frames were not disrupted for *SacB*. Instead of *Xho*I, we used the *Nco*I restriction enzyme recognition site. Using the *Nco*I site instead of *Xho*I adds only one extra amino acid after the initiator methionine (one amino acid at the very start – N-terminus), so that the translated protein has minimal deviation from the natural product of *SacB:* levansucrase [121]. As was done with *PSPTO*, restriction sites attached to ends of primers create restriction sites flanking the insert *SacB* gene, which can then be ligated to corresponding sticky ends of the pBAD vector. Since pairs of restriction enzymes used; *Xho*I/*Xba*I for *PSPTO* and *Nco*I/*Xba*I for *SacB,* have different recognition sites, both genes are inserted in the correct and reverse directions, as dictated by the position of the restriction sites (as in Figure 2-1 and 2-2).

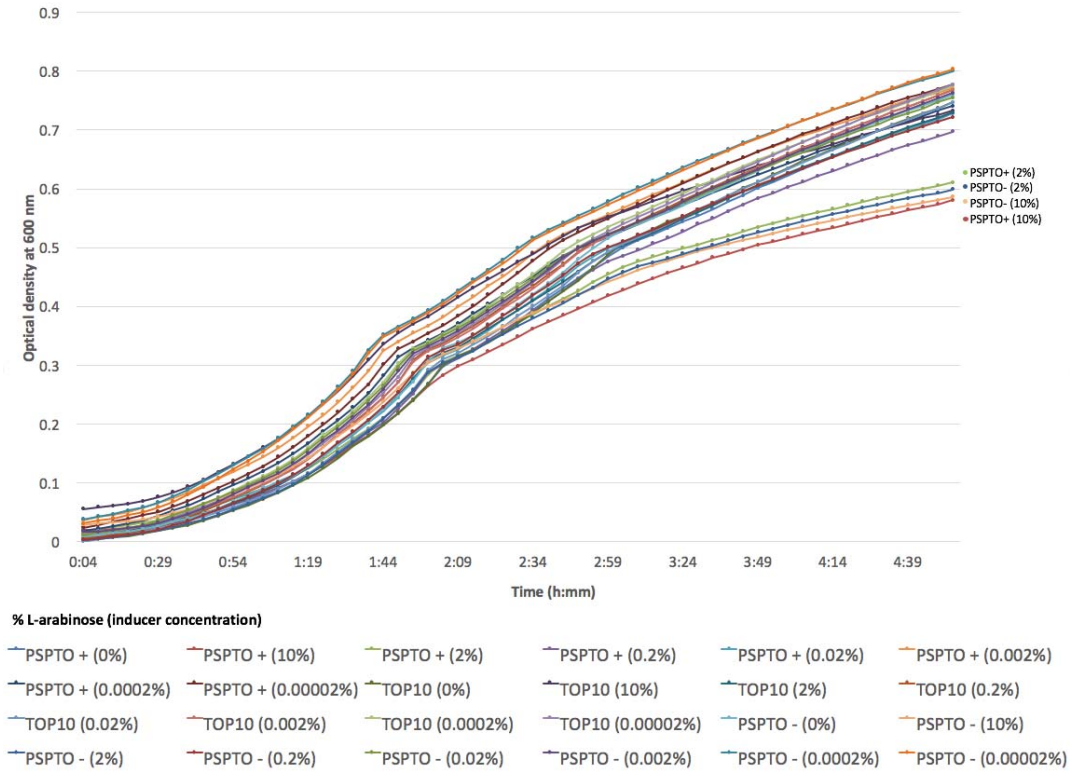

Figure 3-1. PSPTO inserted out-of-frame has no effect on the growth of E. coli TOP10 at various inducer concentrations. 1000-fold dilutions of overnight LB cultures of Escherichia coli TOP10 transformed with pBAD containing the functional PSPTO toxic gene inserted in the correct orientation (PSPTO+), the non-functional PSPTO inserted in the reverse orientation (PSPTO-), and E. coli TOP10 cells with no insert (TOP10) were grown in increasing concentrations of the inducer, L-arabinose (0% to 10%). Optical densities at 600 nm were measured every 5 minutes for 5 hours with constant shaking at 37°C in a plate reader.

3.2 *SacB* is toxic to *E. coli* and expression in pBAD is regulated in a dosage-dependent manner.

SacB [121] was PCR amplified and ligated into pBAD via the TOPO vector, as was done for *PSPTO*. First, we checked that the expression product of the *SacB* gene, levansucrase, confers toxicity to *E. coli* in the presence of sucrose. It had been reported previously in literature, that salt-free LB should be used for higher levels of expression of *SacB* [121]. Indeed, in the absence of salt, and in the presence of sucrose, the

presence of *SacB* inhibited the growth of the host bacterium *E. coli* DH5α (Figure 3-2, Table 3-1).

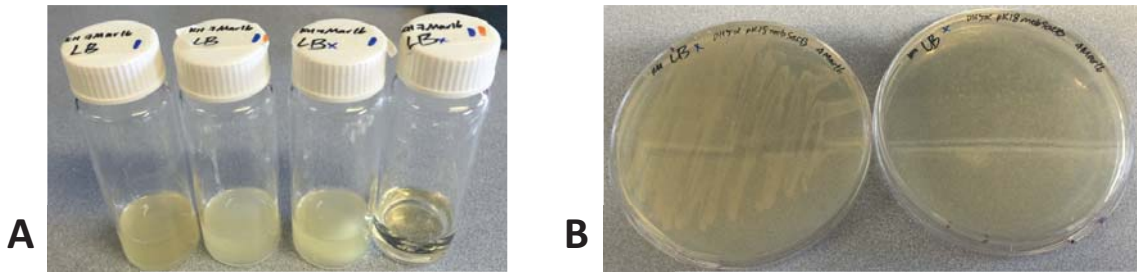

Figure 3-2. Expression of SacB inhibits growth of E. coli in the presence of sucrose. A) 5 mL overnight cultures of E. coli DH5α in (from the left): LB + kanamycin (50 μg/mL), LB + kanamycin (50 μg/mL) + sucrose (10%), salt-free LB + kanamycin (50 μg/mL), salt-free LB + kanamycin (50 μg/mL) + sucrose (10%). **B)** Single colony streaks of E. coli DH5α containing the pK18mobsacB plasmid. Static overnight growth at 37°C.

Media for expression of *SacB* in *E. coli*	OD600 of overnight cultures
LB + Km	2.8207
LB + Km + sucrose	0.3611
Salt-free LB + Km	0.5005
Salt-free LB + Km + sucrose	**0.0000**

Table 3-1. Salt-free LB best supports expression of *SacB* in *E. coli* TOP10. As in Fig. 8A, bacterial growth was measured by optical densities at 600 nm following shaking incubation at 37°C for 18 hours. Overnight culture readings were blanked in the medium used. Km: Kanamycin, 50 μg/mL; sucrose: 10% (w/v).

Following confirmation that *SacB* is toxic when expressed, the gene was amplified via PCR from the pK18mobsacB plasmid in *E. coli* DH5α with *Nco*I and *Xba*I restriction sites flanking the ORF, cloned into PCR8/GW/TOPO, doubly digested with *Nco*I and *Xba*I, ligated into *Nco*I/*Xba*I digested pBAD, and electroporated into *E. coli* TOP10 cells. Expression was induced over a 10-fold range of L-arabinose concentrations, and

effect on growth of *E. coli* TOP10 was measured by measuring the optical density at 600 nm over 18 hours in a plate reader (Figure 3-3). In the absence of the inducer (0% arabinose), growth rate is at maximum, indicating that the $araP_{BAD}$ promoter is repressed without the inducer present. Without arabinose, unintended expression is not detectable, and cells proliferate uninterrupted. Leaky expression would prevent any successful transformants carrying potentially toxic genes from being recovered prior to screening. Only arabinose concentrations above 0.002% had expression levels above our detection threshold, and inhibitory effects on *E. coli* growth increased accordingly with increasing levels of the inducer (Figure 3-3). The tight regulation of the $araP_{BAD}$ promoter, that directs expression of inserted eDNA fragments, makes pBAD a useful expression vector for screening potentially toxic genes, as effects of turning on these genes can be measured in comparison to when expression is switched off in clones of the same organism.

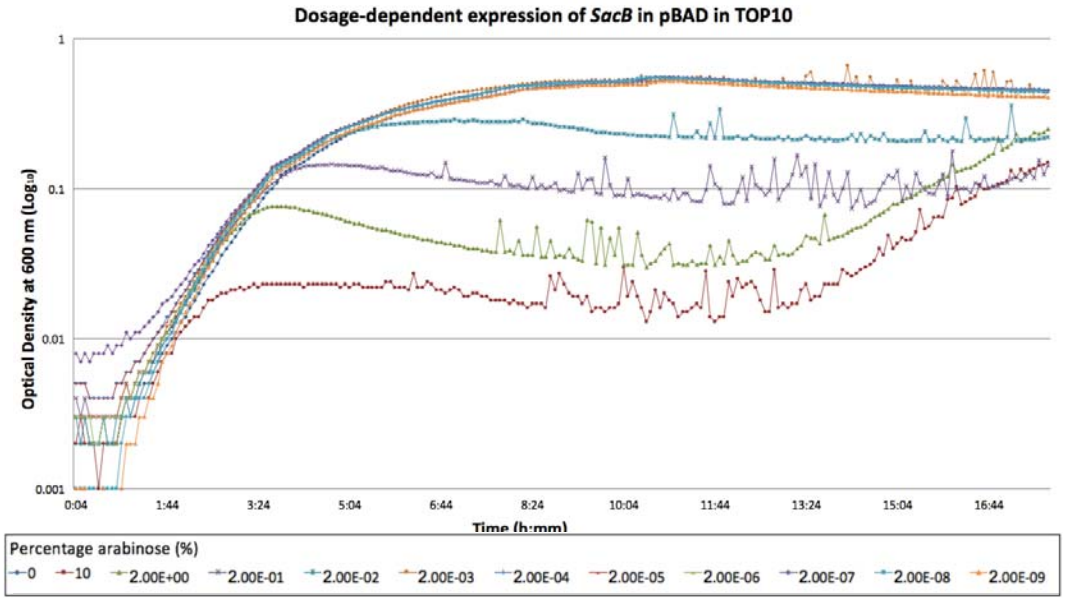

Figure 3-3. pBAD is turned on by arabinose in a dose-dependent manner. At 0% arabinose (no inducer), the growth rate of *E. coli* TOP10 is at maximum. With increasing concentrations of arabinose, the growth rate (optical density/time) decreases, as increasing inducer concentration increases expression of *SacB*, which confers toxicity to *E. coli* in the presence of sucrose.

3.3 Difficulties in extracting high yields of unfragmented DNA from soil

The metagenome was extracted from three different soil samples collected on Massey University Albany campus using the MoBio PowerSoil Kit, as per the manufacturer's recommendations. We reasoned that bacterial abundance and community composition might vary between the geographically separated and physically different types of soil. Accordingly, to capture more bacterial diversity, three types of soils were sampled: muddy soil from a river-bank, slimy soil found adjacent to a pond, and relatively dry soil from outside our lab building.

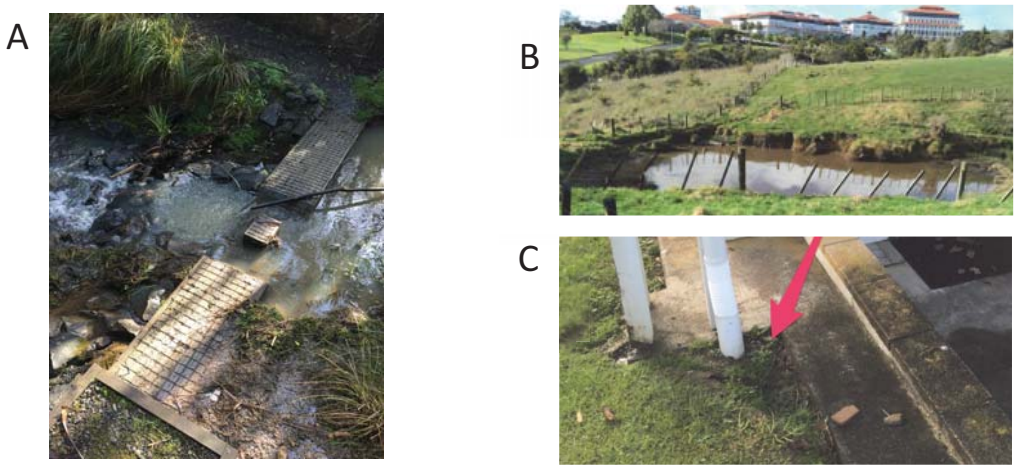

Figure 3-4. Soil samples collected from Massey University Auckland, NZ. A) Wet soil by the bed of the flowing Massey river. Collection area below a dense canopy of trees; damp, moist, and shaded. **B)** Wet soil from a natural reservoir in Fernhill Escarpment. **C)** Dry soil collected from the corner outside our lab building.

Soil type	260/280	260/230	Concentration (ng/µL)
River-bed, moist soil	1.91	2.01	118.5
Pond-side, wet soil	1.90	2.03	124.8
Building-side, dry soil	1.99	2.04	121.6

Table 3-2. Metagenomic DNA extractions from the PowerSoil Kit are similar in quantity and quality for the three different soil samples. The ratio of absorbance at 260 nm and 280 nm indicates purity of the DNA. 260/280 ratios of above 1.8 and 260/230 ratios between 2.0-2.2 are accepted as pure DNA.

As indicated by the 260/280 and 260/230 ratios of the DNA extractions, the PowerSoil Kit yields sufficiently pure DNA (Table 3-2). The 260/280 is the ratio of absorbance at 260 nm and 280 nm, and is used to analyse the purity of nucleic acids. Since DNA absorbs at 260 nm, a low 260/280 ratio indicates the presence of contaminants that absorb light strongly at 280 nm, such as proteins [79]. The 260/230 ratio is a secondary measure of purity, and accounts for contaminants that absorb light strongly at 230 nm, such as EDTA that is often used in DNA isolation [125, 126]. Together, a DNA sample with a 260/280 ratio above 1.8 and a 260/230 ratio between 2.0 – 2.2 is accepted as pure DNA [127]. Although the eDNA recovered from the PowerSoil Kit for all three samples were within these ranges, running them on an agarose gel indicated that DNA was somewhat degraded (Figure 3-5, lane 8). High molecular weight DNA of high concentration is required as short sheared fragments of DNA will mask the correctly restriction-digested ‘sticky end’ fragments. And also, it is of critical importance to note that a large proportion of the DNA is lost during gel extraction of the cut DNA to select a size-range that is suitable for the chosen vector.

Less vigorous vortexing (mechanical lysis) could reduce shearing of DNA. However, this could also lead to a larger proportion of incompletely lysed cells, resulting in lower DNA yields. In hindsight, optimisaiton of the ratio between beating and lysis could

have been determined by testing a soil sample at various vortex speeds and vortexing times and running the extracted DNA on a gel to identify the level of mechanical lysis, in the first step of the extraction procedure, that results in less degradation without compromising DNA yield.

Following cell lysis, DNA becomes exposed to the solution before binding to the column, and at this point, minimising pipetting could also reduce unintentional shearing [128]. Accordingly, more intact high molecular weight DNA may be purified. Wide-bore tips can be used instead of regular narrow-bore tips, and gently swirling or gently tapping tubes instead of pipetting could also mitigate the shearing stress that pipetting causes to DNA in solution.

Both modifications aim to reduce degradation of DNA for repeating extraction procedures involving kits, and should be adopted in our future attempts at library construction. As discussed, high-molecular weight DNA is an advantage for capturing genes or gene clusters flanked by the intended restriction sites.

3.4 Low ligation and transformation efficiencies hinder efforts in generating a large metagenomic library.

To generate a plasmid library, the plasmid vector and eDNA are cut with the same two restriction enzymes, as compatible sticky-ends typically have a higher ligation efficiency in comparison to blunt-end ligation [90]. The *SacB* gene used for the proof-of-principle is 1.4 kilo base pairs (kb), and the ligation reaction with T4 DNA ligase (NEB), as described in the methods, had been optimised for this insert size. Thus,

doubly-digested eDNA between 1-2 kb was selected for, and gel purified, and ligated into doubly-digested pBAD vector (Figure 3-5).

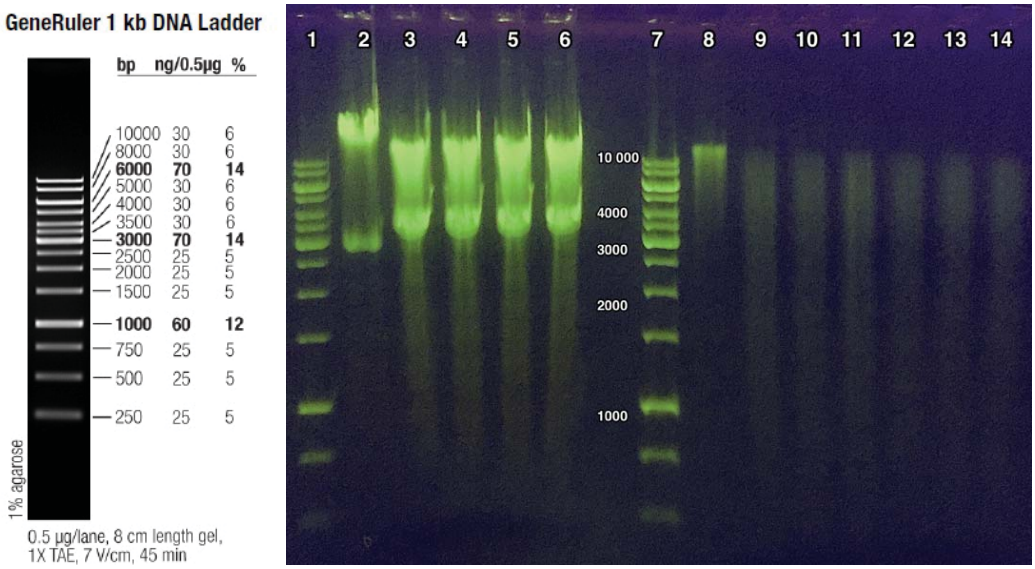

Figure 3-5. Extracting fragments of environmental DNA and cloning vectors with compatible ends. 1.2% agarose gel electrophoresis run at 120 V and 120 A for 90 minutes. Lanes from left: 1) GeneRuler 1 kb DNA ladder, 2) undigested SacB in pBAD vector, 3-6) XbaI/NcoI doubly digested pBAD, 7) GeneRuler 1 kb DNA ladder, 8) undigested eDNA, 9-14) XbaI/NcoI doubly digested eDNA.

In hindsight, using *SacB* in pBAD as a start-point to be able to differentiate uncut and cut vectors was a poor choice as *SacB* is a 1.4 kb fragment, and pBAD is 4.1 kb, so it was presumed that it would be relatively easy to isolate the doubly-digested pBAD vector with the *SacB* cut out. However, this was not the case. Although only the 4.1 kb fragment was gel extracted, undigested vectors were still present in which *SacB* remained intact and functional, resulting in five clones that appeared to carry toxic eDNA inserts subsequently proved to be *SacB* upon sequencing. At this stage, cutting the pBAD vector by itself and dephosphorylating the vector would have been a better alternative that would have yielded a higher number of doubly-digested vectors that would have resulted in a higher ligation efficiency with eDNA. Also, all digested

fragments up to 10 kb should have been extracted, rather than only fragments between 1-2 kb, as there is no logical justification to discarding the doubly digested eDNA fragments outside of this range.

As agarose gel electrophoresis is used to select fragments by size, 100% DNA recovery would be ideal from a gel extraction kit. However, Nanodrop results indicated that less than 20% of the DNA loaded in the gels were being recovered from the Qiagen Quick gel extraction kit on any occasion. Determined to optimise recovery of DNA from gel extractions, a number of modifications were made to the procedure. Firstly, the gel was made by using a minimal amount of agarose (0.6%) and trimmed as much as possible with minimal exposure to UV light. This was to remove excess agarose and limit damage to DNA that impacts clonability, respectively. The MilliQ water used to elute the DNA was heated to 70°C prior to applying it to the column, as higher yields have been credited to this in the past [129]. And since the melting step of gel extraction combines chaotropic salts with heat and denatures DNA, DNA was 're-natured' by warming the eluted DNA to 95°C then slowly cooling it back down to room temperature. None of these modifications made a significant impact on the percentage of DNA able to be recovered from a gel via the Qiagen Quick gel extraction kit. It has also been reported in literature that adding guanosine to agarose gels increased downstream ligation efficiencies 2-3 fold [130]. A 'sunblock' for DNA – this would be interesting to test, in continuation of this project.

Three different kits were tested to optimise the ligation step: NEB T4 DNA Ligase (M0202); Agilent DNA Ligation Kit (Cat# 203003); and Bioline Quick-Stick Ligase (BIO-27028). Different vector to insert ratios were tested, along with a range of various incubation times. Electroporation was also optimised, mainly by making changes to the

preparation of competent *E. coli* TOP10 cells and testing different voltage settings. However, the maximum transformation efficiency for the ligated library we achieved was $4.75x10^3$ transformants per μg of DNA (cfu/μg). Typically, transformation efficiency of ligation products are 10-1000 fold lower than that of supercoiled plasmids, such as pUC19. The transformation efficiency of the pUC19 plasmid, using the same batch of electrocompetent cells, was $1x10^9$ cfu/μg, indicating that my cells were of acceptable quality (as commercially purchased electrocompetent TOP10 are stated to have an expected transformation efficiency of $1x10^9$ cfu/μg) but the transformation efficiency of the ligated library was far too low, as it should have been at least 3 orders of magnitude higher (at least $1x10^6$ cfu/μg, according to Invitrogen cat#C4040). This observation that the same *E. coli* TOP10 competent cells yield a very low number of colonies when the metagenomic library was transformed, in comparison to the pUC19 positive control, is likely due to low DNA concentrations in the ligation. Despite our best efforts to optimise both ligation and transformation steps, the transformation efficiency is far too low to generate a library of the 1 million clones we need in order to identify 100 possible hits.

3.5 Expression screening to identify any clones carrying toxic eDNA fragments

Any clone carrying a gene that confers toxicity to its *E. coli* host is expected to display growth inhibition (low final optical density at 600 nm or a slow doubling-time). A low throughput screening was done on a subset of clones of the library, as described in materials and methods. Each clone of the library of 1504 transformants were transferred to two wells of a 96-well microtitre plate: one with (+ara) and one without (-ara) the inducer. Bacterial growth was monitored for 12 hours by measuring OD_{600}, and the

maximum OD_{600} of each clone was graphed against its maximum doubling time during exponential growth phase (Figure 3-6).

This low throughput, time-consuming, and labour-intensive picking and assaying of single clones, one by one, is eliminated in our proposed screening method where the entire library is grown collectively, as sequencing pooled libraries can be analysed to isolate genes encoding toxic products. However, in conducting this individual assaying of clones, problems in our cloning techniques were identified, and we have explored a few ways in which these issues could be addressed.

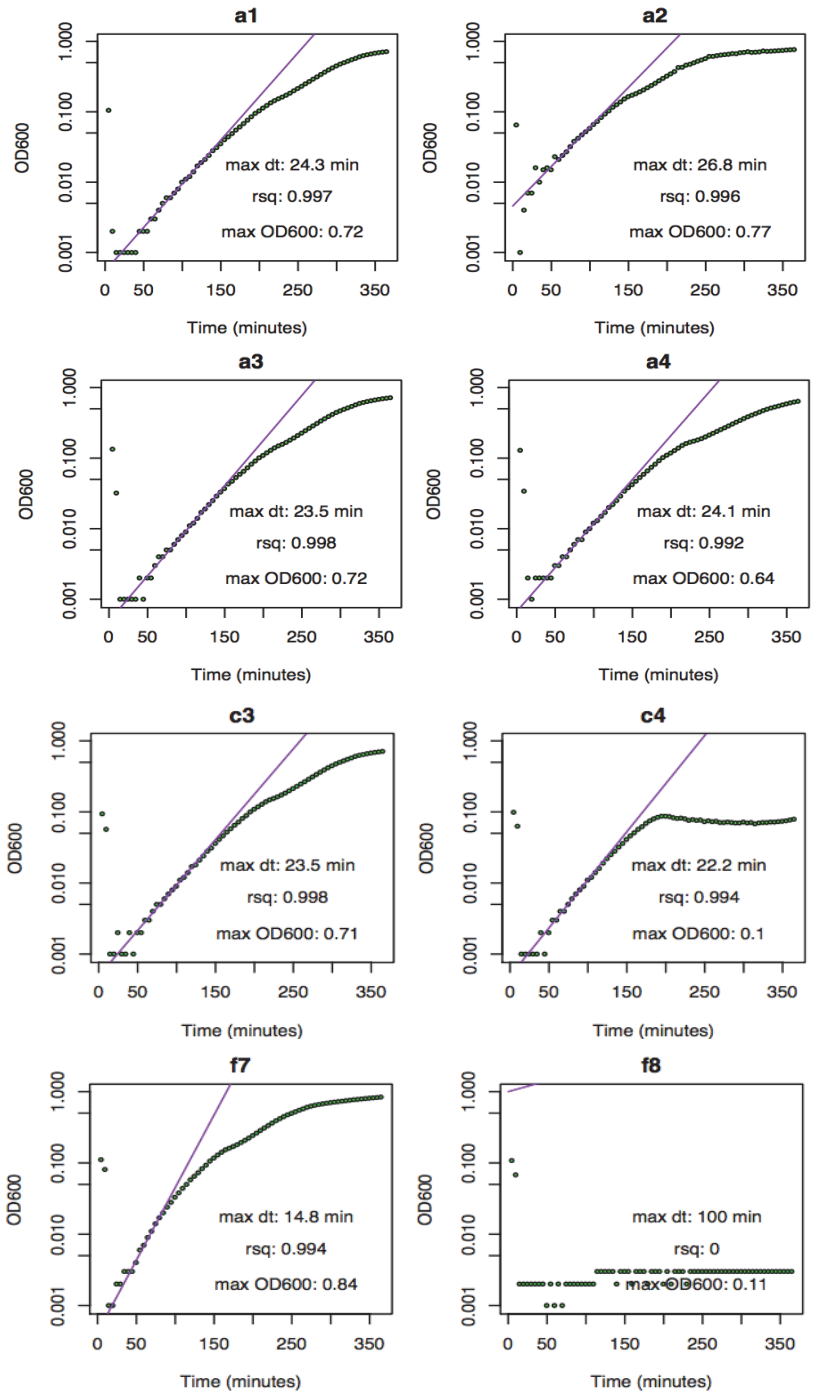

Figure 3-6. Growth assays with and without the inducer present. a1, a2, a3, a4: Clones carrying inserts that do not affect bacterial growth display similar doubling times and grow to similar maximum optical densities at 600nm with (a2, a4) or without (a1, a3) the inducer present. **c3, c4, f7, f8:** Clones carrying possible toxic inserts have a longer doubling time or grow to lower optical densities at 600nm when arabinose is present (c4, f8).

The eDNA inserts of the five out of 1504 clones (Figure 3-7) displaying inhibited growth were PCR amplified using universal pBAD primers and sent to Macrogen (Seoul) for Sanger sequencing.

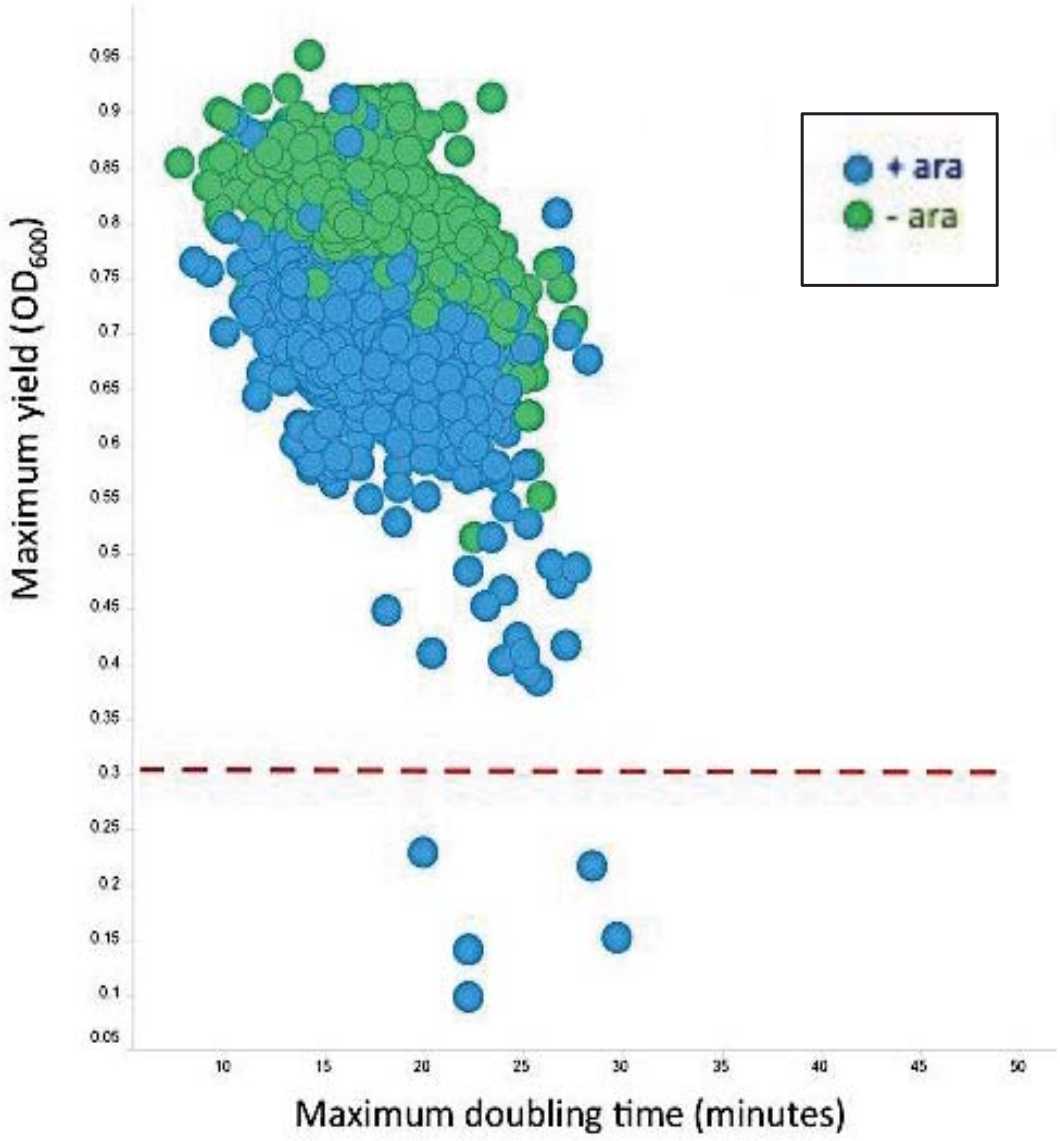

Figure 3-7. Five clones inhibited bacterial growth upon expression of the eDNA insert. N=3008 (a duplication of each of the 1504 clones of the metagenomic library in the presence and absence of arabinose). Each individual clone is represented as a circle. Bacterial growth is analysed in terms of the maximum double time (minutes) and maximum optical density at 600 nm (OD_{600}) based on turbidity of the cell suspension after 12 hours of growth in the absence (-ara, green) and presence (+ara 2%, blue) of *ara*BAD inducer arabinose (ara). Points below the red line indicate clones with low growth rates.

3.6 Analysis of sequences reveal cloning errors

Overall, 5 clones were identified to affect doubling time and final optical density to which their bacterial host grew to (Figure 3-7). Unfortunately, all of these potential "hits" contained the *SacB* stuffer (Figure 3-8), rather than different fragments of eDNA we had anticipated. In hindsight, using *SacB* in pBAD as the starter culture for the pBAD plasmid preparation was not ideal. Instead, pBAD should have been restriction digested then dephosphorylated, rather than trying to 'more easily' separate the doubly digested vector from undigested vector from a gel (Figure 3-5).

Figure 3-8. All five inserts inhibiting growth were SacB. Nucleotide alignment of the forward and reverse sequences for the PCR amplified inserts (T1C3, T2F1 T2F11, T3D3, T4C5) via Geneious R9.1.8. Blasting the consensus sequence resulted in 99% similarity with the *sacB* region of the pK18mobsacB cloning vector.

As for A1 wells from trial 6 (T6A1) to trial 10 (T10A1), we had expected to see random fragments of eDNA inserted into pBAD, from clones that did not display an inhibiting

effect on bacterial growth. The gene *aadA* encodes streptomycin 3"-adenylyltransferase, and mediates bacterial resistance to spectinomycin. Since the PCR8/GW/TOPO vector has the spectinomycin marker, the sequences were mapped onto the TOPO vector sequence as the reference genome (Figure 3-9). Sequences match from bases 950-2079. These regions encode the spectinomycin promoter (bases 930-1063) and resistance gene (1064-2074) (Figure 3-10).

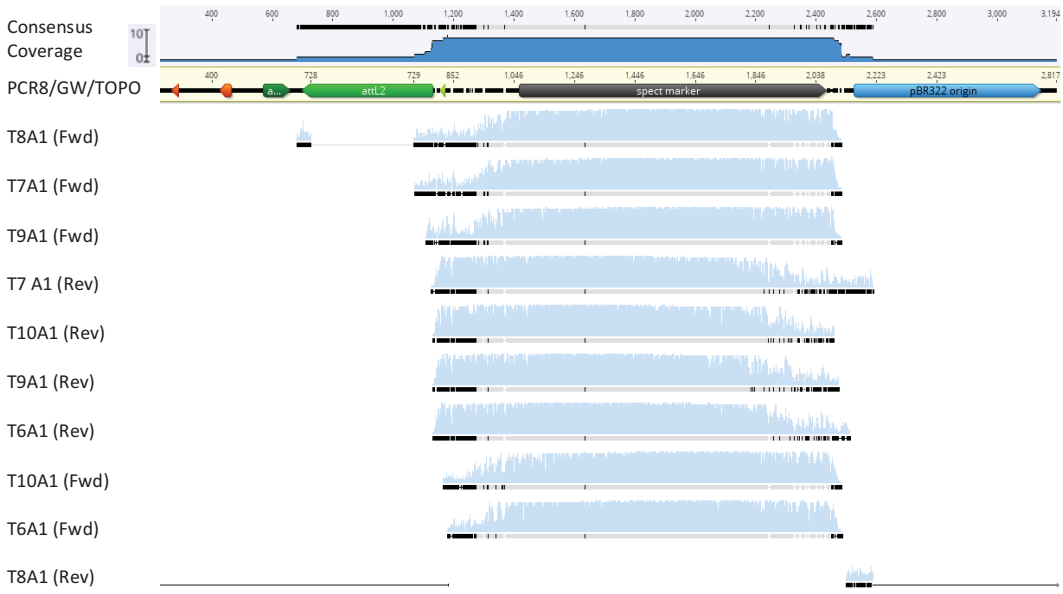

Figure 3-9. T6A1 to T10A1 control inserts were all the same fragment of the TOPO cloning vector. Nucleotide alignment of the forward and reverse sequences for the PCR amplified inserts (T6A1, T7A1, T8A1, T9A1, T10A1) via Geneious R9.1.8. B. Blasting the consensus sequence resulted in 99% similarity with the *aadA* (spectinomycin marker) region of the PCR8/GW/TOPO vector.

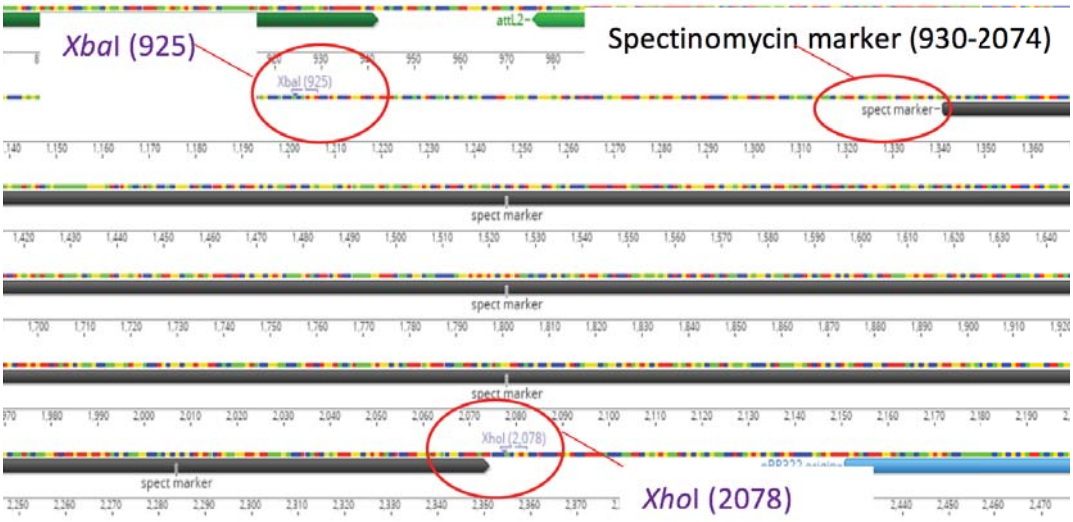

Figure 3-10. *Xba*I and *Xho*I restriction sites flank the spectinomycin marker in the TOPO vector. The spectinomycin marker lies between bases 930-2074 in the PCR8/GW/TOPO (Invitrogen) vector. The *Xba*I restriction site lies just before the spectinomycin promoter region, at 925 bases, and the *Xho*I restriction site lies downstream of the spectinomycin resistance gene, at 2078 bases.

The spectinomycin marker should not have been cut out from the TOPO vector. What should have been cut from the TOPO vector is the *SacB* gene flanked by *Xho*I and *Xba*I restriction sites within the multiple cloning site. There was suspicion that all the clones that had been isolated so far did not contain an eDNA insert. A random 96-well plate from the arabinose assay of eDNA in pBAD in TOP10 was selected for PCR (Figure 3-11). Here, a range of different-sized eDNA inserts are expected to have been ligated into the vector. However, 79/96 of the clones had no inserts as PCR products were the same size as the fragment in the original pBAD vector amplified with the universal pBAD primers, which is 293 base-pairs in length, as also shown in lane 72 with the vector-only control. This indicates that the vector had not been efficiently cut prior to eDNA ligation, or some contamination of uncut vector was present. As for the minority of clones that do contain an insert, we suspect that many are likely not eDNA. From

sanger sequencing of 10 samples, we can see that these inserts are primarily either *sacB* (1.4 kb) or the spectinomycin marker (1.15 kb).

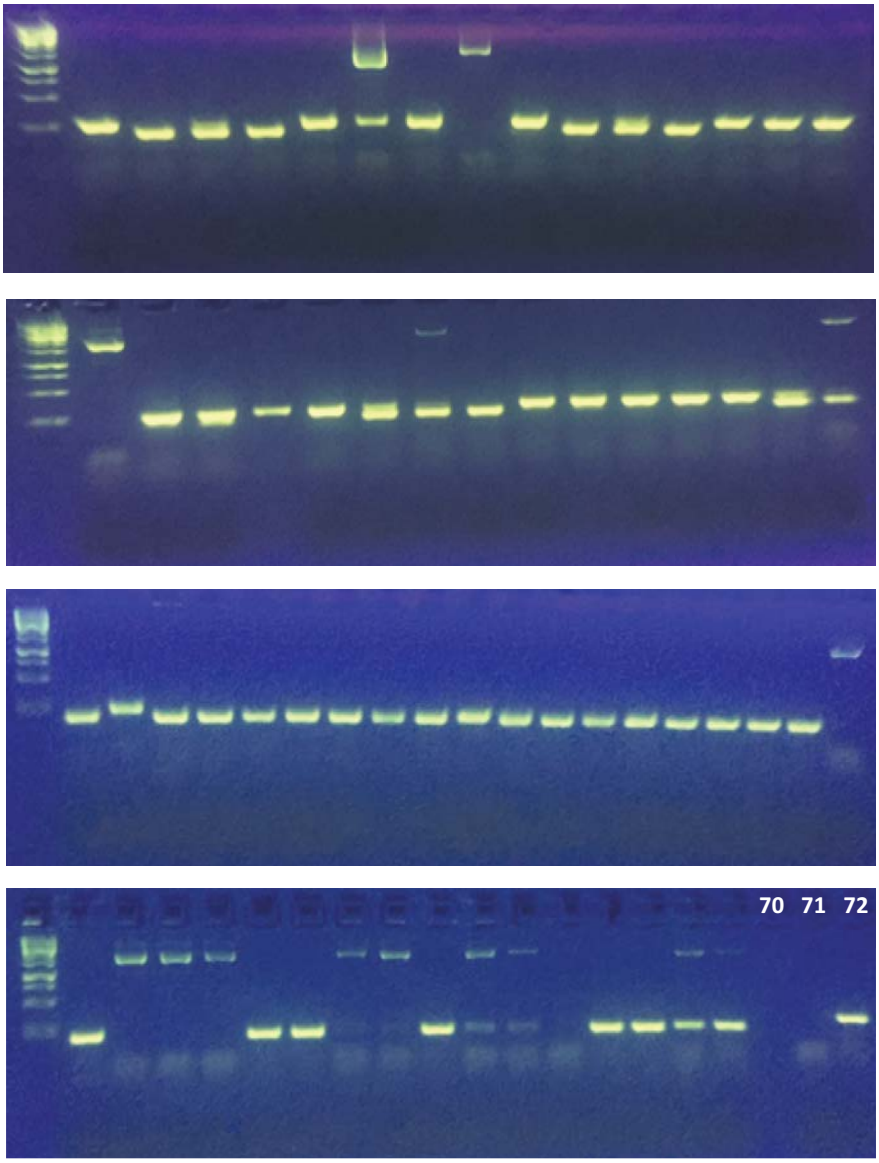

Figure 3-11. Most clones contain an empty vector with no insert. PCR products from trial 2 clones run on a 1.5% agarose gel at 120 V and 120 A for 90 minutes. Lane 70: no DNA template (negative control), 71: *E. coli* TOP10 only, 72: pBAD in *E. coli* TOP10. Bands at 290 base-pairs indicate clones with no insert.

In our workflow, we concurrently had growth-inhibiting clones analysed by Sanger sequencing, while generating more clones. We had not anticipated such a high level of

mis-construction of the plasmid library, and were unaware that most of the retrieved clones had no inserts or contained the *SacB* stuffer in place of the eDNA fragment. This contamination of undesired DNA in our 'metagenomic' plasmid library is likely exacerbated by a low recovery of correctly cut vectors and genomic DNA in the early stages of library construction, especially the large fraction that was lost during gel extraction.

4 Conclusion

Elucidating the functional complexity and diversity of the microbial majority that remains reluctant to culturing can help us to harness novel natural products into discovery pipelines, including the development of new antibiotics. Functional metagenomics can provide new insight into unexplored biosynthetic diversity by directly screening functional capabilities of genes encoded in the genomes of mixed microbial communities.

We have presented a functional metagenomics approach that employs high-throughput screening to reduce much of the time-consuming labour involved in individually screening each clone of a metagenomics library, which can be used to identify genes or gene-clusters that encode products that are toxic to bacteria. Several complicating factors confounded our efforts to generate the metagenomic library consisting of the 10^6 unique clones that we required to achieve an estimated 100 hits of such toxic genes.

4.1 Improvements in DNA extraction methods to be pursued

A metagenomic library can only be as good as the environmental genomic DNA that is recovered. Extraction should be effective in recovering intact DNA from all community members without bias, and would ideally be efficient in removing all contaminants that inhibit cloning. Obtaining large amounts of high quality eDNA after gel extraction proved difficult, and it is likely that the eDNA that we had recovered was not sufficient for ligation, resulting in a library of clones lacking inserts. There are many extraction methods that are available, and the major difference between these is in the balance of compensating between removal of soil contaminants, incomplete cell lysis, DNA degradation, and recovery yield [131]. Thus, the choice of protocol remains a

compromise, or a combination of cultivation techniques could be utilised until a 'gold standard' of metagenomic DNA extraction becomes standardised.

4.2 Optimising ligation and transformation efficiencies

The ligation and transformation efficiencies we achieved were at least 10^3-fold lower than the accepted norm for *E. coli* [132]. Despite making modifications: several steps in the gel-extraction protocol to recover higher yields of cut vector and inserts for ligation; optimising electroporation voltage; making and testing multiple batches of electrocompetent cells with modifications in preparation; and testing various ligation kits, ligation and transformation steps still remain far too low to generate the library of 1 million clones that we would need for an estimated 100 hits. The low efficiencies we saw were likely due to the insufficient DNA recovered from the PowerSoil Kit and subsequent gel extraction after fragmentation.

4.3 A prospective large-scale library

From here, further optimisation and testing of metagenomic DNA extraction protocols and ligation strategies can be employed to produce a large enough library for screening. Alternatively, outsourcing the libarary construction may aid in overcoming this hurdle [133].

After a plasmid library is constructed, two approaches that can be taken from here to increase the rate of identification of genes encoding possible antibiotics: **(i)** employ NGS to enable more high-throughput screening. In such a setup, all the clones in the metagenomic library are collectively grown in flasks with or without the inducer present. All clones are expected to grow to similar densities at similar rates, unless the DNA insert being expressed confers toxicity. After growth, the inserts from the uninduced

and induced library are sequenced. Here, multiple copies of each inserted eDNA fragment will be present in the sequence pool. Comparing the two sequence pools will indicate any fragments of DNA that inhibit the growth of their host bacterium as these fragments will be significantly underrepresented or absent from the sequences after expression is induced with the addition of arabinose (Figure 1-2D and E). Illumina sequencing of pooled libraries will be a valuable means of isolating small molecules that impeded bacterial growth, eliminating the need for individual functional assays for each clone, as done as proof-of-principle. Application of this approach to much bigger libraries that may consist of millions of clones would also contribute to increased rates of discovery output. **(ii)** Using multiple heterologous hosts is another approach we are considering. Currently, functional metagenomic studies have mostly been performed in *E. coli*. However, *E. coli* cannot functionally express all of the biosynthetic diversity present in the metagenome covered by any library. It has been shown that different hosts express the same metagenomic library differently [101] as different hosts have different expression capabilities for the same gene. We have tried to conduct screening in two different strains of *E. coli* (TOP10 and LMG194), and with a correctly constructed plasmid library, different expression phenotypes could have possibly been detected, as this phenomenon has been reported in previous studies [91]. Thus, to maximise the value of the metagenomic library, the same plasmid library will be transformed into multiple host organisms, with appropriate switching of the origins of replication in the plasmid, to express as much of the diversity present in the original DNA extract. The importance of using a diverse range of different expression hosts has been noted and studied [59, 134]. Under- and over-representation of clones are partially due to the seleciton of the host. By screening in alternative host systems, we can utilise alternative transcriptional machinery, regulation, and metabolic networks in the

background to broaden the scope of gene expression and reduce host-related limitations and bias. These two parameters will drive our future efforts to successfully construct a metagenomic library to bioprospect the soil for small molecules that are toxic to bacteria.

Bibliography

1. Margulis L, Sagan D (1997) Microcosmos: four billion years of evolution from our microbial ancestors. Univ of California Press

2. Summit M, Baross JA (1998) Thermophilic subseafloor microorganisms from the 1996 North Gorda Ridge eruption. Deep Sea Research Part II: Topical Studies in Oceanography 45:2751–2766. doi: 10.1016/S0967-0645(98)00092-7

3. Staley JT (1997) Biodiversity: are microbial species threatened? Current Opinion in Biotechnology 8:340–345.

4. Fenchel T (2005) Cosmopolitan microbes and their "cryptic" species. Aquatic Microbial Ecology 41:49–54. doi: 10.3354/ame041049

5. Drake LA, Doblin MA, Dobbs FC (2007) Potential microbial bioinvasions via ships' ballast water, sediment, and biofilm. Marine Pollution Bulletin 55:333–341. doi: 10.1016/j.marpolbul.2006.11.007

6. Inagaki F, Nunoura T, Nakagawa S, et al (2006) Biogeographical distribution and diversity of microbes in methane hydrate-bearing deep marine sediments on the Pacific Ocean Margin. PNAS 103:2815–2820. doi: 10.1073/pnas.0511033103

7. Fox D (2014) Lakes under the ice: Antarctica's secret garden. Nature 512:244–246. doi: 10.1038/512244a

8. Counts JA, Zeldes BM, Lee LL, et al (2017) Physiological, metabolic and biotechnological features of extremely thermophilic microorganisms. Wiley Interdiscip Rev Syst Biol Med 9:e1377. doi: 10.1002/wsbm.1377

9. Boetius A, Anesio AM, Deming JW, et al (2015) Microbial ecology of the cryosphere: sea ice and glacial habitats. Nat Rev Microbiol 13:677–690. doi: 10.1038/nrmicro3522

10. Bennett PM (2008) Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. Br J Pharmacol 153 Suppl 1:S347–57. doi: 10.1038/sj.bjp.0707607

11. Walsh C (2003) Antibiotics. Antibiotics. doi: 10.1128/9781555817886

12. Coughlan LM, Cotter PD, Hill C, Alvarez-Ordonez A (2015) Biotechnological applications of functional metagenomics in the food and pharmaceutical industries. Frontiers in microbiology 6:672. doi: 10.3389/fmicb.2015.00672

13. Newman DJ, Cragg GM (2007) Natural Products as Sources of New Drugs over the Last 25 Years. Journal of Natural Products 70:461–477. doi: 10.1021/np068054v

14. Rappe MS, Giovannoni SJ (2003) The uncultured microbial majority. Annu

Rev Microbiol 57:369–394. doi: 10.1146/annurev.micro.57.030502.090759

15. Stewart EJ (2012) Growing unculturable bacteria. Journal of bacteriology 194:4151–4160. doi: 10.1128/JB.00345-12

16. Torsvik V, Goksoyr J, Daae FL (1990) High diversity in DNA of soil bacteria. Applied and Environmental Microbiology 56:782–787.

17. Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. Microbiology and Molecular Biology Reviews 68:669–685. doi: 10.1128/MMBR.68.4.669-685.2004

18. Handelsman J, Rondon MR, Brady SF, et al (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemistry & biology 5:R245–9.

19. Sleator RD, Shortall C, Hill C (2008) Metagenomics. Letters in applied microbiology 47:361–366. doi: 10.1111/j.1472-765X.2008.02444.x

20. National Research Council (US) Committee on Metagenomics: Challenges, Applications F (2007) Why Metagenomics?

21. Katz M, Hover BM, Brady SF (2016) Culture-independent discovery of natural products from soil metagenomes.

22. Spainhour CB (2005) Natural products. Pharmaceutical Sciences Encyclopedia

23. Kushner D (1981) Extreme Environments: Are There Any Limits to Life? In: Comets and the Origin of Life. Springer Netherlands, Dordrecht, pp 241–248

24. Butler MS, Buss AD (2006) Natural products--the future scaffolds for novel antibiotics? Biochemical pharmacology 71:919–929. doi: 10.1016/j.bcp.2005.10.012

25. Natural Product Reports.

26. Wang GY, Graziani E, Waters B, et al (2000) Novel natural products from soil DNA libraries in a streptomycete host. Org Lett 2:2401–2404.

27. Fleming A (1929) On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of B. influenzæ. British journal of experimental pathology 10:226.

28. Schatz A, Bugie E, Waksman SA (2005) Streptomycin, a substance exhibiting antibiotic activity against gram-positive and gram-negative bacteria. 1944. Clin Orthop Relat Res 3–6.

29. Sakula A (1988) Selman Waksman (1888–1973), discoverer of streptomycin: A centenary review. British Journal of Diseases of the Chest 82:23–31. doi: 10.1016/0007-0971(88)90005-8

30. Zhang MM, Qiao Y, Ang EL, Zhao H (2017) Using natural products for drug

discovery: the impact of the genomics era. Expert Opin Drug Discov 12:475–487. doi: 10.1080/17460441.2017.1303478

31. Churchman GJ, Landa ER (2014) The soil underfoot: Infinite possibilities for a finite resource.

32. Weisburg WG, Barns SM, Pelletier DA, Lane DJ (1991) 16S ribosomal DNA amplification for phylogenetic study. Journal of bacteriology 173:697–703. doi: 10.1128/jb.173.2.697-703.1991

33. J T Staley A, Konopka A (2003) Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats. http://dxdoiorg/101146/annurevmi39100185001541 39:321–346. doi: 10.1146/annurev.mi.39.100185.001541

34. CURTIS T, SLOAN W (2004) Prokaryotic diversity and its limits: microbial community structure in nature and implications for microbial ecology. Current opinion in microbiology 7:221–226. doi: 10.1016/j.mib.2004.04.010

35. and MSR, Giovannoni SJ (2003) The Uncultured Microbial Majority. http://dxdoiorg/101146/annurevmicro57030502090759 57:369–394. doi: 10.1146/annurev.micro.57.030502.090759

36. Torsvik V, Øvreås L (2002) Microbial diversity and function in soil: from genes to ecosystems. Current opinion in microbiology 5:240–245. doi: 10.1016/S1369-5274(02)00324-7

37. Schluenzen F, Tocilj A, Zarivach R, et al (2000) Structure of Functionally Activated Small Ribosomal Subunit at 3.3 Å Resolution. Cell 102:615–623. doi: 10.1016/S0092-8674(00)00084-2

38. Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: The primary kingdoms. PNAS 74:5088–5090. doi: 10.1073/pnas.74.11.5088

39. Case RJ, Boucher Y, Dahllöf I, et al (2007) Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Applied and Environmental Microbiology 73:278–288. doi: 10.1128/AEM.01177-06

40. Allen HK, Moe LA, Rodbumrer J, et al (2008) Functional metagenomics reveals diverse [beta]-lactamases in a remote Alaskan soil. The ISME journal 3:243–251. doi: 10.1038/ismej.2008.86

41. Böhnke S, Perner M (2015) A function-based screen for seeking RubisCO active clones from metagenomes: novel enzymes influencing RubisCO activity. The ISME journal 9:735–745. doi: 10.1038/ismej.2014.163

42. Leis B, Angelov A, Mientus M, et al (2015) Identification of novel esterase-active enzymes from hot environments by use of the host bacterium Thermus thermophilus. Frontiers in microbiology 6:275. doi: 10.3389/fmicb.2015.00275

43. Henne A, Daniel R, Schmitz RA, Gottschalk G (1999) Construction of environmental DNA libraries in Escherichia coli and screening for the presence

of genes conferring utilization of 4-hydroxybutyrate. Applied and Environmental Microbiology 65:3901–3907.

44. Rondon MR, August PR, Bettermann AD, et al (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Applied and Environmental Microbiology 66:2541–2547.

45. Gillespie DE, Brady SF, Bettermann AD, et al (2002) Isolation of antibiotics turbomycin a and B from a metagenomic library of soil microbial DNA. Applied and Environmental Microbiology 68:4301–4306.

46. Lim HK, Chung EJ, Kim J-C, et al (2005) Characterization of a forest soil metagenome clone that confers indirubin and indigo production on Escherichia coli. Applied and Environmental Microbiology 71:7768–7777. doi: 10.1128/AEM.71.12.7768-7777.2005

47. Fierer N, Breitbart M, Nulton J, et al (2007) Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. Applied and Environmental Microbiology 73:7059–7066. doi: 10.1128/AEM.00358-07

48. Berlemont R, Delsaute M, Pipers D, et al (2009) Insights into bacterial cellulose biosynthesis by functional metagenomics on Antarctic soil samples. The ISME journal 3:1070–1081. doi: 10.1038/ismej.2009.48

49. Torres-Cortés G, Millán V, Ramírez-Saad HC, et al (2011) Characterization of novel antibiotic resistance genes identified by functional metagenomics on soil samples. Environmental microbiology 13:1101–1114. doi: 10.1111/j.1462-2920.2010.02422.x

50. Knietsch A, Waschkowitz T, Bowien S, et al (2003) Metagenomes of complex microbial consortia derived from different soils as sources for novel genes conferring formation of carbonyls from short-chain polyols on Escherichia coli. J Mol Microb Biotech 5:46–56. doi: 10.1159/000068724

51. Henne A, Schmitz RA, Bomeke M, et al (2000) Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on Escherichia coli. Applied and Environmental Microbiology 66:3113–3116.

52. Handelsman J (2004) Metagenomics: Application of Genomics to Uncultured Microorganisms. Microbiology and Molecular Biology Reviews 68:669–685. doi: 10.1128/MMBR.68.4.669-685.2004

53. Kimelman A, Levy A, Sberro H, et al (2012) A vast collection of microbial genes that are toxic to bacteria. Genome Res 22:802–809. doi: 10.1101/gr.133850.111

54. Hess M, Sczyrba A, Egan R, et al (2011) Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen. Science 331:463–467. doi: 10.1126/science.1200387

55. Lloyd-Jones G, Hunter DWF (2001) Comparison of rapid DNA extraction methods applied to contrasting New Zealand soils. Soil Biology and Biochemistry 33:2053–2059. doi: 10.1016/S0038-0717(01)00133-X

56. Reavy B, Swanson MM, Cock PJA, et al (2015) Distinct circular single-stranded DNA viruses exist in different soil types. Appl Environ Microb 81:3934–3945. doi: 10.1128/AEM.03878-14

57. Young IM, Crawford JW (2004) Interactions and self-organization in the soil-microbe complex. Science 304:1634–1637. doi: 10.1126/science.1097394

58. Charlop-Powers Z, Milshteyn A, Brady SF (2014) Metagenomic small molecule discovery methods. Current opinion in microbiology 19:70–75. doi: 10.1016/j.mib.2014.05.021

59. Uchiyama T, Miyazaki K (2009) Functional metagenomics for enzyme discovery: Challenges to efficient screening. Curr Opin Biotechnol 20:616–622. doi: 10.1016/j.copbio.2009.09.010

60. Banik JJ, Brady SF (2010) Recent application of metagenomic approaches toward the discovery of antimicrobials and other bioactive small molecules. Current opinion in microbiology 13:603–609. doi: 10.1016/j.mib.2010.08.012

61. Li X, Qin L (2005) Metagenomics-based drug discovery and marine microbial diversity. Trends Biotechnol 23:539–543. doi: 10.1016/j.tibtech.2005.08.006

62. Singh RP, Kumari P, Reddy CR (2015) Antimicrobial compounds from seaweeds-associated bacteria and fungi. Applied microbiology and biotechnology 99:1571–1586. doi: 10.1007/s00253-014-6334-y

63. Ling LL, Schneider T, Peoples AJ, et al (2015) A new antibiotic kills pathogens without detectable resistance. Nature 517:455–459. doi: 10.1038/nature14098

64. Marko D, Schatzle S, Friedel A, et al (2001) Inhibition of cyclin-dependent kinase 1 (CDK1) by indirubin derivatives in human tumour cells. British journal of cancer 84:283–289. doi: 10.1054/bjoc.2000.1546

65. Iqbal HA, Craig JW, Brady SF (2014) Antibacterial enzymes from the functional screening of metagenomic libraries hosted in Ralstonia metallidurans. FEMS microbiology letters 354:19–26. doi: 10.1111/1574-6968.12431

66. Scanlon TC, Dostal SM, Griswold KE (2014) A high-throughput screen for antibiotic drug discovery. Biotechnology and bioengineering 111:232–243. doi: 10.1002/bit.25019

67. Chua KYL, Howden BP, Jiang J-H, et al (2014) Population genetics and the evolution of virulence in Staphylococcus aureus. Infection, Genetics and Evolution 21:554–562.

68. Barriuso J, Valverde JR, Mellado RP (2011) Estimation of bacterial diversity using next generation sequencing of 16S rDNA: a comparison of different workflows. Bmc Bioinformatics. doi: 10.1186/1471-2105-12-473

69. Ekkers DM, Cretoiu MS, Kielak AM, Van Elsas JD (2011) The great screen anomaly—a new frontier in product discovery through functional metagenomics. Applied microbiology and biotechnology 93:1005–1020. doi: 10.1007/s00253-011-3804-3

70. Roh C, Villatte F, Kim B-G, Schmid RD (2006) Comparative Study of Methods for Extraction and Purification of Environmental DNA From Soil and Sludge Samples. Applied Biochemistry and Biotechnology 134:97–112. doi: 10.1385/ABAB:134:2:97

71. Williamson KE, Kan J, Polson SW, Williamson SJ (2011) Optimizing the indirect extraction of prokaryotic DNA from soils. Soil Biology and Biochemistry 43:736–748. doi: 10.1016/j.soilbio.2010.04.017

72. Binga EK, Lasken RS, Neufeld JD (2008) Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. The ISME journal

73. Wilson MC, Piel J (2013) Metagenomic approaches for exploiting uncultivated bacteria as a resource for novel biosynthetic enzymology. Chemistry & biology 20:636–647. doi: 10.1016/j.chembiol.2013.04.011

74. Chiu CY, Tian G (2011) Chemical structure of humic acids in biosolids-amended soils as revealed by NMR spectroscopy. Applied Soil Ecology 49 IS - :76–80.

75. Tsai YL, Olson BH (1992) Rapid method for separation of bacterial DNA from humic substances in sediments for polymerase chain reaction. Applied and Environmental Microbiology 58:2292–2295.

76. Frostegård A, Courtois S, Ramisse V, et al (1999) Quantification of bias related to the extraction of DNA directly from soils. Applied and Environmental Microbiology 65:5409–5420.

77. Cho J-C, Lee D-H, Cho Y-C, et al (1996) Direct Extraction of DNA from Soil for Amplification of 16S rRNA Gene Sequences by Polymerase Chain Reaction. The Journal of Microbiology 34:229–235.

78. Holben WE, Jansson JK, Chelm BK, Tiedje JM (1988) DNA Probe Method for the Detection of Specific Microorganisms in the Soil Bacterial Community. Appl Environ Microb 54:703–711.

79. Tanveer A, Yadav S, Yadav D (2016) Comparative assessment of methods for metagenomic DNA isolation from soils of different crop growing fields. 3 Biotech 6:17. doi: 10.1007/s13205-016-0543-2

80. Jiménez DJ, Andreote FD, Chaves D, et al (2012) Structural and Functional Insights from the Metagenome of an Acidic Hot Spring Microbial Planktonic Community in the Colombian Andes. PLoS ONE 7:e52069. doi: 10.1371/journal.pone.0052069

81. Jung J, Philippot L, Park W (2016) Metagenomic and functional analyses of the

consequences of reduction of bacterial diversity on soil functions and bioremediation in diesel-contaminated microcosms. Scientific Reports 6:58. doi: 10.1038/srep23012

82. Fierer N, Lauber CL, Ramirez KS, et al (2012) Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. The ISME journal 6:1007–1017. doi: 10.1038/ismej.2011.159

83. Wu GD, Lewis JD, Hoffmann C, et al (2010) Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. BMC Microbiology 2010 10:1 10:206. doi: 10.1186/1471-2180-10-206

84. Kennedy NA, Walker AW, Berry SH, et al (2014) The Impact of Different DNA Extraction Kits and Laboratories upon the Assessment of Human Gut Microbiota Composition by 16S rRNA Gene Sequencing. PLoS ONE 9:e88982. doi: 10.1371/journal.pone.0088982

85. Sagova-Mareckova M, Cermak L, Novotna J, et al (2008) Innovative methods for soil DNA purification tested in soils with widely differing characteristics. Appl Environ Microb 74:2902–2907. doi: 10.1128/AEM.02161-07

86. Aric Joneja XH (2009) A device for automated hydrodynamic shearing of genomic DNA. Biotechniques 46:553–556. doi: 10.2144/000113123

87. Oefner PJ, Hunicke-Smith SP, Chiang L (1996) Efficient random subcloning of DNA sheared in a recirculating point-sink flow system. Nucleic acids …

88. Okpodu CM, Robertson D, Boss WF, et al (1994) Rapid isolation of nuclei from carrot suspension culture cells using a BioNebulizer. Biotechniques 16:154–159.

89. Deininger PL (1983) Approaches to rapid DNA sequence analysis. Analytical Biochemistry 135:247–263.

90. Walker A, Taylor J, Rowe D, Summers D (2008) A method for generating sticky-end PCR products which facilitates unidirectional cloning and the one-step assembly of complex DNA constructs. Plasmid 59:155–162. doi: 10.1016/j.plasmid.2008.02.002

91. Guzman LM, Belin D, Carson MJ, Beckwith J (1995) Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. Journal of bacteriology 177:4121–4130.

92. Sundquist A, Ronaghi M, Tang H, et al (2007) Whole-genome sequencing and assembly with high-throughput, short-read technologies. PLoS ONE 2:e484. doi: 10.1371/journal.pone.0000484

93. Montaser R, Luesch H (2011) Marine natural products: a new wave of drugs? Future medicinal chemistry 3:1475–1489. doi: 10.4155/fmc.11.118

94. Kakirde KS, Wild J, Godiska R, et al (2011) Gram negative shuttle BAC vector for heterologous expression of metagenomic libraries. Gene 475:57–62. doi: 10.1016/j.gene.2010.11.004

95. Lam KN, Cheng J, Engel K, et al (2015) Current and future resources for functional metagenomics. Frontiers in microbiology. doi: 10.3389/fmicb.2015.01196

96. Aakvik T, Degnes KF, Dahlsrud R, et al (2009) A plasmid RK2-based broad-host-range cloning vector useful for transfer of metagenomic libraries to a variety of bacterial species. FEMS microbiology letters 296:149–158. doi: 10.1111/j.1574-6968.2009.01639.x

97. Westenberg M, Bamps S, Soedling H, et al (2010) Escherichia coli MW005: lambda Red-mediated recombineering and copy-number induction of oriV-equipped constructs in a single host. BMC biotechnology 10:27. doi: 10.1186/1472-6750-10-27

98. van Elsas JD, Speksnijder AJ, van Overbeek LS (2008) A procedure for the metagenomics exploration of disease-suppressive soils. Journal of Microbiological Methods 75:515–522. doi: 10.1016/j.mimet.2008.08.004

99. Gabor EM, Alkema WB, Janssen DB (2004) Quantifying the accessibility of the metagenome by random expression cloning techniques. Environmental microbiology 6:879–886. doi: 10.1111/j.1462-2920.2004.00640.x

100. Rondon MR, August PR, Bettermann AD, et al (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Applied and Environmental Microbiology 66:2541–2547.

101. Martinez A, Kolvek SJ, Yip CL, et al (2004) Genetically modified bacterial strains and novel bacterial artificial chromosome shuttle vectors for constructing environmental libraries and detecting heterologous natural products in multiple expression hosts. Applied and Environmental Microbiology 70:2452–2463.

102. Williamson LL, Borlee BR, Schloss PD, et al (2005) Intracellular screen to identify metagenomic clones that induce or inhibit a quorum-sensing biosensor. Applied and Environmental Microbiology 71:6335–6344. doi: 10.1128/AEM.71.10.6335-6344.2005

103. Lussier F-X, Chambenoit O, Côté A, et al (2011) Construction and functional screening of a metagenomic library using a T7 RNA polymerase-based expression cosmid vector. Journal of industrial microbiology & biotechnology 38:1321–1328. doi: 10.1007/s10295-010-0915-2

104. Gaida SM, Sandoval NR, Nicolaou SA, et al (2015) Expression of heterologous sigma factors enables functional screening of metagenomic and heterologous genomic libraries. Nat Commun. doi: 10.1038/ncomms8045

105. Laureti L, Song L, Huang S, et al (2011) Identification of a bioactive 51-

membered macrolide complex by activation of a silent polyketide synthase in Streptomyces ambofaciens. PNAS 108:6258–6263. doi: 10.1073/pnas.1019077108

106. van Wezel GP, McDowall KJ (2011) The regulation of the secondary metabolism of Streptomyces: new links and experimental advances. Nat Prod Rep 28:1311–1333. doi: 10.1039/C1NP00003A

107. Carlet J, Collignon P, Goldmann D, et al (2011) Society's failure to protect a precious resource: antibiotics. Lancet 378:369–371. doi: 10.1016/S0140-6736(11)60401-7

108. Spellberg B, Bartlett JG, Gilbert DN (2013) The Future of Antibiotics and Resistance. New England Journal of Medicine 368:299–302. doi: 10.1056/NEJMp1215093

109. Aminov RI (2010) A Brief History of the Antibiotic Era: Lessons Learned and Challenges for the Future. Frontiers in microbiology. doi: 10.3389/fmicb.2010.00134

110. Vollmers J, Wiegand S, Kaster A-K (2017) Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters! PLoS ONE 12:e0169662. doi: 10.1371/journal.pone.0169662

111. Culligan EP, Sleator RD, Marchesi JR, Hill C (2014) Metagenomics and novel gene discovery: promise and potential for novel therapeutics. Virulence 5:399–412. doi: 10.4161/viru.27208

112. Garrido-Cardenas JA, Manzano-Agugliaro F (2017) The metagenomics worldwide research. Curr Genet 29:2253–11. doi: 10.1007/s00294-017-0693-8

113. Turaev D, Rattei T (2016) High definition for systems biology of microbial communities: metagenomics gets genome-centric and strain-resolved. Curr Opin Biotechnol 39:174–181. doi: 10.1016/j.copbio.2016.04.011

114. and JTS, Gosink JJ (2003) Poles Apart: Biodiversity and Biogeography of Sea Ice Bacteria. http://dxdoiorg/101146/annurevmicro531189 53:189–215. doi: 10.1146/annurev.micro.53.1.189

115. Sorek R, Zhu Y, Creevey CJ, et al (2007) Genome-wide experimental determination of barriers to horizontal gene transfer. Science 318:1449–1452. doi: 10.1126/science.1147112

116. Darmon E, Leach DRF (2014) Bacterial genome instability. Microbiol Mol Biol Rev 78:1–39. doi: 10.1128/MMBR.00035-13

117. Lawrence JG (2005) Horizontal and vertical gene transfer: the life history of pathogens. Contributions to microbiology

118. World Health, Organization (2014) Antimicrobial resistance: 2014 global report on surveillance. World Health Organization; Geneva; Switzerland, Geneva, Switzerland

119. Mendelson M, Matsoso MP (2015) The World Health Organization Global Action Plan for antimicrobial resistance. South African Medical Journal 105:325. doi: 10.7196/SAMJ.9644

120. Radhouani H, Silva N, Poeta P, et al (2014) Potential impact of antimicrobial resistance in wildlife, environment and human health. Frontiers in microbiology 5:23. doi: 10.3389/fmicb.2014.00023

121. Pelicic V, Reyrat JM, Gicquel B (1996) Expression of the Bacillus subtilis sacB gene confers sucrose sensitivity on mycobacteria. Journal of bacteriology 178:1197–1199.

122. Casali N (2003) Escherichia coli Host Strains. In: E. coli Plasmid Vectors. Humana Press, New Jersey, pp 27–48

123. Rajninec M, Libantov J, Jopc k M (2016) Optimalisation of expression conditions for production of round-leaf sundew chitinase (Drosera rotundifolia L.) in three E. coli expression strains. Journal of Central European Agriculture 17:1104–1118. doi: 10.5513/JCEA01/17.4.1818

124. Samuelson JC (2010) Recent Developments in Difficult Protein Expression: A Guide to E. coli Strains, Promoters, and Relevant Host Mutations. In: Heterologous Gene Expression in E.coli. Humana Press, Totowa, NJ, pp 195–209

125. Alsante A (1970) Nanodrop Spectrophotometer (ND-1000) for Nucleic Acid v1. protocolsio. doi: 10.17504/protocols.io.id2ca8e

126. Desjardins P, Hansen JB, Allen M (2009) Microvolume Protein Concentration Determination Using the NanoDrop 2000c Spectrophotometer. JoVE (Journal of Visualized Experiments). doi: 10.3791/1610

127. Okamoto T, Okabe S (2000) Ultraviolet absorbance at 260 and 280 nm in RNA measurement is dependent on measurement solution. International Journal of Molecular Medicine. doi: 10.3892/ijmm.5.6.657

128. Kovacic RT, Comal L, Bendich AJ (1995) Protection of megabase DNA from shearing. Nucleic Acids Res 23:3999–4000. doi: 10.1093/nar/23.19.3999

129. Bürgmann H, Pesaro M, Widmer F, Zeyer J (2001) A strategy for optimizing quality and quantity of DNA extracted from soil. Journal of Microbiological Methods 45:7–20. doi: 10.1016/S0167-7012(01)00213-5

130. Sachs M (2008) Faculty of 1000 evaluation for Protection of DNA during preparative agarose gel electrophoresis against damage induced by ultraviolet light. Biotechniques 21:898–903. doi: 10.3410/f.1133854.591946

131. Roose-Amsaleg CL, Garnier-Sillam E, Harry M (2001) Extraction and purification of microbial DNA from soil and sediment samples. Applied Soil Ecology 18:47–60. doi: 10.1016/S0929-1393(01)00149-4

132. Inoue H, Nojima H, Okayama H (1990) High efficiency transformation of

Escherichia coli with plasmids. Gene 96:23–28. doi: 10.1016/0378-1119(90)90336-P

133. Fosmid Libraries | Amplicon Express. In: ampliconexpress.com. http://ampliconexpress.com/fosmid-libraries/?gclid=Cj0KEQjwv_fKBRCG8a3ao-OQuZ8BEiQAvpHp6A7tJPzozESGZO0VPc-TK3uZoj_OATS9uEoaoVCxIKkaAjA28P8HAQ. Accessed 17 Jul 2017

134. Taupp M, Mewis K, Hallam SJ (2011) The art and design of functional metagenomic screens. Curr Opin Biotechnol 22:465–472. doi: 10.1016/j.copbio.2011.02.010

135. Whalen MC, Innes RW, Bent AF, Staskawicz BJ (1991) Identification of Pseudomonas syringae pathogens of Arabidopsis and a bacterial locus determining avirulence on both Arabidopsis and soybean. Plant Cell 3:49–59. doi: 10.1105/tpc.3.1.49

136. Buell CR, Joardar V, Lindeberg M, et al (2003) The complete genome sequence of the Arabidopsis and tomato pathogen Pseudomonas syringae pv. tomato DC3000. PNAS 100:10181–10186. doi: 10.1073/pnas.1731982100

137. Yanisch-Perron C, Vieira J, Messing J (1985) Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mpl8 and pUC19 vectors. Gene 33:103–119. doi: 10.1016/0378-1119(85)90120-9

138. Schäfer A, Tauch A, Jäger W, et al (1994) Small mobilizable multi-purpose cloning vectors derived from the Escherichia coli plasmids pK18 and pK19: selection of defined deletions in the chromosome of Corynebacterium glutamicum. Gene 145:69–73. doi: 10.1016/0378-1119(94)90324-7

Appendix

Name	Genotype	Reference
Escherichia coli OneShot TOP10	F- mcrA Δ(mrr-hsdRMS-mcrBC) ϕ80lacZΔM15 ΔlacX74 recA1 araD139 Δ(araleu)7697 galU galK rpsL (StrR) endA1 nupG	Invitrogen
Escherichia coli LMG194	F- ΔlacX74 gal E thi rpsL ΔphoA (Pvu II) Δara714 leu::Tn10. Please note that this strain is streptomycin and tetracycline resistant.	Invitrogen
Pseudomonas Syringae pathovar tomato DC3000	Contains putative transcriptional regular, locus tag: PSPTO_4315	Whalen *et al.* [135] & Buell *et al.* [136]

Table 0-1. List of strains used in this study.

Name	Characteristics	Reference
pUC19	High copy number *E. coli* cloning vector, Amp^R	Yanisch-Perron *et al.* [137]
pK18mobsacB	Suicide vector containing the sacB gene which is toxic to *E. coli* when expressed, kan^R	Schäfer *et al*. [138]
pBAD/*Myc*-His B	Tightly regulated expression vector, pBR322 origin of replication, low copy number, araBAD promoter, MCS, AraC ORF, Amp^R	Invitrogen

Table 0-2. List of plasmids used in this study.

Name	Sequence 5' to 3'	Target
Primers used to amplify PSPTO in frame		
PSPTO_POS_F	CTCGAGGGCTCCCTGATCA ACGAA	5' of PSPTO
PSPTO_POS_R	GGTACCTTAGCCCAGCAGA GTGGC	3' of PSPTO
Primers used to amplify PSPTO out of frame		
PSPTO_NEG_F	GGTACCATGTCCCTGATCA ACGAA	5' of PSPTO
PSPTO_NEG_R	CTCGAGTTAGCCCAGCAGA GTGGC	3' of PSPTO
Primers used to amplify sacB in frame		
sacB_POS_F	CCATGGGCAACATCAAAAA GTTTGCAAAAC	5' of sacB
sacB_POS_R	TCTAGATTATTTGTTAACTG TTAATTGTCC	3' of sacB
Primers used to amplify sacB out of frame		
sacB_NEG_F	CCATGGATGAACATCAAAAA GTTTGCAAA	5' of sacB
sacB_NEG_R	TCTAGATTATTTGTTAACTG TTAATTGTCC	3' of sacB
Universal primers used for verifying cloned inserts		
pBAD Forward	ATGCCATAGCATTTTTATCC	MCS of pBAD/*Myc*-His B
pBAD Reverse	GATTTAATCTGTATCAGG	

Table 0-3. List of primers used in this study.