Study and exploitation of diverse soil environments for novel natural product discovery using metagenomic approaches

Chiara Borsetto

A thesis submitted to the University of Warwick for the degree of

Doctor of Philosophy

School of Life Sciences,

University of Warwick,

Coventry, CV4 7AL

September 2017

Table of contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my supervisors Prof. Elizabeth Wellington and Prof. David Hodgson at University of Warwick and Dr. Stefano Donadio at KtedoGen s.r.l. (Italy) for their support, guidance and critical discussion on the project.

I would like to thank the Marie Curie funding programme, the EU FP7 (grant agreement N° 289285) and the University of Warwick for providing funding.

I would like to thank Dr. Joseph Christie-Oleza (University of Warwick) and Prof. Svein Valla (Norwegian University of Science and Technology) for donating pJOC100Km and pJB658 vectors respectively.

I would also like to thank present and past members of the Wellington group for their help and support in and out the lab.

Thanks to my advisory panel (Prof. Robin Allaby and Dr. Christophe Corre) for critical discussion on my reports and on the project in general.

Thanks to the TRAIN-ASAP fellows and PIs for great meetings and critical discussion of my progress.

Thanks to all the colleagues at KtedoGen s.r.l. (Milan, Italy). A particular thanks to Dr. Margherita Sosio and Dr. Arianna Tocchetti for constructive discussion on the work performed under their guidance during my secondment at KtedoGen.

Last but not least, I would like to thank in particular Raffaele, my family and Alison for their great support and encouragement during these years.

Declaration

In accordance with the University of Warwick regulations for the degree of Doctor of Philosophy, I certify that this thesis has been written solely by me. No aspect of this work has been submitted to any other institution for any other degree of award.

The work contained in this thesis is my own with the exceptions as follow:
Chapter 2) Assignation of taxonomy to NRPS and PKS reads were performed in collaboration with the European Bioinformatics Institute metagenomics group.
Chapter 3) Undergraduate student Dale Harding performed some of the DNA extractions from soil under my supervision and guidance;
Chapter 5) Due to time constraint, Dr. Ian Lidbury assisted me with the final cloning steps leading to the construction of the vector pBCLI::GSTA and the generation of the 'knock-in' *trfA::Gm* mutant in *Pseudomonas putida* KT2440 and BIRD-1.

The general introduction (Chapter 1) was adapted from the original book chapter "Bioprospecting Soil Metagenomes for Antibiotics" by Borsetto and Wellington published by Springer in the book "Bioprospecting – Success, Potential and Constraints" (Vol.16, 2017). Permission for use was granted by Springer. I provided all the text which was checked by Prof. Wellington.

Part of the results reported in Chapter 5 were already included in the following publication:
Amos GCA*, Borsetto C*, Laskaris P, Krsek M, Berry AE, Newsham KK, Calvo-Bado L, Pearce DA, Vallin C & Wellington EMH (2015) Designing and implementing an assay for the detection of rare and divergent NRPS and PKS clones in European, Antarctic and Cuban Soils. *PLoS ONE* **10**: e0138327.
* AGCA and CB co-authored the article.

Abstract

Natural products with antimicrobial activity have played an important role in the treatment of infection since their discovery. The increasing emergence of pathogens resistant to multiple antibiotics has raised awareness of the urgent need for novel antibiotics. Soil microorganisms are the major source of antibiotics and Actinobacteria in particular have an impressive capacity for production of diverse bioactive secondary metabolites. However, culture-independent studies have shown a greater microbial diversity present in soil with potential for novel chemical structures and these can be explored further using metagenomic approaches capturing genes without the need to cultivate the host.

Different metagenomic tools were used to study and explore microbial secondary metabolite diversity in soil. In particular, amplicon sequencing of 16S rRNA gene, NRPS and PKS biosynthetic genes allowed the identification of novel potential phylogenetic drivers of secondary metabolite diversity in the less characterized phyla Verrucomicrobia and Bacteroidetes and potential geographic hotspots harbouring unique biosynthetic diversity such as Antarctica and Cuba. The exploitation of these hotspots presented some bottlenecks in the form of DNA extraction efficiency, library creation, screening and heterologous expression. These were overcome by comparative analysis of different eDNA extraction methods to optimise fragment size and purity combined with development of new cloning tools for both DNA capture and expression. Modification of the microbial community through the amendment of the soil with chitin, highlighted the beneficial effect of microbial enrichment allowing a higher recovery of eDNA and higher detection of the biosynthetic gene of interest related to secondary metabolite production. Further additions were made to the metagenomic molecular toolbox in the form of BAC vectors (pBCaBAC and pBCkBAC) which were tested with suitable heterologous host systems (*Streptomyces* sp. and the engineered *Pseudomonas putida* species) potentially facilitating heterologous expression.

In conclusion this is the first study to identify the drivers of microbial secondary metabolite diversity *in situ* and provided a comparative analysis of a range of diverse soil types. This approach paired with new developments in metagenomic technologies will make a substantial contribution to improving the likelihood for discovery and exploitation of new drugs for treating multi-resistant pathogenic bacteria.

Abbreviations

A	Adenylation domain
ANOSIM	Analysis of similarity
ANOVA	Analysis of variance
BAC	Bacterial artificial chromosome
BGC	Biosynthetic gene cluster
BLAST	Basic local alignment search tool
BSA	Bovine serum albumin
C	Condensation domain
CBD	Convention on biological diversity
CDS	Coding sequence
CFU	Colony forming unit
CTAB	hexadecyltrimethylammonium bromide
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
E	Epimerization domain
EBI	European bioinformatics institute
EC	Electrical conductivity
eDNA	Environmental deoxyribonucleic acid
EDTA	Ethylenediaminetetraacetic acid
ENA	European nucleotide archive
HMW	High molecular weigh
HSD	Honest significant difference
IMG	Integrated microbial genomes
KS	Ketosynthase domain
MCB	Master cell bank
MES	2-(N-morpholino)ethanesulfonic acid
MG-RAST	Metagenomics rapid annotation using subsystem technology
MOPS	3-(N-morpholino)propanesulfonic acid
NCBI	National center for biotechnology information
NGS	Next-generation sequencing
NRP	Non-ribosomal peptide

NRPS	Non-ribosomal peptide synthetase
ORF	Open reading frame
OTU	Operational taxnomic unit
PCP	Peptide carrier protein
PCoA	Principle coordinate analysis
PCR	Polymerase chain reaction
PFGE	Pulse field gel electrophoresis
PK	Polyketide
PKS	Polyketide synthase
PLG	Phase lock gel
QIIME	Quantitative insights into microbial ecology
qPCR	Quantitative polymerase chain reaction
RDP	Ribosomal database project
rRNA	Ribosomal ribonucleic acid
SDS	Sodium dodecyl sulphate
smCOGs	Secondary metabolism clusters of orthologous groups
TAE	Tris-acetate-EDTA buffer
Te	Thioesterase domain
TE	Tris-EDTA buffer
UK	United Kingdom
WCB	Working cell bank

Chapter 1

Introduction

1.1 Soil biodiversity and exploitation

"We know more about the movement of celestial bodies than about the soil underfoot." ***(Leonardo da Vinci)***

Soil is one of the most challenging environments for ecological understanding of how community structure relates to function (Fierer *et al.*, 2012, Bardgett & van der Putten, 2014). Many studies have reported impressive bacterial diversity within soils; the most significant early breakthroughs were achieved using DNA re-associations techniques, microscopy and molecular analysis targeting 16S rRNA phylogenetic marker gene for bacteria (Fægri *et al.*, 1977, Torsvik *et al.*, 1990, Torsvik *et al.*, 1998, Derakshani *et al.*, 2001). Below-ground diversity is affected by physicochemical properties of the soil but most importantly the plant root systems exert a significant impact on the rhizosphere microbial communities (see review: Bardgett *et al.* (2014)). In addition plant roots and soil fauna present a further habitat for colonisation and bacteria have been found growing in close association with Collembola (Hoffmann *et al.*, 1999) as well as ants and other insects (Kaltenpoth *et al.*, 2006, Kaltenpoth, 2009, Seipke *et al.*, 2013) and as endophytes in a wide variety of plants (Rungin *et al.*, 2012). Soil is a complex environment that changes in texture and composition in relation to the stratification (topsoil, bulk soil, permafrost, etc), edaphic features (e.g. clay content, pH, type of bedrock) but also climate will have a significant impact (Wall *et al.*, 2010). Climate changes can disrupt the normal composition releasing organic matter in soil previously unavailable to the microbial community impacting the carbon cycle (see review: Bardgett *et al.* (2008)). These changes can therefore contribute to major modifications and in some cases to the disappearance of unique microbial diversity (Wall *et al.*, 2010). Pristine environments such as Antarctica whose microbial communities have been shaped by extreme conditions represent a source of novel

functional genes which might be threatened by climate changes and increasing human activities (Pearce *et al.*, 2012, Cowan *et al.*, 2014).

The soil bacterial community has been the source of a large number of useful enzymes which have been applied in biotransformation, catalysis of specific reactions, degradation of pollutants, production of vitamins, pharmaceuticals and many other useful molecules (see review: Adrio & Demain (2014)). Early attempts to exploit the microbial diversity in soil focused on isolation techniques which proved to be highly successful in the discovery of novel antibiotics in the 1960's (Drews, 2000). However, it soon became apparent that different approaches were required to explore greater diversity. The rapid advances in molecular ecology focused on phylogenetic markers, revealed the uncultured majority in soil providing estimates of more than 95% of bacterial diversity had yet to be cultured (Torsvik *et al.*, 1998). The advent of metagenomics has enabled the exploitation of this bacterial diversity via culture-independent routes (see review: Lee & Lee (2013)). A novel salt-tolerant chitobiosidase from a soil metagenomic library represents only one of the many examples of enzymes recovered using metagenomics (Cretoiu *et al.*, 2015). Novel natural products with antimicrobial activities were also isolated and characterized from soil metagenomes (Gillespie *et al.*, 2002, Bauer *et al.*, 2010, Feng *et al.*, 2010). Metagenomics not only opened new horizons to the study and exploitation of microbial diversity but also improved our knowledge of bacterial biogeography. Finlay (2002) argued that the small size of prokaryotes would result in their ubiquitous distribution following the well-known paradigm "*everything is everywhere, but the environment selects"*(Baas Becking, 1934). Therefore the only difference amongst environments was in the relative abundance of each species in relation to the environmental conditions (De Wit & Bouvier, 2006). Biogeographic patterns were also investigated for functional genes such as natural product gene cluster diversity. An initial study based on amplicon sequencing conducted by Charlop-Powers *et al.* (2014) on 96 soil samples revealed that there was a strong correlation between soil type characteristics and secondary metabolite biosynthetic diversity. An extended study on a more global scale was performed by the same research group on 185 soil samples representing different environments such as coastal sediments, deserts, rainforests and temperate forests collected through the project "drugs from dirt". The analysis of the natural product diversity showed that there was a stronger relationship between soil samples collected at a closer physical distance than from similar biome

environments with different geographic locations (Charlop-Powers *et al.*, 2015). Moreover interesting "hotspots" for biomedically relevant compounds and specific gene clusters were identified using eSNaPD analysis (Reddy *et al.*, 2014). Biogeographical studies using metagenomic techniques are likely to become an essential tool to discover novel antibiotics.

1.1.1 Environment exploitation legislation

In the past the prospecting for genetic biodiversity and the exploitation of flora and fauna for discovery and commercialisation of products has been greatly criticised due to the unequal sharing of the benefits of patents and resulting commercial benefits with the source countries. The original explorers collecting novel plant material and other sources of medicinal benefits related to local know-how and folklore, failed to attribute any rights to local people and often this process was referred to as biopiracy (Sheridan, 2005, Mackey & Liang, 2012). Particular concern was focused on pharmaceutical companies' exploitation of genetic and chemical biodiversity for commercial purposes. Agreements and international laws were eventually set in motion with the Convention on Biological Diversity (UN, 1992) signed by 168 countries in 1992 at the Rio Earth Summit (the United Nations Conference on Environment and Development). Three major objectives were set and regulations determined firstly on the access to the genetic resources and technologies, secondly on the fair and equal benefit-sharing within the parties and lastly on the sustainable use of biological diversity. The CBD therefore established the rights and responsibilities amongst the Contracting Parties (sovereign states) to achieve these major goals. Private individuals such as companies or academic institutions were then automatically bound to the CBD depending on national laws. Nowadays the number of countries that signed the CBD reaches a total of 198. Further regulations on the fair and equal sharing of benefits arising from the use of genetic biodiversity were recently set in the Nagoya Protocol (entered into force from 2014), implementing the already existing ones present in the CBD.

1.2 Antibiotics discovery from soil

Since the discovery of penicillin from the fungus *Penicillium chrysogenum* by Alexander Fleming in 1928 (Fleming, 1929), the era of antibiotic discovery started and developed rapidly. The word "antibiotic" was firstly used by Selman Waksman in 1941 to describe small molecules produced by microorganisms with antimicrobial activity (Clardy *et al.*, 2009). Since microorganisms have been identified as the main producers of bioactive compounds with activity for pharmaceutical, industrial and agricultural purposes, the environment (in particular soil) has been explored for novel genes of interest related to the production of these bioactive natural molecules (see review: Keller & Zengler (2004)). During the golden age of antibiotic discovery (1950-1980), novel molecules were discovered from soil bacteria. The most fruitful group of soil-dwelling bacteria producing natural products with relevant medical importance were the Gram-positive Actinobacteria, in particular the genus *Streptomyces* (Watve et al., 2001). However, in the 1960's the rediscovery of the same molecules started to be more frequent in the screening programmes (see review: Fernandes (2006)). A good case was streptomycin which was discovered as a product of the soil isolate *Streptomyces griseus* firstly isolated by Waksman (Schatz & Waksman, 1944) and then rediscovered numerous times in screens of strains closely related to *S. griseus* such as *Streptomyces bikiniensis* (Johnstone & Waksman, 1948). The difficulties in discovering novel compounds and the simultaneous emergence of resistant nosocomial pathogens such as *Staphylococcus aureus* (Jevons, 1961) promoted the development of semi-synthetic antibiotics with better activity, broader spectrum and less susceptibility to the enzyme that conferred resistance (see review: Fernandes (2006)). Chemically synthesized antibiotics such as sulfonamide also extended the chemical diversity of antimicrobial agents even though they were not of natural origins. However, the likely misuse of antibiotics for non-exclusive treatment of infections and the use in agricultural practices increased the selection pressure on antimicrobial resistance genes promoting resistance in common pathogens (see review: Fàbrega *et al.* (2008)). The introduction of the fluoroquinolone antibiotic class initially provided a suitable treatment to these infections. However, in the last two decades resistance to multiple antibiotics has emerged in bacterial pathogens raising awareness of the urgent need for novel antibiotics (Wain *et al.* (1997), Yong *et al.*

(2009); reviews: Levy & Marshall (2004), Jovetic *et al.* (2010)). The isolation of ESKAPE pathogens (*Enterococcus faecium*, *S. aureus*, *Klebsiella pneumonia*, *Acinetobacter baumanii*, *Pseudomonas aeruginosa* and *Enterobacter*) resistant to 3rd generation cephalosporins and carbapenems (Srinivasan *et al.*, 2014)) or treatable only with second-line drugs (WHO, 2014), has induced governments to address antibiotics resistance as a serious and urgent problem and pharmaceutical companies to reinvest in drug discovery programmes (Frearson & Wyatt, 2010, Harbarth *et al.*, 2015). It has been estimated that the absence of novel antibiotics and effective interventions to prevent sepsis will lead to a post-antibiotic era where the level of fatality will return to that experienced in the 1930-1940's (WHO, 2014) and over 10 million people will die from antimicrobial resistant infections by 2050 (O'Neill, 2016).

Recent drug discovery programmes focused not only on the isolation and screening of novel bacterial isolates from soil but also on novel techniques such as functional metagenomics. In this case the bacterial genetic potential present in the soil is captured in suitable cloning vectors to form libraries in heterologous hosts which are screened for novel antimicrobial activities (Handelsman *et al.*, 1998). The use of these techniques has enabled investigation of the uncultured fraction of the microbial communities in different soils and their biosynthetic antimicrobial compounds potential (Reddy *et al.*, 2012, Owen *et al.*, 2013, Charlop-Powers *et al.*, 2014, Charlop-Powers *et al.*, 2015).

1.3. Strain isolation for novel antibiotics

Soil has been one of the major sources of exploitation for strain isolation for the identification of novel compounds with antimicrobial activity (see review: Keller & Zengler (2004)). Different soils have been explored using various isolation conditions with particular interest for the isolation of groups of bacteria that were able to produce compounds with antimicrobial activity (Table 1.1). Actinobacteria and in particular the genus *Streptomyces* for instance, represented one of the major groups producing inhibitory bioactive compounds effective against other bacteria since the beginning of antibiotic discovery programs (Watve *et al.*, 2001, Berdy, 2012). A famous example is streptomycin produced by *Streptomyces griseus* and firstly discovered by Waksman's group which was notable for the broad spectrum of activity of this

aminoglycoside (Schatz & Waksman, 1944). Soils from all over the world have been collected and tested over the last century, leading to the isolation of different Actinobacteria (Table 1.1), some of which produce useful natural products (Table 1.2).

Table 1.1 Examples of species isolated from soil samples collected all over the world.

Species	Soil location for isolation	Reference
Streptomyces venezuelae	Caracas, Venezuelae	(Ehrlich *et al.*, 1948)
Streptomyces bikiniensis	Bikini atoll, Pacific Island	(Johnstone, 1947)
Streptomyces costaricanus	Costa Rica	(Esnard *et al.*, 1995)
Streptomyces beijiangensis	Beijing, China	(Li *et al.*, 2002)
Streptomyces africanus	Cape Town, South Africa	(Meyers *et al.*, 2004)
Amycolatopsis taiwanensis	Yilan county, Taiwan	(Tseng *et al.*, 2006)
Amycolatopsis australiensis	Australia	(Tan *et al.*, 2006)

Table 1.2 Examples of known antibiotics produced by single species isolated from soil samples.

Antibiotic	Class	Species	Soil location for isolation	Reference
Streptomycin	Aminoglycoside	*Streptomyces griseus*	Russia	(Waksman *et al.*, 1948)
Aureomycin	Tetracycline	*Streptomyces aureofaciens*	Missouri, USA	(Duggar, 1948)
Erythromycin	Macrolide	*Saccharopolyspora erythraea*	Philippines	(Anon., 1952)
Vancomycin	Glycopeptide	*Amycolatopsis orientalis*	Orient	(Brigham & Pittenger, 1956)
Daptomycin	Lipopeptide	*Streptomyces roseosporus*		(Allen *et al.*, 1987)
Platensimycin	Diterpenoid - FASII inhibitor	*Streptomyces platensis*	South Africa	(Wang *et al.*, 2006)

A great effort has been made to develop new methods and techniques to access the biological diversity present in the environment. Improved media for isolation of particular group of bacteria (Stevenson *et al.*, 2004, George *et al.*, 2011) in addition to new cultivation techniques (Zengler *et al.*, 2002, Aoi *et al.*, 2009, Nichols *et al.*, 2010, Vester *et al.*, 2015) enabled the isolation of recalcitrant novel bacterial groups. The use of low-nutrient media, long incubation times and different microaerophilic conditions led to the isolation of previously uncultivated members of Acidobacteria

and Verrucomicrobia groups (Stevenson *et al.*, 2004). Development of systems for the encapsulation of single cells in gel microdroplets (Zengler *et al.*, 2002) or diffusion chambers (Nichols *et al.*, 2010) and of hollow-fibre membrane chambers to be inoculated with environmental samples (Aoi *et al.*, 2009) showed an increase in the uncultivated bacteria that were isolated and cultivated in these conditions. A recent success was the isolation of the β-proteobacterium *Eleftheria terrae* producing the novel non-ribosomal peptide teixobactin (Ling *et al.*, 2015) isolated using the isolation chip (iChip) system described by Nichols *et al.* (2010). In this system miniature diffusion chambers were loaded with liquid-agar based media inoculated with diluted soil sample and incubated *in situ* (in direct contact with the soil from where the sample was taken) leading to a recovery of single isolates per channel. Data showed that the system, in comparison to classical isolation in petri dishes, recovers a significantly higher diversity of bacteria with a possibility to cultivate representatives *in vitro* (Nichols *et al.*, 2010). This technique preserved the natural environmental conditions and promoted the isolation of uncultured bacteria. Other available strategies of isolation of novel strains with interesting phenotypes for antimicrobial discovery rely on resistance to particular classes of antibiotics. For instance Thaker *et al.* (2014) used a double selection in the classical Waksman's soil bacterial isolation approach; this included selection for a particular group of soil bacteria, the Actinomycetes, combined with phenotypic selection for antibiotic resistance, in particular glycopeptide resistance. The strains isolated were able to grow on glycopeptide containing plates and contained glycopeptide resistance genes and were producers of related glycopeptides. This screening/isolation technique allowed the identification of derivatives of glycopeptide structures that could have been used to improve activity compared to the existing products. In general, however, strain isolation techniques although successful in the past, still suffer from the limited knowledge of microbial communities in the soil and their physiology (Keller & Zengler, 2004).

1.4 Metagenomics for the capture of novel natural product gene clusters

A different approach is the use of metagenomics to overcome the problem of isolation and study more in depth the structure and function of bacteria within the microbial community (Handelsman, 2004, Allen & Banfield, 2005). Baker *et al.* (2006) pioneered the extraction of genes and partial genomes from environmental DNA (eDNA) to construct a metagenome of the biofilm community from acid mine drainage forming underground at the Richmound Mine at Iron Mountain, California. However, the very first studies on characterisation of uncultured communities using a metagenomic approach were carried out in marine communities where extraction of larger fragments of DNA was possible as illustrated by Stein *et al.* (1996). Community genomics in the oceans provided a paradigm shift in comprehension of uncultured bacterial groups, presenting an improved understanding and knowledge base of marine communities, their metabolisms, biochemical pathways and genomic variability (DeLong, 2005). This culture-independent approach is now a key component of the efforts to exploit bacterial communities in soil. Previous studies reported that only 1 % of the total bacterial community in soil has been accessed with culture techniques (Torsvik *et al.*, 1996), with a remaining 99 % still unexplored and available for further discovery of strains and molecules of interest. Metagenomics allows accessing the uncultured fraction of microbial communities without the need to isolate and cultivate microorganisms in the laboratory conditions. Since the introduction of the concept of metagenomics (Handelsman *et al.*, 1998, Rondon *et al.*, 2000), the technologies involved in this process have developed and improved. Next-Generation Sequencing (NGS) technologies, which allow exploring the DNA diversity present in the environments on a molecular level, are constantly evolving (van Dijk *et al.*, 2014). Since the advent of the first techniques, multiple platforms and systems have been introduced into the market at a rapid pace in order to answer the ever-increasing demand of deeper sequencing with lower error rates and costs (Metzker, 2010, Loman *et al.*, 2012). The 454 pyrosequencing approach was a pioneer in the field of NGS technologies, followed by Illumina (HiSeq and MiSeq), SOLiD, Ion Torrent, Pacific Biosciences and more recently Oxford Nanopore (Branton *et al.*, 2008, Glenn, 2011,

Luo *et al.*, 2012, Quail *et al.*, 2012, Wang *et al.*, 2014). The possibility to sequence amplicons of different lengths, or shotgun sequence DNA samples at different depths, provided new horizons in the exploration of microbial diversity at a molecular level. A number of environments have been used to provide metagenomes which were subsequently annotated and now available to interrogate community diversity at the functional level. An example of a large, deep and annotated metagenome, is the one created from the Global Ocean Sampling Expedition programmes (Rusch *et al.*, 2007). In this project various ocean water samples were collected, deep sequenced and analysed creating a 'reference' metagenome for further studies. A similar project was launched in 2009 for an international soil project termed the "TerraGenome" which aimed at combining the efforts of the global scientific community for the construction (sequencing and annotation) of a 'reference' soil metagenome using Park Grass soil samples, a well characterized soil from Rothamsted Research station (UK) (Vogel *et al.*, 2009). Metagenomic approaches aim to capture community diversity using libraries which are then available for future industrial applications by analyzing the functional genomic diversity. Large metagenomic libraries can constitute an invaluable source of unexplored and unexploited diversity not accessible with classical methods that could be of great interest for green, white and red biotechnology industries. In particular, metagenomic libraries from soil could provide a resource for the discovery of new chemical structures with antimicrobial activity against resistant pathogens such as methicillin resistant *S. aureus* (MRSA) (Kallifidas *et al.*, 2012). The terragines were the first example of novel natural products isolated from soil DNA recombined in a *Streptomyces* heterologous host (Wang *et al.*, 2000). Many further compounds were then isolated and characterized from soil metagenomic library clones, some examples are reported in Table 1.3.

Table 1.3 Examples of antibiotics isolated from metagenomic libraries from soil and expressed in heterologous hosts.

Antibiotic	Class	Soil location	Reference
Utahmycin A and B	2-azaanthraquinone	Utah	(Bauer *et al.*, 2010)*
Fluostatins F, G and H		Anzo Borrego desert, California	(Feng *et al.*, 2010)*
Tetarimycin A	Tetracycline	Arizona desert	(Kallifidas *et al.*, 2012)*
Fasamycin A and B	FASII inhibitor	Arizona desert	(Feng *et al.*, 2012)*
Arixanthomycins A−C	pentangular polyphenol	Arizona desert	(Kang & Brady, 2014)*

**Streptomyces albus* was the heterologous host.

Metagenomic approaches are thus a potent tool to investigate and exploit the microbial diversity in soil for novel natural product gene clusters. However further improvements are needed for both metagenome analysis and functional metagenomics. Shotgun and amplicon metagenomic sequencing are still computational demanding and require constant bioinformatic pipelines development, while functional metagenomics presents a number of bottlenecks in the process of exploiting metagenomic libraries. In the case of antibiotic gene clusters, metagenomic libraries present bias in the extraction efficiency, difficulties in DNA extraction to provide large fragments (essential for antibiotic gene clusters) and heterologous expression of the captured DNA.

1.4.1 Soil sample selection for metagenomic applications in drug discovery

A wide variety of strategies have been used to select soil types for sampling, an initial understanding of the community structure is a useful precursor for site selection. Examples in the past showed that exploring extreme environments allowed the isolation of enzymes with improved characteristics for industrial application with detergents used at low temperature (see review: Struvay & Feller (2012)). In the drug discovery field, soil and more recently marine environments, have been identified as rich sources of diversity (see review: Gerwick & Moore (2012)). Soil is a physically

complex environment and requires sampling strategies enabling characterisation of the vertical and horizontal structure at the macrosite and microsite levels. Differences in the microbial community can depend on the physical factors such as sample depth, due to the presence of different biotic and abiotic stress factors (Fierer *et al.*, 2003).

Soil contains a great variety of habitats with many physicochemical gradients, for example plant roots release exudates providing a nutritional gradient for bacteria to exploit. These gradients will depend on the type of plant and to some extent to the soil structure and climatic effects (Marschner *et al.*, 2001, Fierer *et al.*, 2003). Therefore this strong interaction between plants roots and their microbiomes results in a different community than that present in the surrounding bulk soil (Berg & Smalla, 2009, Berendsen *et al.*, 2012, Bulgarelli *et al.*, 2012).

Sampling strategies are rarely discussed in detail and have received little attention in terms of exploring microbial diversity in soil. Each soil particle might present complex and diverse niches for different microbial adaptation and colonisation strategies (see review: Vos *et al.* (2013)), therefore it is a considerable challenge to devise an appropriate sampling strategy. Considering the heterogeneity, mixing original samples to reduce the variability making it more homogeneous for subsampling could be a valuable option. The sample size constitutes another important parameter in DNA extraction as it might affect microbial group prevalence due to biases in community structure elucidation (Kang & Mills, 2006). Kang & Mills (2006) performed a 16S rRNA gene diversity comparative analysis on DNA extracted from replicate soil samples of sizes from 0.1 g up to 10 g to determine the optimal size for sampling the bacterial population diversity in 0.25 g of soil. However, the authors recommend a size specific study of comparable analysis to obtain the optimal sample size for each study of interest.

1.4.2 DNA extraction for functional metagenomics

DNA extraction is one of the key steps in metagenomics as the diversity and coverage of the metagenome will depend on the efficiency of DNA extraction.

There are multiple DNA extraction protocols available in the literature to treat soil samples that can be mainly classified into direct or indirect methods (Ogram *et al.*, 1988, Frostegard *et al.*, 1999, Krsek & Wellington, 1999). The direct method requires

lysis in situ within the soil during the extraction process, whereas the indirect method involves recovery of cells and their lysis as a separate step (Holben *et al.*, 1988, Lindahl & Bakken, 1995).

The indirect method involves the collection of cellular biomass by separation from soil using various centrifugation steps and a density separation protocol involving the use of the Nycodenz polymer which results in the recovery of a layer of cells at the interface between the polymer and the upper aqueous layer (Bakken & Lindahl, 1995, Berry *et al.*, 2003). Other methods have involved the dispersal of soil using Chelex® 100 prior to centrifugation (Herron & Wellington, 1990). The recovered biomass is then lysed giving clean DNA of a greater size rage than that recovered with direct lysis (Berry *et al.*, 2003). Enzymatic lysis of the recovered biomass in gel plugs was also presented by Liles *et al.* (2008) showing the recovery of high molecular weight (HMW) DNA of 1 Mb and the formation of a metagenomic library with long inserts.

The direct methods instead enable a lysis in situ in the total soil sample with the addition of a lytic mixture which could be based on lytic enzymes or chemical detergents (Krsek & Wellington, 1999). This allows the recovery of HMW DNA, but the purity might be affected by the co-extraction of inhibitors such as humic acids. Depending on the soil type humic compounds can be co-purified with DNA in the extraction process and they can constitute a problem in the following steps of PCR amplification, sequencing and cloning (Tebbe & Vahjen, 1993). Prevention of inhibitor build up can be achieved by pre-treatment of the sample with the addition of $AlNH_4(SO_4)_2$ to remove the inhibitors through chemical flocculation (Braid *et al.*, 2003) or the addition of 1-hexadecyltrimethylammonium bromide (CTAB) and polyvinylpolypyrrolidone (PVPP) to precipitate humic compounds. However, PVPP has been shown to complex with DNA giving a higher loss of recovered DNA then using CTAB (Zhou *et al.*, 1996). Texture and composition of the soil could play a crucial role in the performance of the lysis methods; the content of clay for example can decrease the efficiency of DNA purification with co-extraction of inhibitors (Braid *et al.*, 2003).

Both direct and indirect methods produce bias, but direct methods have been favoured for metagenomics possibly because the efficiency of cell extraction from soil with indirect ones is difficult to evaluate. The average size of DNA is a key parameter for metagenomics approaches. For PCR amplicon-based diversity analysis and shotgun

metagenome sequencing, high yield of good quality DNA with a fragment size in the distribution 10 to 30 kb is usually preferable. Therefore, direct or indirect methods that allow the extraction of inhibitor-free DNA are preferred, such as the commercially available kits FastDNA™ Spin kit for soil (MP Biomedical), or PowerSoil® DNA Isolation Kit (MO BIO). HMW DNA is required for metagenomic library preparation with a size range depending on the vector used to clone the DNA in relation to the size of the genes of interest (from 5 kb for single enzymes up to 200 kb for natural product gene clusters). Gentle methods of extraction are fundamental to obtain high quality DNA to create libraries (e.g. Brady (2007)).

Depending on the efficacy of lysis method all bacterial groups may not be represented equally with DNA from some groups recovered more easily (Bürgmann *et al.*, 2001). For example, many of the Actinobacteria previously favoured for high diversity of natural product gene clusters, produce mycelia and spores, which are often poorly represented in metagenomic libraries because gentle lysis procedures do not lyse spores efficiently and indirect methods do not recover mycelia (Morris *et al.*, 1999).

For metagenomic approaches a compromise between extraction efficiency, DNA size and DNA purity is essential for a good representation of the microbial diversity present in soil and its exploitation.

1.4.3 Bioinformatic tools for sequence analysis and natural product gene diversity study

High-throughput sequencing technologies have developed rapidly since their introduction to the market, allowing the study of microbial communities and ecological relationships amongst microorganisms in more depth. International projects have been established to examine specific habitats such as the TerraGenome project with the focus on soil (Vogel et al., 2009) and the Global Ocean Sampling Expedition (Rusch *et al.*, 2007, Vogel *et al.*, 2009) to sequence and collect metagenomes as references for future study. Amplicon sequencing has been used to study environments for specific genes of interest such as bacterial 16S rRNA gene for phylogenetic diversity (Logares *et al.*, 2013), secondary metabolites biodiversity potential (Charlop-Powers *et al.*, 2014) or antibiotic resistome phylogeny and distribution (Allen *et al.*, 2008, Wright & Poinar, 2012).

Different bioinformatics tools were developed at the same time to enable deep processing of large volumes of sequences generated from NGS platforms. MetaGenomics Rapid Annotation using Subsystems Technology (MG-RAST) (Meyer *et al.*, 2008), European Bioinformatics Institute (EBI) metagenomics portal (Hunter *et al.*, 2014) and Integrated Microbial Genomes/Metagenomic analysis (IMG/M) platform (Markowitz *et al.*, 2014) are some of the main platforms freely available for the analysis of shotgun metagenomes and in the case of MG-RAST and EBI metagenomics also for amplicon sequence analysis. Quantitative insights into microbial ecology (QIIME) software-pipeline (Caporaso *et al.*, 2010) was devised to provide a versatile workflow from raw sequences to data analysis and interpretation. Different algorithms such as USEARCH (Edgar, 2010), PANDAseq (Masella *et al.*, 2012) and Trimmomatic (Bolger *et al.*, 2014) to filter and align the row amplicon reads have been developed during the years and they can be easily integrated in the QIIME pipeline.

Common databases used for phylogenetic analysis and taxonomy assignment are, for example, Greengenes (DeSantis *et al.*, 2006) and the Ribosomal Database Project (RDP) (Cole *et al.*, 2014). However improvements in the analysis of data are continuously being made and reflect the extensive application of methods for classification of 16S rRNA gene sequences (Vinje *et al.*, 2015).

For functional annotation, a variety of databases are available according to the gene of interest. National center for biotechnology information (NCBI) and InterPro (Finn *et al.*, 2017) are two of the most comprehensive ones for metagenomes annotation. Algorithms to identify key signatures of functional genes and analyze metagenomic data have also been developed (Jones *et al.*, 2014).

Specific analysis platforms and databases have been developed for identification and characterization of key functional groups such as the antibiotic gene clusters. Some of the main software to identify secondary metabolite gene clusters include Antibiotics & Secondary Metabolite Analysis Shell (AntiSMASH) (Medema *et al.*, 2011), Natural Product Domain Seeker (NaPDoS) (Ziemert *et al.*, 2012), Prediction Informatics for Secondary Metabolomes (PRISM) (Skinnider *et al.*, 2015) and Environmental Surveyor of Natural Product Diversity (eSNaPD) (Reddy *et al.*, 2014). Constant updates are included in these pipelines and new functions are added to identify orphan gene clusters (Dejong *et al.*, 2016) or potentially new antibiotic resistance genes included in antibiotic gene clusters (Alanjary *et al.*, 2017). In addition to each database

related to the pipelines described above, two main databases for secondary metabolites are also available. The IMG Atlas of Biosynthetic gene Clusters (IMG/ABC) provides a collection of experimentally characterized and computational predicted secondary metabolite gene clusters available from single genomes or metagenomes present in the IMG database (Hadjithomas *et al.*, 2015). The Minimum Information about a Biosynthetic Gene cluster (MIBiG) not only provides a collection of annotated experimentally characterized biosynthetic gene clusters (BGCs) but also provides guidance of minimum information standards required to describe a biosynthetic cluster (Medema *et al.*, 2015).

1.4.4 Construction and screening of functional metagenomic libraries for novel natural product gene clusters

To explore environmental biodiversity one of the key tools is metagenomic libraries. The creation of these allows analysis and capture of genes for industrial or pharmaceutical use. Different type of libraries can be created in a variety of vectors depending on the genes of interest: plasmids libraries allow inserts of up to 10 kb so HMW DNA is not required and they are usually used to capture single genes such as those coding for enzymes with industrial application or related to antimicrobial resistance (McGarvey *et al.*, 2012). Cluster genes such as those related to natural product biosynthesis often require larger inserts which can be captured with cosmid or fosmid vectors (up to 40 kb) or bacterial artificial chromosome (BAC) vectors (up to 300 kb).

Cosmids contain *Escherichia coli* bacteriophage lambda *cos* sites which allow the recognition of the vector from the lambda bacteriophage head particles during *in vitro* reactions. This vector has been used for classical cloning with great success (Collins & Hohn, 1978) and has been more recently applied in single strain genomic and metagenomic library preparation (Brady, 2007). Lambda bacteriophage headful packaging can incorporate approximately 50 kb, therefore metagenomic libraries with insert size of up to 40-45 kb per clone can be created. The cosmid pWEB (Epicentre) for example was recently used to prepare metagenomic libraries for novel antimicrobial activity screening (Brady, 2007, Owen *et al.*, 2013).

Fosmid vectors were developed by engineering a cosmid vector in order to overcome instability problems caused by rearrangements of cosmids (Kim *et al.*, 1992). Fosmids contain the bacterial F-plasmid replicon system that maintains a single copy of the vector per cell, therefore they are more stable. Optimization of the original pFOS1 fosmid vector (Kim *et al.*, 1992) was performed in subsequent years and engineered fosmids are currently available for the creation of metagenomic libraries.

BACs are vectors that can be used to insert up to 300 kb of DNA and were first engineered in the early 1990s (Shizuya *et al.*, 1992). These vectors can be transferred through transformation of *E. coli* but there is a lower efficacy which causes a bottleneck in the preparation of large libraries. Nevertheless, the BAC vectors can insert up to 300-350 kb (even though the average size is ~140kb) (Kim *et al.*, 1996) and offer an alternative for the capture of larger gene clusters beyond 50 kb. Other possible vectors similar to BACs are the phage P1-derived vectors such as P1-derivate Artificial Chromosome (PAC) which also enable insertions up to 300 kb with an average insert size of 130-150 kb (Ioannou *et al.*, 1994). Engineered BAC vectors such as the *E. coli – Streptomyces* Artificial Chromosome (ESAC) vectors (e.g pPAC-S1) have been developed and used for the creation of genomic libraries that could be easily replicated and manipulated in *E. coli* and stably maintained in a second host (e.g. *Streptomyces coelicolor*) for the expression of the inserted DNA (Sosio *et al.*, 2000). These types of BACs have proved useful for the manipulation of antibiotic gene clusters and had been used to create metagenomic libraries for this purpose (Berry *et al.*, 2003). These vectors can autonomously replicate in *E. coli* and integrate stably into *Streptomyces* chromosomes in a specific site of recombination recognising phage *ΦC31 attB* sites. This ability makes these vectors useful for library construction as the positive clones containing the genes of interest can be identified in *E. coli* by PCR screening and then transferred using intergeneric conjugation to different *Streptomyces* spp. for heterologous expression and characterization.

The choice of the vector used for the creation of the metagenomic library can influence the number of clones recovered according to the efficiency of ligation and transformation of the vector chosen. Plasmids, fosmids and cosmids usually have a better efficiency than BAC vectors in terms of number of clones recovered but because the insert is on average considerably smaller, the total amount of DNA captured could be potentially equivalent. Although large libraries can be created using fosmid/cosmid

vectors, the size of the library remains a fundamental issue as it represents the depth of analysis and success in recovery targeted genes or gene clusters of interest. To provide a prospective of the enormous diversity present within the soil, Daniel (2005) estimated that approximately 10^{11} BAC clones would be needed to achieve complete coverage of the bacterial genomes present within a gram of soil. However, metagenomic libraries of this size would present issues related to storage, handling and screening. Although innovative systems relying on semi-liquid media, where clones grow individually even though they are pooled, were introduced with great success (Elsaesser & Paysan, 2004, Hrvatin & Piel, 2007) further improvements would be needed for larger metagenomic libraries.

The size of metagenomic libraries dictates the screening approach and high-throughput methods are required for the majority of targets. Screening methods can be divided into two main categories: target directed screening or whole cell screening. In the first method, the inhibition of specific targets (usually enzymes) is investigated, while in the second case, growth inhibitory biological activity (i.e. unspecific target) is detected. A combination of the two methods (target-based and whole-cell screening) can also results in the identification of novel bioactivity. An example of this resulted in the identification of platensimycin, a diterpenoid produced by *Streptomyces platensis* that inhibits fatty acid biosynthesis (Wang *et al.*, 2006). The target of the inhibitory activity FabF (fatty acid synthase type II) was confirmed by reducing the level of this enzyme in a *S. aureus* strain using antisense DNA thus providing a supersensitive strain for screening purposes using growth as an indicator (Young *et al.*, 2006). The exposure of the sensitive *S. aureus* strain to platensimycin showed high level of growth inhibition confirming the target of the compound (Wang *et al.*, 2006). Phenotypic and activity screening for drug discovery usually rely on the modification of a target that can be detected as inhibition of sensitive bacterial cells, for example growth inhibition or a change in the morphology of the expressing clone. Heterologous hosts used for expression of metagenomic library DNA provide an ideal background for detection of novel antibiotics as any new bioactive product could be detected by de novo activity against whole cell bioassay of a known pathogen (Brady, 2007). These types of screening usually involve considerable efforts for processing large number of clones, so automated assays have been developed for high-throughput screening programmes (Moy *et al.*, 2009, Nybond *et al.*, 2013, Rajamuthiah *et al.*, 2014).

Another available approach based on molecular similarity is PCR screening for genes of interest, often performed in tandem with phenotypic and activity screening methods, where host expression system incompatibility occurs as evidenced by Gabor *et al.* (2004). For instance, *E. coli*, which is the usual host for genomic and metagenomic libraries, is not an optimal heterologous host as it is able to express only 40 % of the total eDNA captured in the library due to the absence of primary substrates or differences in the translation system (Gabor *et al.*, 2004). Therefore, a PCR screening approach is usually performed for the identification of natural product gene clusters. There are a few examples of degenerate primers that can be used for metagenomic library screening to discover natural product gene clusters. These include primers targeting genes such as polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS) (Ayuso-Sacido & Genilloud, 2005, Schirmer *et al.*, 2005, Laskaris, 2009, Parsley *et al.*, 2011, Owen *et al.*, 2013) and conserved enzymes characteristic of specific groups of natural products (Banik & Brady, 2008). In particular, PKS and NRPS are involved in the biosynthesis of a large proportion of known secondary metabolites. Their genes are responsible for encoding multi-complex enzymes that are organized in modules with conserved domains such as the adenylation (A) and ketosynthase (KS) domains in NRPS and PKS respectively (Challis *et al.*, 2000, Donadio *et al.*, 2007, Strieker *et al.*, 2010), that can be targeted with degenerate primers. Conserved enzyme coding genes present in particular classes of antibiotics of interest such as the OxyC-like enzyme, strictly involved in a specific C-C bond formation in the aglycone of vancomycin- and teicoplanin-like glycopeptides (Banik & Brady, 2008) are other potential targets for degenerate primers design. The presence of conserved domains or genes that can be targeted with degenerate primers is the basis of a PCR screening assay for metagenomic libraries aiming to identify gene clusters that could potentially lead to novel bioactive structures.

Different heterologous hosts have been used for expression, recovery and characterization of enzymes with a commercial use in industry, and examples include cellulases from *Zymomonas mobilis* for ethanol production (Linger *et al.*, 2010) and gene clusters coding for antimicrobial metabolites isolated from a single strain (Zhang *et al.*, 2010). Few heterologous super-hosts have been engineered over the last decades especially for secondary metabolite expression. One of the most widely used heterologous hosts is *S. coelicolor* with all its derivate strains such as M1152 and

M1154 (Gomez-Escribano & Bibb, 2011). The main characteristic of these strains is that being a secondary metabolites producer, *S. coelicolor* has all the necessary machinery for secondary metabolites production. This host was further modified by removing known antibiotic gene clusters from the chromosome in order to obtain a lower secondary metabolite background production and to increase the availability of substrates for production of metabolites from introduced gene clusters. Other valuable heterologous hosts belonging to the same genus are *Streptomyces lividans* and *Streptomyces albus* J1074 (Gullon *et al.*, 2006, Lombo *et al.*, 2006). These heterologous hosts have been employed to express silent gene clusters and to move gene clusters from genetically intractable producers into a more readily controlled host background (Baltz, 2010). These engineered *Streptomyces* hosts are usually the first choice when expressing antibiotic clusters. However the expression of eDNA (clusters or single genes) will demand a much wider range of hosts due to the high diversity of soil bacteria with very different physiologies. Craig *et al.* (2010) tested seven different heterologous hosts belonging to the Proteobacteria phylum (*Agrobacterium tumefaciens, Burkholderia graminis*, *Caulobacter vibrioides*, *E. coli*, *Pseudomonas putida* and *Ralstonia metallidurans*) to screen a metagenomic library from soil for novel antimicrobial compounds. The results showed that the use of different hosts from the main bacterial phyla populating the soil microbial communities increased the detection of novel phenotypes presenting antimicrobial activity.

Hypothesis and aims of the project

The main hypothesis of this projects is that soils in extreme environments harbour unusual microbiota which can be more easily accessed through culture-independent techniques such as functional metagenomics. The main aim was to study and capture antibiotic gene clusters from soil metagenomes using different metagenomic approaches. Actinobacteria have always been the main known producers of secondary metabolites with antimicrobial activity, however, genome mining approaches have shown that other phyla more recalcitrant to isolation and laboratory conditions might have the genetic potential for natural products biosynthesis.

The aims of this project were:

- To investigate and optimise eDNA extraction from soil to obtain HMW DNA with the goal of capturing antibiotic gene clusters in functional metagenomic libraries from the identified hotspots.
- To investigate soil amendments for a higher recovery of DNA and therefore increasing the chances of capturing natural product gene clusters.
- To investigate phylogenetic drivers for NRPS and PKS secondary metabolite diversity and identify potential hotspots for novel natural products.
- To investigate some of the bottlenecks of metagenomic library screening and engineer novel tools for library construction, screening and heterologous expression.

Chapter 2

Extraction of HMW DNA from soil for metagenomic library applications

2.1 Introduction and aims

DNA extraction is one of the key steps for microbial community studies and metagenomic applications. Different types of extraction can be used to obtained the required grade of purity and size of the DNA from soil. A great number of protocols have been developed over the years for the extraction of eDNA from different environments such as soil (Ogram *et al.*, 1988, Selenska & Klingmuller, 1991, Picard *et al.*, 1992, Frostegard *et al.*, 1999, Krsek & Wellington, 1999). Extraction of nucleic acids can be mainly classified in two categories: direct and indirect methods (Figure 2.1). Direct methods apply an *in situ* lysis of cells and purification of DNA directly in soil, while indirect ones separate the cellular biomass first and extract the DNA from this pellet in a separate step (Holben *et al.*, 1988, Steffan *et al.*, 1988, Bakken & Lindahl, 1995, Lindahl & Bakken, 1995). In general, the recovery of the cellular biomass has been reported to reduce the yield of DNA extracted (Steffan *et al.*, 1988) however, there are reduced molecular inhibitors and the DNA is therefore purer (Tebbe & Vahjen, 1993). Moreover, with indirect methods, the microbial diversity extracted is different from the diversity observed from direct methods of extraction (Courtois *et al.*, 2001). On the contrary, direct methods extract higher quantity of DNA using less laborious steps, but more humic acids and other molecular compounds are also co-extracted more easily. It has been shown that these compounds can constitute a problem for molecular steps such as transformation, PCR amplification and restriction enzyme digestion of the eDNA extracted (Tebbe & Vahjen, 1993, Schrader *et al.*, 2012).

Different lysis approaches have been tested and compared in multiple studies with each method having advantages and disadvantages (Zhou *et al.*, 1996, Krsek & Wellington, 1999). For instance, in direct methods, physical lysis showed high lysis

efficiency on spore forming bacteria which are usually difficult to lyse (Moré *et al.*, 1994); chemical lysis method extracted a low amount of DNA with humic acids contaminant but with high molecular weight (Krsek & Wellington, 1999). The presence of humic acid chelating agents such as hexadecylmethylammonium bromide (CTAB) or polyvinylpolypyrrolidone (PVPP) in chemical lysis treatment has been proved to improve the purity of the DNA recovered reducing the quantity of humic acids co-extracted. In particular, between both chemicals, CTAB showed a lower loss of DNA than PVPP on the final yield recovered (Zhou *et al.*, 1996).

The aims of the current study were firstly to investigate and optimise different methods of DNA extraction from soil in order to recover HMW eDNA to create metagenomic libraries (Figure 2.1) and secondly to optimise a quantitation assay for NRPS and PKS using quantitative PCR (qPCR).

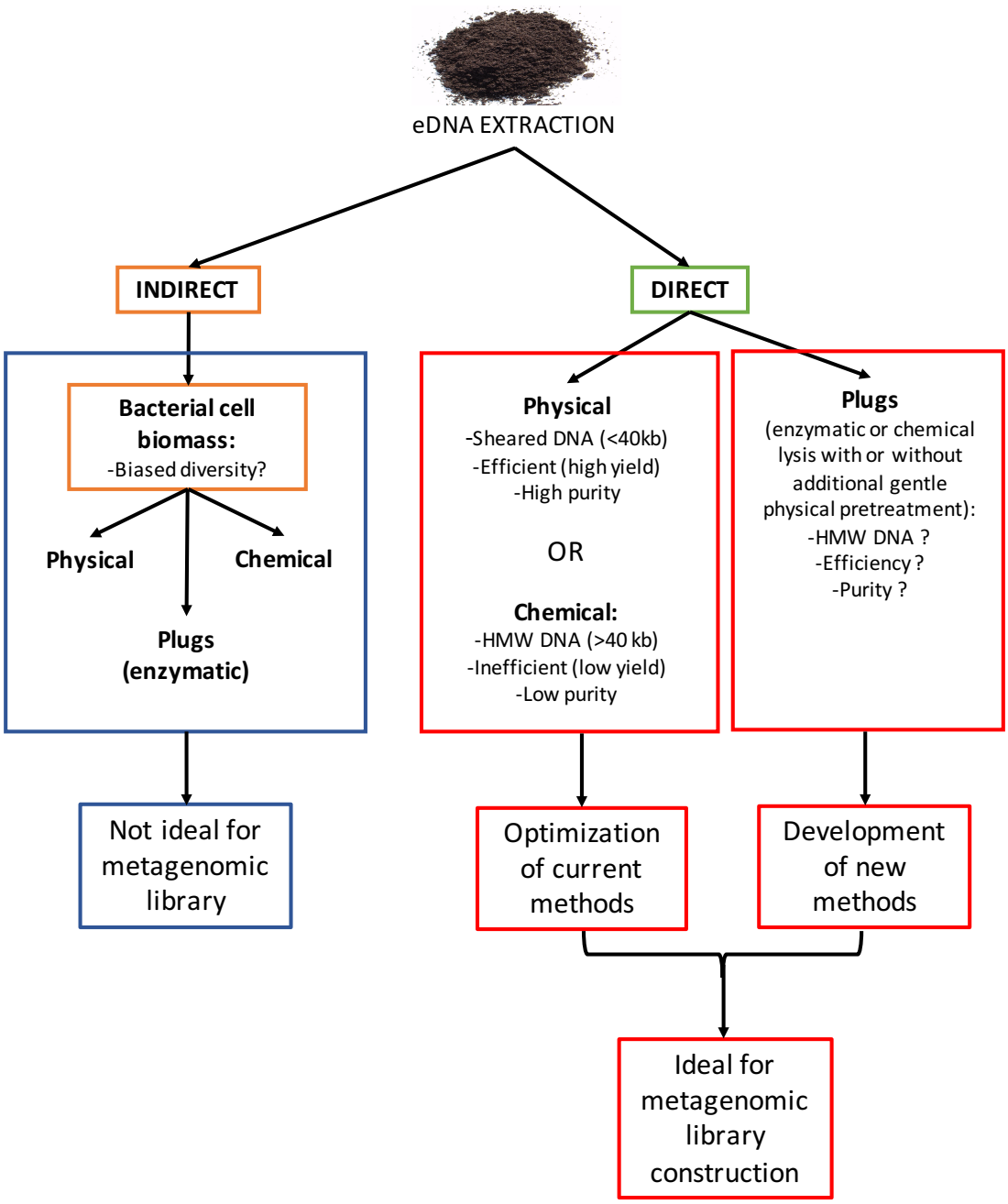


Figure 2.1 Main methods of eDNA extraction from soil investigated in the current study. Direct and indirect methods of lysis are reported in orange and green boxes. The blue box represents methods which are not ideal for metagenomic library construction, while red boxes indicates indirect methods which were optimized in the current study for future metagenomic library creation.

Therefore in this study, the achievement of HMW DNA was an essential requirement for the creation of metagenomic library from soil with large DNA inserts. Indeed, antibiotic biosynthetic gene clusters usually do not only contain biosynthetic genes but also resistance and regulatory genes, reaching in size to over 60 kb up to 100 kb

(Ikeda *et al.*, 1999, Du *et al.*, 2000, Bentley *et al.*, 2002, Sosio *et al.*, 2004, Gomez-Escribano *et al.*, 2012). Therefore, in order to capture whole gene clusters HMW DNA is required to include large insert of DNA in vectors of the library.

Moreover, a rapid assay based on qPCR to determine the potential biosynthetic richness of a specific soil for both NRPS and PKS genes, which are two of the main biosynthetic system for antimicrobial compounds, would help to discriminate between soils richer in one functional gene type or the other and it would be of great interest to target the exploitation of specific environments.

2.2 Material and methods

2.2.1 Soil samples

Two soil samples were selected to test 15 types of DNA extraction methods. The samples included in the study were collected from Warwick (Cryfield, Coventry) and Reading and stored at -20 °C until usage (Table 3.1).

Table 2.1 Summary of characteristics of soils used in this study. TOC = Total Organic Carbon (%); TN = Total Nitrogen (%); EC = Electrical conductivity (µS/cm).

Soil sample	TOC (%)	TN (%)	pH	EC (µS/cm)	Sand (%)	Silt (%)	Clay (%)	Soil texture
Reading	1.1	0.107	6.8	1400	72	19	9	Sandy Loam
Warwick (Cryfield)	1.4	0.146	6.9	2050	34	49	17	Sandy Silt Loam

2.2.2 DNA extraction methods

DNA was extracted in triplicate for each method from both soils. Extraction protocols were retrieved from the literature, modified and optimised for the purpose. Different types of lysis were tested on samples which were subsampled in different quantities (100g, 10g and 0.5g) according to the protocol tested (Table 2.2). These methods included different treatments such as chemical lysis, physical lysis, physical-chemical lysis and/or pre-treatment of samples with detergents or chelating agents before the cell lysis step.

Table 2.2 Summary of main characteristics of DNA extractions methods tested on Warwick and Reading soils.

Method	Method ID	Type of lysis	Soil used per each replica (g)	Elution volume (mL)	References
Brady100	B100	Chemical	100	5	(Brady, 2007)
Brady	A1	Chemical	10	5	Modified (Brady, 2007)
Brady (2X)	A2	Chemical	10	5	This study
Blending + Brady	B	Physical and Chemical	10	5	This study
TPNE	C	Chemical	10	5	Modified (Krsek & Wellington, 1999)
Phosphate buffer	D	Chemical	10	5	Modified (Selenska & Klingmuller, 1991)
Freeze-boil	E	Physical	10	5	Modified (Moré *et al.*, 1994)
FastDNA™ Spin Kit for soil	F	Physical and Chemical	0.5	0.1	MPBio
Chelex® + Brady	G	Chemical	10	5	Modified (Herron & Wellington, 1990, Turpin *et al.*, 1993)
Sodium deoxycholate +Brady	H	Chemical	10	5	
PEG6000 + Brady	I	Chemical	10	5	
Plugs (enzymatic lysis)	J	Enzymatic	1	0.1	Modified (Liles *et al.*, 2008)
Plugs (Brady)	K	Chemical	1	0.1	This study
Plugs (blending + Brady)	L	Physical and Chemical	1	0.1	This study
Plugs (phosphate buffer)	M	Chemical	1	0.1	This study

In Method B100 (Table 2.2) a gentle chemical lysis was performed on 100g of soil per replicate as described by Brady (2007). The pre-warmed lysis buffer (100 mM Tris-HCl, 100 mM Na EDTA, 1.5 M NaCl, 1% (w/v) cetyl trimethyl ammonium bromide (CTAB), 2% (w/v) SDS, pH 8.0) was added to the soil and incubated at 70 °C for 2 h with gentle resuspension every 30 min. After the lysis step was completed, the soil particles were separated from the supernatant by two steps of centrifugation at 5500 rpm (>3600 x g) for 10-20 min using a Beckman centrifuge (rotor JA25.50). Once the soil particles were removed, the DNA was precipitated using 100% isopropanol and washed with 70 % ethanol. The DNA pellet recovered was then resuspended in 5 mL of TE buffer (10 mM Tris, 1 mM EDTA, pH 8) and stored at 4 °C.
The scaled-down version of Method B100 was optimized for future work with soil samples available in limited quantities. Method A1 started from a smaller amount of soil (10 g instead of 100 g) and kept the same protocol as Method B100. Method A2 lysed a second time the soil pellet obtained from Method A1 after the lysis step; fresh lysis buffer was added to the extracted soil pellet and incubated at 70 °C for an additional 2 h. The rest of the protocols was the same as Method A1.

Physical pre-treatment to Method A1 were also tested. In Method B, after the addition of the Brady lysis buffer, a gentle blending step was performed for 1 min using the handheld rotor-stator homogenizer TissueRuptor® (QIAGEN) at the lowest speed setting. Before use, the probe was washed twice with sterile deionized water and sterilized in 70 % ethanol, then washed again in sterile deionized water to avoid ethanol contaminations.

In addition to the optimization of the Brady method, different lysis buffers were also tested in direct lysis methods: in Method C 50 mL of TNPE buffer (50 mM Tris-HCl, 100 mM NaCl, 1% (w/v) polyvinylpolypyrolidone (PVPP), 10 mM EDTA pH 8.0) (Krsek & Wellington, 1999) and in Method D 30 mL of 0.1 M sodium phosphate buffer pH 8 (Selenska & Klingmuller, 1991) were added to 10 g of soil and incubated for 2 h at 70 °C with regular resuspension of the samples every 20 min. All the centrifugation and precipitation steps were kept the same as in the Brady method (A1) in order not to introduce many variables when comparing the lysis steps.

Two physical lysis methods were also tested. The FastDNA™ Spin kit for soil (MPBio) (Method F, Table 2.2, page 28) combines a homogenization and lysis of the sample using different sizes of glass beads and recovery of DNA through a binding matrix. In the freezing-boiling method (Method E, Table 2.2, page 28), samples were snap frozen using super-cooled ethanol for 10 min then moved to a 65 °C water bath and incubated for 10 min. This process was repeated 5 times. After the last freezing-boiling step, samples were gently resuspended in 15 mL of room temperature Brady lysis buffer in order to recover the DNA extracted following the DNA precipitation protocol of Method A1. In this case, the Brady lysis buffer was used only to collect DNA extracted through the physical lysis, therefore no incubation time was performed. The use of this buffer did not introduce additional variables to the DNA recovery process.

In indirect Methods G, H and I (Table 2.2, page 28), cells were recovered by first adding 20 mL of Chelex® 100 mix (10% (w/v) Chelex® 100, 2.5 % (w/v) polyethylene glycol, 0.1 % (w/v) sodium deoxycholate), or polyethylene glycol 6000 2.5 % (w/v) or sodium deoxycholate 0.1 % (w/v) to the soil. Samples were incubated at 4 °C for 2 h with regular resuspension every 20 min, then centrifuge for 30 s at 1000 x g and supernatant was decanted through a sterile gauze. The pellet was then resuspended in 20 mL of fresh solution (Chelex® 100 mix, PEG6000 or sodium deoxycholate) and incubated at 4 °C with regular resuspension for 1.5 h. The samples were then centrifuged and supernatant was filtered through a sterile gauze and collected together with the first supernatant. The Brady lysis buffer (15 mL) was then added to the treated supernatant and the same lysis and DNA precipitation steps were followed as in Method A1.

Different ways to preserve the integrity of HMW DNA during extraction were developed. In Methods J, K, L and M (Table 2.2, page 28), soil was included in agarose plugs and cells were lysed *in situ*. Plugs were created using 1 g of soil mixed with 0.5 mL of TE buffer pH 8 or disruption buffer (0.2M NaCl, 50 mM Tris-HCl, pH 8) (Berry *et al.*, 2003), vortexed for 30 s and then mixed with 0.5 mL of 1.5 % low-melting point agarose dissolved in TE25SUC (25 mM Tris-HCl pH8, 25 mM EDTA pH 8, 0.3 M Sucrose). From each soil-agarose mix, approximately 20 plugs were created using the

PFGE plug molds with 80-100 µL of the mixture per well. Plugs were left to cool and solidify at 4 °C for 1 h. For each treatment tested 10-20 plugs in triplicate were prepared. The first method tested on plugs was the enzymatic digestions (Method J, Table 2.2, page 29). Plugs prepared using the disruption buffer were collected and incubated at 37 °C for 2 h into 2 mL of lysis buffer (1 % sarcosyl, 1 % sodium deoxycholate, 1 mg/mL lysozyme, 10 mM Tris-HCl pH 8, 0.2 mM EDTA pH 8 and 50 mM NaCl) (Liles *et al.*, 2008). After the lysis was completed, plugs were transferred into 4 mL of ESP buffer (1 % sarcosyl, 1 mg/mL proteinase K, 0.5 M EDTA pH8) and incubated at 55 °C for 16 h (Liles *et al.*, 2008). At the end of the incubation, 5 µL of 100 µM phenylmethylsulfonyl fluoride (PMSF) in DMSO were added in each tube and incubated at room temperature for 1h. Plugs were then washed three times with TE25 (25 mM Tris-HCl pH8, 25 mM EDTA pH 8) and stored at 4 °C in 2 mL of TE25 buffer until DNA was recovered.

For all the other treatments (Methods K, L and M, Table 2.2, page 28), plugs were prepared using TE buffer, then incubated for 2 h at 37 °C in 2 mL of pre-heated Brady lysis buffer (at 70 °C) or sodium phosphate buffer. The blending pretreatment in Method L was performed before the formation of the plugs after the addition of TE buffer to the soil samples. Resuspended soils were blended with the TissueRuptor® (QIAGEN) for 45 s in ice and then mixed with low-melting point agarose to create plugs. After the lysis incubation plugs were collected and DNA was recovered using Phase Lock Gel (PLG) Light (Eppendorf) tubes. The PLG tubes manufacturer's protocol was modified to enable the recovery of DNA from soil plugs. PLG light tubes were centrifuged at 1500 x g for 2 min immediately prior to use, plugs were then transferred into the tubes with 0.5 mL of TE buffer (10 mM Tris-HCl, 1 mM EDTA pH 8) and melted at 65 °C for 5-10 min until completely dissolved. Tris-saturated buffer phenol pH 8 (2 mL) was added to each sample and gently mixed without vortex. Phases were separated through centrifugation at 1500 x g for 5 min. The aqueous phase of each sample was recovered and transfer to a new PLG tube with 1 mL of room temperature phenol:chlorophorm:isoamyl alcohol (PCI 25:24:1). After 5 min at 1500 x g, aqueous layers were recovered and transferred to normal 2 mL tubes and gently mixed with 0.5 mL of chlorophorm:isoamyl alcohol (CI 24:1). Phases were separated again through centrifugation at 1500 x g for 5 min at 4 °C and the recovered aqueous layers were mixed with 0.5 mL ice-cold isopropanol in 2 mL tubes. Samples were

incubated at room temperature for 5 min, then DNA was pelleted for 5 min at 1500 x g and washed with 0.5 mL of 70 % ice-cold ethanol. The DNA pellet was then recovered by centrifugation and air-dried at room temperature for 5 min and at 37 °C for 3 min. DNA was resuspended in 100 µL of TE buffer pH 8.

All eDNA samples concentration were tested using the spectrophotometer NanoDrop™ 2000 and gel electrophoresis (see Section 2.2.3).

2.2.3 Gel electrophoresis

All eDNA samples were run in a 0.6 % (w/v) agarose gel (TAE 1X buffer) at room temperature at 50 V for 5 h, then at 30 V overnight. All samples were prepared by aliquoting 10 µL of DNA per sample and adding 2 µL of loading dye 6X using wide bore tips to prevent shearing of the DNA. Lambda HindIII marker was used as ladder and CopyControl™ Fosmid DNA (Epicentre®) as concentration marker (2 µL of 100 ng/µL stock).
A selection of qPCR products (standards and eDNA amplicons) were run in 1 % (w/v) agarose gel (TAE 1X buffer) at room temperature at 110 V for 1 h with 2-log ladder (NEB).

2.2.4 Quantification of bacterial DNA extracted using qPCR assay

Bacterial 16S rRNA, NRPS and PKS gene copy numbers were determined in each extracted sample using qPCR in order to evaluate the efficiency of extraction.
In particular, the general 16S rRNA gene was quantified to detect the general bacterial abundance present in the samples, while specific primers for actinobacterial and verrucomicrobial 16S rRNA genes were used to detect specific groups potentially involved in natural product diversity as shown in Chapter 4.

The 7500 Fast Real-Time PCR System (Applied Biosystem) was used to perform all the qPCR assays. QPCR reaction for general bacterial 16S rRNA gene, actinobacterial 16S rRNA gene, NRPS and PKS genes (total volume 25 µL): 12.5 µL SYBR® Green

master mix (Invitrogen), primer forward 0.4 µM, primer reverse 0.4 µM, 0.5 mg/mL BSA, 1 µL DNA template.

QPCR reaction for verrucomicrobial 16S rRNA (total volume 50 µL): 25 µL SYBR® Green master mix (Invitrogen), primer forward 0.2 µM, primer reverse 0.2 µM, 0.5 mg/mL BSA, 1 µL DNA template.

All primers used in qPCR reactions are shown in Table 3.3.

Conditions for qPCR (general and actinobacterial 16S rRNA gene detection): 1. initial denaturation 95 °C 10 min, 2. denaturation 95 °C 15 s, 3. annealing 60 °C 1 min; Steps 2 and 3 were repeated for 40 cycles. Melt curves were registered at the end of each cycle. The same conditions were applied for NRPS and PKS genes detection with the exclusion of the annealing steps: NRPS annealing was performed at 62 °C for 1.5 min, while PKS at 60 °C for 1.5 min.

Conditions for the verrucomicrobial 16S rRNA gene quantification were optimized for the system available on the basis of the conditions reported by Ranjan *et al.* (2015): 1. initial denaturation 95 °C 3 min, 2. denaturation 95 °C 30 s, 3. annealing 50 °C 30 s, 4. extension 68 °C 1 min; Steps 2 to 4 were repeated for 40 cycles. Melt curves were registered at the end of each cycle.

QPCR standards for the quantitation of all the genes tested were prepared. Genomic DNA from *E. coli* EPI300™ and *S. coelicolor* M145 were extracted and used for amplification of general bacterial 16S rRNA gene and actinobacterial 16S rRNA gene, NRPS and PKS genes respectively. For the verrucomicrobial 16S rRNA gene standard a synthetic double-stranded oligonucleotide of the *Verrucomicrobium spinosum* strain DSM 4136 (NR_026266.1) 16S rRNA gene was order from IDT oligo using the gBlocks® Gene Fragments technology for the synthesis.

The concentration of PCR products was measured using the spectrophotometer NanoDrop™ 2000 (ThermoScientific™). The number of amplicon copies present in the purified PCR products or the synthetic DNA was calculated according to Equation 2.1 (http://cels.uri.edu/gsc/cndna.html).

$$number\ of\ copies = \frac{amount\ [ng] * 6.022x10^{23}[number/mole]}{lenght\ [bp] * 1x10^{9}[ng/g] * 650[g/mole\ of\ bp]}$$

Equation 2.1 Calculation for the number of copies of the gene of interest in the PCR amplicon sample used for the preparation of standard curves for qPCR assay. The equation takes into consideration the amount of DNA measured (ng), the length of the gene (PCR product - bp) and the average weight of a single nucleotide base.

QPCR standard curves were created amplifying serial dilutions of the PCR product/synthetized oligo using six points (from 10^7 to 10^2 copies) measured in triplicate. Standard curves for each target were calculated plotting the log quantities of the standards (*x* axis) against the C_t values obtained from the amplification of these samples (*y* axis). A linear regression was fitted and the equation was reported with the coefficient of determination (R^2). The efficiency of amplification of each assay was calculated as Equation 2.2 (Bustin *et al.*, 2009).

$$Efficiency\ (\%) = (10^{-1/slope} - 1) * 100$$

Equation 2.2 Calculation of qPCR assay efficiency percentage.

Table 2.3 Primers for qPCR detection of bacterial 16S rRNA gene, NRPS and PKS genes. * Optimised annealing temperature for qPCR assays.

Gene targeted	Name	Sequence	Annealing T (°C)	Amplicon size (bp)	Reference
General 16S rRNA gene	16S1369F 16S1492R	CGGTGAATACGTTCYCGG GGWTACCTTGTTACGACTT	60	123	(Suzuki *et al.*, 2000, Nandi *et al.*, 2004)
Actinobacte-rial 16S rRNA gene	F243 Eub518R	GGATGAGCCCGCGGCCTA ATTACCGCGGCTGCTGG	60	275	(Heuer *et al.*, 1997) (Muyzer *et al.*, 1993)
Verrucomicro-bial 16S rRNA gene	VER_37F VER_673R	TGGCGGCGTGGWTAAGA TGCTACACCGWGAATTC	68	636	(Ranjan *et al.*, 2015)
NRPS	NRPS_F2 NRPS_R	CGCGCGCATGTACTGGACNGGNGAYYT GGAGTGGCCGCCCARNYBRAARAA	62*	480	(Amos *et al.*, 2015): this study
PKS	PKS_F PKS_R	GGCAACGCCTACCACATGCANGNNYT GGTCCGCGGGACGTARTCNARRTC	60*	350	(Amos *et al.*, 2015): this study

A selection of qPCR products (standards and eDNA samples) for each target tested was also checked through gel electrophoresis (see Section 2.2.3) to assess the size of the amplicon and the specificity of the assays.

2.2.5 Statistical analysis

Data recovered from qPCR assays were normalized according to the elution volume used to resuspend the DNA at the end of the extraction and the quantity of material used for the DNA extraction in order to be able to compare all methods. Each 16S rRNA gene quantity was then normalized to genome equivalents according to the copy number reported on the ribosomal RNA database (rrnDB-4.4.4 -

https://rrndb.umms.med.umich.edu) for general Bacteria (2.5 copies of 16S rRNA gene per genome), Actinobacteria (3.12 copies of 16S rRNA gene per genome) and Verrucomicrobia (1.6 copies of 16S rRNA gene per genome). For NRPS and PKS genes data the copy number normalization step was not possible as the number of domains in each secondary metabolite cluster is variable and an average number is not available.

Quantified data were then Log10 transformed to normalize the qPCR count. One-way analysis of variance (ANOVA), two-ways ANOVA and post-hoc testing Tukey's HSD were performed using R (version 0.99.902) in order to test variance within and between groups of treatments and soils.

2.3 Results

2.3.1 Plugs extraction optimisation

The plugs system (Figure 2.2) for *in situ* lysis of cells present in soils was more effective when an enzymatic lysis was applied instead of a chemical lysis.

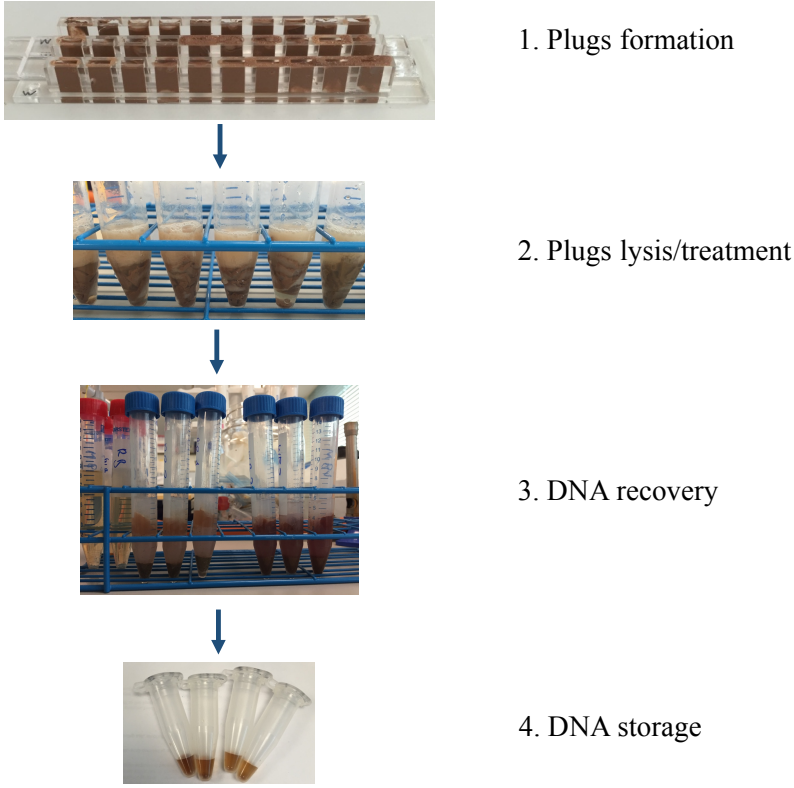


Figure 2.2 Schematic summary of soil plugs system for HMW eDNA recovery directly from soil. Plugs are created in disposable well molds (1) and lysed in 15 mL tubes with the buffer of choice (2). DNA is recovered through phenol:chlorophorm:isoamyl alcohol extraction in Phase Lock Gel tubes (3) and stored at 4 °C (4).

The optimised protocol of recovery of eDNA from plugs using PLG tubes was easier and more efficient when phosphate buffer or enzymatic lysis were performed on plugs instead of the chemical lysis using Brady lysis buffer. In particular, in phosphate buffer or enzymatic lysis, the aqueous solution containing the eDNA was recovered as per manufacturer' instructions above the gel phase which separates the aqueous phase from the organic phase (phenol:chlorophorm:isoamyl alcohol). On the contrary, phase separation proved more complex in the presence of Brady lysis buffer as the aqueous phase containing eDNA was blocked between the gel phase and the organic phase (Figure 2.3).

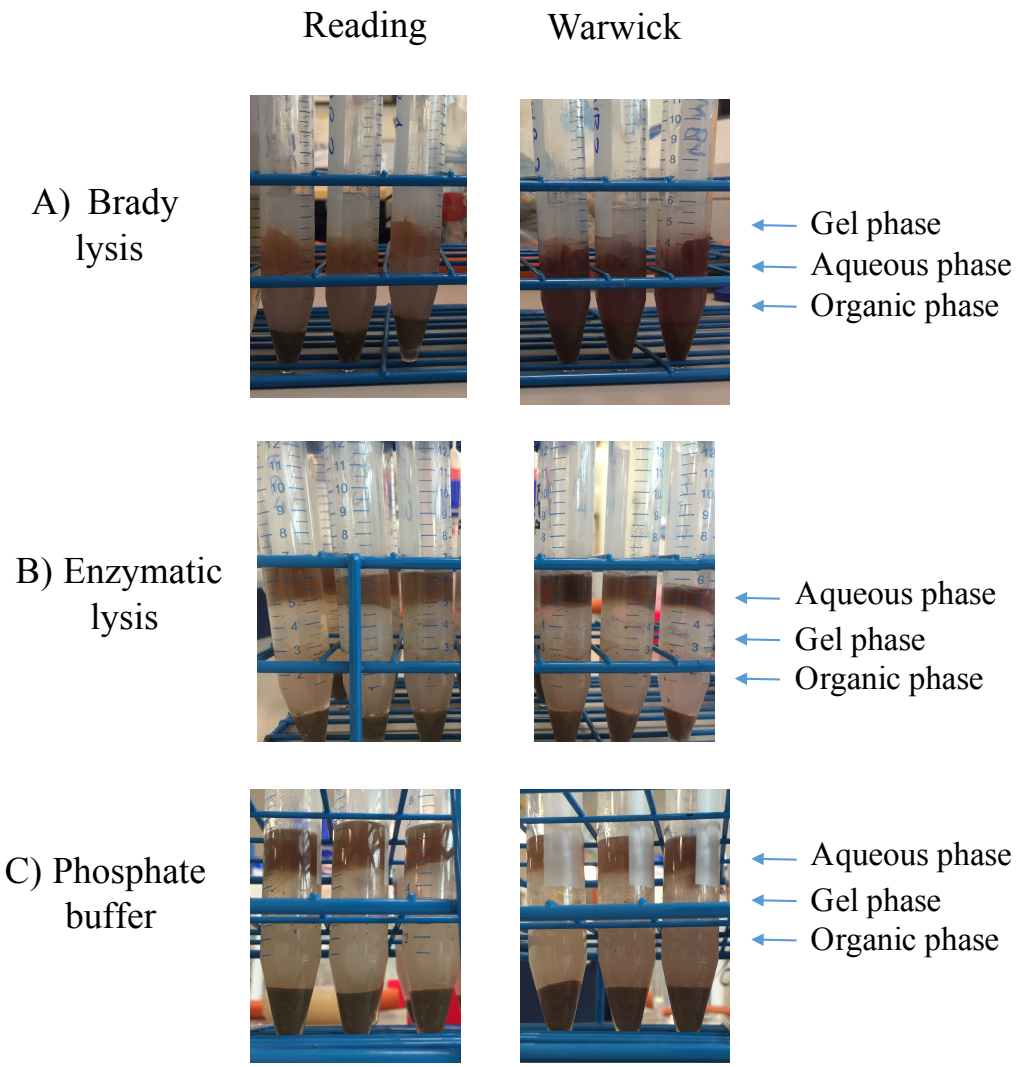


Figure 2.3 Recovery of HMW eDNA from soil plugs using the Phase Lock Gel tubes. Soil plugs were treated with different lysis methods for both soil (Reading on the left and Warwick on the right). Lysis methods tested: A) Brady lysis buffer (Method K); B) Enzymatic lysis (Method J); C) Phosphate lysis buffer (Method M).

2.3.2 Quality and concentration of eDNA

The concentration of all DNA samples extracted with methods described in Section 2.2.2 were measured using the NanoDrop™ 2000 (Thermos scientific) (Table S2.1). However, the presence of humic acids which absorbs at the same wavelength of DNA (260 nm) and other co-extracted compounds (Figures S2.1 and S2.2) biased the reading for most of the samples, therefore the spectrophotometer reads were not reliable for a comparable quantitation of the DNA recovered. Overnight agarose gels were run in order to estimate the concentration and control the size of the DNA extracted (Figure 2.4).

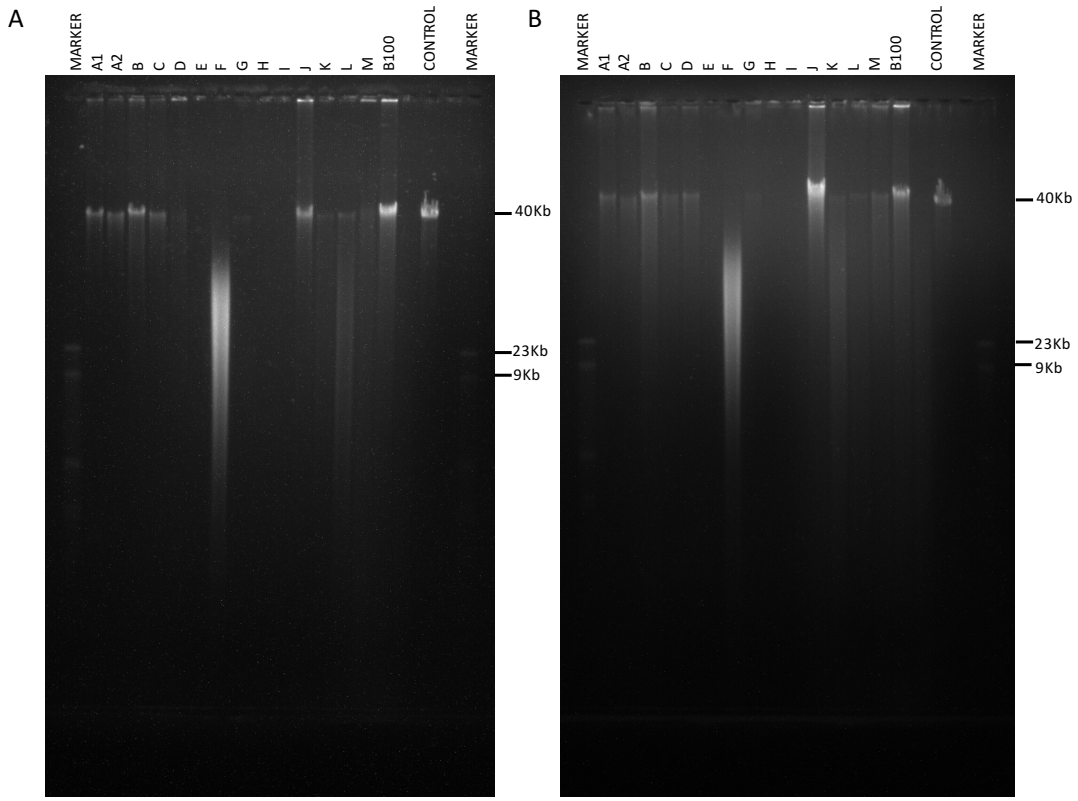


Figure 2.4 Gel electrophoresis of all eDNA samples extracted from A) Reading and B) Warwick soils using different lysis methods (Methods A1-M). CopyControl™ Fosmid DNA from Epicentre® (40Kb) was used as control and the marker used was lambda-HindIII DNA (Promega).

The estimates of DNA concentration obtained with the free software Image Studio™ Lite (LI-COR Biosciences) did not show any similarity with data retrieved from the NanoDrop™ reads and was difficult to estimate accurate concentrations.

The FastDNA™ kit allowed the recovery of concentrated DNA at low molecular weight, while the other methods recovered less concentrated DNA but HMW (> 40 kb) comparable to the control sample (Figure 2.4).

2.3.3 QPCR assay optimization

2.3.3.1 Verrucomicrobial 16S rRNA gene

Different conditions of reaction were tested in order to improve the amplification of the verrucomicrobial 16S rRNA gene. The general run conditions were kept as described by Ranjan *et al.* (2015), but different final reaction volumes (50 µL and 25 µL) were tested producing comparable results in terms of efficiency of amplification

(88 % in 50 μL reaction, 90 % in 25 μL reaction) and specificity (melt curves obtained by samples were similar to the one produced by the standards) (Figures 2.5 and 2.6). The amplification plots however, showed a difference in the C_t values at which standards were detected. In the 50 μL reaction, the top standard was detected at C_t 18 and the standard curve had a slope of -3.652, while in the 25 μL reaction, the slope was -3.586 with the highest standard appearing at C_t 21. The shift of the standards detection by 3 cycles is approximately equal to a factor 10 difference in the quantitation. Therefore, the conditions with the lowest C_t value for the lowest standard (i.e. 50 μL reaction) represented the best condition for a lower detection limit for environmental samples.

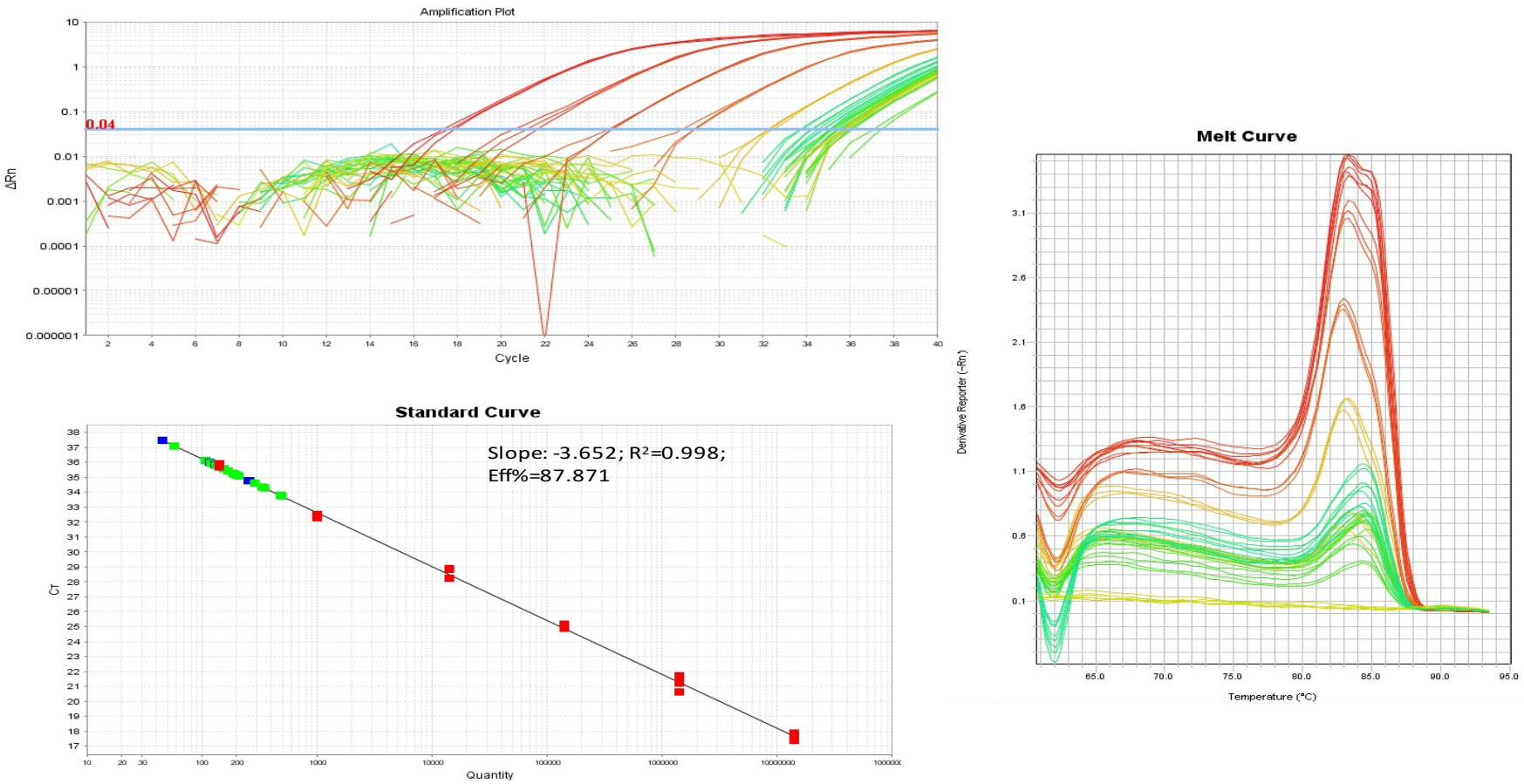


Figure 2.5 Optimization of the verrucomicrobial qPCR assay using 0.2 µM primers in 50 µL reaction. Amplification plots, melt curves and standard curves are reported for standards and some of the eDNA extracted in this study. Amplification and melt curves: standards (red) and samples (green and yellow); Standard curves: standard (red), unknown samples (blue and green).

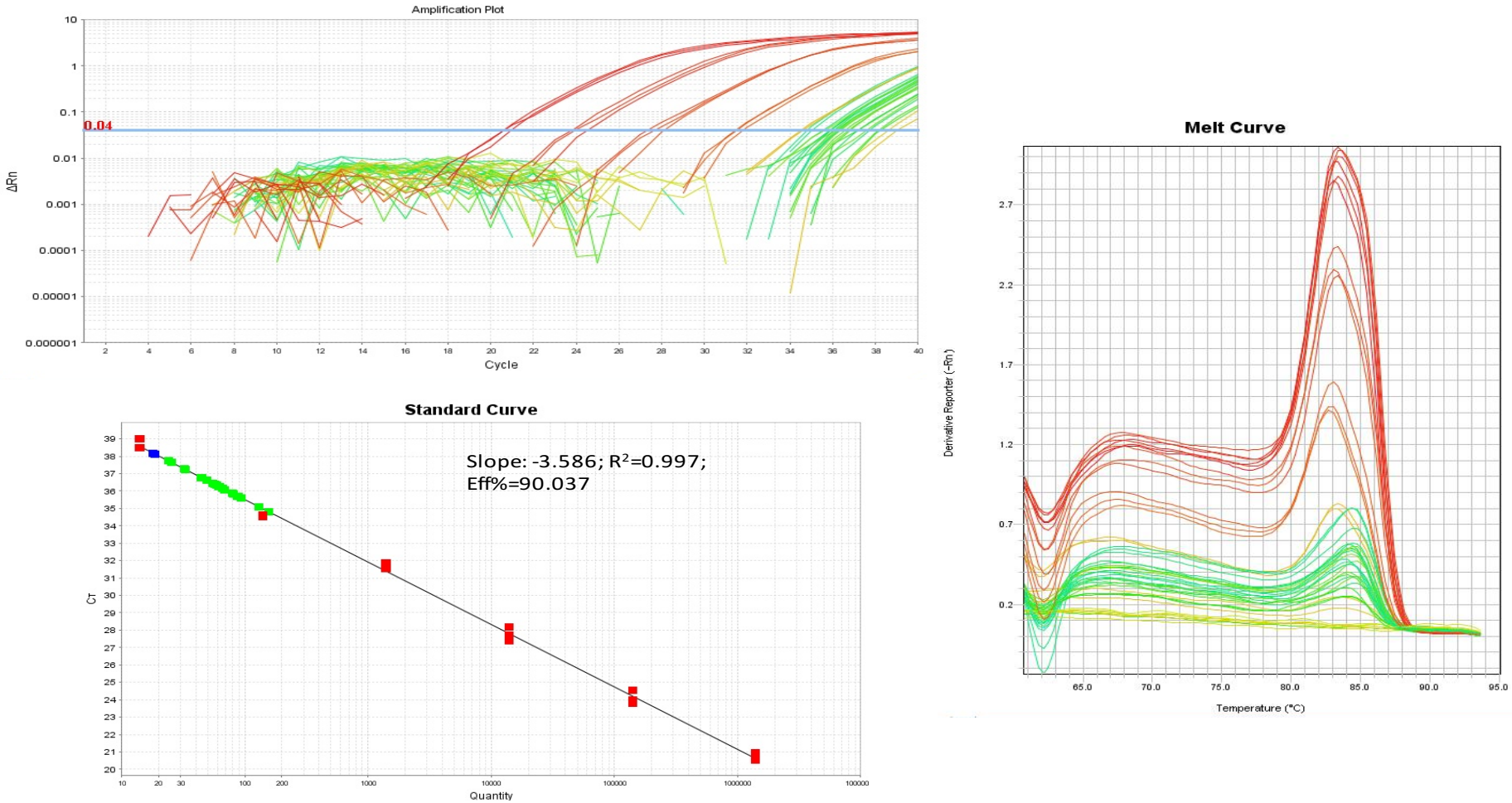


Figure 2.6 Optimization of the verrucomicrobial qPCR assay using 0.2 µM primers in 25 µL reaction. Amplification plots, melt curves and standard curves are reported for standards and some of the eDNA extracted in this study. Amplification and melt curves: standards (red) and samples (green and yellow); Standard curves: standard (red), unknown samples (blue and green).

2.3.3.2 NRPS and PKS

The first condition tested for NRPS and PKS was the same reaction and run protocol used for general bacterial 16S rRNA gene amplification with a different annealing/elongation condition, which was performed at either 60 °C or 62 °C for 1.5 min. Results showed that both primers proved specific for the target and NRPS amplified better at 62 °C while PKS at 60 °C. However, the efficiency of amplification was only 63 % and 69 % for NRPS and PKS respectively (Figure 2.7).

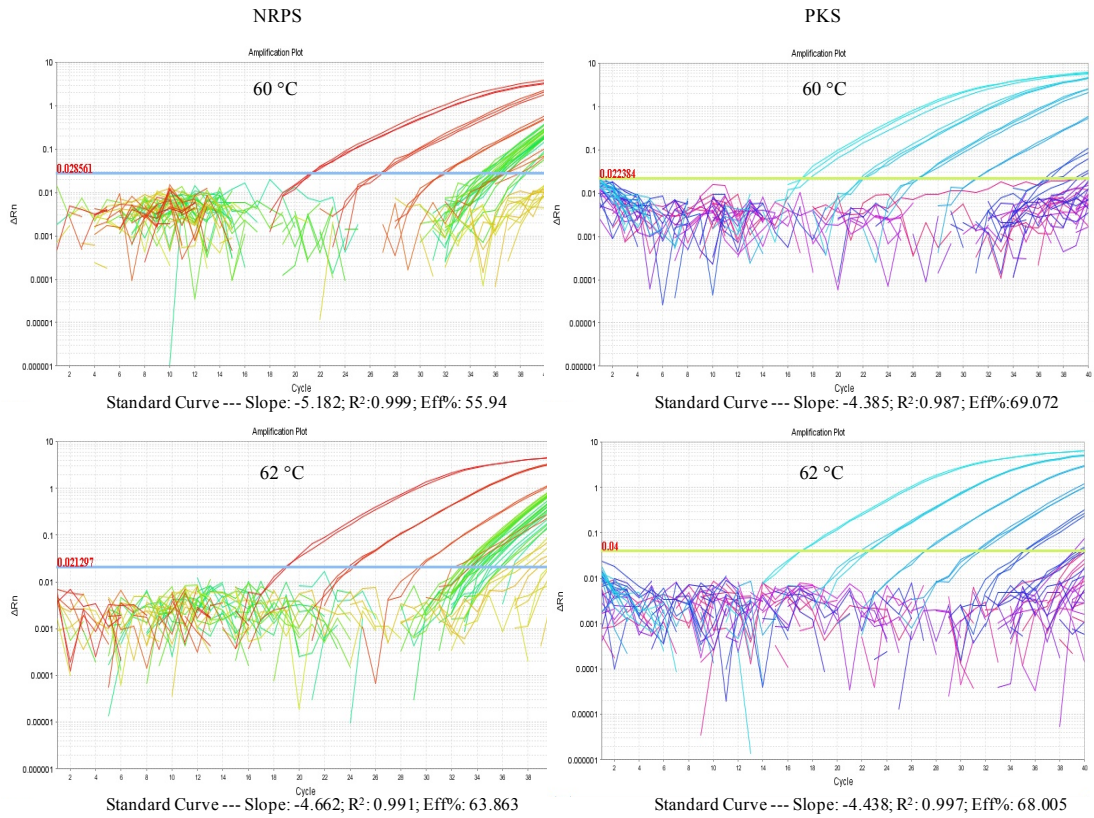
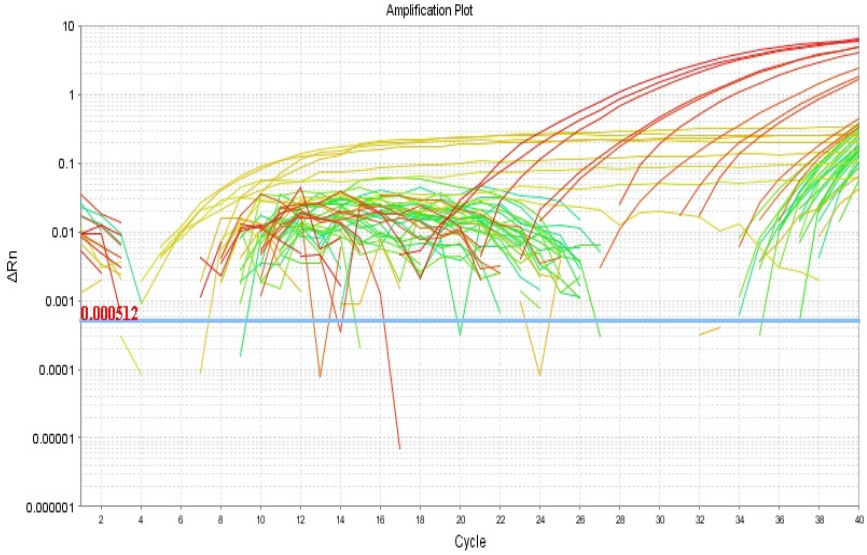


Figure 2.7 Optimization of the annealing temperature of the NRPS and PKS qPCR assay. Amplification plots for NRPS and PKS standards (red or light blue) and few eDNA samples (yellow/green and pink/purple) at 60°C and 62 °C with relative efficiency (slope and eff%) and specificity (R^2) parameters.

In order to improve the efficiency, bigger reaction volumes (up to 50 µL with 0.2 µM primers) were tested to exclude the possibility of reagents limiting performance of the reaction. Results showed that 50 µL reaction did not improve the assay but on the contrary the amplification between replicates was less robust and environmental samples did not amplify correctly (Figure 2.8)

NRPS (62 ºC)



PKS (60 ºC)

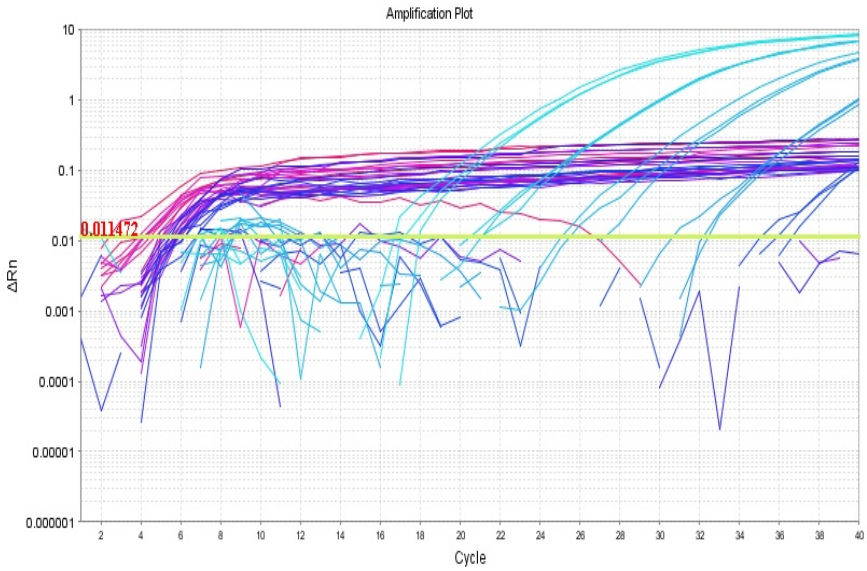
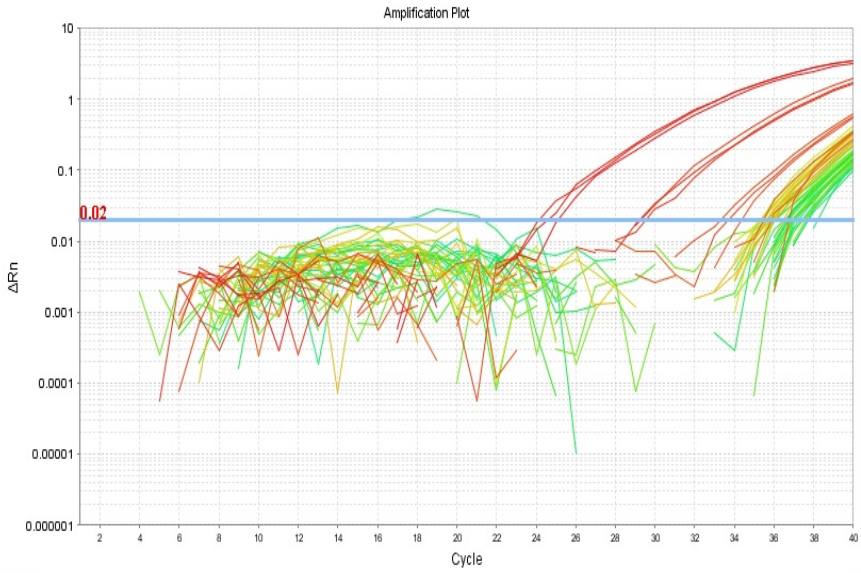


Figure 2.8 Optimization of the reaction volume of the NRPS and PKS qPCR assay. Amplification plots for NRPS and PKS standards (red and light blue) and few eDNA samples (yellow/green and pink/purple) using up to 50 µL reaction with already tested run conditions.

The reaction conditions were therefore maintained at 25 µL and the run conditions were changed to those used for the verrucomicrobial 16S rRNA gene assay. However, results did not show an improvement in the efficiency of amplification and also in this case the environmental samples did not amplify correctly as previously observed for 50 µL reactions (Figure 2.9).

NRPS (62 ºC)



PKS (60 ºC)

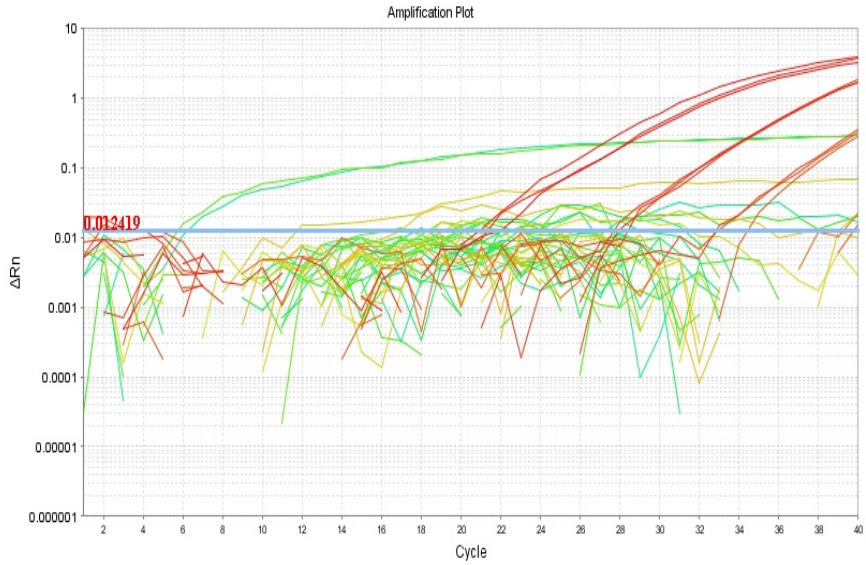


Figure 2.9 Optimization of the NRPS and PKS qPCR assay running protocol. Amplification plots for NRPS and PKS standards (red) and few eDNA samples (green) using verrucomicrobial 16S rRNA gene run conditions.

Even though the first conditions tested for NRPS and PKS amplification did not have a high efficiency of amplification, they were more specific for the target and more consistent than the other conditions, therefore they were chosen for detection and relative quantitation of NRPS and PKS genes.

2.3.4 DNA extraction method comparison

2.3.4.1 Variation within replica of extraction

The original Brady method (Method B100) showed a wider variability between replicates than all the other extraction methods for both soils and for all phylogenetic targets tested (Figures 2.10, 2.11 and 2.12). All methods of extraction were more robust in Reading soil than Warwick soil, showing less variation between replicates for the same method. In particular, Warwick soil showed a greater variability between replicates for extraction from plugs treated with Brady lysis buffer (Method K), (Figures 2.10, 2.11 and 2.12).

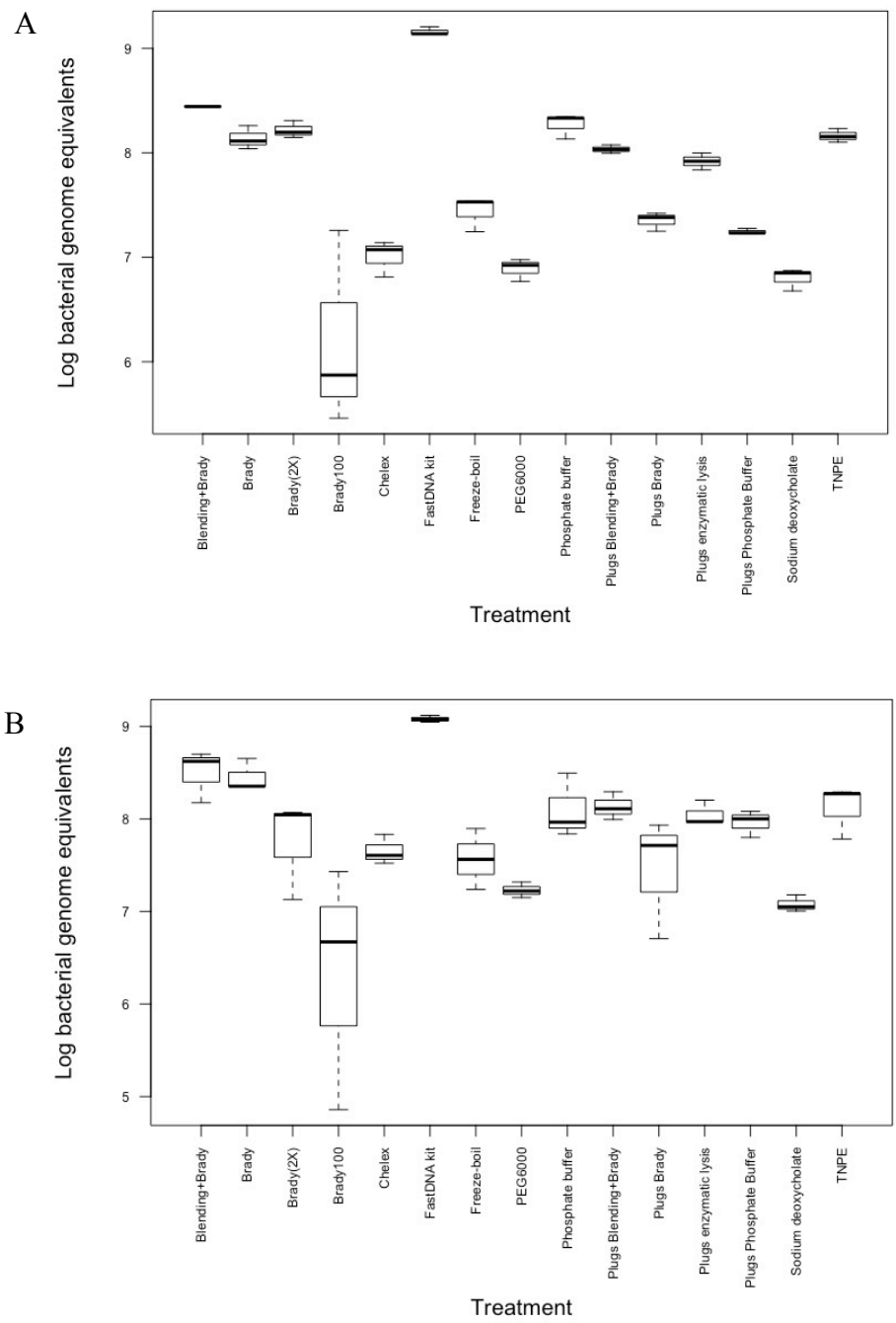


Figure 2.10 Variation of general bacterial genome equivalents in Reading (A) and Warwick (B) soil within each extraction method tested.

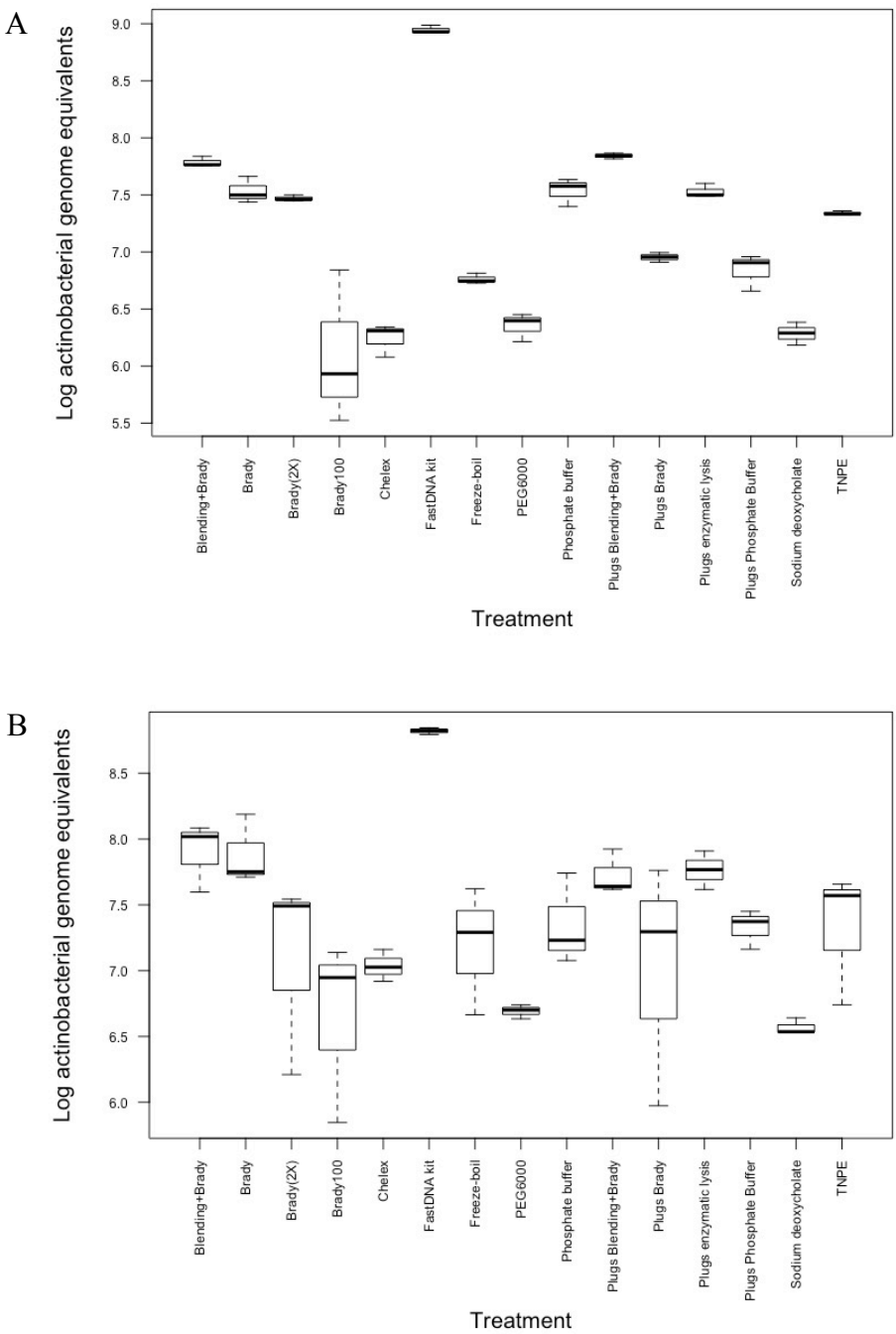


Figure 2.11 Variation of actinobacterial genome equivalents in Reading (A) and Warwick (B) soil within each extraction method tested.

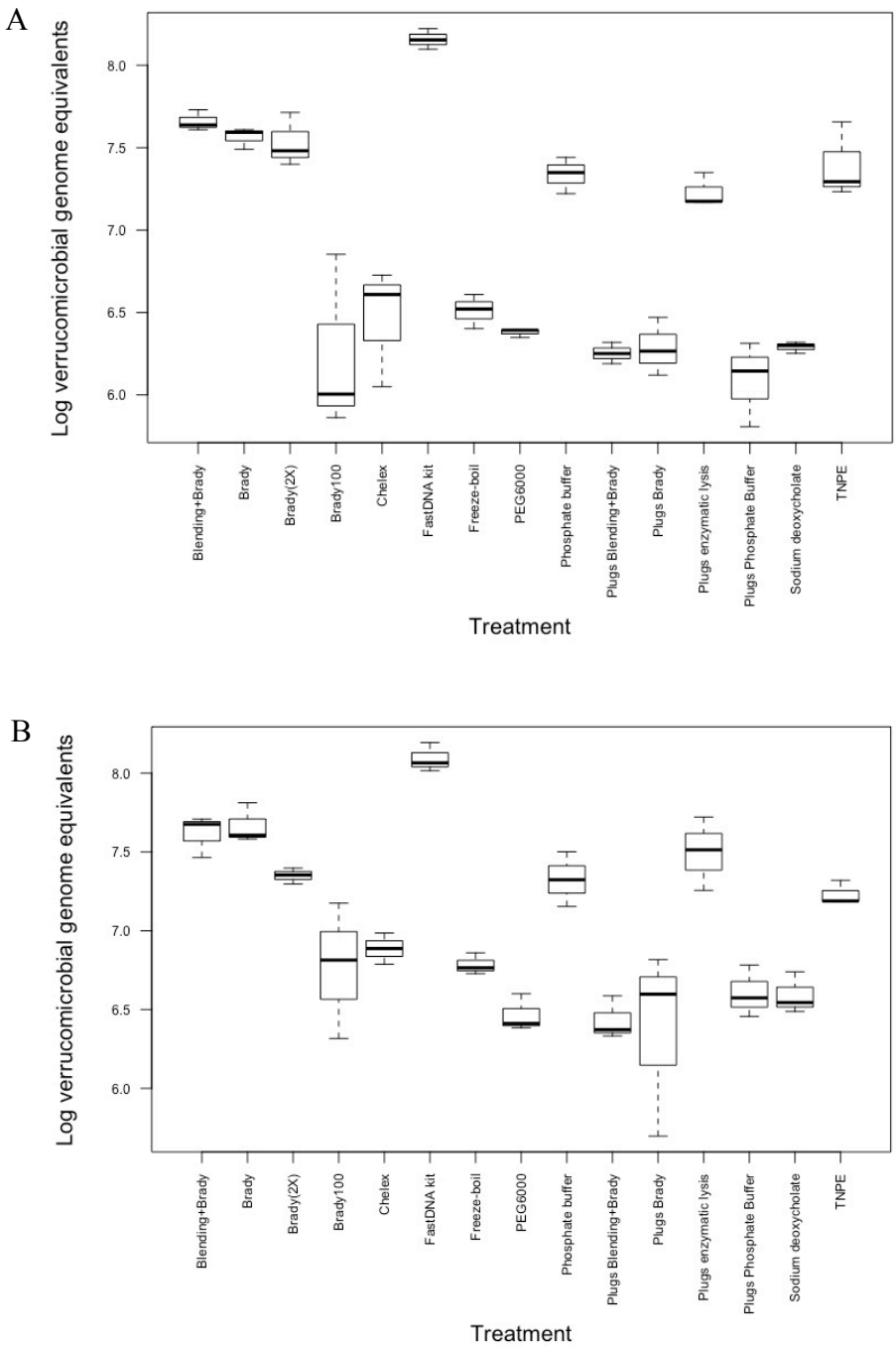


Figure 2.12 Variation of verrucomicrobial genome equivalents in Reading (A) and Warwick (B) soil within each extraction method tested.

The same variability pattern within replicate was observed for NRPS data (Figure 2.13). Extractions from Reading soil were less variable than from Warwick soil. The original Brady method (Method B100) and plugs treated with the Brady lysis buffer (Method K) were the two methods that showed a greater variability, especially in Warwick soil (Figure 2.14).

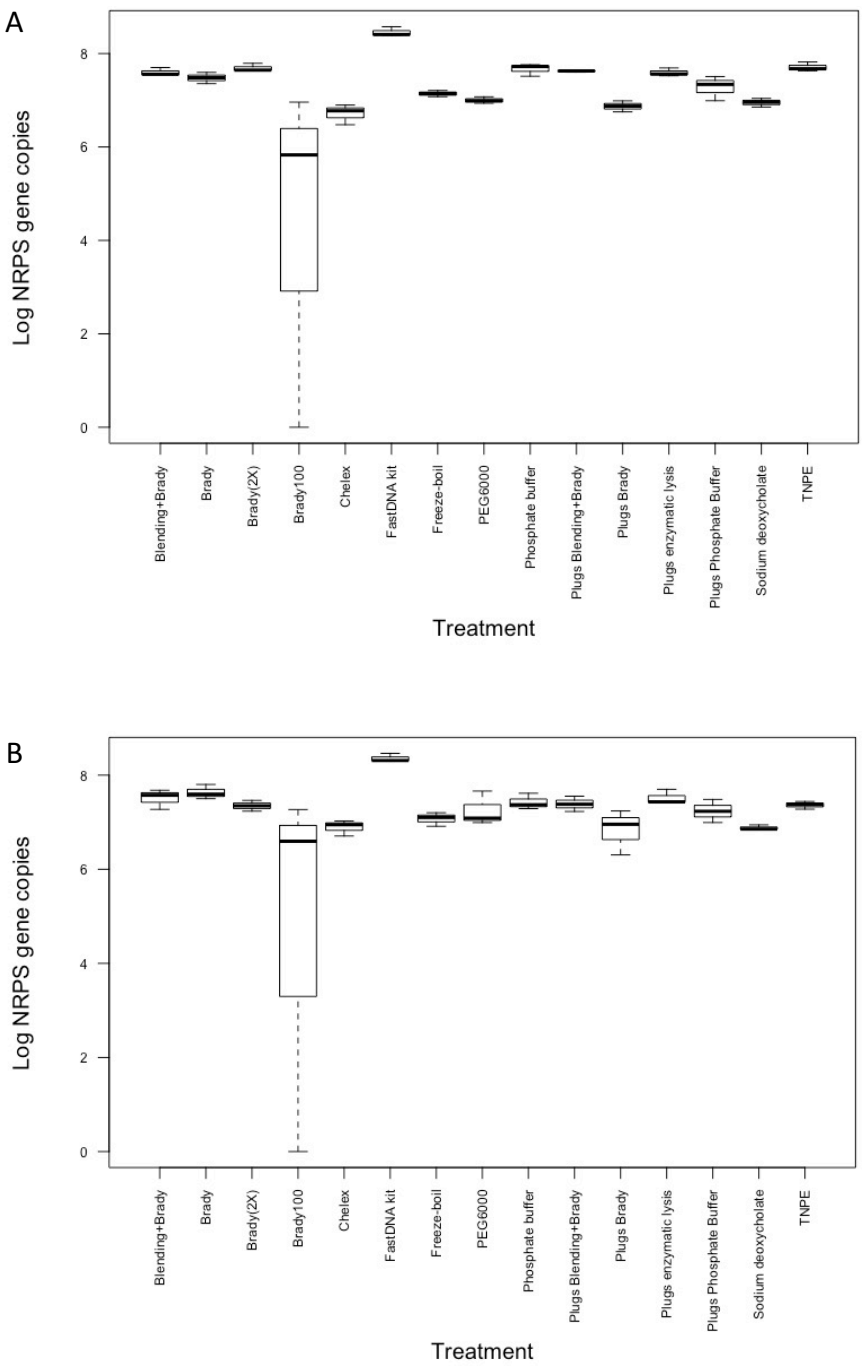


Figure 2.13 Variation of NRPS gene copies in Reading (A) and Warwick (B) soils within each extraction method tested.

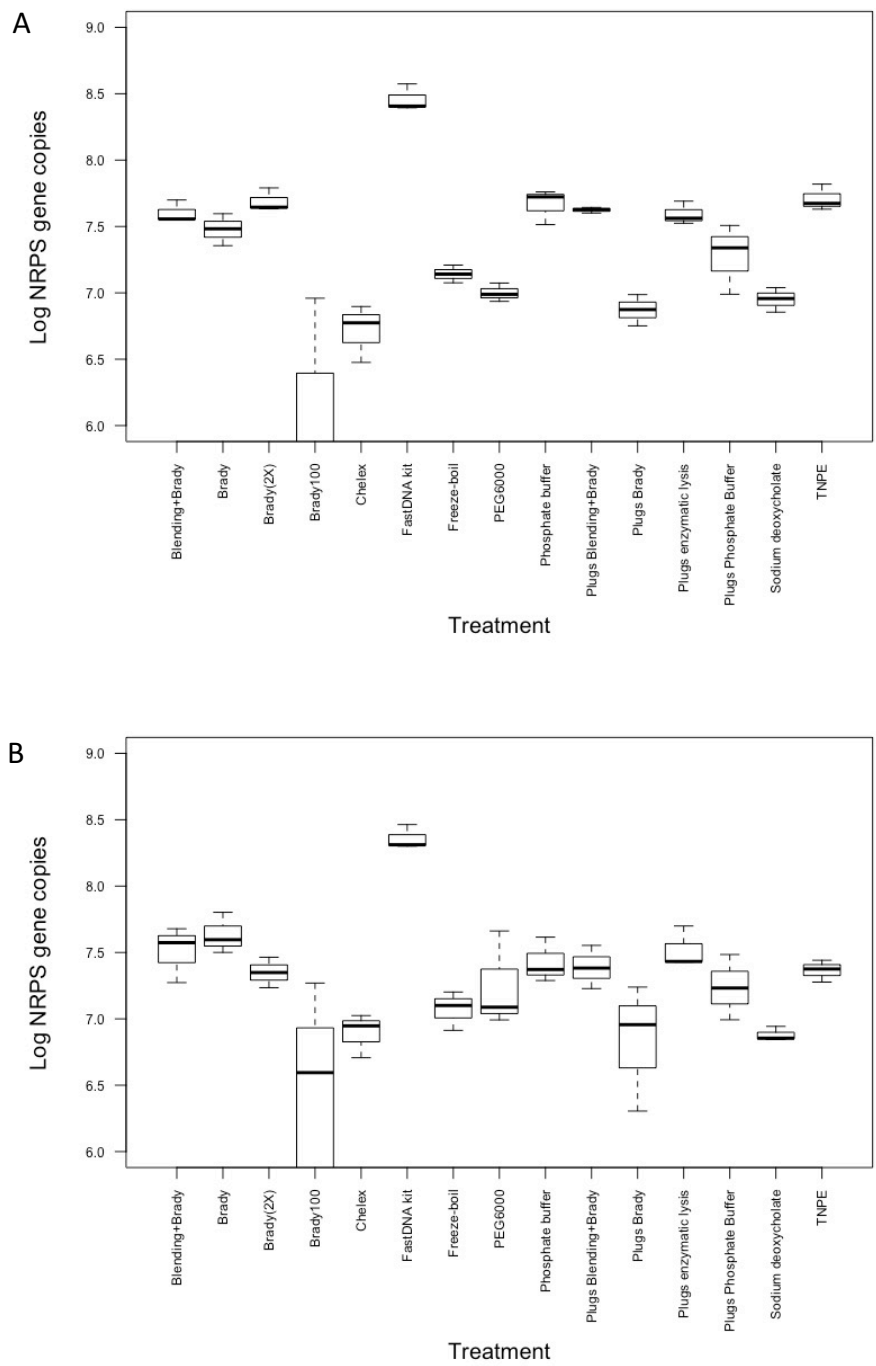


Figure 2.14 A zoom in on Figure 2.13 to better visualise variation differences between samples; Reading (A) and Warwick (B).

For PKS gene amplification, the majority of samples did not amplify and for the few positive samples, an appropriate number of positive replicates was not detected to perform statistical analysis with exception for Method F (Figure 2.15). FastDNA™ Spin kit (Method F) was the only method with reproducible results between all replicates and with a limited range of variability (Figure 2.15).

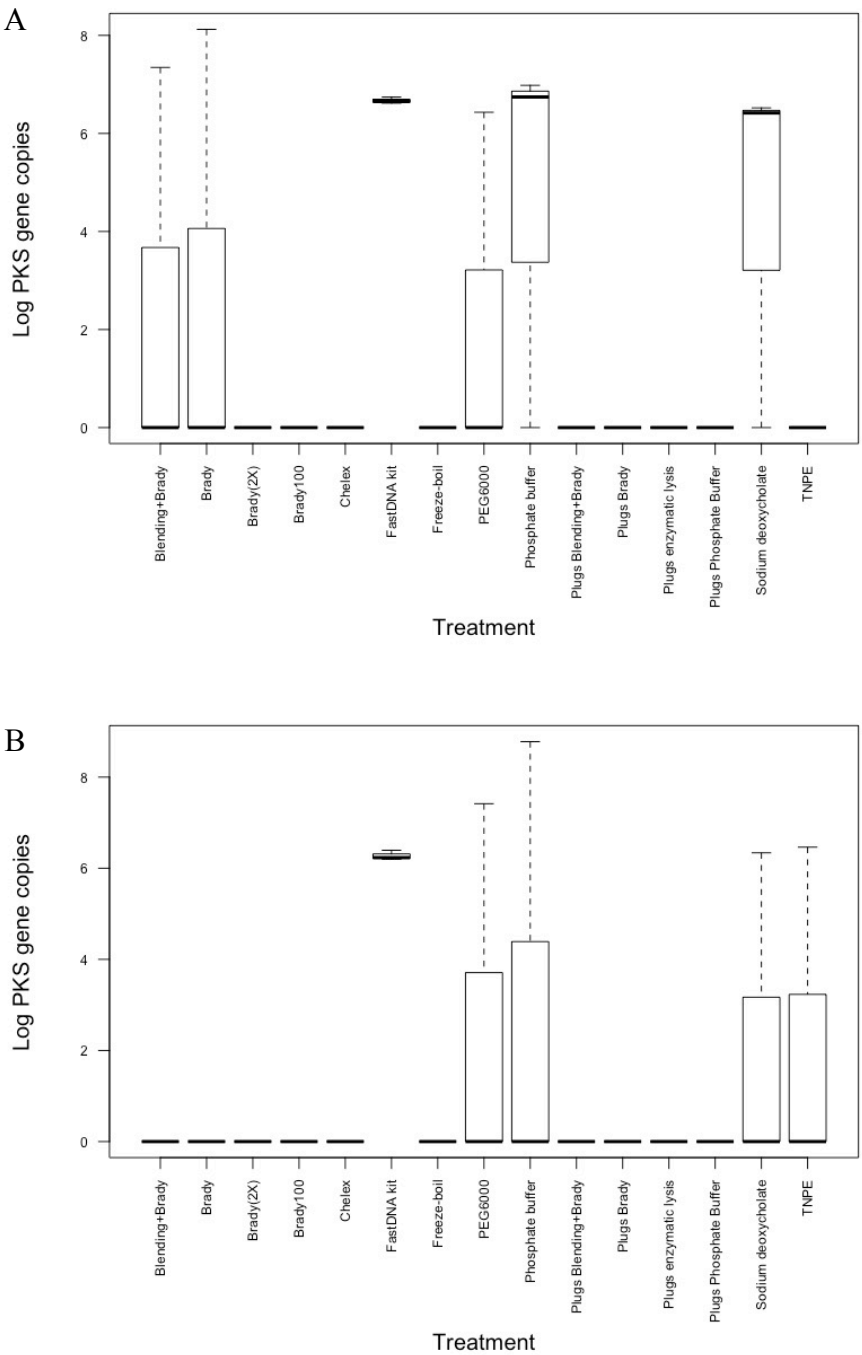


Figure 2.15 Variation of PKS gene copies in Reading (A) and Warwick (B) soils within each extraction method tested.

2.3.4.1 Variation between extraction methods

There was a significant difference in the mean copy numbers detected by qPCR between methods used to extract DNA from both soils for all the phylogenetic markers

16S rRNA genes tested (Table 2.4). A significant difference in extraction methods was also identified for NRPS genes detected in DNA extracted from Reading soil (One-way ANOVA $F_{14,30}=2.71$, $p < 0.05$*) but not from Warwick soil (One-way ANOVA $F_{14,30}=1.692$, $p=0.111$) (Table 2.4). The ANOVA was not calculated for PKS as not enough data were detected.

Table 2.4 Comparison of DNA extraction methods in each soil (Reading and Warwick) for each target (general bacterial 16S rRNA, actinobacterial 16s rRNA, verrucomicrobia 16S rRNA and NRPS genes). (One-way ANOVA; significance code recorded in brackets (***) p=0.001, (*) p=0.05).

Soil	QPCR target	Degrees of freedom	Residuals	F statistic	P
Reading	General bacterial 16S rRNA	14	30	26.5	p < 0.001 (***)
	Actinobacterial 16S rRNA	14	30	46.7	p < 0.001 (***)
	Verrucomicrobial 16S rRNA	14	30	31.94	p < 0.001 (***)
	NRPS	14	30	2.71	p < 0.05 (*)
Warwick	General bacterial 16S rRNA	14	30	6.621	p < 0.001 (***)
	Actinobacterial 16S rRNA	14	30	5.714	p < 0.001 (***)
	Verrucomicrobial 16S rRNA	14	30	16.61	p < 0.001 (***)
	NRPS	14	30	1.692	p > 0.05

2.3.4.3 Reading soil: variation in general bacterial 16S rRNA extraction

For general bacterial 16S rRNA gene extraction from Reading soil, the original Brady100 method (Method B100) was significantly different (Tukey's HSD $p < 0.05$) to all methods a part from the extractions performed with sodium deoxycholate (Method H) and PEG6000 (Method I) (Table 2.5). The blending+Brady method (Method B) was the only method that did not show significant difference (Tukey's HSD $p > 0.05$) from the FastDNA™ Spin Kit (Method F), while all the other methods were significantly different (Tukey's HSD $p < 0.05$) from the kit (Table 2.5). The introduction of the blending step before the chemical gentle lysis of the Brady lysis

buffer (Method B) was significantly different (Tukey's HSD $p < 0.05$) to the physical lysis method of freezing-boiling (Method E) but not to the FastDNA™ Spin kit (Method F) and the methods containing the anionic detergents (Method H), the dispersant resin (Method G) and the protein precipitant PEG6000 (Method I) (Tukey's HSD $p > 0.05$). The general 16S rRNA gene counts from DNA extracted from blended soil (Method B) did not result in a significant difference (Tukey's HSD $p > 0.05$) from the DNA extracted from the same blended soil encapsulated in agarose plugs (Method L, Table 2.5). However, these plugs showed a significant difference (Tukey's HSD $p<0.05$) with Brady100 (Method B100), Chelex® 100 (Method G), sodium deoxycholate (Method H) and PEG6000 (Method I). The same pattern was observed for plugs treated with an enzymatic lysis (Method J). Soil plugs extracted with Brady lysis buffer (Method K) did not show a significant difference (Tukey's HSD $p > 0.05$) from the Brady method applied directly on 10g of soil (Method A1). However, plugs treated with phosphate buffer (Method M) showed a significant difference (Tukey's HSD $p < 0.05$) with soil directly treated with the same buffer (Method D).

Table 2.5 Identification of significantly different methods of extraction for Reading soil general bacterial 16S rRNA genes detection. Tukey's HSD; Dark Blue = same method; light blue = no significant difference between methods $p > 0.05$, green = borderline no significant difference $0.050 \leq p \leq 0.055$, red = significant difference between methods $p < 0.05$.

READING GENERAL 16S rRNA		B100	A1	A2	B	C	D	E	F	G	H	I	J	K	L	M
		Brady100	Brady	Brady (2X)	Blending + Brady	TPNE	Phosphate buffer	Freeze-boil	FastDNA Spin Kit for soil	Chelex + Brady	Sodium deoxycholate +Brady	PEG6000 + Brady	Plugs (enzymatic lysis)	Plugs (Brady)	Plugs (blending + Brady)	Plugs (phosphate buffer)
B100	Brady100															
A1	Brady															
A2	Brady (2X)															
B	Blending + Brady															
C	TPNE															
D	Phosphate buffer															
E	Freeze-boil															
F	FastDNA Spin Kit for soil															
G	Chelex + Brady															
H	Sodium deoxycholate +Brady															
I	PEG6000 + Brady															
J	Plugs (enzymatic lysis)															
K	Plugs (Brady)															
L	Plugs (blending + Brady)															
M	Plugs (phosphate buffer)															

2.3.4.4 Reading soil: variation in actinobacterial 16S rRNA gene extraction

For actinobacerial 16S rRNA gene from Reading soil, the commercial kit FastDNA™ Spin kit was significantly different from all methods (Tukey's HSD $p < 0.05$) (Table 2.6). The Brady 100 (Method B100) was comparable only to methods using Chelex® 100 (Method G), sodium deoxycholate (Method H) and PEG6000 (Method I)

(Tukey's HSD $p > 0.05$). The use of the Brady lysis buffer and the blending pre-treatment did not impact the extraction significantly when used on plugs or directly on soil (Tukey's HSD $p > 0.05$). Phosphate buffer however, had a significant effect (Tukey's HSD $p < 0.05$) when applied to soil plugs instead of free soil (Tukey's HSD $p > 0.05$).

Table 2.6 Identification of significantly different methods of extraction for Reading soil actinobacterial 16S rRNA genes detection. Tukey's HSD; Dark Blue = same method; light blue = no significant difference between methods $p > 0.05$, green = borderline no significant difference $0.050 \leq p \leq 0.055$, red = significant difference between methods $p < 0.05$.

READING ACTINOBACTERIA 16S rRNA		B100	A1	A2	B	C	D	E	F	G	H	I	J	K	L	M
		Brady100	Brady	Brady (2X)	Blending + Brady	TPNE	Phosphate buffer	Freeze-boil	FastDNA Spin Kit for soil	Chelex + Brady	Sodium deoxycholate +Brady	PEG6000 + Brady	Plugs (enzymatic lysis)	Plugs (Brady)	Plugs (blending + Brady)	Plugs (phosphate buffer)
B100	Brady100															
A1	Brady															
A2	Brady (2X)															
B	Blending + Brady															
C	TPNE															
D	Phosphate buffer															
E	Freeze-boil															
F	FastDNA Spin Kit for soil															
G	Chelex + Brady															
H	Sodium deoxycholate															
I	PEG6000 + Brady															
J	Plugs (enzymatic lysis)															
K	Plugs (Brady)															
L	Plugs (blending + Brady)															
M	Plugs (phosphate buffer)															

2.3.4.5 Reading soil: variation in verrucomicrobial 16S rRNA gene extraction

For verrucomicrobial 16S rRNA gene from Reading soil, the optimized Brady protocol and the addition of the blending step were not significantly different (Tukey's HSD $p > 0.05$) from the physical lysis obtained using the FastDNA™ spin kit (Table 2.7). The Brady method performed on 10g of soil was significantly different (Tukey's HSD $p < 0.05$) from the original method (Method B100) as well as most of the treatments involving a pre-treatment either chemical or physical followed by a Brady buffer based lysis.

Plugs treated with Brady, blending+Brady and phosphate lysis buffers were significantly different (Tukey's HSD $p < 0.05$) from the use of their respective treatment on free soil and also from plugs treated with an enzymatic reaction (Table 2.7).

Table 2.7 Identification of significantly different methods of extraction for Reading soil verrucomicrobial 16S rRNA genes detection. Tukey's HSD; Dark Blue = same method, not tested; light blue = no significant difference between methods p > 0.05, red = significant difference between methods p < 0.05.

READING VERRUCOMICROBIA 16S rRNA		B100	A1	A2	B	C	D	E	F	G	H	I	J	K	L	M
		Brady100	Brady	Brady (2X)	Blending + Brady	TPNE	Phosphate buffer	Freeze-boil	FastDNA Spin Kit for soil	Chelex + Brady	Sodium deoxycholate +Brady	PEG6000 + Brady	Plugs (enzymatic lysis)	Plugs (Brady)	Plugs (blending + Brady)	Plugs (phosphate buffer)
B100	Brady100															
A1	Brady															
A2	Brady (2X)															
B	Blending + Brady															
C	TPNE															
D	Phosphate buffer															
E	Freeze-boil															
F	FastDNA Spin Kit for soil															
G	Chelex + Brady															
H	Sodium deoxycholate															
I	PEG6000 + Brady															
J	Plugs (enzymatic lysis)															
K	Plugs (Brady)															
L	Plugs (blending + Brady)															
M	Plugs (phosphate buffer)															

2.3.4.6 Reading soil: variation in NRPS extraction

For NRPS, the majority of methods varied significantly from Brady100 (Method B100) (Tukey's HSD $p < 0.05$) (Table 2.8). Methods using Chelex® 100, sodium deoxycolate, PEG6000 as a pre-treatment on free soil (Methods G, H, I) and an enzymatic lysis on soil plugs (Method K) were the only methods that were not significantly different (Tukey's HSD $p > 0.05$) from Method B100.

Table 2.8 Identification of significantly different methods of extraction for Reading soil NRPS genes detection. Tukey's HSD; Dark Blue = same method; light blue = no significant difference between methods p > 0.05, green = borderline no significant difference 0.050≤p≤0.055, red = significant difference between methods p < 0.05.

READING NRPS		B100	A1	A2	B	C	D	E	F	G	H	I	J	K	L	M
		Brady100	Brady	Brady (2X)	Blending + Brady	TPNE	Phosphate buffer	Freeze-boil	FastDNA Spin Kit for soil	Chelex + Brady	Sodium deoxycholate +Brady	PEG6000 + Brady	Plugs (enzymatic lysis)	Plugs (Brady)	Plugs (blending + Brady)	Plugs (phosphate buffer)
B100	Brady100															
A1	Brady															
A2	Brady (2X)															
B	Blending + Brady															
C	TPNE															
D	Phosphate buffer															
E	Freeze-boil															
F	FastDNA Spin Kit for soil															
G	Chelex + Brady															
H	Sodium deoxycholate															
I	PEG6000 + Brady															
J	Plugs (enzymatic lysis)															
K	Plugs (Brady)															
L	Plugs (blending + Brady)															
M	Plugs (phosphate buffer)															

2.3.4.7 Warwick soil: variation in general bacterial 16S rRNA extraction

For general bacterial 16S rRNA genes in Warwick soil, Brady 100 and FastDNA™ Spin kit (Methods B100 and F) were the two methods that showed a significant difference (Tukey's HSD $p < 0.05$) to almost all the other methods. In particular, the FastDNA™ Spin kit did not produce significantly different results (Tukey's HSD $p > 0.05$) from the majority of the methods using the optimized Brady method (Methods A1, A2, B, L), TNPE (Method C), phosphate buffer (Methods D and M) and enzymatic lysis (Method J) (Table 2.9).

Table 2.9 Identification of significantly different methods of extraction for Warwick soil general bacterial 16S rRNA genes detection. Tukey's HSD; Dark Blue = same method; light blue = no significant difference $p > 0.05$, green = borderline no significant difference $0.050 \leq p \leq 0.055$, red = significant difference $p < 0.05$.

WARWICK GENERAL 16S rRNA		B100	A1	A2	B	C	D	E	F	G	H	I	J	K	L	M
		Brady100	Brady	Brady (2X)	Blending + Brady	TPNE	Phosphate buffer	Freeze-boil	FastDNA Spin Kit for soil	Chelex + Brady	Sodium deoxycholate +Brady	PEG6000 + Brady	Plugs (enzymatic lysis)	Plugs (Brady)	Plugs (blending + Brady)	Plugs (phosphate buffer)
B100	Brady100															
A1	Brady															
A2	Brady (2X)															
B	Blending + Brady															
C	TPNE															
D	Phosphate buffer															
E	Freeze-boil															
F	FastDNA Spin Kit for soil															
G	Chelex + Brady															
H	Sodium deoxycholate															
I	PEG6000 + Brady															
J	Plugs (enzymatic lysis)															
K	Plugs (Brady)															
L	Plugs (blending + Brady)															
M	Plugs (phosphate buffer)															

2.3.4.8 Warwick soil: variation in Actinobacteria 16S rRNA extraction

For Actinobacteria 16S from Warwick soil, FastDNA™ Spin kit (Method F) showed a significant difference with most of the other methods (Tukey's HSD $p < 0.05$) (Table 2.10). Interestingly the optimized Brady method on 10g of soil (Method A1), the blending pre-treatment to the soil before the lysis protocol on soil for direct extraction (Method B) or plugs formation prior extraction (Method L) and the enzymatic lysis on soil plugs (Method J) were the only treatments which did not show a significant difference to the commercial kit (Tukey's HSD $p > 0.05$) (Table 2.10).

Table 2.10 Identification of significantly different methods of extraction for Warwick soil actinobacterial 16S rRNA genes detection. Tukey's HSD; Dark Blue = same method; light blue = no significant difference between methods p > 0.05, green = borderline no significant difference 0.050≤p≤0.055, red = significant difference between methods p < 0.05.

WARWICK ACTINOBACTERIA 16S rRNA		B100	A1	A2	B	C	D	E	F	G	H	I	J	K	L	M
		Brady100	Brady	Brady (2X)	Blending + Brady	TPNE	Phosphate buffer	Freeze-boil	FastDNA Spin Kit for soil	Chelex + Brady	Sodium deoxycholate +Brady	PEG6000 + Brady	Plugs (enzymatic lysis)	Plugs (Brady)	Plugs (blending + Brady)	Plugs (phosphate buffer)
B100	Brady100															
A1	Brady															
A2	Brady (2X)															
B	Blending + Brady															
C	TPNE															
D	Phosphate buffer															
E	Freeze-boil															
F	FastDNA Spin Kit for soil															
G	Chelex + Brady															
H	Sodium deoxycholate															
I	PEG6000 + Brady															
J	Plugs (enzymatic lysis)															
K	Plugs (Brady)															
L	Plugs (blending + Brady)															
M	Plugs (phosphate buffer)															

2.3.4.9 Warwick soil: variation in Verrucomicrobia 16S rRNA extraction

For Verrucomicrobia 16S from Warwick soil, the FastDNA™ Spin kit showed a significant difference with all methods (Tukey's HSD $p < 0.05$) excluding the optimized Brady protocol (Method A1), the blending pre-treatment (Method B) and the enzymatic lysis performed on plugs (Method J) (Tukey's HSD $p > 0.05$) (Table 2.11). The Brady100 method results were comparable to most of the methods (no significant variation, Tukey's HSD $p > 0.05$), apart from Brady (Method A1), blending+Brady (Method B), FastDNA™ Spin kit (Method F) and plugs enzymatic lysis methods (Method J) (Tukey's HSD $p < 0.05$). The use of the anionic detergent sodium deoxycolate (Method H), polymer PEG6000 (Method I) and ion exchange resin Chelex® 100 (Method G) showed a significant difference (Tukey's HSD $p < 0.05$) with the commercial kit and with most of the methods using the Brady or the phosphate lysis buffers directly on soil but not when applied to soil in plugs (Tukey's HSD $p > 0.05$) (Table 2.11).

Table 2.11 Identification of significantly different methods of extraction for Warwick soil verrucomicrobial 16S rRNA genes detection. Tukey's HSD; Dark Blue = same method; light blue = no significant difference between methods p > 0.05, red = significant difference between methods p < 0.05.

WARWICK VERRUCOMICROBIA 16S rRNA		B100	A1	A2	B	C	D	E	F	G	H	I	J	K	L	M
		Brady100	Brady	Brady (2X)	Blending + Brady	TPNE	Phosphate buffer	Freeze-boil	FastDNA Spin Kit for soil	Chelex + Brady	Sodium deoxycholate +Brady	PEG6000 + Brady	Plugs (enzymatic lysis)	Plugs (Brady)	Plugs (blending + Brady)	Plugs (phosphate buffer)
B100	Brady100															
A1	Brady															
A2	Brady (2X)															
B	Blending + Brady															
C	TPNE															
D	Phosphate buffer															
E	Freeze-boil															
F	FastDNA Spin Kit for soil															
G	Chelex + Brady															
H	Sodium deoxycholate															
I	PEG6000 + Brady															
J	Plugs (enzymatic lysis)															
K	Plugs (Brady)															
L	Plugs (blending + Brady)															
M	Plugs (phosphate buffer)															

2.3.4.10 Interaction of soil type and extraction method

The Two-Way ANOVA results for general bacterial 16S rRNA gene and NRPS showed that overall there was not a significant difference between soil types but there was between methods of extraction (Table 2.12). The interaction between soil type and method of extraction was not significant for the three targets (Table 2.12). Actinobacterial and verrucomicrobial 16S rRNA gene showed a significant difference for soil type and methods but the interaction of the two independent variables was not significant (Table 2.12).

Table 2.12 Comparison of soil types, methods of DNA extraction and their interaction with the general bacterial population (general 16S rRNA gene), actinobacterial 16S rRNA gene, verrucomicrobial 16S rRNA gene and NRPS genes. (Two-ways ANOVA, significance code recorded in brackets (***) p=0.001, (**) p=0.01, (*) p=0.05).

QPCR target	Soil/method	Degrees of freedom	Residuals	F statistic	P
General bacterial 16S rRNA	Soil type	1	60	3.453	p > 0.05
	Method	14	60	22.393	p < 0.001 (***)
	Soil type:Method	14	60	0.994	p > 0.05
Actinobacterial 16S rRNA	Soil type	1	60	7.064	p < 0.1 (*)
	Method	14	60	24.045	p < 0.001 (***)
	Soil type:Method	14	60	1.375	p > 0.05
Verrucomicrobial 16S rRNA	Soil type	1	60	11.194	p < 0.01 (**)
	Method	14	60	45.582	p < 0.001 (***)
	Soil type:Method	14	60	1.671	p > 0.05
NRPS	Soil type	1	60	0.048	p > 0.05
	Method	14	60	4.254	p < 0.001 (***)
	Soil type:Method	14	60	0.062	p > 0.05

In particular, for general bacterial 16S rRNA gene (Table 2.13) the original Brady method on 100 g of soil (Method B100) and the FastDNA™ Spin kit on 0.5 g of soil (Method F) were significantly different to most of the other methods (Tukey's HSD $p < 0.05$) (Table 3.13). The additional chemical pre-treatments (e.g. Chelex® 100 and PEG6000) to the scaled down Brady method on 10 g of soil were significantly different from the Brady method (Method A1) and the phosphate buffer treatment (Method D) (Tukey's HSD $p < 0.05$). The lysis methods applied to plugs were not significantly different between each other and relative equal treatment on soil not encapsulated in agarose plugs (Tukey's HSD $p > 0.05$) (Table 2.13).

Table 2.13 Identification of significantly different methods of extraction for general bacterial 16S rRNA gene from both soils. Tukey's HSD; Dark Blue = same method; light blue = no significant difference between methods p > 0.05, red = significant difference between methods p < 0.05.

GENERAL BACTERIAL 16S rRNA		B100	A1	A2	B	C	D	E	F	G	H	I	J	K	L	M
		Brady100	Brady	Brady (2X)	Blending + Brady	TPNE	Phosphate buffer	Freeze-boil	FastDNA Spin Kit for soil	Chelex + Brady	Sodium deoxycholate +Brady	PEG6000 + Brady	Plugs (enzymatic lysis)	Plugs (Brady)	Plugs (blending + Brady)	Plugs (phosphate buffer)
B100	Brady100															
A1	Brady															
A2	Brady (2X)															
B	Blending + Brady															
C	TPNE															
D	Phosphate buffer															
E	Freeze-boil															
F	FastDNA Spin Kit for soil															
G	Chelex + Brady															
H	Sodium deoxycholate +Brady															
I	PEG6000 + Brady															
J	Plugs (enzymatic lysis)															
K	Plugs (Brady)															
L	Plugs (blending + Brady)															
M	Plugs (phosphate buffer)															

For actinobacterial 16S rRNA gene the FastDNA™ Spin kit results were significantly different from all the methods (Tukey's HSD $p < 0.05$) (Method A2; Table 2.14). The scaled down Brady method (Method A1) resulted significantly different (Tukey's HSD $p < 0.05$) to both physical lysis Methods F and E (i.e. the commercial kit and freezing-boiling). It was also significantly different (Tukey's HSD $p < 0.05$) from methods which included a chemical pre-treatment (Methods G, H and I) and inclusion of soil in plugs (Method K) followed by the main lysis step performed with the same buffer (Table 2.14).

Table 2.14 Identification of significantly different methods of extraction for actinobacterial 16S rRNA gene from both soils. Tukey's HSD; Dark Blue = same method; light blue = no significant difference between methods p > 0.05, red = significant difference between methods p < 0.05.

ACTINOBACTERIA 16S rRNA		B100	A1	A2	B	C	D	E	F	G	H	I	J	K	L	M
		Brady100	Brady	Brady (2X)	Blending + Brady	TPNE	Phosphate buffer	Freeze-boil	FastDNA Spin Kit for soil	Chelex + Brady	Sodium deoxycholate +Brady	PEG6000 + Brady	Plugs (enzymatic lysis)	Plugs (Brady)	Plugs (blending + Brady)	Plugs (phosphate buffer)
B100	Brady100															
A1	Brady															
A2	Brady (2X)															
B	Blending + Brady															
C	TPNE															
D	Phosphate buffer															
E	Freeze-boil															
F	FastDNA Spin Kit for soil															
G	Chelex + Brady															
H	Sodium deoxycholate +Brady															
I	PEG6000 + Brady															
J	Plugs (enzymatic lysis)															
K	Plugs (Brady)															
L	Plugs (blending + Brady)															
M	Plugs (phosphate buffer)															

For verrucomicrobial 16S rRNA gene, the original Brady method (Method B100) was significantly different (Tukey's HSD $p < 0.05$) from the optimised one on smaller amount of starting material (Method A1) and also from methods using different

chemical lysis buffer (Methods C, D and M) or physical (Methods A2, F, and L) and enzymatic lysis (Method J) (Table 2.15). The introduction of the blending step before lysis was mainly significantly different (Tukey's HSD $p < 0.05$) to the two physical lysis methods (Methods E and F) and the chemical pre-treatments to the Brady lysis (Methods G, H, I). TPNE and phosphate buffers did not produce results significantly different (Tukey's HSD $p > 0.05$) from the Brady method (Method A1) and the enzymatic digestion of soil in plugs. Interestingly, the phosphate buffer results were significantly different (Tukey's HSD $p < 0.05$) depending on whether used on soil constrained in plugs or on free soil directly.

Table 2.15 Identification of significantly different methods of extraction for verrucomicrobial 16S rRNA gene from both soils. Tukey's HSD; Dark Blue = same method; light blue = no significant difference between methods p > 0.05, red = significant difference between methods p < 0.05.

VERRUCOMICROBIA 16S rRNA		B100	A1	A2	B	C	D	E	F	G	H	I	J	K	L	M
		Brady100	Brady	Brady (2X)	Blending + Brady	TPNE	Phosphate buffer	Freeze-boil	FastDNA Spin Kit for soil	Chelex + Brady	Sodium deoxycholate +Brady	PEG6000 + Brady	Plugs (enzymatic lysis)	Plugs (Brady)	Plugs (blending + Brady)	Plugs (phosphate buffer)
B100	Brady100															
A1	Brady															
A2	Brady (2X)															
B	Blending + Brady															
C	TPNE															
D	Phosphate buffer															
E	Freeze-boil															
F	FastDNA Spin Kit for soil															
G	Chelex + Brady															
H	Sodium deoxycholate +Brady															
I	PEG6000 + Brady															
J	Plugs (enzymatic lysis)															
K	Plugs (Brady)															
L	Plugs (blending + Brady)															
M	Plugs (phosphate buffer)															

For NRPS extraction, the original method (Method B100) was the only one that showed a significant difference (Tukey's HSD $p < 0.05$) with all the other methods (Table 2.16). All the other methods did not show a significant difference (Tukey's HSD $p > 0.05$) between them, therefore they can be considered comparable.

Table 2.16 Identification of significantly different methods of extraction for NRPS gene from both soils. Tukey's HSD; Dark Blue = same method; light blue = no significant difference between methods p > 0.05, red = significant difference between methods p < 0.05.

NRPS		B100	A1	A2	B	C	D	E	F	G	H	I	J	K	L	M
		Brady100	Brady	Brady (2X)	Blending + Brady	TPNE	Phosphate buffer	Freeze-boil	FastDNA Spin Kit for soil	Chelex + Brady	Sodium deoxycholate	PEG6000 + Brady	Plugs (enzymatic lysis)	Plugs (Brady)	Plugs (blending + Brady)	Plugs (phosphate buffer)
B100	Brady100															
A1	Brady															
A2	Brady (2X)															
B	Blending + Brady															
C	TPNE															
D	Phosphate buffer															
E	Freeze-boil															
F	FastDNA Spin Kit for soil															
G	Chelex + Brady															
H	Sodium deoxycholate +Brady															
I	PEG6000 + Brady															
J	Plugs (enzymatic lysis)															
K	Plugs (Brady)															
L	Plugs (blending + Brady)															
M	Plugs (phosphate buffer)															

2.3.4.11 Efficiency of extraction of bacterial populations

FastDNA™ Spin kit (Method F) had the highest quantitation data. Therefore, it was considered as the reference for the maximum quantity of bacterial DNA that could be extracted to evaluate the efficiency of extraction of the other methods. The blending+Brady method (Method B) showed the highest efficiency of extraction for general bacterial DNA after Method F: in Reading soil 19 % of the total bacterial community was detected and 29.6 % in Warwick (Figure 2.16). The second most efficient method was either phosphate buffer (Method D, 13 %) or Brady (Method A1, 25 %) for Reading and Warwick soils respectively. Sodium deoxycholate (Method H), PEG6000 (Method I) and Brady100 (Method B100) showed a very poor efficiency (<1% for each method) (Figure 2.16). A similar pattern was identified for Actinobacteria 16S rRNA recovery. In this case, the method blending+Brady (Method B) could only recover 7 % and 13 % of the actinobacterial population when applied directly on Reading and Warwick soil respectively and 8% for both soils when using plugs. However, for Warwick soil, the blending pre-treatment did not increase the efficiency of the extraction compared to Method A1 (13 % in both cases). Even in the case of verrucomicrobial population, the methods that extracted the most after FastDNA™ are Brady (Method A) with 25.5 % and 38 % and blending+Brady (Method B) with 32 % and 34 % for Reading and Warwick soils respectively. TNPE and phosphate buffer also achieved 12-18 % efficiency and enzymatic treatment of soil plugs 12 % in Reading and 27.5 % in Warwick soils.

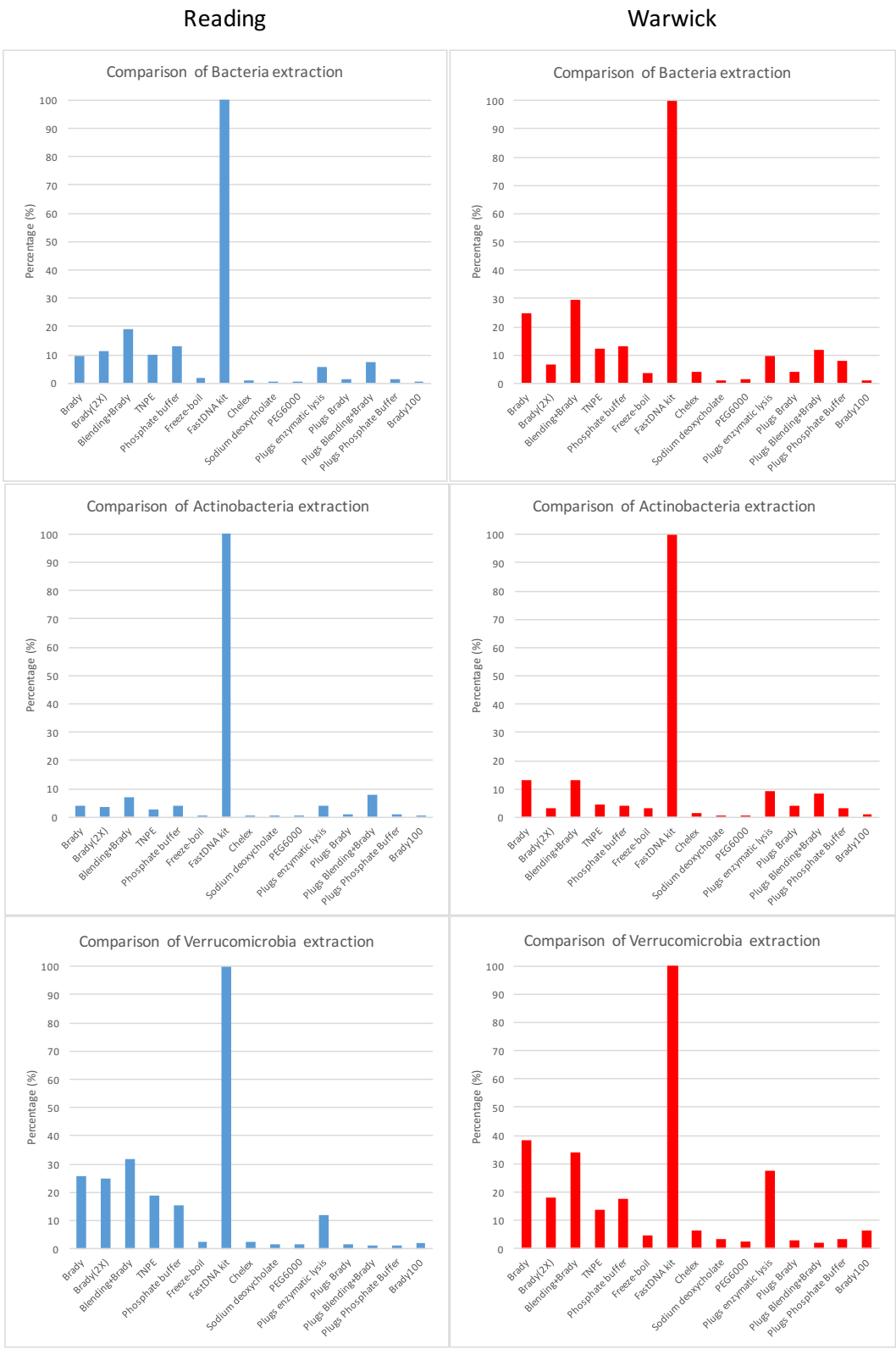


Figure 2.16 Comparison of extraction methods for recovery of Bacterial population and in particular Actinobacteria and Verrucomicrobia phyla. The amplification obtained for samples extracted with FastDNA™ Spin kit was considered as the maximum amplification (100%) and all the other methods were calculated as a percentage.

For each treatment, the relative abundance (%) of each phylum group was calculated against the general bacterial 16S rRNA gene counts for the same treatment. For both soils, Actinobacteria and Verrucomicrobia together represented an average of 51-56% of the total community recovered (Figure 2.17). Method B100 was excluded from this analysis as the general bacterial 16S rRNA gene detection failed.

Figure 2.17 Relative abundance (%) of Actinobacteria (orange), Verrucomicrobia (yellow) and other bacteria (green) for each DNA extraction method in Reading and Warwick soils. 100 % represent the general bacterial 16S rRNA genome equivalents obtained for each specific method of extraction.

3.3.2.14 Efficiency of extraction of NRPS genes

The FastDNA™ Spin kit (Method F) was the method with the highest counts for NRPS gene, therefore it was considered as a reference for the maximum amount that could be extracted from soil. The majority of the methods tested extracted an average of 10-20 % of the total NRPS present in the soil (Figure 2.18). Freezing-boiling (Method E), chemical pre-treatments such as Chelex® 100 (Method G) and sodium deoxycholate (Method H) and inclusion of soil in plugs treated with Brady lysis buffer (Method K) only extracted between 2 % and 5 % of the functional gene of interest.

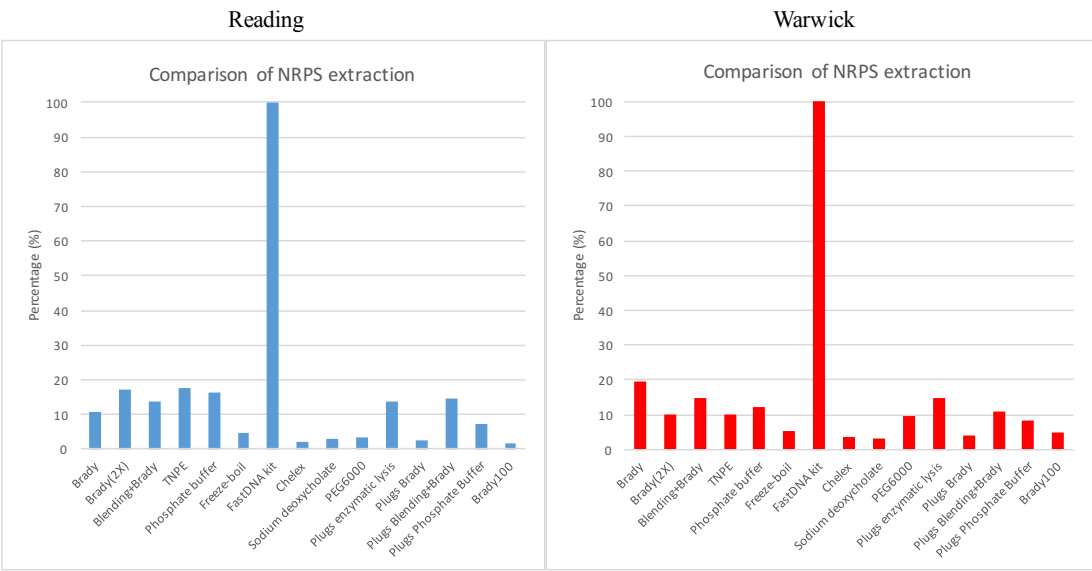


Figure 2.18 Comparison of extraction methods for recovery of NRPS genes. The amplification obtained for samples extracted with FastDNA™ Spin kit was considered as the maximum amplification (100%) and all the other methods were calculated as a percentage.

3.4 Discussion

The aim of the current study was to optimise and compare DNA extraction and employ a range of analyses to study the impact of extraction techniques. DNA extraction methods were compared and evaluated not only on the basis of the DNA yield but also on the quality of the DNA recovered in terms of size and purity. The comparison of the extracted eDNA samples showed that physical methods such as FastDNA™ Spin kit (Method F) allowed the recovery of medium-high DNA yield with a low concentration of molecular inhibitors such as humic acids (Figures S2.1 and S2.2). However, the DNA size was not sufficient for the final aim of this study as it was sheared (< 40 kb) and could not be used for metagenomic library preparation. All chemical methods of extraction instead resulted in the recovery of higher molecular weight DNA (> 40 Kb) but with a lower lysis efficiency and a higher concentration of molecular inhibitors (Table S2.1 and Figures S2.1 and S2.2). The co-extraction of humic acids and other molecular inhibitors is an important parameter to consider when preparing eDNA for metagenomic library preparation as those compounds could cause difficulties in cloning the eDNA into the vector and transforming the host (see review: Schrader *et al.* (2012) and Tebbe & Vahjen (1993)).

Comparison of general bacterial population recovery enabled evaluation of differences in the DNA extraction efficiency and potential inhibition caused by the co-extracted inhibitors. FastDNA™ spin kit was the most efficient and reproducible method in recovering bacterial DNA, while the original Brady (Method B100), which is currently used to prepare eDNA for metagenomic library to capture antibiotic clusters (Brady, 2007), was the most variable and the least efficient. Method B100 also presented the highest concentration of humic acids, which affected the amplification and detection of bacterial 16S rRNA genes underestimating the size of the bacterial population in both soils. The optimisation of the scaled-down version of this method (Method A1 and related variants) gave an improvement in both reproducibility and sensitivity. The presence of less starting material allowed a better mixing and diffusion of the lysis buffer inside the soil samples increasing the lysis and recovery of purer eDNA. The other chemical lysis methods that were tested (Methods C and D) achieved comparable efficiency results to Method A1. TNPE and phosphate buffers used in Method C and D were previously reported to result in low-medium purity and medium-high yield of extraction of DNA because of the chelating effect of EDTA and sodium ions present in those buffers which protect the DNA from adsorption to soil particles and reduce the efficiency of restriction enzymes (Selenska & Klingmuller, 1991, Paget *et al.*, 1992, Krsek & Wellington, 1999). In the current study, Method D showed medium-high yield but low-medium purity (which were comparable to method A1), while Method C showed low yield but medium-high purity. Different studies showed that physical lysis is a more efficient lysis than chemical treatment especially for spores (Moré *et al.*, 1994, Krsek & Wellington, 1999, Robe *et al.*, 2003), therefore allowing the recovery and detection of a higher proportion of the bacterial population. The freezing-boiling method (Method E) recovered pure DNA preserving the size but was less efficient. The introduction of a gentle physical lysis pre-treatment such as a short blending step combined to a chemical lysis (Method B) also proved to be efficient together preserving the size of the DNA recovered. The DNA yield of this combined method was higher than Method A1 and the purity was comparable.

Method G and derivative H and I recovered DNA from part of the bacterial population and potentially relic DNA, proving to be less efficient than the other methods, extracting very low amount of DNA and high concentration of other compounds. Carini *et al.* (2016) showed the biases related to relic DNA, therefore DNA related to dead cells, in the estimation of bacterial population diversity. However, for the

purpose of the current study the source of DNA (either from dead or living cells) was not important as even relic DNA can contribute to the identification of novel functional genes of interest related to secondary metabolite diversity.

The plug system for *in situ* lysis of cells was previously developed as an indirect method of extraction and performed on cells isolated from soil (Liles *et al.*, 2008). HMW DNA ranging from 20 kb to 1 Mb was recovered from the cellular fraction and cloned into BAC vectors creating a metagenomic library (Liles *et al.*, 2008). The current study adapted the enzymatic lysis performed by Liles *et al.* (2008) and tested other different lysis methods on soil plugs for the first time. Results showed that on plug-encapsulated soil, the most efficient lysis approach was the enzymatic one. Krsek & Wellington (1999) previously observed that lysozyme not only acted as a lysing agent on the cells' membrane, but it also interacted with humic acids forming a precipitate. Recent studies on extracellular protein activity associated with organic colloids and clay in soil also showed that humic acids can encapsulate lysozyme inactivating its binding site under certain pH and ionic conditions promoting electrostatic attraction (Li *et al.*, 2013). In the current study, a prolonged incubation of the plugs with the lysozyme solution might have contributed to a better diffusion inside the plugs increasing the extraction efficiency. Amongst the chemical lysis based extractions, the blending pre-treatment before the formation of the plugs improved the extraction efficiency compared to the sole chemical lysis. For all eDNA extraction methods in plugs a medium-high yield and a low-medium purity were observed.

In general, a difference in the chemical extraction efficiency was observed between soils likely due to different characteristics affecting the lysis steps. Indeed, amongst the chemical lysis methods only Method B100 showed a significantly different efficiency in Warwick soil, while for Reading soil differences were detected between most of the chemical lysis protocols. Differences in the soil texture (Warwick is classified as sandy silt loam, while Reading as sandy loam) might contribute to different interactions with chemicals affecting the DNA extraction in terms of lysis and co-extraction of inhibitors as already described by Zhou *et al.* (1996).

Potential bias towards particular groups of bacteria such as Actinobacteria and Verrucomicrobia which were identified as key drivers for secondary metabolites (Chapter 4) were tested. The detection of the actinobacterial population in Warwick soil was not significantly different between chemical lysis while the detection of the verrucomicrobial population showed significant differences according to the chemical

lysis performed. In the case of Reading soil, both populations were affected by the chemical lysis applied. The physical characteristic of the soil might have influenced the release of cells from soil particles or aggregates. In general, the combination of chemical and physical lysis increased the efficiency for Actinobacteria extraction, whose spores are usually recalcitrant to gentle lysis methods (Frostegard *et al.*, 1999). All the optimized chemical extraction methods showed an average of up to 25 % of Verrucomicrobia in the total population, which is in accordance with recent studies presented by Bergmann *et al.* (2011). These results showed that although chemical lysis methods did not achieve the maximum DNA extraction efficiency, the representation of the bacterial population was not biased against the taxa of interest.

Also presented here were novel qPCR assays to quantify NRPS and PKS genes in soils using degenerate primers which had been used previously to screen metagenomic libraries (Chapter 5) and to assess the biosynthetic potential diversity of different environments (Chapter 4). The low efficiency of the optimised assays for NRPS and PKS primers could be related to a combination of multiple factors that might affect the qPCR reaction kinetic such as primer degeneracy, primer mismatches and amplicon size (Bru *et al.*, 2008, Brankatschk *et al.*, 2012, Debode *et al.*, 2017). Both primer sets contained multiple degenerate bases and amplified products whose size is longer than the optimal one for qPCR assay of approximately 150 bp. The amplicon size might cause a dissociation of the amplicon strand from the template before they are completely copied reducing the available template for the next cycle preventing denaturation of the double-strand and impeding primer binding. Although the optimized conditions showed a low efficiency, both assays still produced reproducible standard curves with a constant interval of cycles between amplification curves in multiple runs. The amplicons of the standards presented the correct size and with no amplification of unspecific target or primer dimers (Figures S2.3 and S2.4).
Potential co-extraction of inhibitors was reported for most of the eDNA samples obtained with chemical lysis methods (Table S2.1), possibly affecting the amplification, particularly in the PKS assay. It is known that molecular inhibitors present in environmental samples as well as the possible introduction of inhibitors from the reagents used during the extraction could affect the PCR amplification interfering with the DNA target, binding the polymerase or restricting access to its co-factors (see review: Schrader *et al.* (2012)). In the particular case of the current PKS

assay on eDNA samples, the most likely effect of the inhibitors was likely to be related to interference with primer binding. Additionally, for both NRPS and PKS genes only an absolute abundance of gene copies could be calculated as no reference value is available to normalise count data to quantify the number of clusters present in the soil as the number of domains present in biosynthetic clusters varies. These assays could then be used to estimate the biosynthetic potential richness of different environments helping with the identification of potential hotspots for mining of novel antimicrobials. For this purpose, it is important to consider the lower prevalence of PKS genes in microbial genomes (Wang *et al.*, 2014), therefore in the environment, in comparison to NRPS genes. The current study aligned with PKS absolute abundance values previously reported in tropical soil by Le *et al.* (2014) using PKS degenerate primers specifically designed for qPCR assay. In Le *et al.* (2014) study, PKS type I ranged from 1.7×10^6 to 4.6×10^6 and type II from 2.4×10^5 to 1.5×10^6 copies per gram of soil, while in the current study the PKS absolute abundance ranged from 4.1×10^6 to 5.9×10^6 copies per gram of soil in Reading and 1.36×10^6 to 2.7×10^6 copies per gram of soil in Warwick soils. The current NRPS assay instead detected a higher number of NRPS genes than PKS genes in soil, with 2.3×10^8 to 4.1×10^8 NRPS copies per gram of soil in Reading and 1.9×10^8 to 3.7×10^8 NRPS copies per gram of soil in Warwick soil.

In conclusion, this study provided the development of novel quantitation assays for NRPS and PKS genes and a comparison between optimised methods to recover HMW for the preparation of metagenomic libraries for the capturing of large biosynthetic clusters. The scaled down version of the already widely used method developed by Brady (2007) resulted in the optimal recovery of DNA for construction of metagenomic libraries. This study has provided a method which is both suitable for library preparation but has not compromised diversity and efficiency of extraction compared to previous methods. The introduction of the gentle physical pre-treatment to method A1 and the encapsulation of soil in plugs could also help increase the DNA availability without compromising the DNA size or purity.

The identification of optimised methods to extract HMW DNA from different soils and to evaluate NRPS and PKS richness of soil using qPCR assay allowed the exploration of amendment conditions to improve the recovery of the genes of interest

(NRPS and PKS) for the creation of enriched metagenomic library. These amendment conditions were tested as described in Chapter 3.

Chapter 3

Amendment of Cuban soil to improve the recovery of NRPS and PKS gene clusters

3.1 Introduction and aims

Actinobacteria are the best characterised natural antimicrobial producers, in particular *Streptomyces* account for 80 % of known antimicrobials produced by Actinomycetes (Berdy, 1974, Watve *et al.*, 2001). The life cycle of these bacteria typically involves the formation of spores and germination when environmental conditions are favourable (Mayfield *et al.*, 1972, Kieser *et al.*, 2000). Studies have shown that DNA extraction from bacterial spores is not efficient and physical lysis usually recovers higher yields of DNA (Herron & Wellington, 1990, Frostegard *et al.*, 1999). Different strategies to promote germination and bacterial growth have been studied, for example, enrichment of soil with various chitin-rich compounds can modify the bacterial community composition increasing some groups such as Actinobacteria (Vionis *et al.*, 1996, Metcalfe *et al.*, 2002, Jacquiod *et al.*, 2013, Johnson-Rollings *et al.*, 2014). Chitin is one of the most abundant biopolymers on earth and it is an essential source of carbon and nitrogen in the ocean (Beier *et al.*, 2011) and soil (reviews: (Souza *et al.*, 2011, Beier & Bertilsson, 2013). The natural degradation of this polymer is usually performed in soil by secreted chitin-degrading enzymes (chitinases) of bacterial or fungal origins (Metcalfe *et al.*, 2002, Lindahl & Finlay, 2006, Hjort *et al.*, 2010). The degradation of this linear polysaccharide chain of N-acetyl-D-glucosamine moieties linked by β-1,4-glucosidic bonds (GlcNAc) usually produces smaller molecules of glucose, acetate, glucosamine and N-acetylglucosamine which can then be incorporated directly into the microbial biomass or mineralised and reintegrated in the carbon or nitrogen cycles (Gooday, 1990, Olander & Vitousek, 2000). Soil enrichment strategies have been used to promote the recovery of biotechnologically interesting novel enzymes such as chitinases (Kielak *et al.*, 2013) (Cretoiu *et al.*, 2015) and to recover rare microbial genomes related to tolerance to heavy metals, salt or high temperature conditions (Delmont *et al.*, 2015).

The aim of this Chapter was to evaluate the effect of chitin amendment for the recovery of HMW DNA for future metagenomic library creation. In particular, the assessment of the effectiveness of the chitin amendment in the recovery of NRPS and PKS genes from Cuban soil for future potential discovery of novel antimicrobial compounds was of interest. The hypothesis was that the introduction of nutritional sources, resulting from chitin degradation, would promote germination of spores which are usually difficult to lyse with chemical lysis method for HMW DNA recovery. This will result in more accessible DNA from antimicrobial-producing groups, such as Actinobacteria, consequently leading to a higher abundance of NRPS and PKS genes. Cuban soil has been reported to be a potential hotspot for NRPS and PKS diversity (Chapter 4) and it would be interesting to try to exploit its diversity using functional metagenomics.

3.2 Material and methods

3.2.1 Soil sample properties

The soil used in the study was previously collected from Cayo Blanco, Cuba (latitude: 23.2033; longitude: -81.0396) and stored at -20 °C until usage (Table 3.1).

Table 3.1 Summary of characteristics of soils used in this study. TOC = Total Organic Carbon (%); TN = Total Nitrogen (%); EC = Electrical conductivity (µS/cm).

Soil sample	TC (%)	TOC (%)	TN (%)	pH	EC (µS/cm)	Sand (%)	Silt (%)	Clay (%)	Soil texture
Cayo Blanco (Fir-Shrub)	11.1	2.7	0.068	8.2	6360	87.72	11.54	0.74	Sand

3.2.2 Microcosm conditions

Two conditions were tested (chitin amended and unamended) with nine replicate microcosms prepared in 500 mL glass bottles containing 100g each of soil. Chitin powder was added to the amended microcosms at a final concentration of 1 % (w/w). The initial moisture content (25 % w/w) was calculated according to the Equation 3.1 and then adjusted with the addition of 8 mL of deionized sterile water to avoid severe changes of moisture during the incubation.

$$MC = \left(1 - \frac{DS}{WS}\right) * 100$$

Equation 3.3 Moisture content (MC) calculation. DS = soil dry weight (g); WS = soil wet weight (g).

All microcosms were incubated as described by Johnson-Rollings (2012) at 28 °C for seven days in the dark in the presence of wet paper to prevent high evaporation in the microcosms. Every two days, the microcosms were gently mixed by hand to promote aeration of soil. After the incubation, microcosms were collected and snap-frozen using dry ice and stored at -20 °C until DNA extraction was performed.

3.2.3 DNA extraction

Three microcosms were pooled to obtain bigger samples for different DNA extraction methods for future metagenomic and metaproteomic analysis. Three samples were obtained for each condition (chitin amended and unamended soil) and DNA was extracted in singular replica using FastDNA™ Spin kit (0.5 g of soil), the scaled-down optimised Brady method on 10 g of soil and the phosphate buffer method also on 10 g of soil (Figure 4.1). DNA extraction methods (method IDs A1, D and F respectively for Brady, phosphate buffer and FastDNA™ Spin kit) are described in detail in Chapter 2 Section 2.2.2.

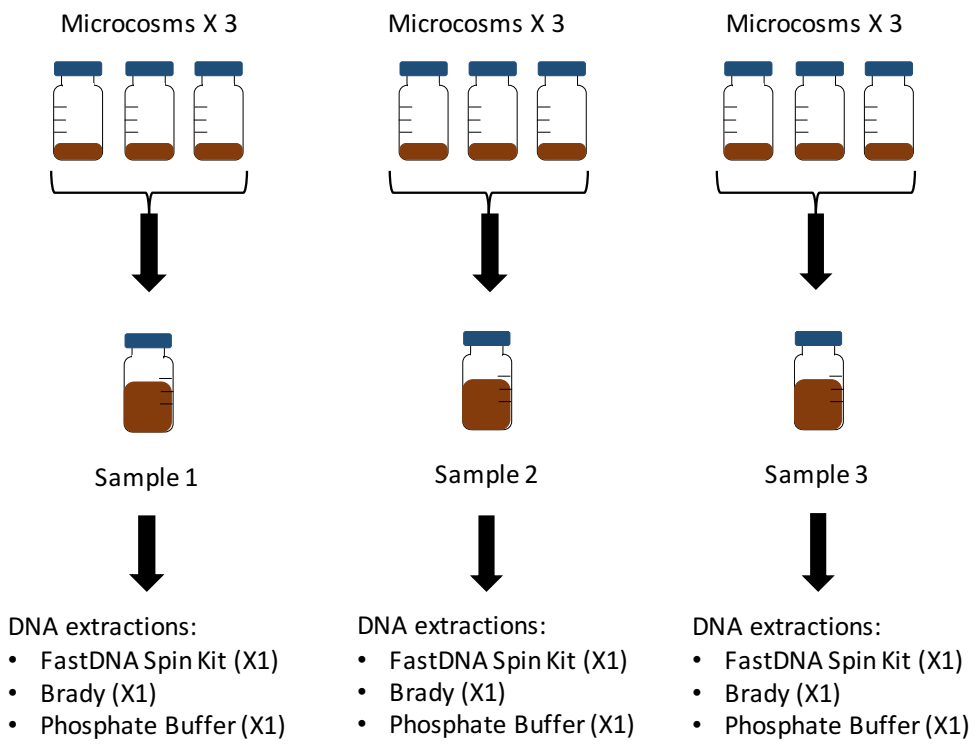


Figure 3.1 Schematic representation of microcosms set up and DNA extraction for each condition tested (chitin amended and unamended). Three microcosms were mixed to form a sample. Three samples were obtained and extracted using the commercial kit FastDNA™ Spin kit, the Brady extraction and the phosphate buffer lysis extraction.

3.2.4 Quantification of DNA extracted using qPCR

Microcosm bacterial diversity was investigated using qPCR assays as described in Chapter 2 Section 2.1.4. General bacterial 16S rRNA genes as well as phyla-specific 16S rRNA genes (i.e. Actinobacteria and Verrucomicrobia) were amplified using the optimised conditions described in Chapter 2 in order to evaluate the different abundance of each target between the chitin amended and unamended soils. NRPS and PKS assays were also performed according to the optimised conditions described in Chapter 2.

3.2.5 Statistical analysis

Data recovered from qPCR assays were normalised as described in Chapter 2 Section 2.1.5 according to elution volume, weight of soil used for the extraction and the average copy number of 16S rRNA genes for each group.

Quantified data were then Log10 transformed to normalize the qPCR counts. ANOVA and Tukey's HSD tests were performed in Rstudio (version 0.99.902) in order to test variance within and between groups of extraction and amendment.

3.3 Results

3.3.1 Comparison between extraction methods in each condition

In both conditions tested (unamended and chitin amended), extraction methods were all significantly different in the mean qPCR count (one-way ANOVA, Tables 3.2 and 3.3). The three methods were all significantly different for all the 16S rRNA genes detected in both conditions (Tukey's HSD $p<0.05$). For NRPS genes, Brady and phosphate buffer methods were significantly different in chitin amended soil but not in unamended soil. For PKS genes, phosphate and Brady extraction methods were not significantly different in both conditions.

Table 3.2 Comparison of DNA extraction methods in unamended soil for each target (general bacterial 16S rRNA gene, actinobacterial 16S rRNA gene, verrucomicrobia 16S rRNA gene, NRPS and PKS genes). (One-way ANOVA; significance code recorded in brackets (***) p=0.001).

QPCR target	Degrees of freedom	Residuals	F statistic	P
General bacterial 16S rRNA gene	2	6	190.8	p<0.001 (***)
Actinobacterial 16S rRNA gene	2	6	252.9	p<0.001 (***)
Verrucomicrobial 16S rRNA gene	2	6	279.4	p<0.001 (***)
NRPS	2	6	73532	p<0.001 (***)
PKS	2	6	960.2	p<0.001 (***)

Table 3.3 Comparison of DNA extraction methods in chitin amended soil for each target (general bacterial 16S rRNA gene, actinobacterial 16S rRNA gene, verrucomicrobia 16S rRNA gene, NRPS and PKS genes). (One-way ANOVA, significance code recorded in brackets (***) p=0.001, (**) p=0.01).

QPCR target	Degrees of freedom	Residuals	F statistic	P
General bacterial 16S rRNA gene	2	6	148.2	p<0.001 (***)
Actinobacterial 16S rRNA gene	2	6	214.4	p<0.001 (***)
Verrucomicrobial 16S rRNA gene	2	6	56.87	p<0.001 (***)
NRPS	2	6	21.74	p<0.05 (**)
PKS	2	6	20404	p<0.001 (***)

3.3.3 Interaction of enrichment conditions and extraction method

For bacterial 16S rRNA gene detection (Figure 3.2) there was a significant difference between all methods (two-way ANOVA $F_{2,12} = 337.59$, $p < 0.001$*** and Tukey's HSD $p < 0.05$). The counts obtained for the chitin amended microcosms were significantly different from the ones obtained for the unamended condition (two-way ANOVA $F_{1,12} = 312.54$, $p < 0.001$***). The interaction between the method of extraction and the treatment on the sample (chitin amendment) was also significantly different (two-way ANOVA $F_{2,12} = 11.94$, $p < 0.05$ **). Interestingly, the only two pairs that were not significantly different were FastDNA™ Spin kit:unamended-Brady:chitin amended (Tukey's HSD $p = 0.99$) and Brady:unamended-Phosphate:chitin amended (Tukey's HSD $p = 0.13$).

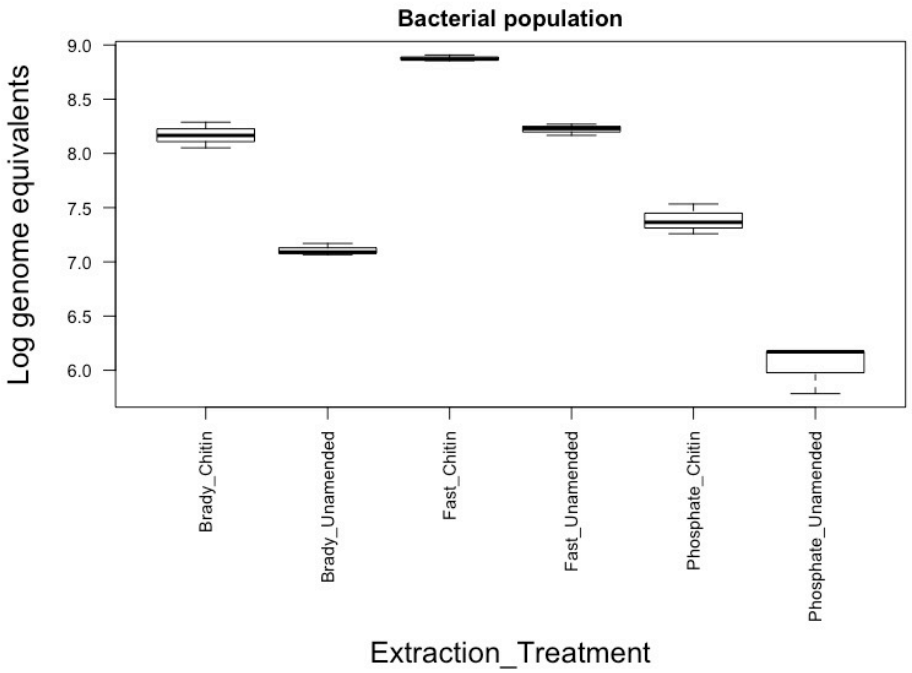


Figure 3.2 Variation of general bacterial 16S rRNA gene genome equivalents in Cuban soil according to the amendment and extraction method tested. Extraction methods: Brady, FastDNA™ Spin kit and Phosphate buffer; Treatment: chitin amended and unamended.

For Actinobacterial 16S rRNA gene (Figure 3.3), there was a significant difference between all methods of extraction (two-way ANOVA $F_{2,12} = 467.1$, $p < 0.001$*** and Tukey's HSD $p < 0.05$). The treatments were significantly different (two-way ANOVA $F_{1,12} = 7.06$, $p<0.05$*). The interaction between methods of extraction and treatments was also significantly different (two-way ANOVA $F_{2,12} = 4.34$, $p<0.05$*), even though the same method applied in different conditions was not significantly different (e.g. Brady on unamended or chitin amended soil was not significantly different, Tukey's HSD $p > 0.05$).

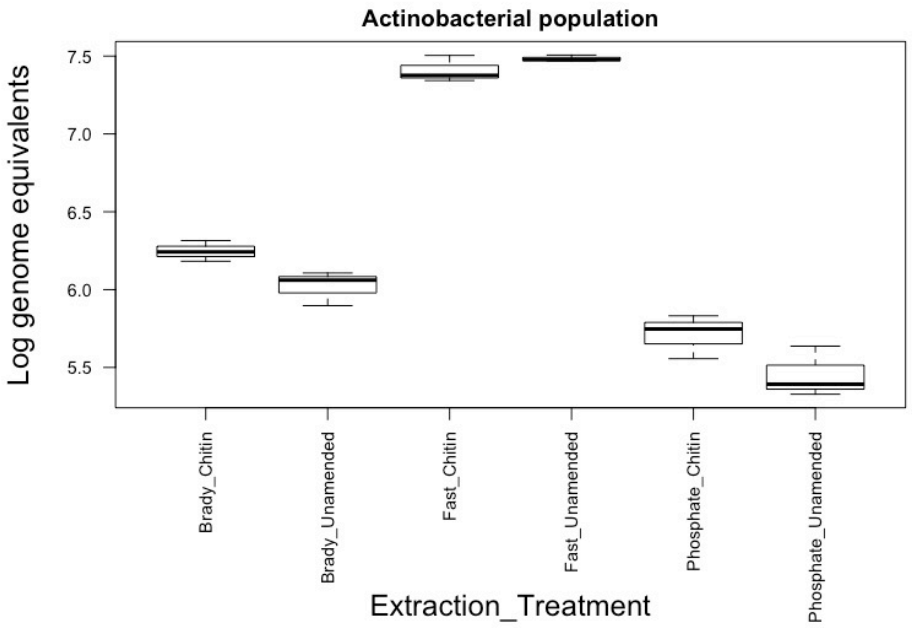


Figure 3.3 Variation of Actinobacterial 16S rRNA genome equivalents in Cuban soil according to the amendment and extraction method tested. Extraction methods: Brady, FastDNA™ Spin kit and Phosphate buffer; Treatment: chitin amended and unamended.

For Verrucomicrobial 16S rRNA (Figure 3.4), there was a significant difference between the two conditions tested (two-way ANOVA $F_{1,12} = 55.8$, $p < 0.001$***) and the methods (two-way ANOVA $F_{2,12} = 249.5$, $p < 0.001$*** and Tukey's HSD $p < 0.05$). The interaction between the treatments and the extraction methods was also significantly different (two-way ANOVA $F_{2,12} = 12.7$, $p < 0.05$**). In particular, all pairs of extraction:treatment were significantly different (Tukey's HSD $p < 0.05$) except the FastDNA™ kit which was not significantly different in both conditions (unamended and chitin amended) (Tukey's HSD $p = 0.99$) and the Brady method performed on the unamended soil in comparison to the phosphate method on chitin amended soil which showed comparable results (Tukey's HSD $p = 0.23$).

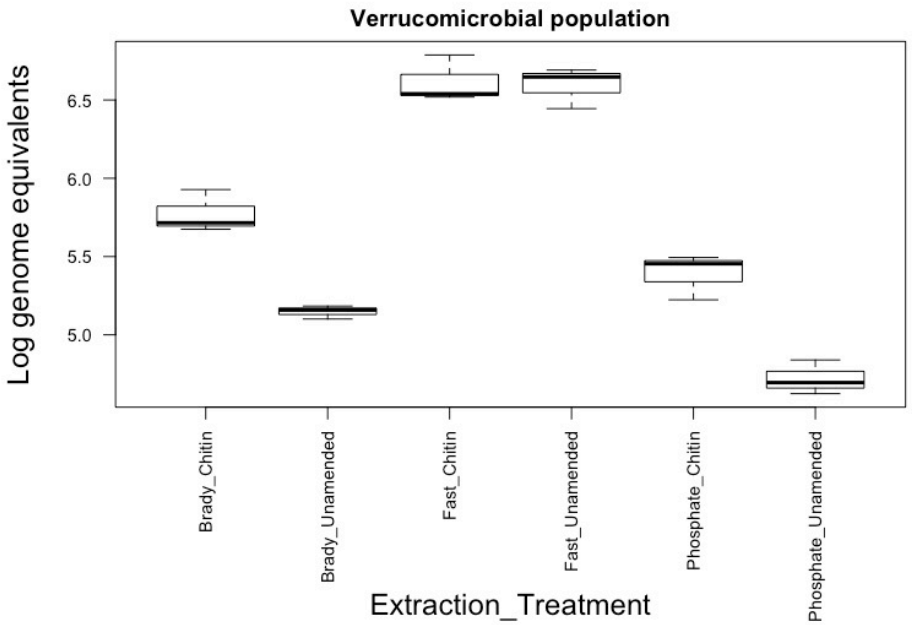


Figure 3.4 Variation of Verrucomicrobial 16S rRNA genome equivalents in Cuban soil according to the amendment and extraction method tested. Extraction methods: Brady, FastDNA™ Spin kit and Phosphate buffer; Treatment: chitin amended and unamended.

For NRPS gene detection (Figure 3.5), all methods of extraction were significantly different (two-way ANOVA $F_{2,12} = 910$, $p < 0.001$*** and Tukey's HSD $p<0.05$). The chitin amendment showed an impact in detection of NRPS genes in comparison to the unamended soil (two-way ANOVA $F_{1,12} = 2664.4$, $p < 0.001$***). The interaction of method and treatment was also significantly different (two-way ANOVA $F_{2,12} = 489.6$, $p < 0.001$***) for the majority of pairs. Interestingly, the Brady extraction from amended soil showed comparable results to the FastDNA™ Spin kit performed on the unamended soil (Tukey's HSD $p = 0.99$).

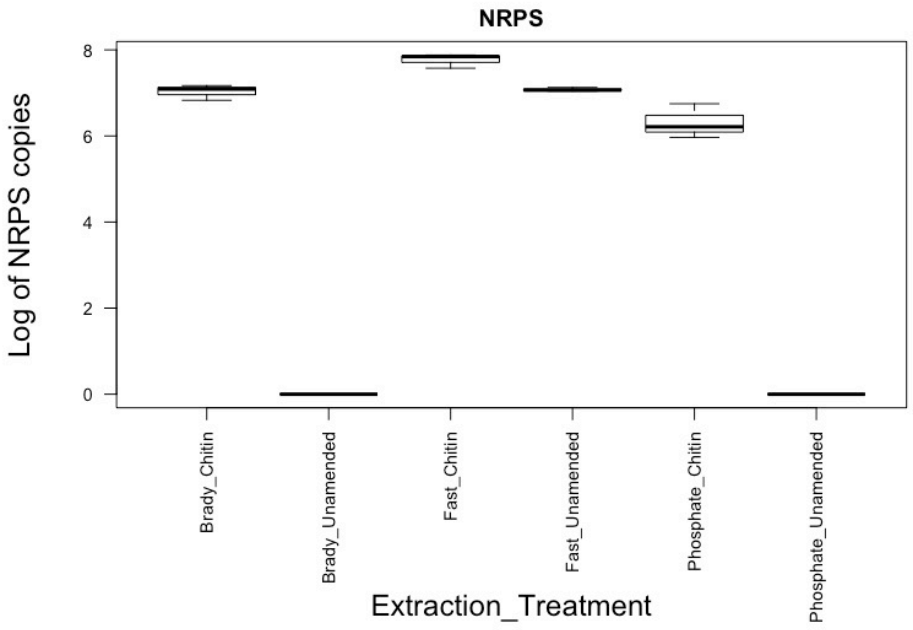


Figure 3.5 Variation of NRPS gene qPCR counts in Cuban soi according to the amendment and extraction method tested. Extraction methods: Brady, FastDNA™ Spin kit and Phosphate buffer; Treatment: chitin amended and unamended.

For PKS gene detection (Figure 3.6), only the methods of extraction contributed significantly (two-way ANOVA $F_{2,12} = 3647.9$, $p < 0.001$***), with FastDNA™ spin kit significantly different (Tukey's HSD $p < 0.05$) to the comparable Brady and phosphate extraction method (Tukey's HSD $p = 1$).

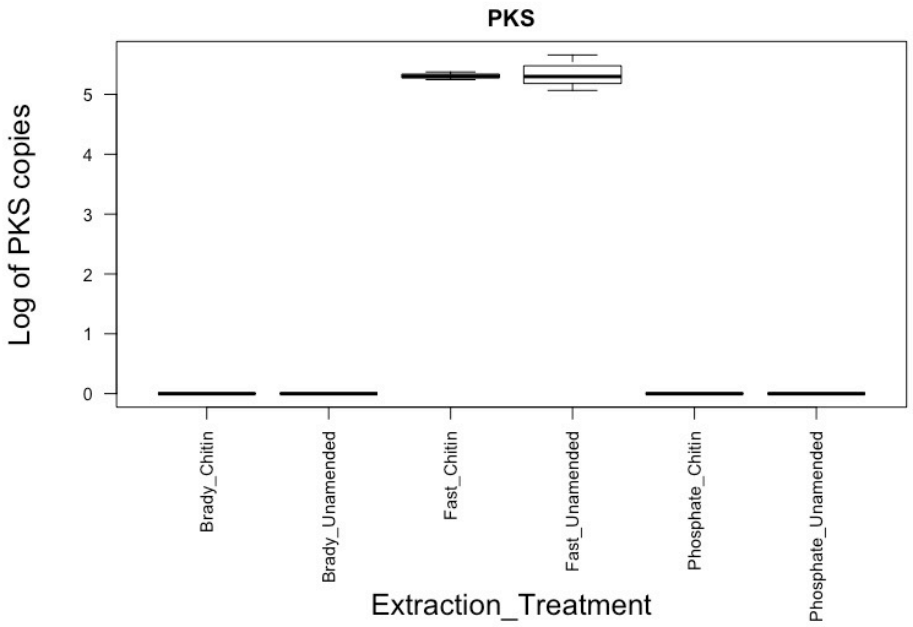


Figure 3.6 Variation of PKS gene qPCR counts in Cuban soil according to the amendment and extraction method tested. Extraction methods: Brady, FastDNA™ Spin kit and Phosphate buffer; Treatment: chitin amended and unamended.

3.3.4 Efficiency of extraction and detection of bacterial populations and functional genes

FastDNA™ Spin kit was the most efficient method in recovering bacterial DNA and functional genes of interest (Table 3.4) in both conditions (unamended and chitin amended). For each treatment, Brady and phosphate buffer methods were compared to the extraction kit which was considered as the reference maximum amount of DNA that could be extracted for each target (Table 3.5). This comparison showed that in the unamended soil the Brady method extracted only 7.7 % of the general bacterial population and an even smaller amount of Actinobacteria and Verrucomicrobia (3.5 % each). Phosphate buffer was even less efficient, extracting only 0.7 % of the general bacterial fraction, 1 % of Actinobacteria and 1.3 % of Verrucomicrobia. NRPS and PKS genes were not detected from any unamended samples extracted with either of these gentle chemical lysis methods.

In comparison to FastDNA™ Spin kit, the Brady method on chitin amended samples extracted 20 % of the total bacterial population, 6.9 % of Actinobacteria, 14.3 % of Verrucomicrobia and 18.4 % of the NRPS genes. Phosphate buffer instead, extracted 3.3 % of the total Bacteria, 2.1 % of Actinobacteria, 5.9 % of Verrucomicrobia and 4.5 % of NRPS genes in presence of chitin. Both methods did not extract any PKS genes even after addition of chitin (Table 3.5).

Table 3.4 Average qPCR counts for all gene targets tested. G16S = General bacterial 16S rRNA gene; A16S = Actinobacterial 16S rRNA gene; V16S = Verrucomicrobial 16S rRNA gene. (16S rRNA data were corrected for copies number).

Method ID	Treatment	Average count G16S	Average count A16S	Average count V16S	Average count PKS	Average count NRPS
FastDNA™ Spin kit	Unamended	1.68×10^8	3.05×10^7	4.04×10^6	2.57×10^5	1.20×10^7
Brady	Unamended	1.29×10^7	1.07×10^6	1.41×10^5	0	0
Phosphate buffer	Unamended	1.20×10^6	2.98×10^5	5.36×10^4	0	0
FastDNA™ Spin kit	Chitin amended	7.56×10^8	2.59×10^7	4.30×10^6	2.05×10^5	6.10×10^7
Brady	Chitin amended	1.51×10^8	1.78×10^6	6.13×10^5	0	1.12×10^7
Phosphate buffer	Chitin amended	2.52×10^7	5.32×10^5	2.55×10^5	0	2.72×10^6

Table 3.5 Efficiency of DNA extraction method in comparison to FastDNA™ Spin kit for chitin amended and unamended soil. G16S = General bacterial 16S rRNA gene; A16S = Actinobacteria 16S rRNA gene; V16S = Verrucomicrobia 16S rRNA gene.

Method ID	Treatment	G16S (%)	A16S (%)	V16S (%)	PKS (%)	NRPS (%)
FastDNA™ Spin kit	Unamended	100.0	100.0	100.0	100.0	100.0
Brady	Unamended	7.7	3.5	3.5	0.0	0.0
Phosphate buffer	Unamended	0.7	1.0	1.3	0.0	0.0
FastDNA™ Spin kit	Chitin amended	100.0	100.0	100.0	100.0	100.0
Brady	Chitin amended	20.0	6.9	14.3	0.0	18.4
Phosphate buffer	Chitin amended	3.3	2.1	5.9	0.0	4.5

The chitin amendment showed an improvement in the detection of different bacterial groups and functional genes of interest for all methods of extraction tested. In particular, the FastDNA™ method in the presence of chitin allowed the recovery of 3.5X more bacteria and 4.1X more of NRPS genes comparing to the same method performed on the unamended samples.

The Brady method performed on chitin amended soil improved the detection of Bacteria (10.7X higher), Verrucomicrobia (3.3X higher) and NRPS genes (from no detection to 10^7 copies) than in the absence of chitin.

Phosphate buffer extraction was also more efficient on samples amended with chitin, recovering 20X more Bacteria and 3.8X more Verrucomicrobia than the same method applied on the unamended samples. Method D also recovered NRPS in chitin amended samples (from no detection to 10^6 copies).

4.4 Discussion

In this study, the newly optimised NRPS and PKS qPCR assays previously described in Chapter 2 were applied to investigate the potential of Cuban soil for biosynthetic genes richness. Comparison of the DNA extraction methods showed that the physical lysis performed using the commercial kit was the most efficient one and between the two chemical lysis methods, the Brady lysis was more efficient than phosphate buffer as already proved in Chapter 2.

The introduction of chitin in soil with incubation showed a positive effect in enhancing the number of bacteria detectable through qPCR assay when less efficient DNA extraction methods were used. This is of particular interest when HMW DNA has to

be recovered for the creation of functional metagenomic libraries as gentle chemical lysis protocols are needed to prepare high quality HMW DNA. Chitin is a natural biopolymer that can be degraded in soil by bacterial enzymes to increase the available nitrogen and carbon sources to support the living microbial community (Olander & Vitousek, 2000). The introduction of additional nutrients to the soil was proven to promote the growth of bacterial cells and the germination of spores (Mayfield *et al.*, 1972). Jacquiod *et al.* (2013) showed that the introduction of chitin shifted the microbial community diversity increasing chitin degrading taxa such as Actinobacteria, Beta-proteobacteria and Alpha-proteobacteria. Interestingly it had a beneficial effect promoting individuals belonging to actinobacterial genera (e.g. *Williamsia*, *Thermoleophilum* and *Rhodocista*) usually associated to the soil rare biosphere which are generally represented by a low relative abundance in soil metagenomes (Jacquiod *et al.*, 2013). Johnson-Rollings *et al.* (2014) also observed an increase of Actinobacteria, Proteobacteria and Bacteroidetes in chitin amended microcosms detecting more diversity in amended conditions. The same three phyla presented a higher relative abundance also in the rhizosphere of lettuce pots amended with chitin, which also showed a decrease in the Verrucomicrobia relative abundance in the amended conditions (Debode *et al.*, 2016). The current study did not show a shift in the prevalence of Actinobacteria or Verrucomicrobia phyla. However, a general increase of the bacterial population was observed, suggesting that chitin has an effect on the growth of different bacterial groups even though not specifically on the two phyla of interest of this study. This effect could be directly related to bacteria capable of degrading chitin or to cheaters which exploit nutrients made available by chitin degraders as suggested by Jacquiod *et al.* (2013). Interestingly, in the current study the presence of chitin promoted the recovery of a significantly higher number of NRPS genes, but not of PKS genes, suggesting that the chitin amendment had an impact on the microbial population fraction carrying NRPS genes but not PKS genes. In particular, for PKS genes, only eDNA samples recovered using the commercial kit showed a comparable target amplification in both conditions (amended and unamended soil), but did not present any amplification in samples extracted with gentle chemical lysis. It is unclear the reason why gentle chemical lysis presented difficulties in the PKS amplification assay. Possible co-extraction of inhibitors or residual compounds used during the extraction could interfere with either the DNA or the PKS primers not allowing amplification as observed also in Chapter 2. For NRPS

primers instead, a significant increase in the detection of NRPS genes was observed not only in samples extracted with the physical lysis method but also with gentle chemical lysis methods. The ability to increase the availability of NRPS genes, in particular when recovering HMW DNA with gentle chemical lysis methods for further applications such as metagenomic library preparation is of great interest for novel drug discovery programmes.

In conclusion, this study has demonstrated the merit of enhancing microcosms to promote changes in the bacterial functional prevalence to unveil the potential antimicrobial biosynthetic richness of the microbial population. The data presented here proved that incubation with an amendment is useful and improves the likelihood of a successful extraction of DNA even with gentle chemical lysis which are used for HMW DNA recovery for further metagenomic library preparation. In particular, the potential application of amendments to identified hotspots for specific bacterial phyla correlating with NRPS and PKS diversity would prove useful to enrich the recovered DNA for the gene of interest for metagenomic library exploitation.

Chapter 4

NRPS and PKS genes diversity in soil

4.1 Introduction and aims

The introduction of molecular analysis *in situ* and subsequentially NGS-based approaches opened new frontiers to genomics and in particular to metagenomic studies showing that a large proportion of the microbial community remains elusive to cultivation (Torsvik *et al.*, 1990, Handelsman *et al.*, 1998). Most of the known antibiotics are produced by the soil-dwelling Gram-positive bacteria belonging to the phylum Actinobacteria and in particular the genus *Streptomyces* (reviews: Watve *et al.* (2001), Berdy (2012)). However, recent genomic studies showed that also other bacteria belonging to Proteobacteria and less characterised phyla such as Bacteroidetes, Verrucomicrobia and Planctomycetes have the genetic potential for secondary metabolites biosynthesis in soil (Letzel *et al.*, 2013, Esmaeel *et al.*, 2016, Graça *et al.*, 2016). The majority of these potential clusters have not been studied and characterized experimentally yet, therefore there is still a limited knowledge of the real biosynthetic capacities of these groups.

The majority of the metagenomic studies related to secondary metabolites discovery in soil focused on the identification of potential hotspots for novel compounds on the basis of their richness in two of the main biosynthetic systems, PKS and NRPS (Reddy *et al.*, 2012, Charlop-Powers *et al.*, 2014, Charlop-Powers *et al.*, 2015). Soils with different characteristics were collected, DNA extracted and the biosynthetic potential was investigated through analysis of 454 pyrosequencing of adenylation (for NRPS) and ketosynthase domains (for PKS) amplified directly from the extracted eDNA (Charlop-Powers *et al.*, 2014, Charlop-Powers *et al.*, 2015) or from the eDNA captured in metagenomic libraries (Reddy *et al.*, 2012). These studies showed that both the geographic location and the soil characteristics play an important role in shaping and selecting the secondary metabolite biosynthetic potential in the soil microbiome in agreement with the Baas-Becking's hypothesis of global microbial

distribution "*everything is everywhere*, but *the environment selects*" (Baas Becking, 1934, De Wit & Bouvier, 2006).

The aim of this study was to use metagenomic approaches to investigate different natural soil environments for potential novel natural product genes for future targeted drug discovery screening programs. In particular, correlation studies between the functional genes related to the biosynthesis of polyketide (PK) and nonribosomal peptide (NRP) compounds and the phylogenetic marker 16S rRNA gene were investigated in order to provide insights into the microbial phylogenetic drivers of secondary metabolite diversity.

The selection of soil sites was influenced by our collaborative research work with local scientists who shared an interest in secondary metabolite biosynthetic diversity. In all cases permission from local governments were obtained for sampling and where necessary a license for collecting sensitive samples was obtained (Cuba). The diversity of soil characteristics presented in the method section reveals that a wide range of edaphic features were obtained with pH varying from 4.5 to 8.6. In addition, a preference was made for the selection of sandy dry soil with an alkaline nature but from a wide geographic location. This choice was based on previous work and results from Charlop-Powers *et al.* (2014) which reported an increase in the biosynthetic richness in samples with a dry biome due to the higher prevalence of Actinomycetes.

4.2 Materials and methods

4.2.1 Soil samples

Soil samples were collected from 13 sites across different countries from either bulk or rhizosphere soil (Table 4.1; Figure 4.1). Samples were stored at -20 °C until use and thawed at room temperature before extraction. Chemical and physical analyses of the soils were performed by YARA Analytical Service, LanCrop Laboratories, Grimsby, UK (Table 4.2).

Table 4.1 Soil samples locations (latitude and longitude GPS coordinates)

Soil ID n.	Soil sample	Latitude	Longitude	Country
1	Algerian B3	34.85	5.733333	Algeria
2	Algerian KI	33.3713397	6.8479682	Algeria
3	Algerian KII	33.3713397	6.8479682	Algeria
4	Antarctica Mars Oasis	-71.886	-68.260775	Antarctica
5	Cayo Blanco (Fir-Shrub)	23.2033	-81.0396	Cuba
6	Cayo Blanco (Shrub)	23.2047	-81.0398	Cuba
7	Iceland	64.257271	-21.144291	Iceland
8	Kilkenny	52.88614	-7.50723	Ireland
9	Sourhope	55.47	-2.2313	UK
10	South Tyrol	46.4982953	11.3547582	Italy
11	Trinidad	21.7960343	-79.9808143	Cuba
12	Tuscany	43.5333333	10.71666667	Italy
13	Warwick (Cryfield)	52.37622467	-1.569414139	UK

Table 4.2 Summary of characteristics of soils used in this study. TC = Total Carbon (organic + inorganic) (%); TOC = Total Organic Carbon (%); TN = Total Nitrogen (%); EC = Electrical conductivity (µS/cm);

Soil sample	TC (%)	TOC (%)	TN (%)	pH	EC (µS/cm)	Sand (%)	Silt (%)	Clay (%)	Soil texture
Algerian B3	1.6	0.8	0.024	8.1	2050	98.49	1.33	0.18	Sand
Algerian KI	1.7	0.9	0.040	8.1	2660	95.26	4.14	0.6	Sand
Algerian KII	2.1	0.7	0.021	8.1	2280	98.31	1.68	0.01	Sand
Antarctica Mars Oasis	0.6	0.6	0.052	8.1	2210	57.3	27.72	14.98	Sandy Loam
Cayo Blanco (Fir-Shrub)	11.1	2.7	0.068	8.2	6360	87.72	11.54	0.74	Sand
Cayo Blanco (Shrub)	9.9	1.2	0.009	8.6	5170	97.03	2.64	0.33	Sand
Iceland	1.3	1.3	0.089	6.8	2170	79.27	20.31	0.41	Loamy Sand
Kilkenny (Ireland)	3.3	3.3	0.347	7.1	2050	70.65	25.96	3.39	Sandy Loam
Sourhope (UK)	8.5	8.3	0.729	4.5	82	68.35	25.15	6.50	Sandy Loam
South Tyrol (Italy)	10.3	4.2	0.231	7.5	2060	80.78	17.64	1.58	Loamy Sand
Trinidad (Cuba)	7.6	5.3	0.385	8.0	2580	48.94	39.54	11.52	Sandy Silt Loam
Tuscany (Italy)	3.2	3.2	0.275	8.0	2240	54.65	33.92	11.43	Sandy Loam
Warwick (Cryfield-UK)	1.4	1.4	0.146	6.9	250	34.41	48.80	16.79	Sandy Silt Loam

Figure 4.1 Maps of soil samples locations: A) Italy and Algeria, B) United Kingdom, C) Iceland, D) Antarctica, E) Cuba. Maps created using the web application EasyMapMaker (http://www.easymapmaker.com)

4.2.2 DNA extraction and sequencing

Environmental DNA was extracted using FastDNA® SPIN Kit for Soil (MP Biomedicals) from each sample in triplicate using 0.5 g of soil per extraction.

MiSeq library were prepared using Illumina® Nextera XT library preparation kit following the protocol optimized for 16S rRNA gene by Illumina® (Anon., 2014). Amplicons libraries were prepared for 16S rRNA gene targeting the V3-V4 region of the 16S rRNA gene, NRPS adenylation domain and PKS ketosynthase domain. Details on the design of the NRPS and PKS degenerate primers are reported in Chapter 5. Library preparation optimization has been performed for NRPS and PKS amplicons using customized oligonucleotides designed in accordance with Illumina MiSeq instructions for genes of interest different from 16S rRNA. The NRPS_F/R and PKS_F/R primers were adapted to Illumina Miseq sequencing with the addition of the adaptor sequences provided by Illumina (Table 4.3). Optimization of the PCR reaction was performed; the first PCR step, where Illumina-tags are added to the primers, were performed with Promega 2X PCR mastermix instead of NEB high-fidelity Phusion DNA polymerase (used for 16S rRNA gene amplification) and 40 cycles instead of 25 were run in order to obtain amplicons for the next step. All the other steps were followed as described in the 16S rRNA protocol (Anon., 2014) for all genes.

Illumina® MiSeq run was performed following manufacturer's instructions for multiplexed 96 samples, 2 x 300 bp paired-end reads.

Table 4.3 Primers for amplicon sequencing using Illumina MiSeq technology. All the oligonucleotides have the adaptor sequences suggested by Illumina and the customised primers for the gene of interest.

Gene targeted	Name	Sequence	Annealing T (°C)	Amplicon size (bp)	Reference
16S rRNA (V3-V4 region)	16S_ill_F 16S_ill_R	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC	60	460	Modified (Klindworth *et al.*, 2013)
NRPS	NRPS_Ill_F NRPS_ill_R	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGCGCGCATGTACTGGACNGGNGAYYT GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGAGTGGCCGCCCARNYBRAARAA	63	480	This study
PKS	PKS_ill_F PKS_ill_R	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGCAACGCCTACCACATGCANGNNYT GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGTCCGCGGGACGTARTCNARRTC	61.9	350	This study

4.2.3 Data processing: manual pipeline

Illumina 2x300 bp paired-end sequences were initially checked for quality using FastQC (Andrews, 2010). After assessing the general quality of the reads, the reads forward and reverse were assembled using PANDAseq assembler (Masella *et al.*, 2012) setting 10 bp as minimum overlap between forward and reverse reads to increase the quality of the assembly and 300 bp, 350 bp and 250 bp respectively for 16S rRNA gene, NRPS and PKS sequences as minimum length for a sequence to be kept after removing the primers. All assembled sequences were then dereplicated and sorted by size discarding singletons using USEARCH v8.1.1861 (Edgar & Flyvbjerg, 2015). 16S rRNA gene sequences were also filtered for chimera using as reference database the RDP gold database. All filtered sequences were then clustered at 97 % similarity using the UPARSE algorithm (Edgar, 2013) to form operational taxonomic units (OTUs). The most abundant sequence per OTU was selected to be the representative sequence of that specific OTU. OTU tables were created for the three genes of interest and summarized in biom tables. Taxonomy was then assigned to 16S rRNA gene OTU using QIIME (Caporaso *et al.*, 2010) script *assign_taxonomy.py* using the RDP classifier and the Greengenes database.

4.2.4 Data processing: automatic pipeline (EBI metagenomics)

All MiSeq raw reads were also submitted to ENA (project PRJEB11689) and processed using the EBI metagenomics pipeline version 3 (Project ERP013097) (Mitchell *et al.*, 2016). This pipeline (Figure 4.2) used SeqPrep to assemble raw reads, Trimmomatric and Biopython to trim and remove short sequences and HMMER to mask non-coding RNA sequences. Processed reads were then run either into QIIME to organise reads with recognised rRNA sequences (in particular 16S rRNA gene) into OTU and assign taxonomic lineage, or into FragGeneScan to predict open reading frame (ORF) in short sequences. In this last case, the predicted coding sequences (CDS) are then processed through InterProScan to provide a functional annotation.

Figure 4.2 EBI metagenomics pipeline v.3 (30 June 2016) used for the analysis of Illumina MiSeq pair-end reads (https://www.ebi.ac.uk/metagenomics/about).

For this study, further analyses were processed in collaboration with the EBI metagenomics group. A taxonomic lineage of the functional reads was provided using the Unipept software, which is usually used in metaproteomic study to match peptide sequences to taxonomy. Taxonomic lineage resulting from Unipept for the functional genes and QIIME for the 16S rRNA gene marker were then unified to a common taxonomy based on NCBI taxonomy criteria in order to compare functional genes and microbial communities' composition.

Matrixes containing reads counts for 16S rRNA, NRPS and PKS sequences with assigned taxonomy were manually analysed to compare the drivers of diversity of PKS and NRPS and the microbial community they belonged to. Reads could not be rarefied in this case, but relative abundance was calculated on the total number of reads obtained per each sample. Further manual analysis up to genus level were conducted for the more interesting phyla for NRPS and PKS diversity. In this case percentage of each species was calculated on the total number of reads of that particular phylum and then added together according to higher level on the taxonomy classification such as genus, family, order or class.

Corresponding location and IDs used in the two different pipelines (manual and automatic) for the same sample are reported in Table 4.4.

Table 4.4 Location and sample_ID used in the analyses performed manually (same ID for 16S rRNA, NRPS and PKS genes) and the Run_ID used in the EBI pipeline.

Sample_ID (manual pipeline)	Location	RUN_ID_16S (EBI)	RUN_ID_NRPS (EBI)	RUN_ID_PKS (EBI)
S1	Iceland	ERR1111102	ERR1527888	ERR1527849
S2	Iceland	ERR1111103	ERR1527889	ERR1527850
S3	Iceland	ERR1111104	ERR1527890	ERR1527851
S4	Trinidad	ERR1111105	ERR1527891	ERR1527852
S5	Trinidad	ERR1111106	ERR1527892	ERR1527853
S6	Trinidad	ERR1111107	ERR1527893	ERR1527854
S7	Sourhope	ERR1111108	ERR1527894	ERR1527855
S8	Sourhope	ERR1111109	ERR1527895	ERR1527856
S9	Sourhope	ERR1111110	ERR1527896	ERR1527857
S10	Tuscany	ERR1111111	----------	ERR1527858
S11	Tuscany	ERR1111112	ERR1527897	ERR1527859
S12	Tuscany	ERR1111113	ERR1527898	ERR1527860
S13	South_Tyrol	ERR1111114	ERR1527899	ERR1527861
S14	South_Tyrol	ERR1111115	ERR1527900	ERR1527862
S15	South_Tyrol	ERR1111116	ERR1527901	ERR1527863
S16	Kilkenny	ERR1111117	ERR1527902	ERR1527864
S17	Kilkenny	ERR1111118	ERR1527903	ERR1527865
S18	Kilkenny	ERR1111119	ERR1527904	ERR1527866
S19	Alg_KII	ERR1111120	ERR1527905	ERR1527867
S20	Alg_KII	ERR1111121	ERR1527906	ERR1527868
S21	Alg_KII	ERR1111122	ERR1527907	ERR1527869
S22	Alg_B3	ERR1111123	ERR1527908	ERR1527870
S23	Alg_B3	ERR1111124	ERR1527909	ERR1527871
S24	Alg_B3	ERR1111125	ERR1527910	ERR1527872
S25	Alg_KI	ERR1111126	ERR1527911	ERR1527873
S26	Alg_KI	ERR1111127	ERR1527912	ERR1527874
S27	Alg_KI	ERR1111128	ERR1527913	ERR1527875
S28	Cuba_Fir	ERR1111129	ERR1527914	ERR1527876
S29	Cuba_Fir	ERR1111130	ERR1527915	ERR1527877
S30	Cuba_Fir	ERR1111131	ERR1527916	ERR1527878
S31	Antarctic	ERR1111132	ERR1527917	ERR1527879
S32	Antarctic	ERR1111133	ERR1527918	ERR1527880

Table 4.4 Continuation

S33	Antarctic	ERR1111134	ERR1527919	ERR1527881
S34	Cuba_Sand	ERR1111135	ERR1527920	ERR1527882
S35	Cuba_Sand	ERR1111136	ERR1527921	ERR1527883
S36	Cuba_Sand	ERR1111137	ERR1527922	ERR1527884
S37	Warwick	ERR1111138	ERR1527923	ERR1527885
S38	Warwick	ERR1111139	ERR1527924	ERR1527886
S39	Warwick	ERR1111140	ERR1527925	ERR1527887

4.2.5 Statistical analyses and visualization software

QIIME v. 1.9.1-20150604 and Rstudio v. 0.99.902 were used for all main analysis and statistical tests. Alpha and beta diversity were tested for each group of sequences using *core_diversity.py*, *alpha_diversity.py*, *alpha_rarefaction.py*, *beta_diversity.py*, *group_significance.py* scripts in QIIME and paired T-test in Rstudio.

Random sequences resampling (rarefaction) was performed on each sample and diversity was calculated and tested before and after rarefaction to control that diversity was not affected.

Alpha diversity index Simpson inverse was calculated according to Equation 4.1 using QIIME and compared in Rstudio for statistically significant differences. The Simposon inverse index calculates the diversity of the sample considering the richness and evenness of the species present in the sample; the higher is the index, the more diverse is the sample.

$$\frac{1}{D} = \frac{N(N-1)}{\sum n(n-1)}$$

Equation 4.1 Simpson inverse index (1/D) formula. N = total number of organisms of all species; n = total number of organisms of one species.

Beta diversity analyses were performed using various distance metrics including unweighted/weighted unifrac. Unweighted unifrac accounts for the presence/absence of OTU, whereas weighted unifrac measures relative OTU abundance. Both weighted/unweighted unifrac distance metrics also account for phylogenetic distances between observed microorganisms (Lozupone & Knight, 2005, Lozupone *et al.*, 2006,

Lozupone *et al.*, 2011). In addition to this, the Bray Curtis dissimilarity index was used to quantify the compositional dissimilarity in terms of OTU abundance between samples according to Equation 4.2 without taking into consideration the phylogeny. The Bray-Curtis index has a range between 0 and 1, where 1 represent no similarity between samples.

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

Equation 4.2 Bray-Curtis dissimilarity index equation. C represent the sum of the lesser species in common between samples, S represents the sum of all species in the sample and i and j represent the two samples to be compared.

OTU networks of functional gene OTUs were created using QIIME script *make_otu_network.py* and visualized using the free software Cytoscape. iTOL (Interactive Tree Of Life - http://itol.embl.de) and Emperor were used to visualize phylogenetic trees and PCoA respectively.

The correlation between 16S rRNA gene diversity and the two functional genes diversity was investigated using a Procrustes superimposition of PCoA plots based on Bray-Curtis dissimilarity matrix with a Mantel permutation (999 permutations) of the second PCoA (functional gene) over the reference one (phyla 16S rRNA gene PCoA). This is known as Procrustean randomization test (PROTEST) and gives a goodness-of-fit (M^2) value for each superimposition performed. The null hypothesis of the test is that two data sets are no more correlated than random data sets would be and the M^2 value represents the dissimilarity between the PCoA plots used in the superimposition. Therefore, the lower the M^2 value, the more similar are the PCoA plots and the better is the correlation between the data sets. The Procrustes analysis were performed in QIIME.

Further correlations were investigated between the 16S rRNA, PKS and NRPS genes diversity and environmental variables using Pearson's correlation and linear model fitting in Rstudio.

4.3 Results

4.3.1 Manual pipeline

4.3.1.1 Reads processing and alpha diversity

Raw sequences for 16S rRNA, NRPS and PKS reads were assembled, filtered and clustered in OTUs using PANDAseq and QIIME. After the first steps of assembly, filtering and clustering for each type of sequences, the following summarizing data characterizing OTU tables were collected (Table 4.5).

Table 4.5 Summary of sequence counts per samples and total observed OTU processed with PANDAseq, USEARCH and UPARSE.

Target gene	N. of samples	N. of observation (OTU)	Total count (amplicon reads)	Amplicon reads counts/sample summary				
				Min	Max	Median	Mean	Std. dev
16S rRNA	39	12236	3422219	17457	222975	79936	87749.205	53609.175
NRPS	38	5834	2227137	22	398724	46399.5	58608.868	67983.691
PKS	39	9625	1694806	16	113981	427228	43456.564	27791.191

Sample S10 in the NRPS amplicon sequencing failed and no sequences were collected. The sample with the highest number of sequences for 16S rRNA gene had 222975 sequences, while for NRPS gene 398724 sequences and for PKS gene 113981 sequences. The average number of sequence per sample was 87749 for 16S rRNA gene, 58609 for NRPS and 43457 for PKS. The highest number of OTUs observed was 3375 for 16S rRNA gene, 1063 for PKS and 850 for NRPS. All curves for observed OTU according to location reached a plateau for NRPS and PKS genes showing that the primers used covered the diversity of these genes in those environments. On the contrary for 16S rRNA gene samples, four (Iceland, Tuscany, South Tyrol and Trinidad) out of 13 locations did not reach a plateau, therefore part of the potential diversity might have not been covered in those samples (Figures 4.3, 4.4 and 4.5).

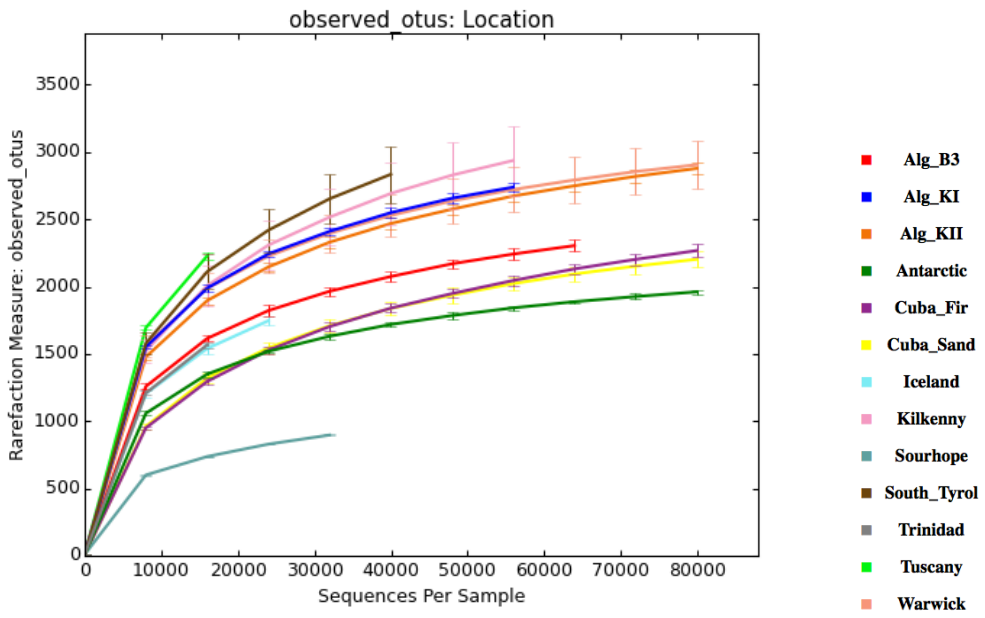


Figure 4.3 Rarefaction curves for 16S rRNA gene amplicons based on observed OTUs. Samples are coloured according to location.

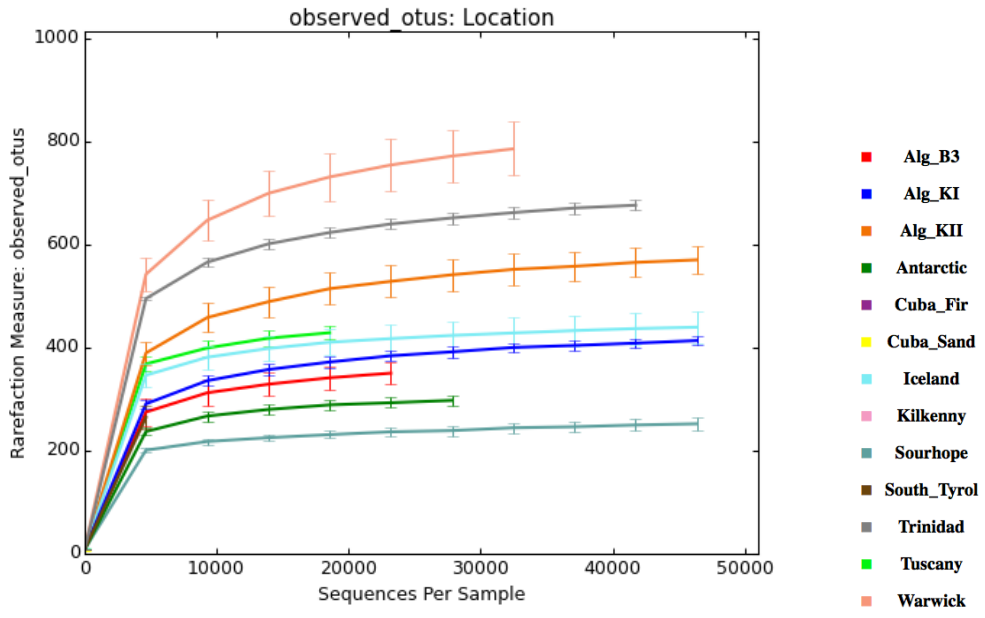


Figure 4.4 Rarefaction curves for NRPS gene amplicons based on observed OTUs. Samples are coloured according to location.

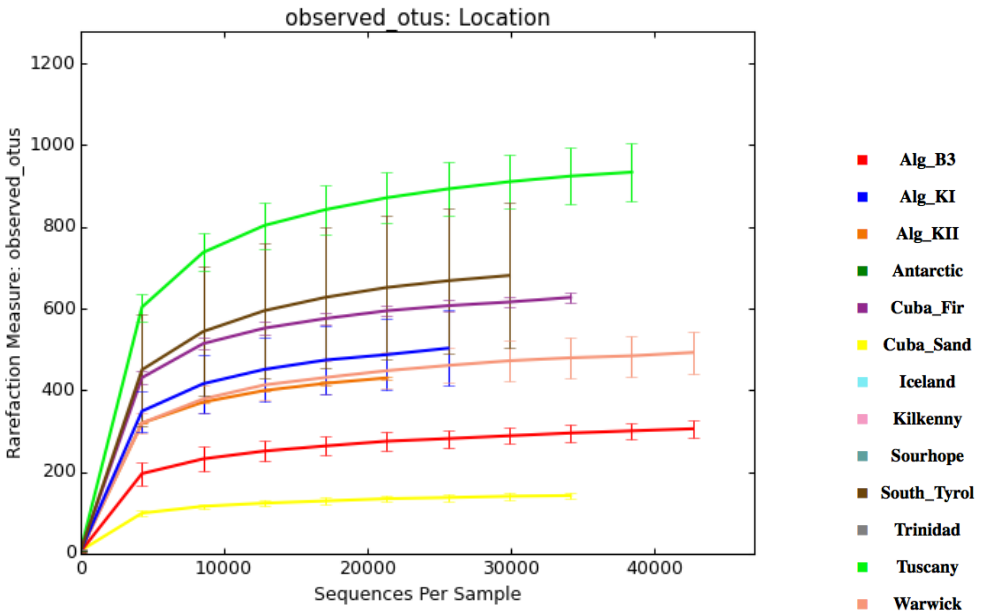


Figure 4.5 Rarefaction curves for PKS gene amplicons based on observed OTUs. Samples are coloured according to location.

Rarefaction was used to standardise all samples to the same number of sequence for comparison between samples. The rarefaction cut-off values were chosen according to the point where the majority of the curves (Figures 4.3, 4.4 and 4.5) reached a plateau without removing too many samples from future comparison analysis. Rarefaction was performed at different cut-offs for each group of sequences: 16S rRNA gene sequences were rarefied at 17000, NRPS at 8500 and PKS at 2600 sequences. The rarefaction step did not remove any sample from 16S rRNA gene sequences, but excluded four samples (S18, S28, S29 and S36) from NRPS and five samples (S2, S5, S9, S17 and S33) from PKS as they had a considerably low number of sequences.

The alpha diversity Simpson inverse index's results showed that there was not a significant difference in samples diversity for 16S rRNA gene before and after rarefaction (Paired T-test, $t(38)=1.128$, $p=0.266$) and between samples that were present in both rarefied and not rarefied tables for PKS sequences (Paired T-test, $t(33)=0.997$, $p=0.326$). For NRPS genes, the Simpson inverse index calculated on the 34 remaining samples showed a significant difference between rarefied and not rarefied samples (Paired T-test, $t(33)=3.67$, $p=0.001$). However, rarefying NRPS

samples at a higher threshold would have excluded more samples excluding completely some environments from future analysis.

For 16S rRNA gene sequences, the Simpson inverse index ranged from 22.2 to 397.4 and there was a significant difference between location diversity (ANOVA F=37.3, p=2.72e-13) (Figure 4.6). In particular, Tuscany and Warwick had a significantly higher diversity to all other soils (Tukey's honest significant difference (HSD) p<0.05). Algeria KI, Algeria KII, Iceland, Kilkenny, South Tyrol and Trinidad also resulted significantly different to Algeria B3, Antarctic, Cuba fir, Cuba sand and Sourhope (Tukey's HSD p<0.05).

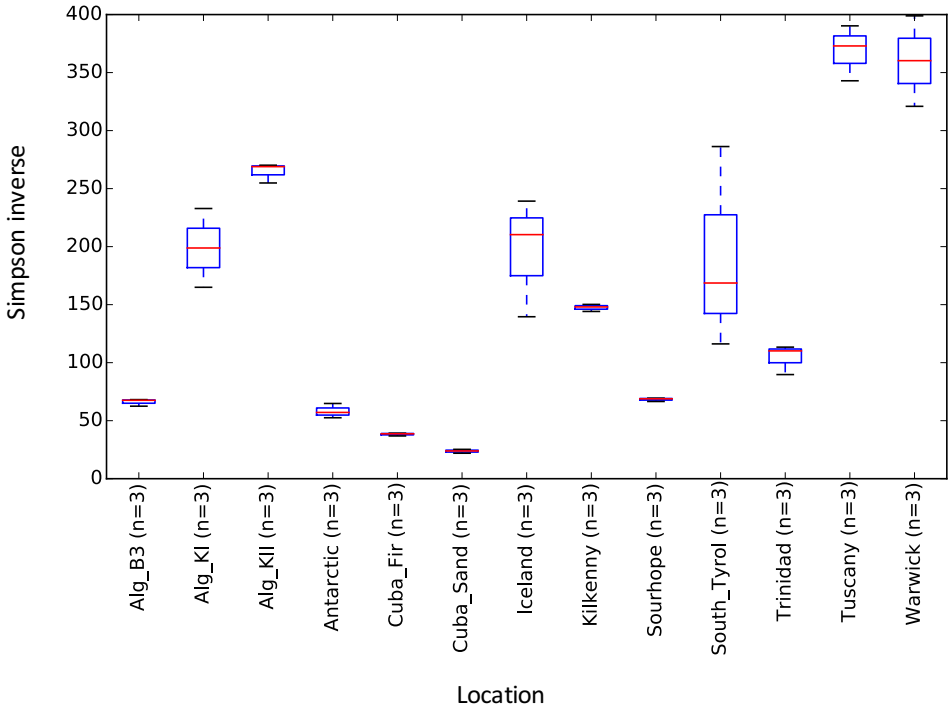


Figure 4.6 Range analysis of alpha diversity of 16S rRNA gene diversity according to the Simpson inverse index for each sample location.

For NRPS genes, the Simpson inverse index ranged from 4.4 to 128.5 and there was a significant difference between locations (ANOVA F=5.1, p=0.0006). There was a significantly higher diversity in Tuscany than in all the other samples with exclusion for Iceland, Kilkenny and South Tyrol (Tukey's HSD p<0.05). Iceland also had a significantly different diversity to Cuba (sand), while all the other locations diversity was not significantly different between each other (Tukey's HSD p>0.05) (Figure 4.7).

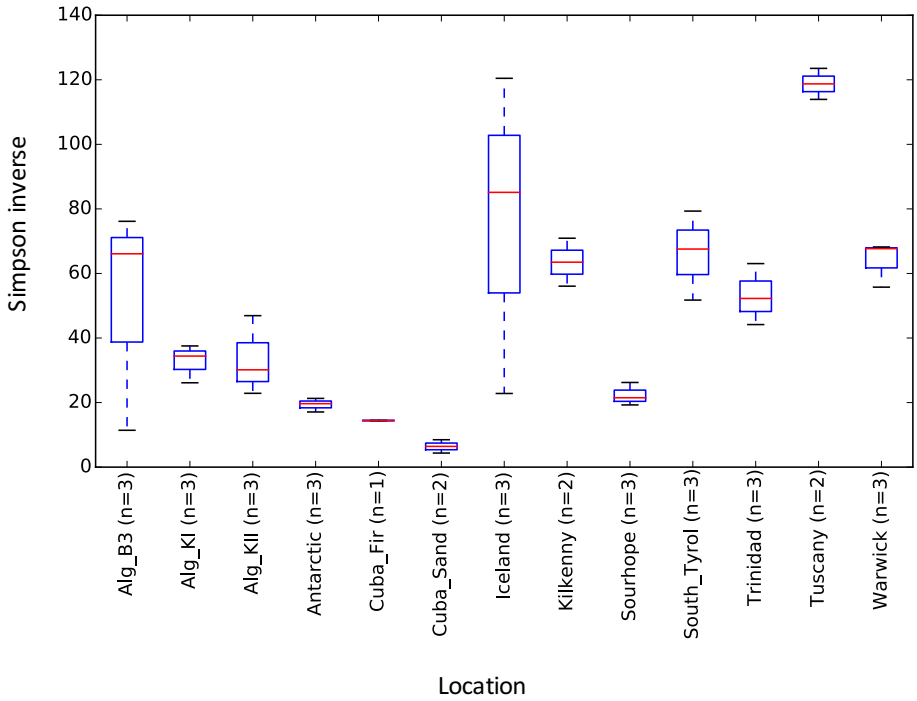


Figure 4.7 Range analysis of alpha diversity of NRPS gene sequences according to the Simpson inverse index for each sample location.

For PKS sequences, the Simpson inverse index ranged from 2.8 to 144.1 across samples and there was a significant difference between locations (ANOVA F=5.5, p=0.0004). In particular, Tuscan soil diversity was significantly higher than all other samples (Tukey's HSD $p<0.05$), which were not significantly different between each other (Tukey's HSD $p>0.05$) (Figure 4.8).

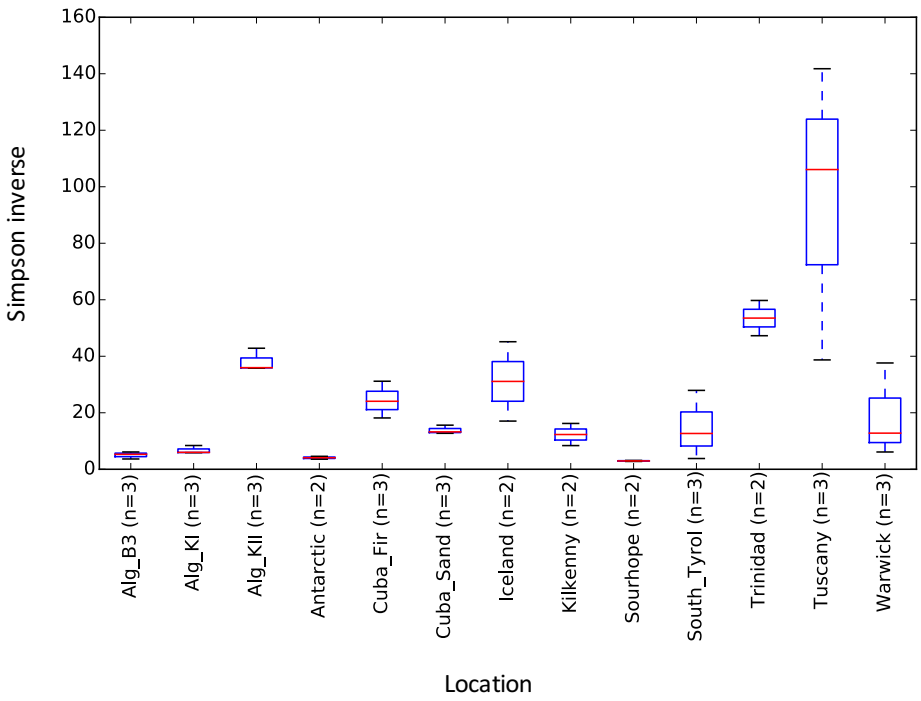


Figure 4.8 Range analysis of alpha diversity of PKS gene sequences according to the Simpson inverse index for each sample location.

4.3.1.2 Beta diversity

Microbial community composition was investigated at different taxonomic levels. At phylum level, all microbial communities in the study showed a high abundance of Proteobacteria (average 25%) and Actinobacteria (average 24%), followed then by less evenly distributed Firmicutes (average 12%) and Acidobacteria (average 10%). European soils appeared very similar in microbial community composition, even though Sourhope showed the highest proportion of Acidobacteria (average 28.2 %). Algerian desert soils showed a higher abundance of unassigned taxa (2-5 %) than the other samples (less than 1 %). Cuban samples as well as Antarctic soil showed a high abundance of Bacteroidetes (6-13 %). Antarctic samples also showed a distinct bacterial community composition with a higher presence of Cyanobacteria (average 25.8 %). All samples had an evenly distributed proportion of Planctomycetes and Chloroflexi phyla and 24 out of 39 samples (in particular European and Antarctic soils) showed a high abundance of Verrucomicrobia (average 2.9 % - in particular Antarctic samples were between 5-6.6 %) (Figure 4.9).

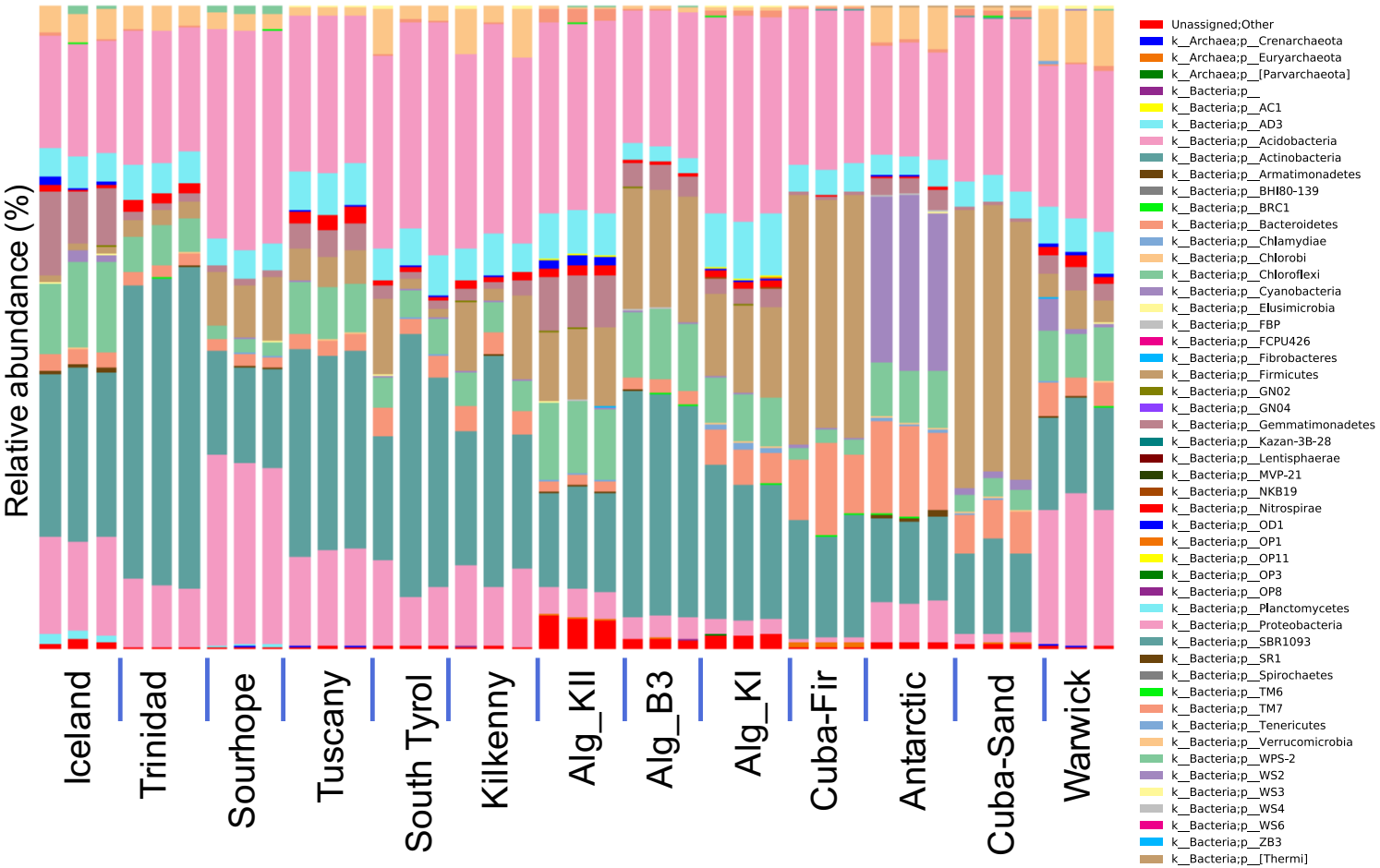


Figure 4.9 Community composition at phylum level of each sample based on 16S rRNA gene annotation using GreenGenes database.

At a lower taxonomic level, Alphaproteobacteria was the most abundant Proteobacteria class (average 13.5 %), followed by Gammaproteobacteria in particular in Sourhope, Algeria KII, Algeria KI, Cuba Sand and Cuba Fir. Betaproteobacteria and Deltaproteobacteria resulted more evenly distributed in the samples but were less abundant than the other two classes (average 3% each). For Actinobacteria, the Actinobacteria class resulted in the most abundant one (average 11.9%). Firmicutes showed a high proportion of Bacilli, in particular in Cuban sand (3.6%), Cuban fir (3.7%) and Algerian soils (average 1.3%). Bacteroidetes were represented by Flavobacteriia class in the Cuban samples and Cytophagia class in Antarctic soil.

Beta diversity analysis based on Bray-Curtis dissimilarity matrix identified six main groups: Antarctic, Iceland, Sourhope, Cuban, Algerian and European soils (Figure 4.10). Those six groups were also identified in principal coordinate analysis (PCoA) based on the same metrics and groups were statistically checked using analysis of similarities (ANOSIM, R=0.96 p=0.001) (Figure 4.11).

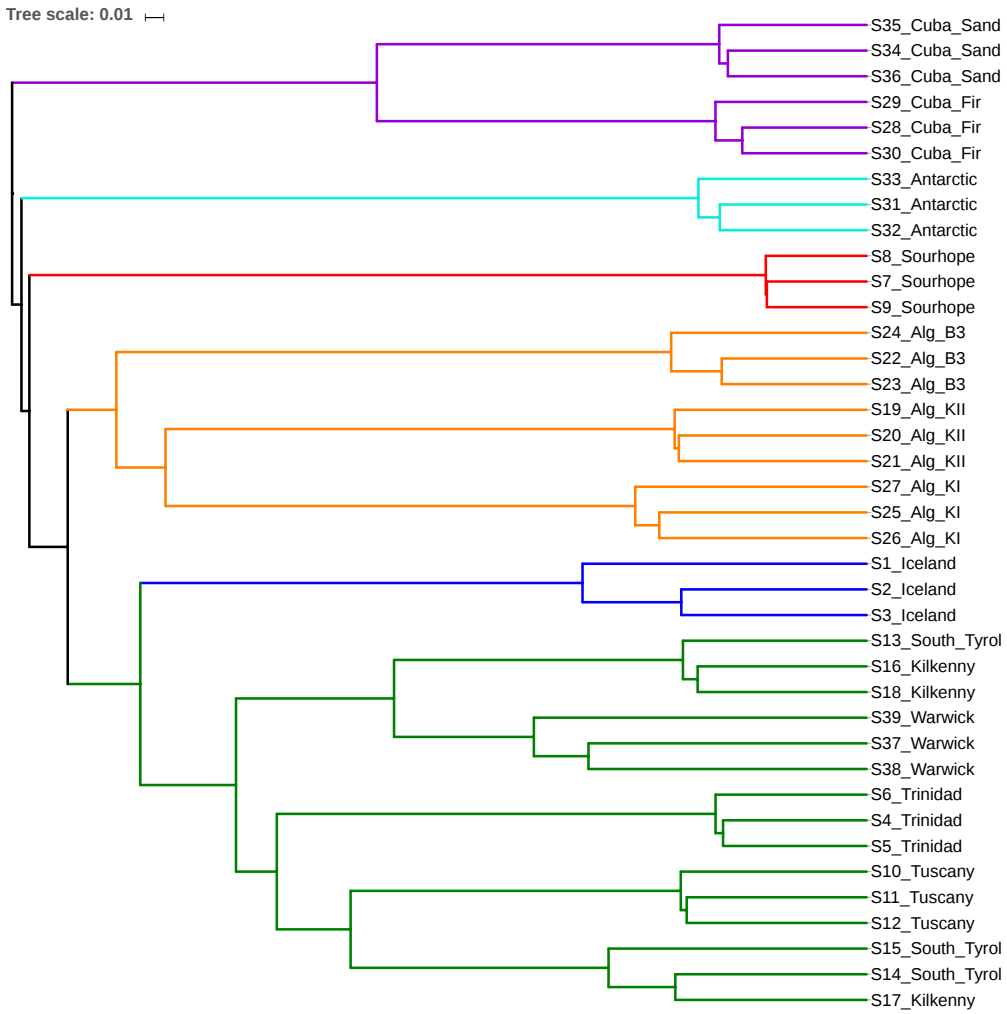


Figure 4.10 Hierarchical grouping of samples according to 16S rRNA genes diversity based on Bray-Curtis dissimilarity matrix. In the UPMGA tree, the group in green includes samples from different geographic locations which have been impacted by agricultural treatment as opposite to the other samples which were pristine environments.

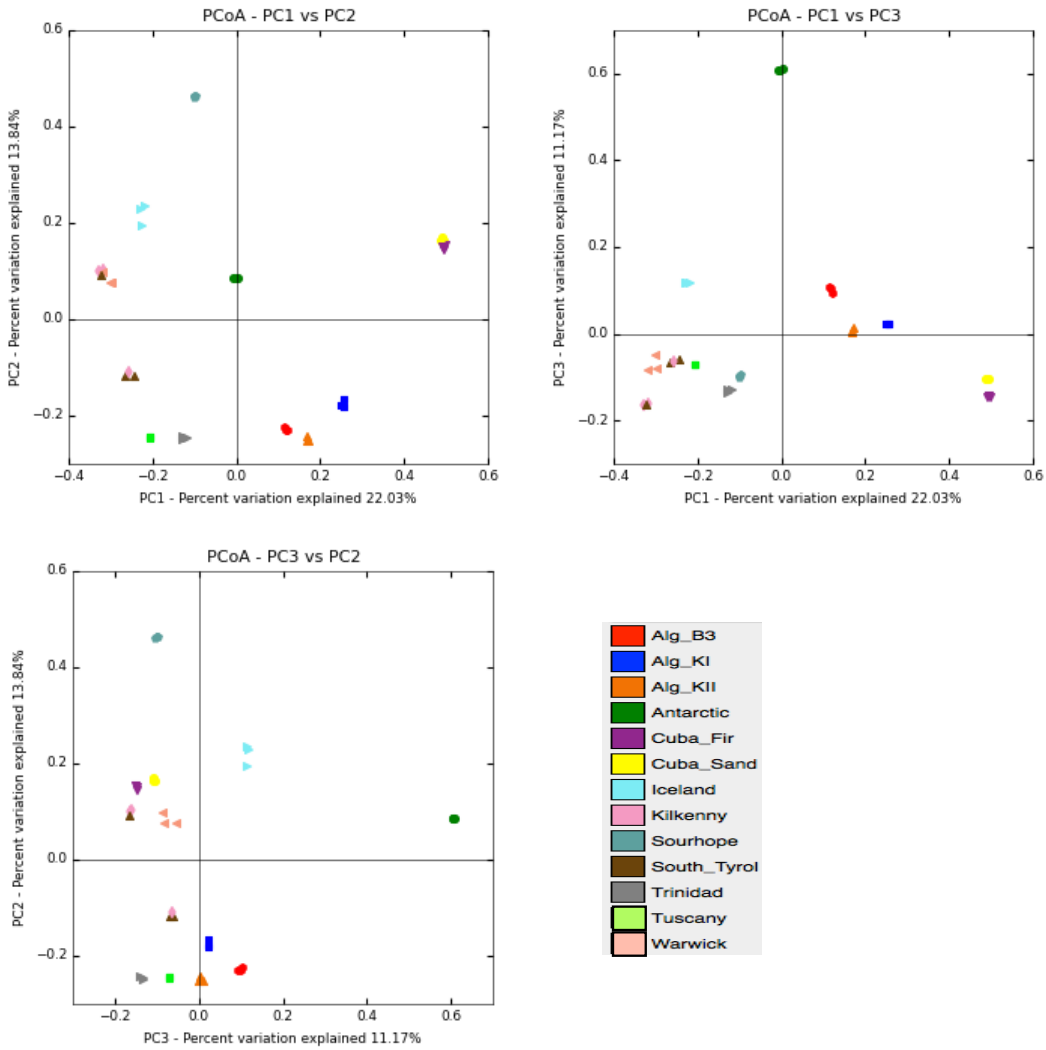


Figure 4.11 Grouping of soil based on the principal components of diversity in two-dimension (PCoA). Analysis of 16S rRNA genes based on Bray-Curtis dissimilarity matrices and coloured by location.

For NRPS genes, a distinct grouping of European and Cuban soils was observed, while Antarctic, Sourhope and Trinidad soil separated individually (ANOSIM, R=0.66 p=0.001) (Figure 4.12).

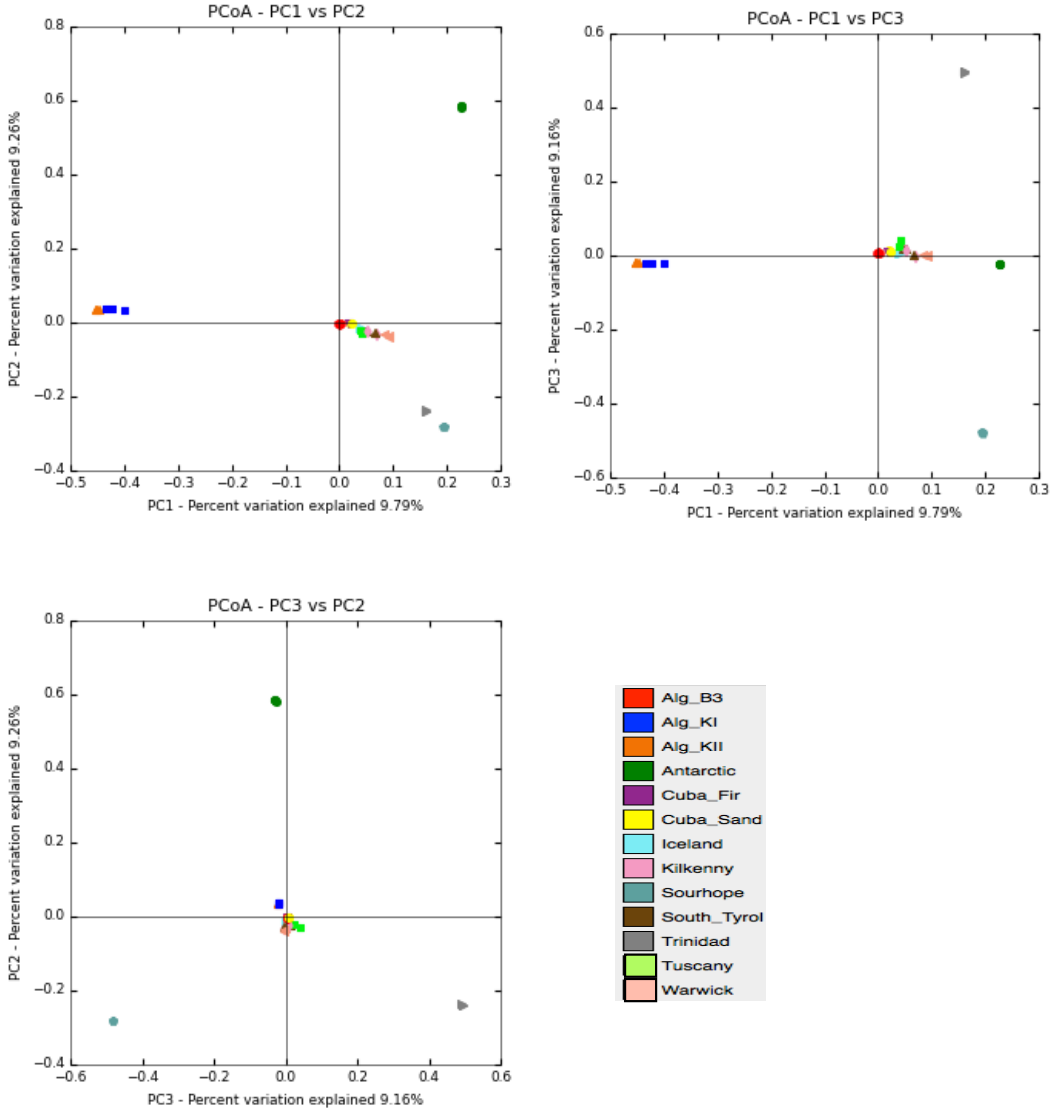


Figure 4.12 PCoA of NRPS genes of all samples based on Bray-Curtis dissimilarity matrixes and coloured by location.

For PKS genes Sourhope, Algerian and Cuban soils grouped as observed for the other genes studied, while Antarctic and European soils grouped differently: Antarctica, Trinidad, Iceland and South Tyrol grouped together, leaving the remaining European soils as a separate group (ANOSIM, R=0.66 P=0.001) (Figure 4.13).

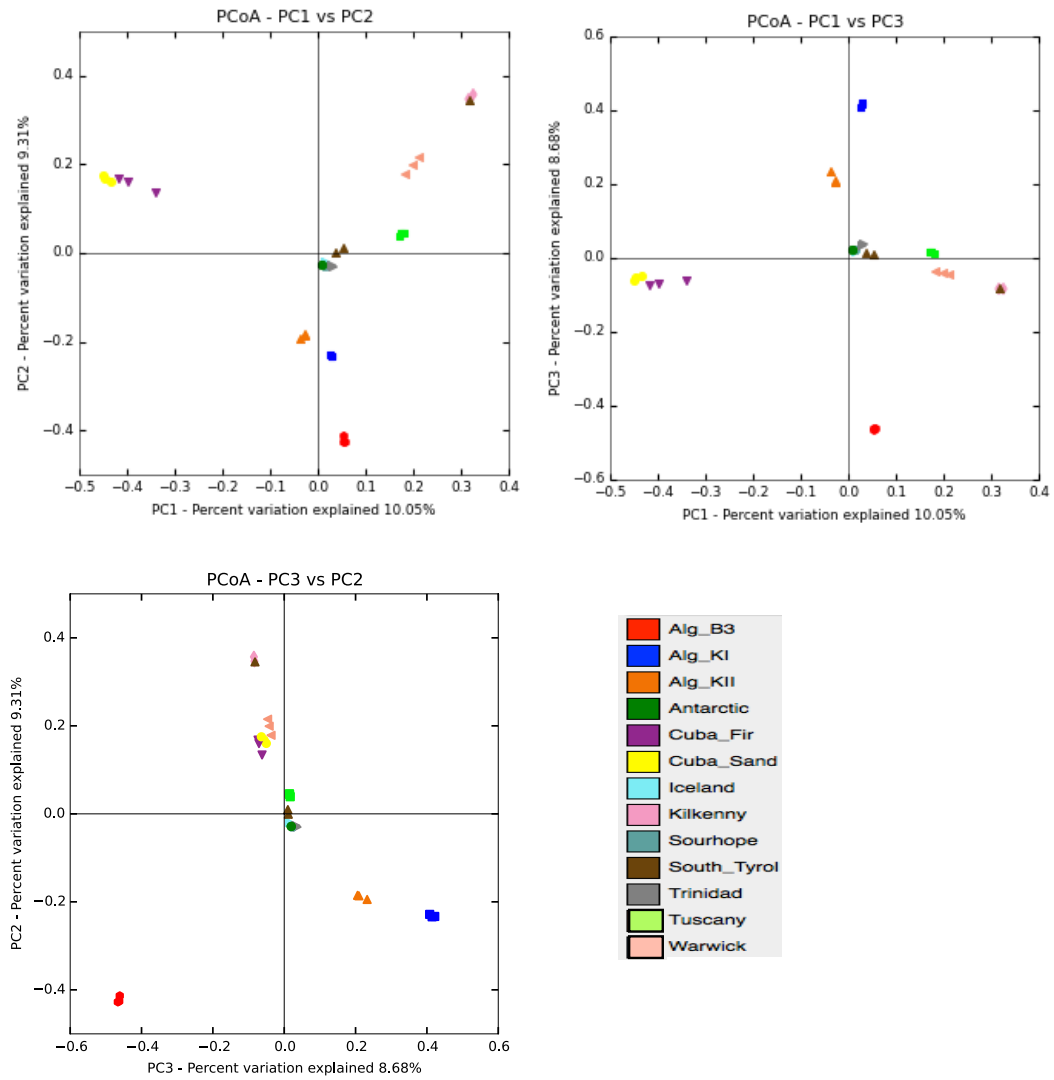


Figure 4.13 PCoA of PKS genes of all samples based on Bray-Curtis dissimilarity matrixes and coloured by location.

4.3.1.3 OTU networks

The OTU produced diversity of sequence type and were analysed on the base of their similarity.

NRPS and PKS OTU networks showed the same pattern of grouping of samples as already observed in beta diversity analysis for the two sets of genes: European soils clustered together and Iceland, Trinidad, Sourhope, Algerian, Cuban and Antarctic soils clustered individually (Figures 4.14 and 4.15).

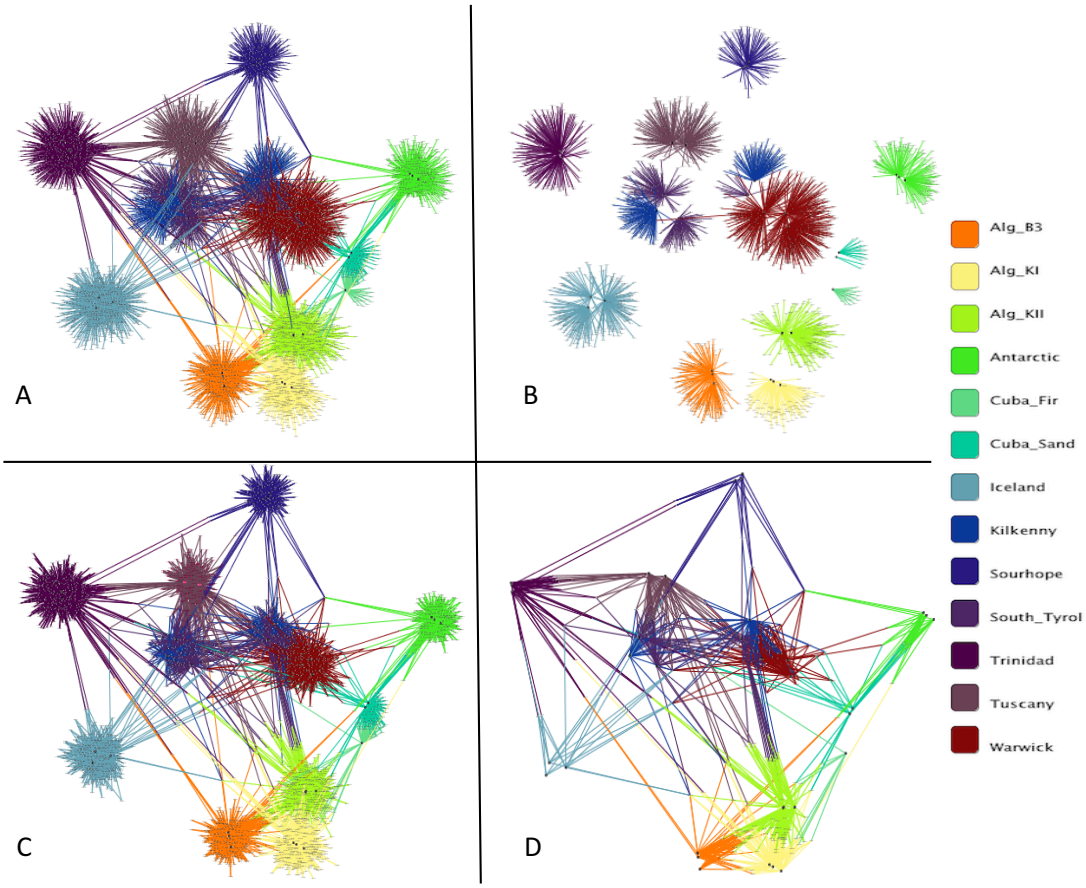


Figure 4.14 Representation of NRPS diversity between samples using OTU network of NRPS sequences. Main nodes (black dots) represent soil samples, nodes at the end of edges (grey dots) represent single OTUs. The distance between main nodes is equal to the similarity between samples. Edges connect OTU nodes to sample nodes. A) All nodes are represented for each sample. B) Only unique OTUs belonging to each sample. C) All shared OTUs between samples were represented and all unique OTUs were removed. D) Only OTUs shared between different locations (not between the same location) were represented (node degree≥4).

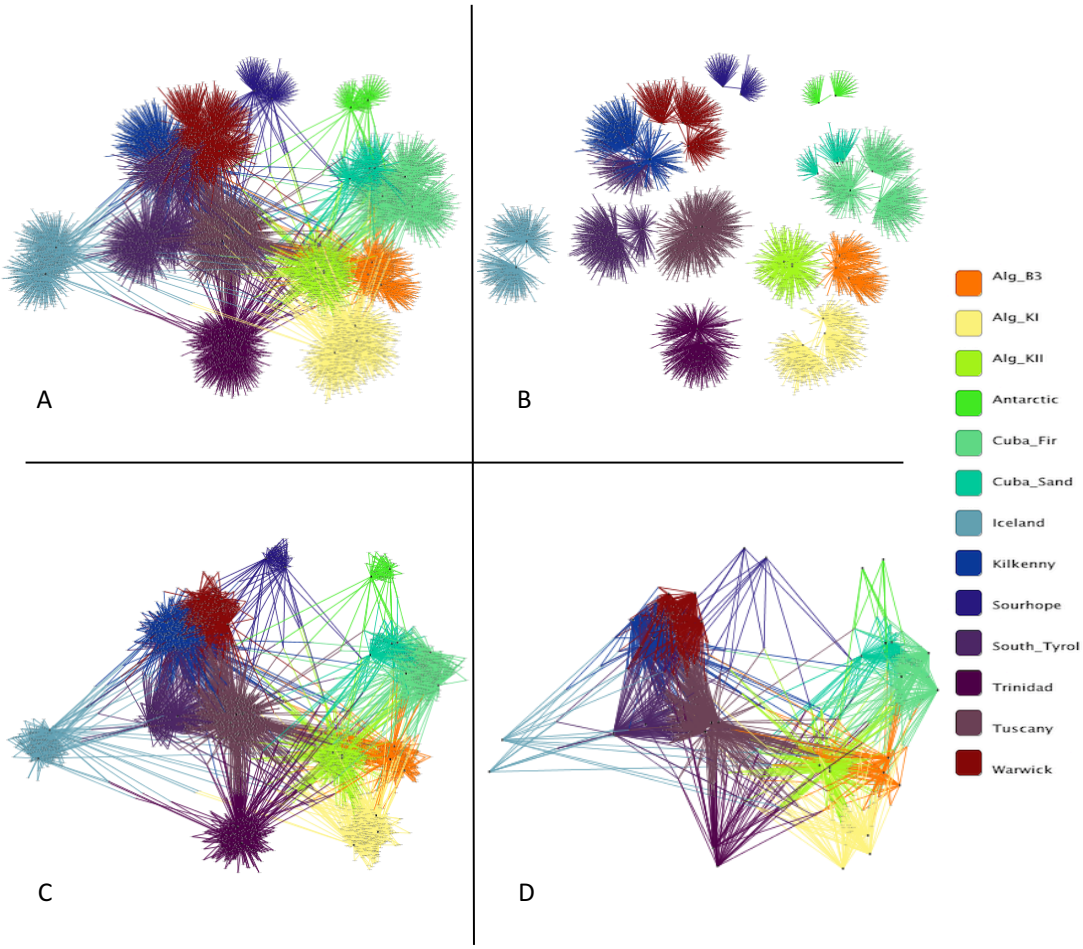


Figure 4.15 Representation of PKS diversity between samples using OTU network of PKS sequences. Main nodes (black dots) represent soil samples, nodes at the end of edges (grey dots) represent single OTUs. The distance between main nodes is equal to the similarity between samples. Edges connect OTU nodes to sample nodes. A) All nodes are represented for each sample. B) Only unique OTUs belonging to each sample. C) All shared OTUs between samples were represented and all unique OTUs were removed. D) Only OTUs shared between different locations (not between the same location) were represented (node degree≥4).

All soils showed unique OTUs and some locations shared less OTUs with other locations than others. For example, for NRPS sequences, Antarctica, Sourhope and Cuba had less shared OTUs with other locations than South Tyrol (Figure 4.16), while for PKS, Antarctic, Sourhope and Iceland soils had less connections with other environments than Tuscany (Figure 4.17).

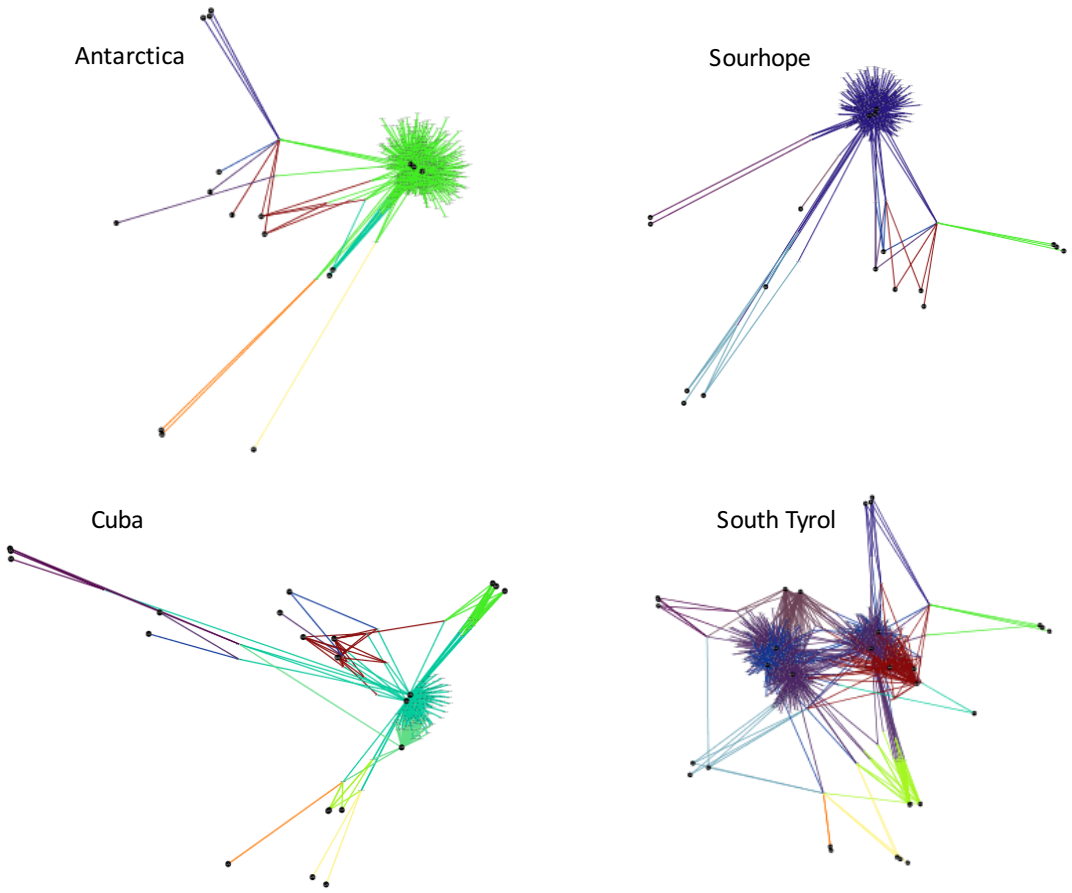


Figure 4.16 Selected OTU networks of NRPS sequences according to location (Antarctica, Sourhope, Cuba and South Tyrol). Main nodes (black dots) represent soil samples, nodes at the end of edges (grey dots) represent single OTUs. The distance between main nodes is equal to the similarity between samples. Edges connect OTU nodes to sample nodes.

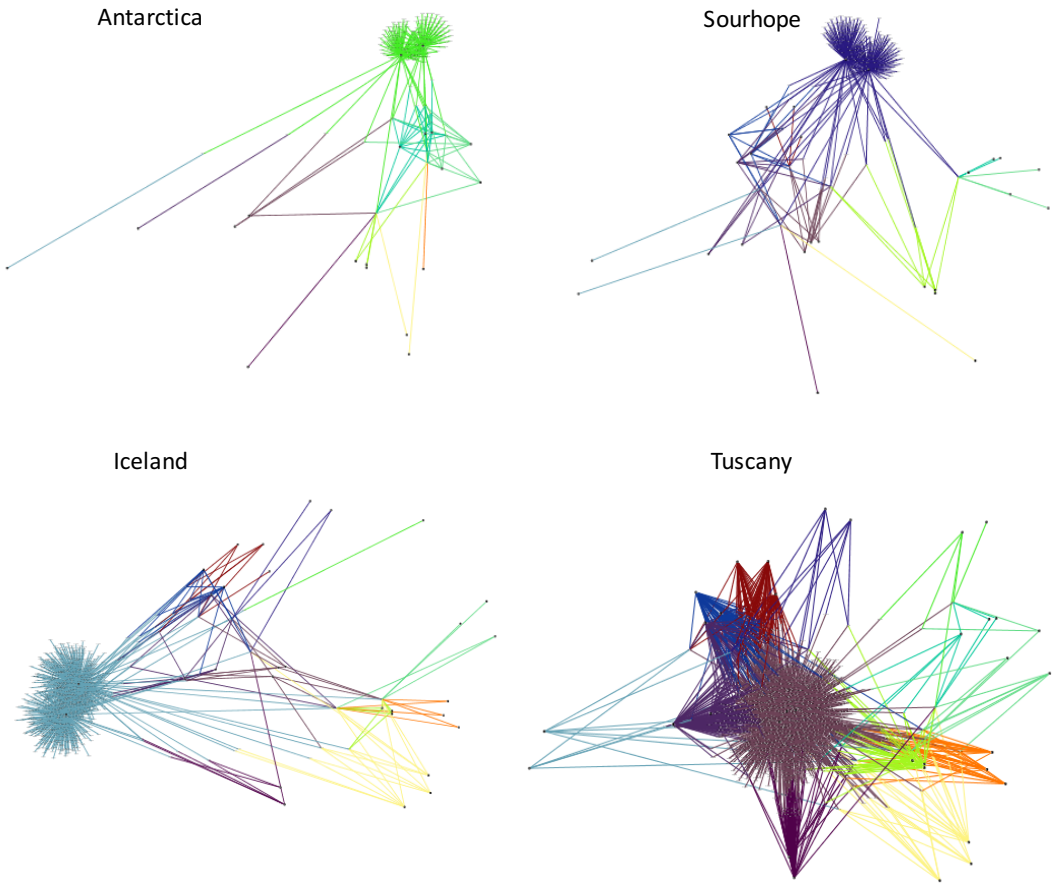


Figure 4.17 Selected OTU networks of PKS sequences according to location (Antarctica, Sourhope, Iceland and Tuscany). Main nodes (black dots) represent soil samples, nodes at the end of edges (grey dots) represent single OTUs. The distance between main nodes is equal to the similarity between samples. Edges connect OTU nodes to sample nodes.

4.3.1.4 Phylogenetic correlation of functional gene diversity

Mantel correlation over 999 permutations between main phyla present in the microbial community (16S rRNA gene) and functional genes (NRPS and PKS) showed a positive significant correlation with all main phyla (Tables 4.6 and 4.7). In addition to Actinobacteria which is the phylum already known for a great NRPS and PKS diversity, an even stronger correlation was identified for Bacteroidetes and Verrucomicrobia for both functional genes.

Table 4.6 Correlation between phyla (16S rRNA gene diversity) and NRPS gene diversity in all samples (Mantel correlation). The R values were statistically significant with a p-value=0.001 (**).

Phyla 16S rRNA-NRPS	
	R (**)
Verrucomicrobia	0.82042
Bacteroidetes	0.81641
Actinobacteria	0.81329
Acidobacteria	0.78534
Chloroflexi	0.78155
Proteobacteria	0.77487
Planctomycetes	0.75403
Cyanobacteria	0.75134
Firmicutes	0.61803

Table 4.7 Correlation between phyla (16S rRNA gene diversity) and PKS gene diversity in all samples (Mantel correlation). The R values were statistically significant with a p-value=0.001 (**).

Phyla 16S rRNA-PKS	
	R (**)
Bacteroidetes	0.82899
Actinobacteria	0.80631
Verrucomicrobia	0.79326
Proteobacteria	0.78704
Chloroflexi	0.78091
Acidobacteria	0.77819
Planctomycetes	0.74839
Cyanobacteria	0.71288
Firmicutes	0.6985

To identify hotspots where the observed correlations between specific phyla and NRPS or PKS genes diversity played an important role, a Procrustes analysis was performed. The Procrustes superimposition of PCoA based on Bray-Curtis

dissimilarity matrix were all statistically significant according to the Procrustean randomization test (PROTEST) (Tables 4.8 and 4.9). The best goodness-of-fit (lowest M^2 value) was obtained between Actinobacteria 16S rRNA gene PCoA and NRPS PCoA (M^2=0.356) and between Bacteroidetes 16S rRNA gene PCoA and PKS PCoA (M^2=0.225). As the M^2 value represents the grade of dissimilarity between PCoA, Actinobacteria and Bacteroidetes PCoA resulted the most similar to NRPS and PKS PCoA respectively. Bacteroidetes also obtained a good M^2 value for NRPS-16S rRNA gene PCoA superimposition even though Verrucomicrobia and Acidobacteria had a higher M^2 value than Bacteroidetes. Indeed, Verrucomicrobia (M^2=0.389) showed the second-best value for Procrustes correlation with NRPS gene after Actinobacteria (M^2=0.398) and Acidobacteria the third one (M^2=0.398). For PKS genes instead, a better superimposition correlation was identified in Proteobacteria (M^2=0.248) and Planctomycetes (M^2=0.317) instead of Actinobacteria (M^2=0.450) and Verrucomicrobia (M^2=0.5).

Table 4.8 PROTEST of phyla 16S rRNA gene diversity and NRPS gene diversity in all samples. The M^2 values were statistically significant with a p-value<0.001 (**) over 999 permutations.

Phyla 16S rRNA-NRPS	
	$M^{2\,(**)}$
Actinobacteria	0.356
Verrucomicrobia	0.389
Acidobacteria	0.398
Bacteroidetes	0.418
Proteobacteria	0.417
Planctomycetes	0.479
Chloroflexi	0.591
Cyanobacteria	0.610
Firmicutes	0.619

Table 4.9 PROTEST of phyla 16S rRNA gene diversity and PKS gene diversity in all samples. The M^2 values were statistically significant with a p-value<0.001 (**) over 999 permutations.

Phyla 16S rRNA-PKS	
	$M^{2\,(**)}$
Bacteroidetes	0.225
Proteobacteria	0.248
Planctomycetes	0.317
Chloroflexi	0.443
Actinobacteria	0.450
Acidobacteria	0.474
Cyanobacteria	0.499
Verrucomicrobia	0.500
Firmicutes	0.588

Procrustes plots for NRPS data in relationship to Actinobacteria, Verrucomicrobia, Proteobacteria, Acidobacteria and Cyanobacteria showed a strong correlation present in Antarctic, Algerian KI, Algeria KII and Sourhope. For Bacteroidetes the correlation with NRPS genes was mainly relevant in Antarctic soil. Planctomycetes resulted relevant in NRPS correlation in Antarctic, Sourhope and all three Algerian soil, while for Chloroflexi a strong correlation was not shown in any soil studied (Figures 4.18 and 4.19).

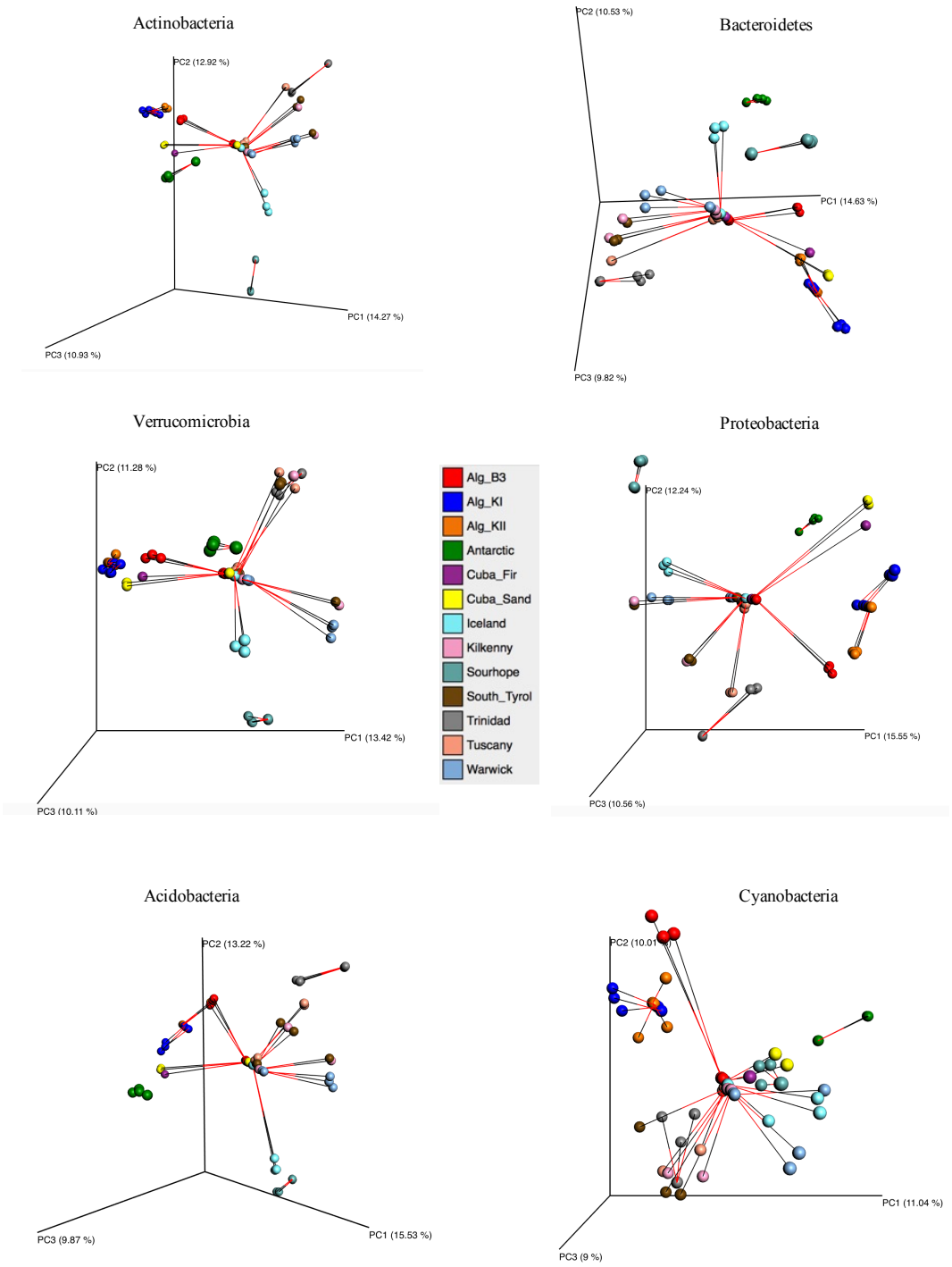


Figure 4.18 Rotated Procrustes plots of the correlations between phylum 16S rRNA gene (Actinobacteria, Bacteroidetes, Verrucomicrobia, Proteobacteria, Acidobacteria and Cyanobacteria) and NRPS genes according to location.

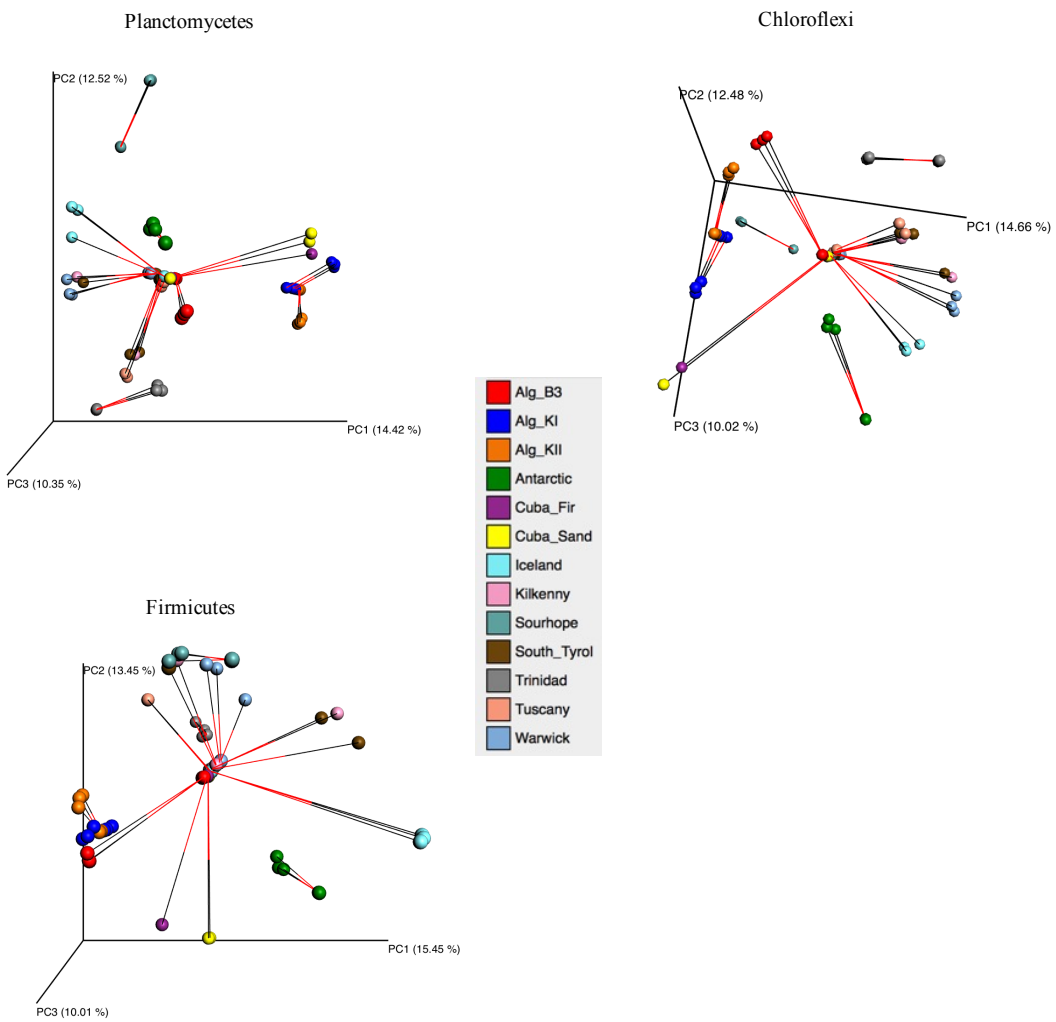


Figure 4.19 Rotated Procrustes plots of the correlations between phylum 16S rRNA gene (Planctomycetes, Chloroflexi and Firmicutes) and NRPS genes according to location.

For PKS genes a different pattern of association was shown by Procrustes plots with the main phyla. Actinobacteria correlated to PKS genes mainly in Algeria KII and Kilkenny soils, while Bacteroidetes showed a strong correlation with PKS genes in Antarctic, Cuban, Warwick, Sourhope, Iceland and Kilkenny soils. Proteobacteria correlated with PKS mainly in Cuba, Algeria B3, Algeria KII, Antarctica, Kilkenny, Warwick and South Tyrol. Antarctic soil also showed a strong correlation between PKS and Verrucomicrobia. Acidobacteria and Cyanobacteria showed a low correlation with PKS in all samples a part from Cuban samples which correlated strongly in the case of Cyanobacteria (Figure 4.20).

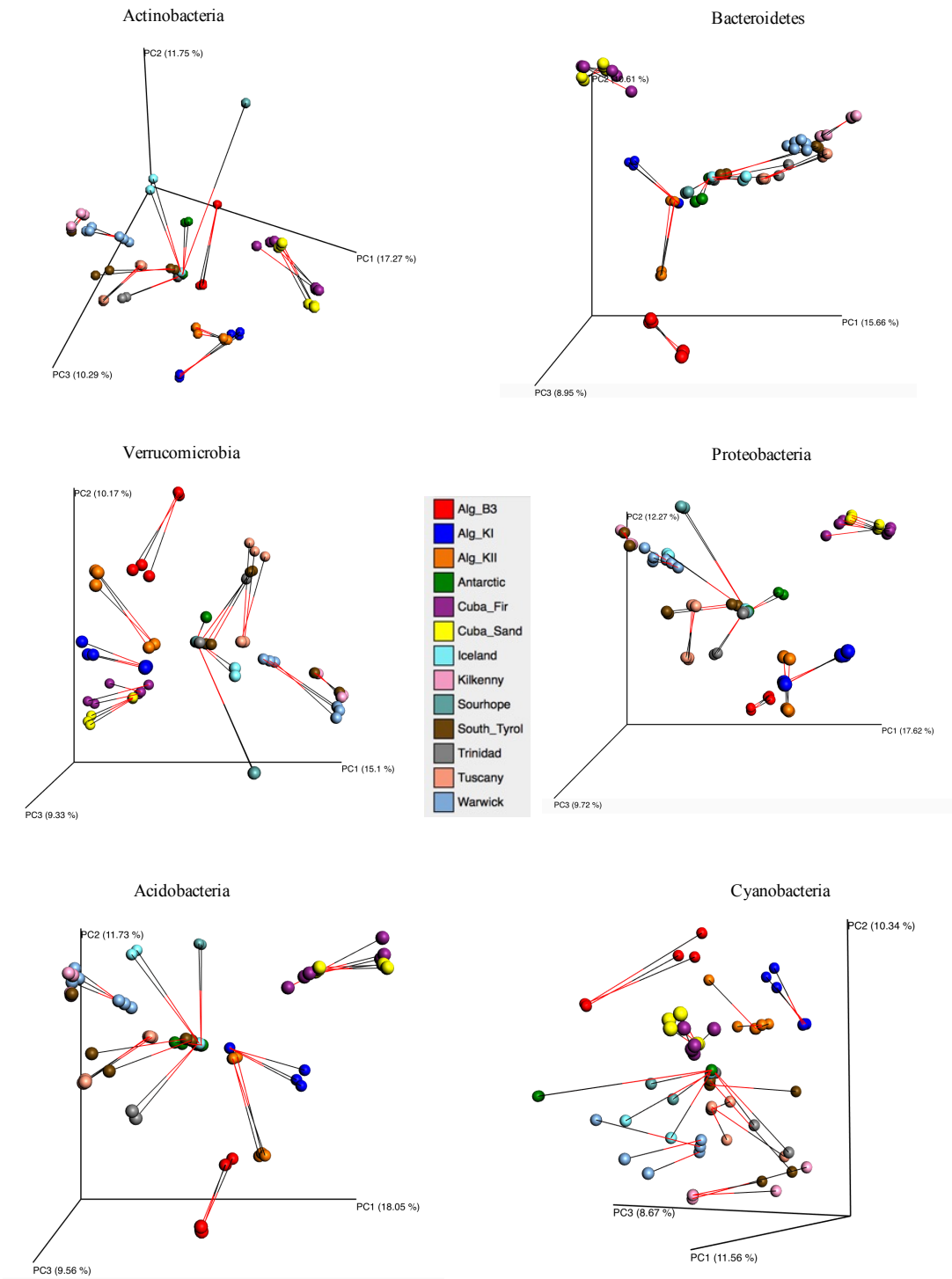


Figure 4.20 Rotated Procrustes plots of the correlations between phylum 16S rRNA gene (Actinobacteria, Bacteroidetes, Verrucomicrobia, Proteobacteria, Acidobacteria and Cyanobacteria) and PKS genes according to location.

Planctomycetes showed a strong association to PKS in Algeria B3, Kilkenny and Warwick samples, followed by Antarctic, Algeria KII, Cuba, South Tyrol and Tuscany. Chloroflexi showed a strong correlation in particular in Antarctica and

Algeria B3, while Firmicutes in particular in Algeria KII, Cuba, Tuscany and Warwick (Figure 4.21).

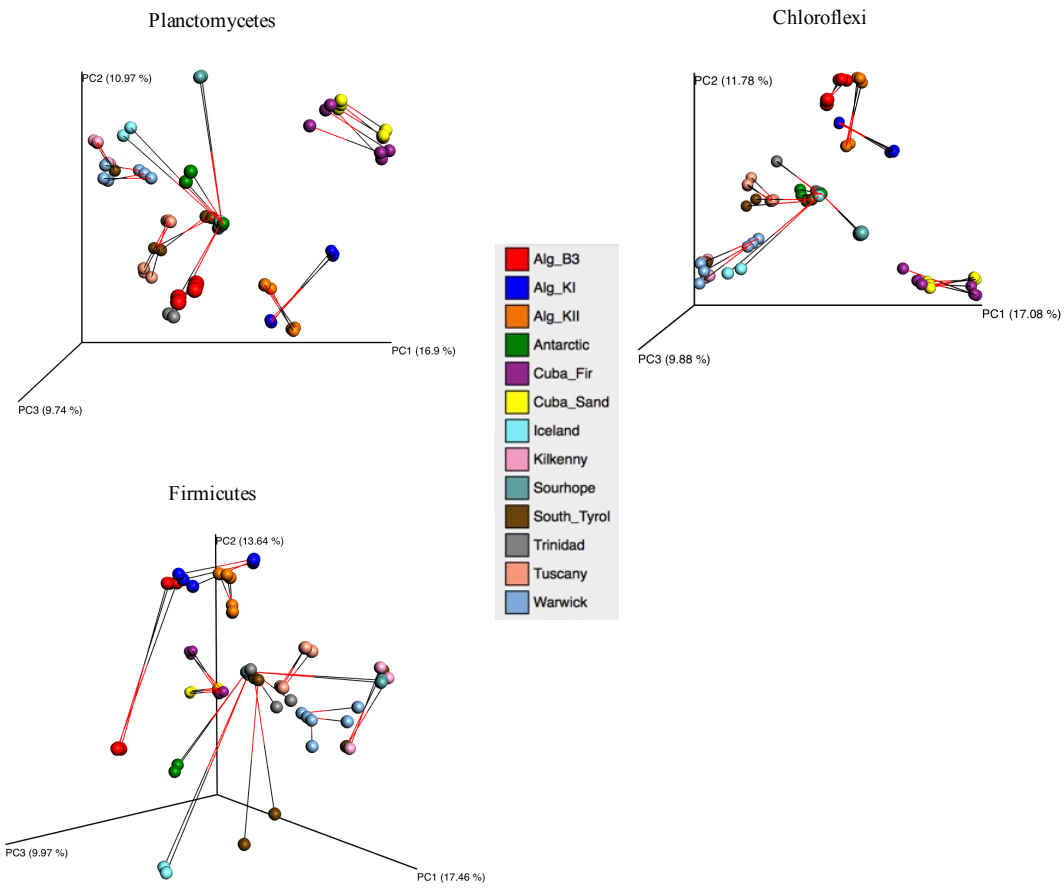


Figure 4.21 Rotated Procrustes plots of the correlations between phylum 16S rRNA gene (Planctomycetes, Chloroflexi and Firmicutes) and PKS genes according to location.

4.3.1.4 Environmental variables correlation with phylogenetic and functional gene diversity

Correlation plots between 16S rRNA gene diversity (the mean values of the three replica for each location were used) and environmental variables (pH, salinity, longitude and distance from equator) showed a scattered distribution for most of the data sets (Figure 4.22).

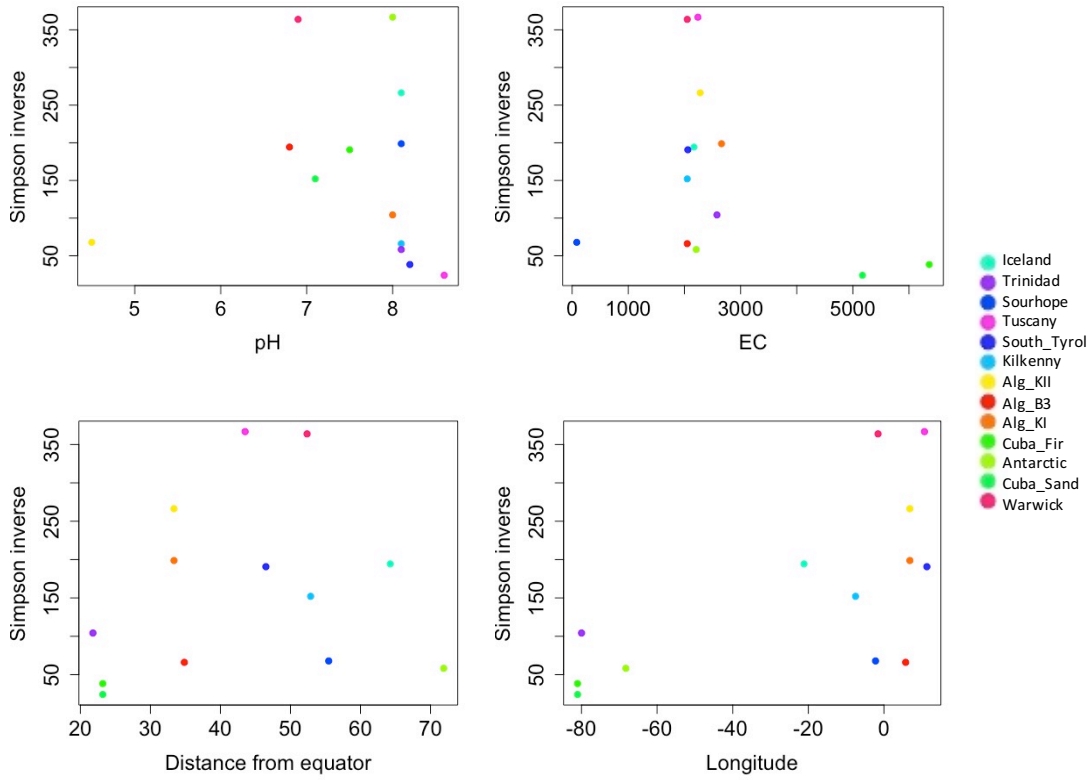


Figure 4.22 Correlation plots of 16S rRNA gene diversity (Simpson inverse index) against the environmental variables pH, salinity (EC), longitude and distance from the equator.

Pearson's correlation and linear fitted regression models were tested for each environmental variable. All correlations and models for pH, salinity and distance from equator predictors resulted not statistically significant. Only for longitude a moderate positive correlation (cor=0.64, p= 0.0185) and a significant linear model (F(1,11)=7.627, p=0.0185) were observed. The linear model was fitted as follow:

Simpson inverse index = [205.4625 $^{(***)}$]x[Longitude value] + 1.9228 $^{(*)}$

where $^{(***)}$ and $^{(*)}$ correspond respectively to 0 and 0.01 significance values. The adjusted R-squared value was 0.36, therefore according to this model longitude explain 36 % of the variability of the diversity data. Hierarchical models were also fitted, but the addition of multiple environmental variables did not have any important impact that explained more of the variability than already explained by longitude alone.

Similar results were observed for NRPS diversity (Figure 4.23). A moderate positive borderline statistically significant Pearson's correlation was identified for longitude against NRPS diversity (cor= 0.517; p= 0.07044), while all the other environmental variables resulted not statistically significant correlated. The only fitted model that

resulted statistically significant was the one designed for longitude (F(1,11)=4.012, p=0.07044) as follow:

Simpson inverse index = [57.7170 $^{(***)}$]x[Longitude value] + 0.4190 $^{(.)}$

where $^{(***)}$ and $^{(.)}$ correspond respectively to 0 and 0.1 significance values. Hierarchical models were also fitted but they all resulted not statistically significant.

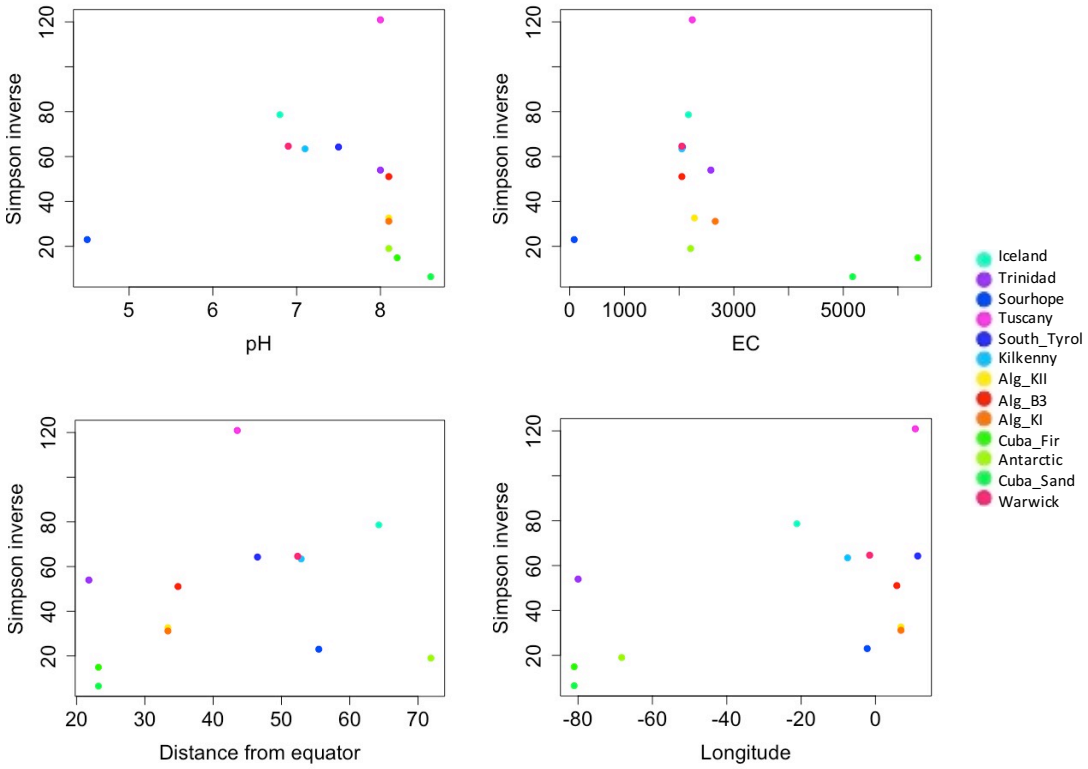


Figure 4.23 Correlation plots of NRPS diversity (Simpson inverse index) against the environmental variables pH, salinity (EC), longitude and latitude.

For PKS genes, as observed from the scatter plots (Figure 4.24), no statistically significant correlations were identified with any of the environmental variables tested. Linear models with one or multiple predictors were also fitted but no statistically significant results were obtained.

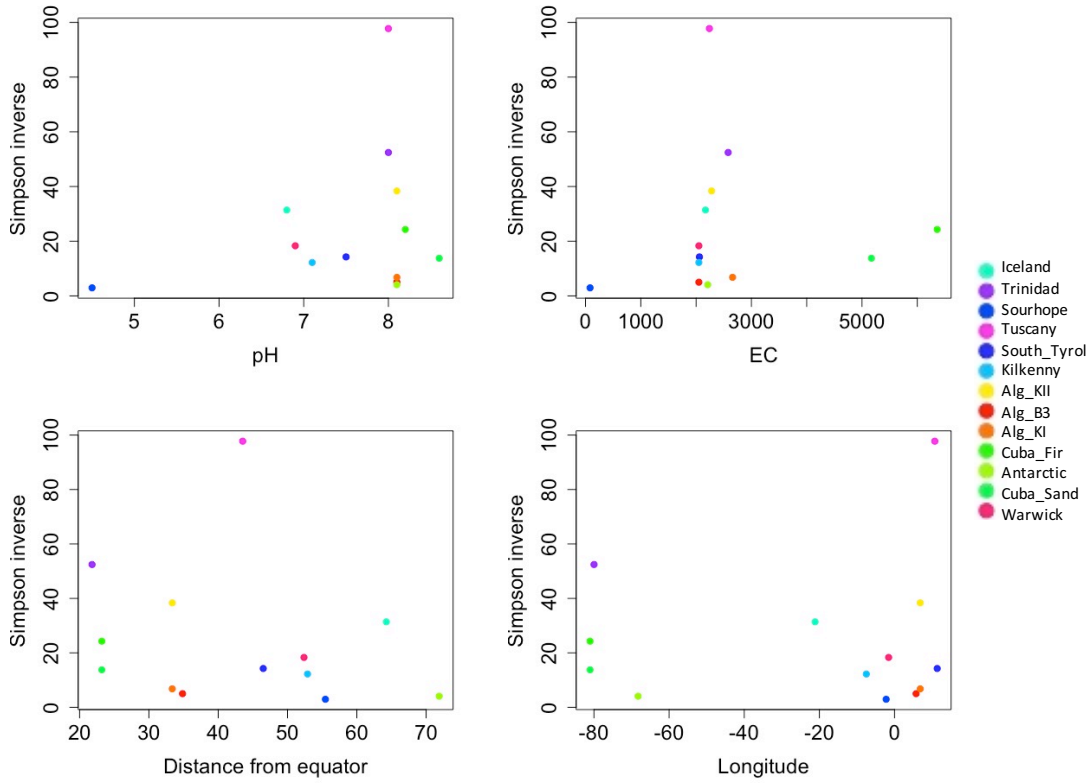


Figure 4.24 Correlation plots of PKS diversity (Simpson inverse index) against pH, salinity (EC), longitude and latitude environmental variables.

4.3.2 Automatic pipeline (EBI metagenomics)

4.3.2.1 Reads processing and alpha and beta diversity

From the automatic pipeline, single OTU tables for 16S rRNA gene were obtained and merged together. The merged OTU table of all samples gave the counts reported in Table 4.10.

Table 4.10 Summary of raw sequences and counts per samples processed through the EBI pipeline.

Target gene	N. of samples	N. of observation	Total count	Counts/sample summary				
				Min	Max	Median	Mean	Std. dev
16S rRNA	39	31622	4348283	31126	289648	100822	111494.436	60640.735

For PKS and NRPS genes, an average of 0.5 % or less of the trimmed and filtered reads were incorrectly recognised as 16S rRNA gene and the related taxonomy assigned. All the remaining reads related to the functional genes were translated into

proteins for function identification using InterPro and taxonomic assignment using Unipept. OTU tables of the nucleotide sequences were not available for NRPS and PKS genes. However, reads abundance matrix for all genes (16S rRNA, PKS and NRPS) with the NCBI unified taxonomy were provided for further analysis.

The alpha diversity index (Simpson inverse) calculated on rarefied samples to 31100 sequences per sample (Figure 4.25) resulted significantly different from the values obtained from the manual pipeline (Paired-T test, T=3.9297, p=0.00035). Higher diversity was estimated for reads filtered through the automatic pipeline than the manual one. However, the general diversity composition of samples resulted the same in both pipelines.

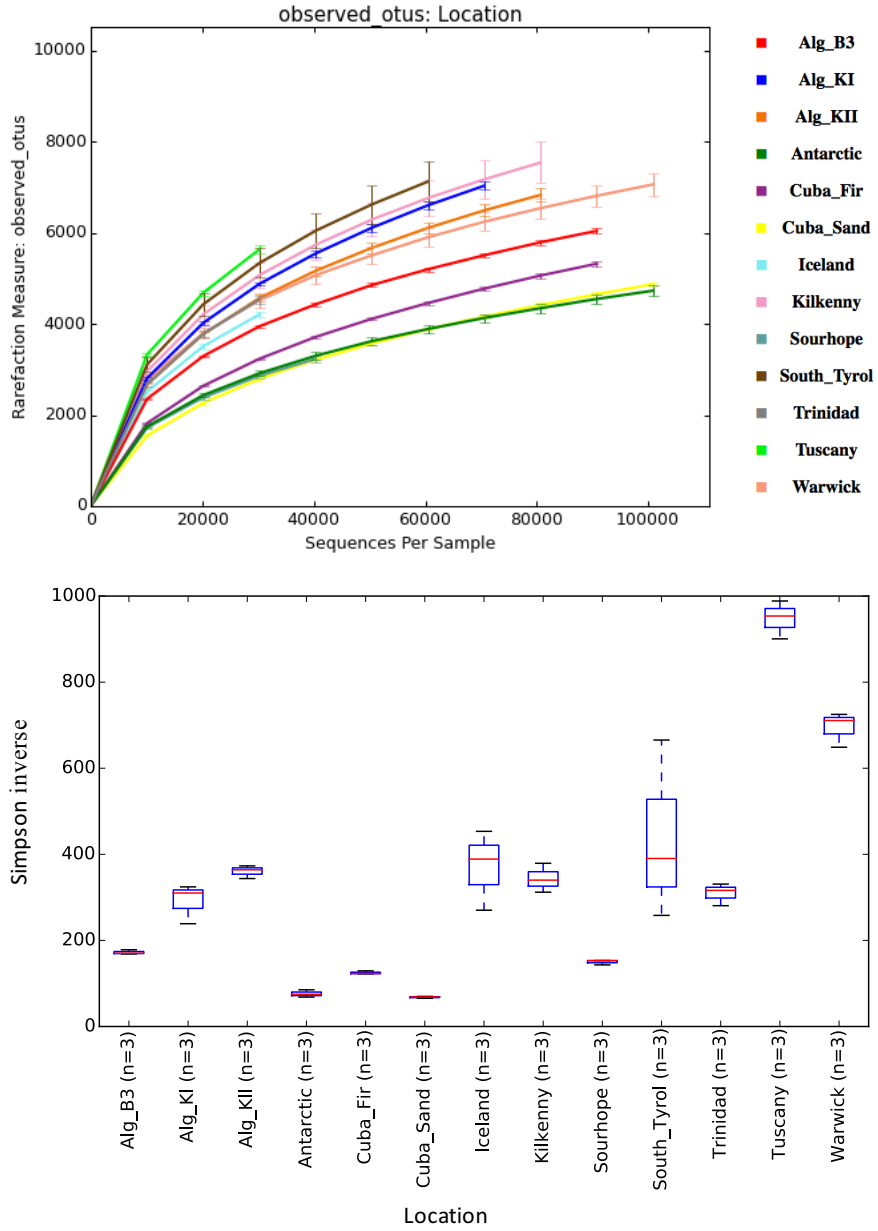


Figure 4.25 Rarefaction curves based on observed OTUs for not rarefied 16S rRNA gene sequences and range analysis of alpha diversity (Simpson inverse) of rarefied 16S rRNA gene sequences, arranged according to sample location.

Taxa plots based on reads relative abundance showed that the main taxa composition of samples was similar to the one obtained from the manual pipeline: Actinobacteria, Acidobacteria, Proteobacteria, Firmicutes, Bacteroidetes, Verrucomicrobia, Planctomycetes, Chloroflexi and Cyanobacteria were identified as the predominant phyla in the microbial communities (Figure 4.26). Actinobacteria and Proteobacteria were the main phyla represented in all communities (14-41% for Actinobacteria and 17-37% for Proteobacteria), while the other phyla had a more variable distribution

according to location. Algeria KI and Cuban soils resulted low in Acidobacteria (0.6-2.1%) comparing to the other samples (average 11.1%), while Sourhope and Warwick had a higher abundance of the same phylum (average 29.9% and 24.4% respectively). Bacteroidetes and Cyanobacteria resulted higher only in Antarctic soil (average 11.1% and 28.6%), while Verrucomicrobia were high not only in Antarctica (average 6.2%) but also in Warwick (average 9.6%) and Kilkenny (average 6.6%) soils. Planctomycetes resulted high in Algeria KI and KII (4.9-5.9%) and low in Algeria B3 (average 2%). Iceland resulted richer in Chloroflexi phylum (average 9.6%) than all the other sample, while Sourhope and Cuban soils had a lower abundance of the same group (between 0.9%-1.8%). However, Cuban soils resulted particularly rich in Firmicutes (average 49.2%). A relative high abundance of Firmicutes was also detected in Algerian soils (average 21.8%). Interestingly in the automatic pipeline less unassigned taxa were identified than in the manual pipeline.

Distinctive microbial communities could be grouped in five main groups according to location: Antarctic, Cuban sand and fir, Algeria KI and KII, Algeria B3 and all the European soils (with the addition of Trinidad soil).

Beta diversity analysis on rarefied 16S rRNA gene OTU table showed similar results to the ones obtained from the manual pipeline (Figure 4.27).

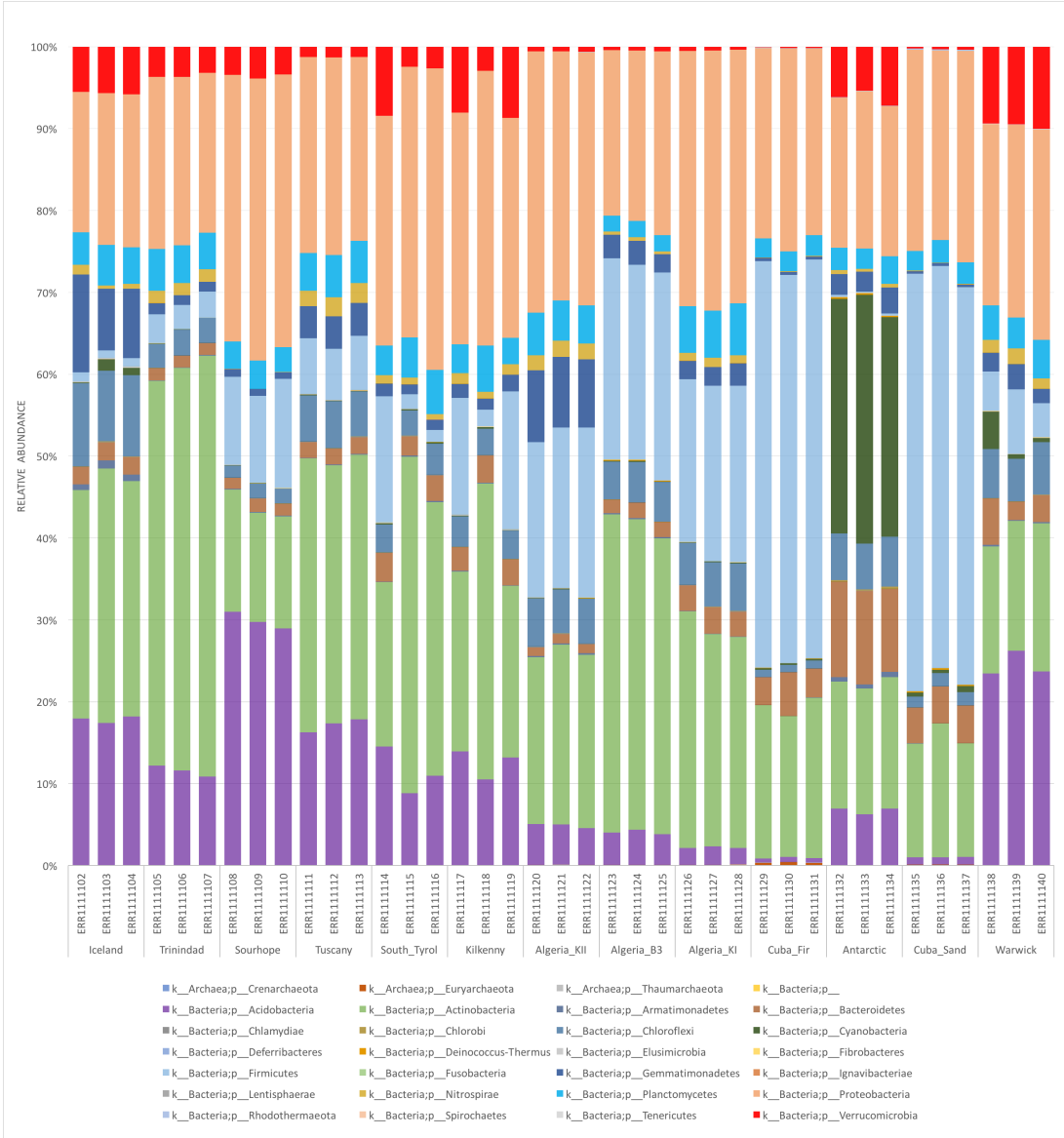


Figure 4.26 Microbial community composition at phylum level of all samples based on unified taxonomy (NBCI based) assigned to 16S rRNA gene reads (automatic pipeline).

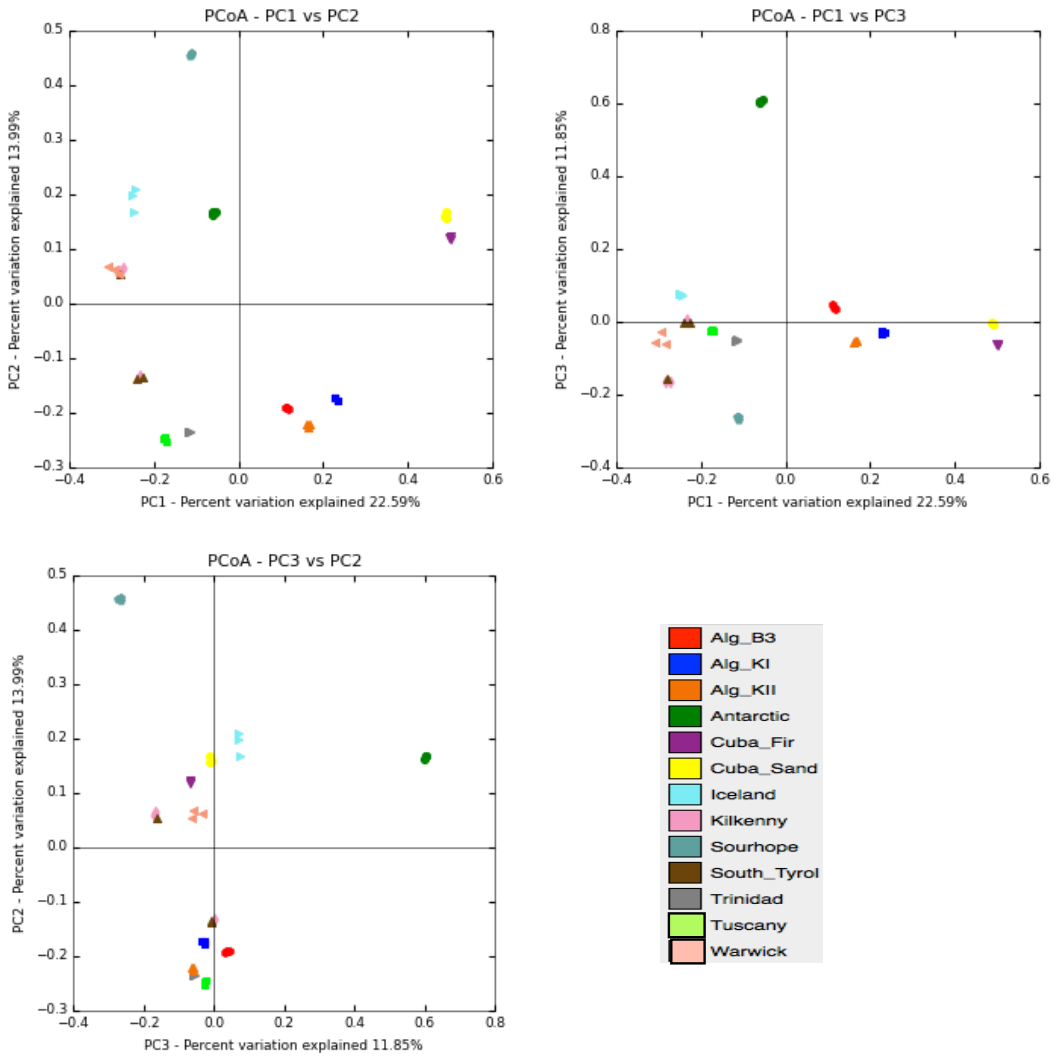


Figure 4.27 PCoA of 16S rRNA gene diversity of all samples based on Bray-Curtis dissimilarity matrix and coloured by location (automatic pipeline).

4.3.2.2 Functional genes annotation and taxonomy assignation

The annotation of NRPS and PKS reads through the InterPro software showed that the NRPS annotation resulted more specifically to secondary metabolite pathways than PKS reads. More than 98% of each NRPS reads per sample were assigned to AMP-dependent synthetase/ligase (IPR000873), phosphopantetheine binding ACP domain (IPR009081) and AMP-binding enzyme (IPR002510), while only an average of 70-80% of PKS reads per sample were assigned to either thiolase-like (IPR016039) or beta-ketoacyl synthase (IPR013030 and IPR013031) domains. These families' domains were closely related to adenylation/thiolase and ketosynthase domains respectively, which were the targets of the primer sets used for the sequences

amplification. The remaining reads were assigned to other proteins not directly related to domains involved in secondary metabolite biosynthesis.

The unified taxonomy for functional reads, assigned most of the NRPS reads to Bacteria with an average in each sample greater than 95% of the total number of reads. The remaining reads were assigned either to Archaea or Eukaryota. In particular, within Eukaryota, the dominant represented phylum belonging to the kingdom of Fungi was Ascomycota in 34 out of 38 samples.

The main three bacterial phyla associated to NRPS reads were Actinobacteria (average 43.8%), Proteobacteria (average 36.5%) and Firmicutes (average 11.7%), followed then at a lower percentage by Cyanobacteria (average 4.5%), Verrucomicrobia (average 1.6%) and Bacteroidetes (average 1.1%) (Figure 4.28). In particular, for Actinobacteria, the most abundant order was Streptomycetales (average 44.4%), followed by Corynebacteriales, Actinomycetales, Pseudonocardiales and Micromonosporales. At a genus level, reads belonging to *Actinomyces*, *Nocardia*, *Rhodococcus*, *Amycolatopsis*, *Actinoplanes* and *Streptomyces* genera were present in all samples. For the Proteobacteria phylum, the most abundant class observed was Gammaproteobacteria (average 40.7%), followed by Deltaproteobacteria (27%), Betaproteobacteria (14.2%) and Alphaproteobacteria (6.5%). Within the Deltaproteobacteria, known secondary metabolite producers belonging to the genus *Myxococcus* were identified, in particular in Antarctic samples. *Stenotrophomonas* genus resulted higher in Cuba (sand) (19.3% instead of 2.4%) and *Pseudomonas* genus in Algeria KI soil (54.4% instead of average 19.1%). Antarctic showed a lower concentration of *Pseudomonas* genus (average 4.4%) comparing to the average percentage in the other samples. The genus *Burkholderia* was also observed in all samples. Within the Alphaproteobacteria class, 68.8% of reads were classified as belonging to the important plant-microbiome interaction order of Rhizobiales. For the Firmicutes phylum, an average of 82% of the reads were classified as belonging to the Bacilli class with Paenibacillaceae family as the most represented one. The main represented classes for Cyanobacteria phylum were Oscillatoriales, Nostocales and Synechococcales. For the Verrucomicrobia phylum, the majority of reads were assigned to the classes Opitutales and Verrucmicrobiae, while for Bacteroidetes phylum to Flavobacteria and Sphingobacteriia. Within these last two classes, *Flavobacterium*, *Chryseobacterium* and *Pedobacter* genera were identified.

In the fungal fraction, the known genera for natural products *Penicillum* and *Aspergillus* were also identified in the majority of the samples.

The main phyla followed the same relative distribution in all samples with exception for Sourhope, Algeria KI and Antarctic soils, which showed a different functional taxonomic composition comparing to all the other soils. However, biological replicates within the same location were consistent in most cases; only few samples (e.g. Iceland and Warwick) showed variation between replicates.

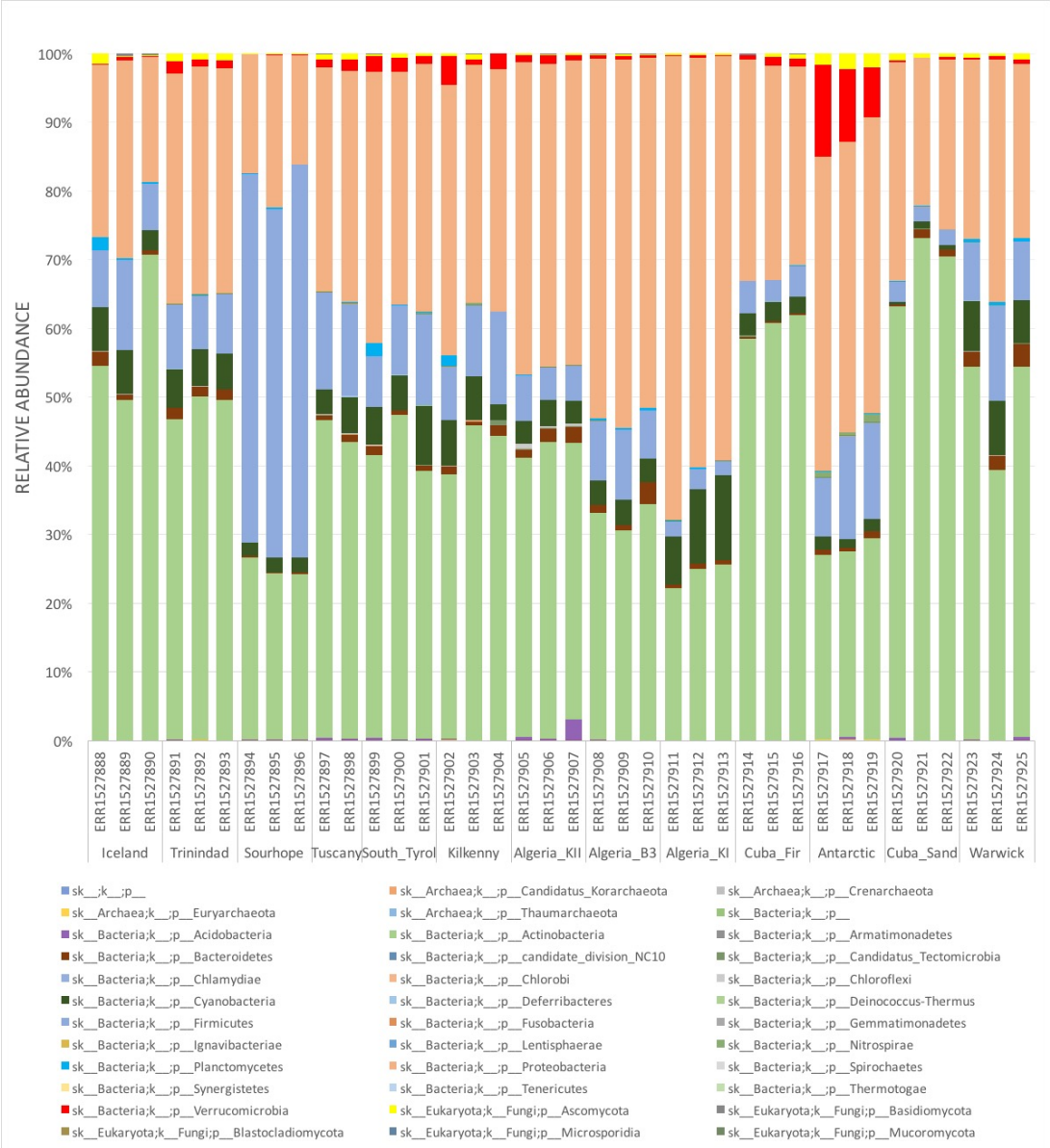


Figure 4.28 Representation at a phylum level of the taxonomic lineage of NRPS sequences assigned according to NCBI nomenclature.

PKS reads were assigned mainly to Bacteria (average 99.1%) and only a small fraction to Eukaryota and Archaea. In all samples the most represented bacterial phyla were Actinobacteria (average 81.7%) and Proteobacteria (average 11.2%). A smaller percentage of reads were also assigned to Firmicutes (1.8%), Planctomycetes (1.6%), Bacteroidetes (1.4%), Verrucomicrobia (0.5%) and Cyanobacteria (0.4%) (Figure 4.29).

In the Actinobacteria phylum, *Streptomyces* genus was the most abundant in all samples (average 61.8%), with the exception of Sourhope and Algeria KI where the *Catenulispora* and *Saccharomonospora* genera were the most abundant, respectively (average >70%). Even in the case of PKS genes as well as NRPS genes described above, most of the known secondary metabolites producers' genera such as *Streptomyces*, *Planobispora*, *Amycolatopsis* and *Micromonospora* were observed in the taxonomic table of PKS reads. For Proteobacteria, the most represented class was Alphaproteobacteria (average 41.8%), followed then by Gammaproteobacteria (average 33.7%), Betaproteobacteria (average 10%) and Deltaproteobacteria (average 5.6%). The orders Rhizobiales, Burkholderiales, Myxococcales and Pseudomonadales were the most abundant represented in these classes. For Firmicutes, the Bacillus class had a higher abundance than Clostridia (average of 59.3% instead of 15.4%), with *Bacillus* genus showing a higher prevalence than *Paenibacillus*. Planctomycetes phylum was entirely represented by the Planctomycetales order and in particular the Planctomycetaceae and Isosphaeraceae families. For the closely related Verrucomicrobia phylum, the majority of the reads were either classified only at the phylum level or belonging to the Spartobacteria, Opitutae and Verrucomicrobiae classes. For the Bacteroidetes phylum, the Cytophagia class was the major representative in Antarctic soil, while Sourhope and Algeria KII showed the highest abundance of yet unclassified Bacteroidetes. All samples had similar distribution of all phyla with the exception of Cuban fir and Antarctic soils. All replicates within the same geographic site were similar to each other, showing a low variability in most cases.

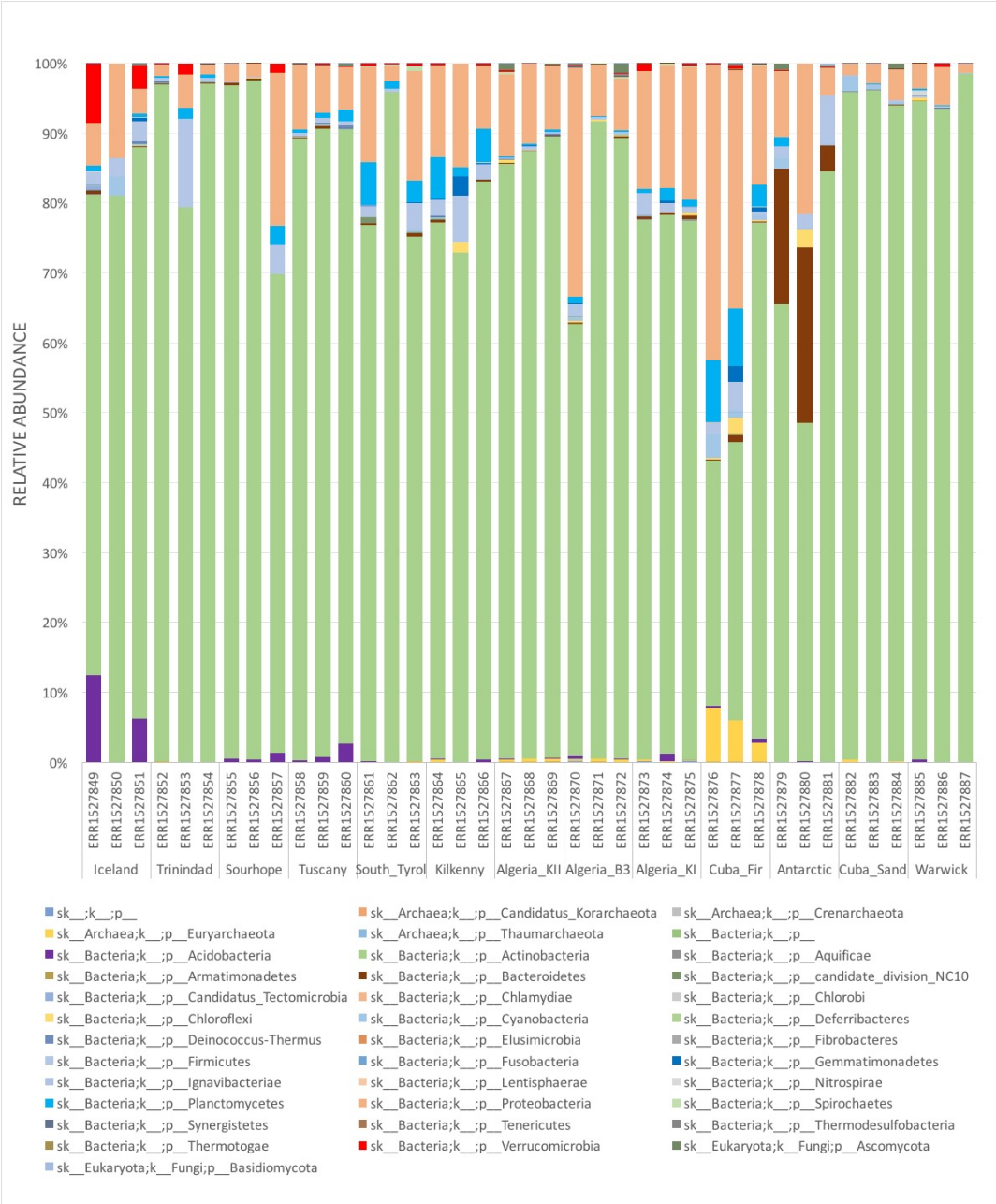


Figure 4.29 Representation at a phylum level of the taxonomic lineage of PKS sequences assigned according to NCBI nomenclature.

4.4 Discussion

The work presented in this Chapter provides an insight into the main phyla related to polyketide and nonribosomal peptide secondary metabolites diversity in soil environments. The study of secondary metabolite gene diversity in diverse soil samples can facilitate metagenomic approaches as illustrated in this work by identifying hotspots for novelty. More importantly the phyla correlating with NRPS

and PKS diversity were identified by association analysis which would also facilitate a selective isolation approach, enrichment and compatible cloning procedures.

Whilst others have reported diversity in NRPS and PKS in the microbial community (Reddy *et al.*, 2012, Charlop-Powers *et al.*, 2014, Charlop-Powers *et al.*, 2015), no studies were performed in elucidating the taxonomic drivers. Correlation analysis with environmental variables and biomes from the same geographic location were presented by Charlop-Powers *et al.* (2014) showing a linear correlation between biosynthetic genes richness and pH or moisture content. Geographic distance was also identified as a possible variable in shaping the microbial community and therefore the biosynthetic diversity (Charlop-Powers *et al.*, 2015). In the current study, however, no correlation was identified for either pH, salinity or latitude for both biosynthetic gene domains. A moderate correlation was identified for NRPS diversity and longitude, but not for PKS. A different biome related to a costal or inland location might impact the diversity of the microbial population and therefor the secondary metabolite genes related to it. The identification of variables that affects the microbial community diversity in large scale is usually quite difficult because of the intrinsic heterogeneity of the soil matrix and all the abiotic and biotic variables that constitute the environment at a micro-scale (see review: Vos *et al.* (2013)). However, it has been shown that even the microorganisms as well as macroorganisms (plants and animals) experience environmental variables and geographic location in driving biogeographic patterns (see reviews: Horner-Devine *et al.* (2004), Martiny *et al.* (2006), Martiny *et al.* (2011), Hanson *et al.* (2012)). The current study showed that the environmental conditions as well as the geographic distance played an important role in shaping the microbial community of the different soils tested affecting the biosynthetic diversity. In particular, environments with more extreme conditions and a bigger geographic distant between each other, such as Antarctica, Cuba, Sourhope or Algeria desert, showed different diversity and distribution of OTUs than European soils. This was not only observed for the microbial community but also for PKS and NRPS diversity in accordance with previous studies by Charlop-Powers *et al.* (2015). Taxa composition demonstrated that in those extreme environments a different bacterial community or different relative abundance of the same microorganisms were detected compared to the European soils which were more similar to each other. These results may indicate that the environment selects as proposed by Baas Becking (1934) in the well-known paradigm "*everything is everywhere, but the environment selects*", but not necessarily

that "*everything is everywhere*". OTU networks of PKS and NRPS ketosynthase and adenylation domains showed that in the environments with a different distribution of the microbial diversity (Antarctic, Cuban and Algerian samples), there was also a unique selection of diverse amplicons in the form of OTUs representing the enzymes involved in metabolites production.

Correlation analysis showed that Actinobacteria, Proteobacteria, Firmicutes and Cyanobacteria which are the known main phyla for natural products diversity (Donadio *et al.*, 2007, Micallef *et al.*, 2015, Esmaeel *et al.*, 2016), might not be the only drivers of secondary metabolites belonging to polyketide and nonribosomal peptide families. Interestingly, high correlations were identified for the phyla Verrucomicrobia and Bacteroidetes respectively for NRPS and PKS genes. BGCs from human-associated Bacteroidetes were recently identified (Donia *et al.*, 2014) and only a few genomes of isolated Verrucomicrobia strains have been sequenced (Letzel *et al.*, 2013). Bioinformatic analysis on the genomes available have recently identified potential in Verrucomicrobia strains for novel NRPS and PKS clusters (Letzel *et al.*, 2013) and the analysis presented in the current work emphasize the potential of this phylum in driving the diversity particularly of NRPS. Taxonomic lineage assignation of NRPS and PKS reads showed that only a small fraction of the domains sequenced were associated to Verrucomicrobia and Bacteroidetes. This lower abundance could be due, not only to the general lower abundance of these phyla in the microbial community, but also to a lower representation in the database available for the assignation providing fewer matches to these groups for sequenced reads. Most of the NRPS and PKS reads were indeed assigned to Actinobacteria, Proteobacteria and Firmicutes, which are better characterised groups and more information is available on different databases. For both types of genes, already-known producers were well represented in the majority of samples. The *Streptomyces* genus was the most abundant of the Actinobacteria phylum and within this genus NRPS and PKS genes belonging to some of the medically and commercially relevant natural product producers such as *Streptomyces venezuelae* (producer of the NRP chloramphenicol), *Streptomyces cattleya* (NRP - cephamycin), *Streptomyces hygroscopicus* (NRP/PK - rapamycin), *Streptomyces griseus* (PK - fredericamycin), *Streptomyces viridochromogenes* (NRP - laspartomycin), *Streptomyces albus* (NRP/PK - oxazolomycin), *Streptomyces flavogriseus* (PK - xantholipin) and *Streptomyces peucetius* (PK - doxorubicin) were present in high abundance. In addition, other

relevant species for medically important secondary metabolites production such as *Saccharopolyspora erythraea* (producer of the PK erythromycin), *Amycolatopsis orientalis* (NRP – vancomycin) and *Amycolatopsis balhimycin* (NRP – balhimycin) were assigned to NRPS and PKS reads. The ability to assign a taxon to nucleotide reads of functional genes related to natural products enabled the investigation of biosynthetic potential within the soils studied identifying possible microbial groups involved in the biosynthesis. However, some limitations must be taken into consideration. Although the advent of cheap genome sequencing has allowed the study of more producer strains and the identification of novel BGCs (Donadio *et al.*, 2007, Letzel *et al.*, 2013), the ability to annotate sequences produced through NGS as in the current study is still limited by the databases available (Brewer *et al.*, 2016). Therefore, for some of the phyla which are more reluctant to cultivation in laboratory conditions, such as Verrucomicrobia and Bacteroidetes, the information available in most databases is likely not to be fully representative of the real potential of these microbial groups. Nevertheless, this study provided interesting correlations with novel candidate drivers of PK and NRP natural products supported by the amplifying through NGS technology of some of the genes belonging to those phyla.

Identification of potential hotspots for particular bacterial groups driving secondary metabolite diversity was achieved through Procrustes superimposition analyses. This approach is of great interest for either culture-based or metagenomic future exploitation of targeted environments. In particular, Antarctic soil proved to be a hotspot not only for NRPS and PKS in Actinobacteria and Proteobacteria, but also for the newly identified secondary metabolite diversity drivers Verrucomicrobia and Bacteroidetes phyla. Previous studies showed that the Antarctica is a potential hotspot for microbial diversity, in particular for Actinobacteria and Proteobacteria, even though not entirely accessible through the available culture techniques (Babalola *et al.*, 2009). Algerian and Cuban soils were the other two environments identified as possible hotspots for diversity in those phylogenetic drivers. Algerian soils in particular were previously identified as a source of new Actinobacteria species with potential for novel drugs discovery (Selama *et al.*, 2014). The Cuban soils have not been previously studied for natural product diversity but proved to be hotspots for enzymes of biotechnological impact such as chitinase mainly recovered by Actinobacteria and Proteobacteria (Johnson-Rollings *et al.*, 2014).

This current study has pioneered methods to identify members of the community driving the diversity of secondary metabolites in soils. This information will facilitate improved exploration of these soils either by a focused isolation programme on the new phyla and/or modifying existing metagenomic pipelines with more compatible heterologous hosts for these groups to facilitate natural product biosynthesis.

This study introduced a new application for the primers used to amplified NRPS and PKS sequences, whose design and characterization are described more in detail in Chapter 5. According to InterPro functions and Unipept taxonomy, the primers delivered good specificity for the target of interest and an extensive range of phylogenetic diversity was captured. Therefore, the use of these primers in NGS approaches to evaluate the richness of different environments and so increase the chances to identify novel hotspots for antibiotic discovery programs. Moreover, the integration of two different bioinformatic pipelines allowed the study of secondary metabolites taxonomy and the correlation with the general microbial community. In conclusion, this is the first study to identify taxonomic groups capable of driving unique diversity within secondary metabolite gene clusters. This is important not only in ecological terms but also in improving our ability to develop targeted drug discovery initiatives according to the location where there is a higher BGC richness of interest. Alternately, it could be argued that going for the less studied microbial taxa such Verrucomicrobia could lead to improve discovery campaigns as was for the rare Actinobacteria (Lazzarini *et al.*, 2000, Pozzi *et al.*, 2011, Tocchetti *et al.*, 2015).

The identification of hotspots for potential unique diversity of PKS and NRPS genes was then further exploited using metagenomic libraries as described in Chapter 5.

Chapter 5

New tools for metagenomic library exploitation

5.1 Introduction and aims

The advent of metagenomics has opened new horizons in the exploitation of biological diversity, in particular for novel natural product discovery (Handelsman *et al.*, 1998, Rondon *et al.*, 2000). The ability to capture eDNA in metagenomic libraries for novel natural products has proved to be a successful strategy, discovering novel compounds such as turbomycin A and B (Gillespie *et al.*, 2002), utahmycins A and B (Bauer *et al.*, 2010), tetarimycin A (Kallifidas *et al.*, 2012), fasamycins A and B (Feng *et al.*, 2012) and aryxanthomycins A-C (Kang & Brady, 2014).

Different challenges and issues can be related to functional metagenomics starting from the creation of the library in a suitable vector and host, the screening methods to identify the clones of interest and ultimately the expression of genes with unknown origins in different heterologous hosts (see review: Lam *et al.* (2015), Borsetto & Wellington (2017)).

The ability to capture large fragments of eDNA is an important requirement when performing functional metagenomics to screen for natural products. Fosmid and cosmid vectors have been commonly used for the preparation of large metagenomic libraries. With the aid of phage packaging systems to create the libraries, fosmid and cosmid vectors can usually capture approximately 40 kb of eDNA per vector (Kim *et al.*, 1992, Brady, 2007, Owen *et al.*, 2013). Natural product biosynthetic clusters could readily exceed the maximum potential size of DNA packaged into the fosmid and cosmid vectors, due to the presence of regulatory genes, resistance genes and tailoring reaction genes in addition to the core biosynthetic genes. For example, the compound avermectin or teicoplanin-like clusters which are respectively 82 kb and 73 kb (Ikeda *et al.*, 1999, Sosio *et al.*, 2004), are subsequently too large to be recovered in a single fosmid or cosmid vector. Recombination systems such as the transformation-associated recombination (TAR) system in *Saccharomyces cerevisiae* have been developed to recombine partial biosynthetic clusters contained in multiple clones from the same metagenomic library (Feng *et al.*, 2010, Kim *et al.*, 2010, Kallifidas & Brady,

2012) and bacterial artificial chromosome (BACs) vectors have been engineered to capture up to 300 kb of DNA (Shizuya *et al.*, 1992, Ioannou *et al.*, 1994). In particular, BAC vectors able to transfer to different heterologous hosts have been developed to help the expression of novel compounds in multiple hosts (Sosio *et al.*, 2000, Martinez *et al.*, 2004, Aakvik *et al.*, 2009). Different studies demonstrated the importance of using diverse hosts to increase the possibility of expressing different compounds allowing the detection of a higher number of clones of interest according to different phenotypes and antibacterial activity (Martinez *et al.*, 2004, Craig *et al.*, 2009, Baltz, 2010, Craig *et al.*, 2010, Gomez-Escribano & Bibb, 2011). Phenotypic screening could also provide a useful tool to enrich the metagenomic library for clones containing the genes of interest. Charlop-Powers *et al.* (2013) showed a phenotypic selection on iron limited medium for clones containing environmental genes able to restore phosphopantetheine transferase (PPTase) activity in the host allowing the use of iron to sustain growth. Usually PPTase genes are present in the NRPS/PKS secondary metabolite gene clusters, therefore the selection enriched for these types of clusters. The other common screening method is an amplicon based PCR screening for genes of interest. The approach uses primers targeting genes either related to a specific class of antibiotics such as OxyC-like genes, which are usually found in glycopeptide gene clusters (Banik & Brady, 2008), or targeting more general conserved domains of the biosynthetic genes, such as the adenylation domain for NRPS and the ketosynthase domain for PKS (Bauer *et al.*, 2010). NRPS and PKS thiotemplate systems are usually structured in modules (initiation, elongation and termination) (Figure 5.1) and along each module essential domains perform the same type of activity (Table 5.1) (for further details see reviews: Shen (2003), Jenke-Kodama *et al.* (2005), Fischbach & Walsh (2006), Donadio *et al.* (2007), Weissman (2014)).

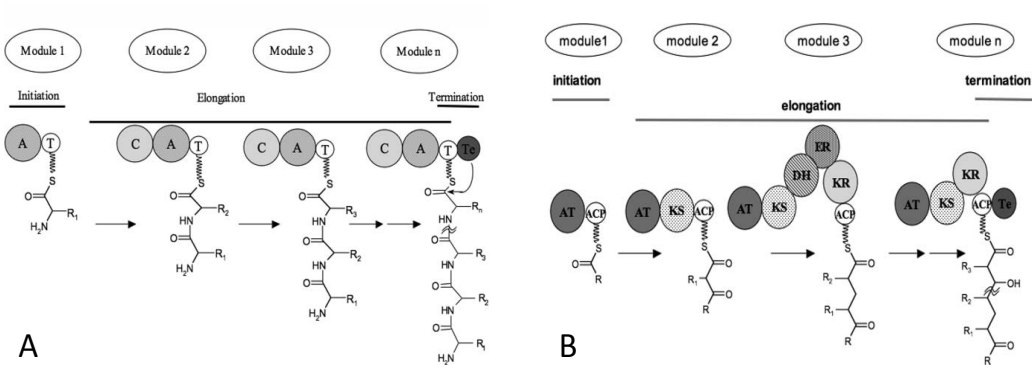


Figure 5.1 Schematic representation of: A) NRPS structure; B) PKS structure (Donadio *et al.*, 2007).

Table 5.1 Brief description of main domains present in NRPS and PKS systems and their functions.

Domain function	NRPS domain	PKS domain
Recognition of specific starter unit and activation	Adenylation (A): adenylation	Acyltransferase (AT): phosphopantetheinylation
Binding extender activated units	PCP	ACP
Elongation by addition of an activated unit	Condensation (C): Peptide bond	Ketosynthase (KS): decarboxylative condensation (C-C bond)
Termination via hydrolysis or macrocyclization	Thioesterase (TE)	Thioesterase (TE)
Additional functional domains present in elongation modules	Amino acid epimerization (E)	β-ketoreductase (KR)
	Methylation (M)	Dehydratase (DH)
	Reduction (R)	Enoylreductase (ER)

Over the past two decades, the conserved regions in the adenylation and ketosynthase domains of NRPS and PKS genes have been the targets for the design of multiple degenerate probes (Ayuso-Sacido & Genilloud, 2005, Schirmer *et al.*, 2005, Reddy *et al.*, 2012, Amos *et al.*, 2015). Owen *et al.* (2013) demonstrated the potential of some of these probes in the high-throughput screening of a metagenomic library, which recovered multiple clones containing secondary metabolites gene clusters. However, comparative studies investigating the suitability of primers targeting the same domain have not been reported before, therefore the contribution of each primer set to the identification of novel clusters remains unknown.

The main aims of the current study were firstly to further characterize degenerate primers previously designed in the Wellington group at the University of Warwick and compare them with the already available primer sets used for novel natural product discovery reported in the literature; secondly to recover and express potential

secondary metabolites clusters from metagenomic libraries and thirdly to engineer novel vectors and heterologous hosts for future metagenomic libraries.

5.2 Material and methods

5.2.1 Bacterial strains

Various bacterial strains were used to perform molecular cloning, extraction of genomic and plasmid DNA, metagenomic libraries storage and screening for antimicrobial clusters (Table 5.2).

Table 5.2 Strains used in this study

Strain	Genotype and comments	Incubation T (°C)	References
E. coli EPI300™-T1R	*F⁻ mcrA Δ(mrr-hsdRMS-mcrBC) Φ80dlacZΔM15 ΔlacX74 recA1 endA1 araD139 Δ(ara, leu)7697 galU galK λ⁻ rpsL (*Str^R*) nupG trfA tonA*	37	Epicentre®
E. coli BW25113/pKD20	K-12 derivative (Δ*araBAD*, Δ*rhaBAD*) carrying plasmid pKD20	30	(Datsenko & Wanner, 2000)
E. coli ET12567/pUZ8002	*dam-13:: Tn9 dcm-6 hsdM* Chl^R, carrying helper plasmid pUZ8002	37	(Paget *et al.*, 1999)
E. coli ET12567	*dam-13:: Tn9 dcm-6 hsdM* Chl^R	37	ATCC
E. coli TOP10 pR9406	*Dam+, dcm+* carrying the helper plasmid pR9406	37	
E. coli JM109	*endA1, recA1, gyrA96, thi, hsdR17 (rk–, mk+), relA1, supE44, Δ(lac-proAB), [F′ traD36, proAB, laqIqZΔM15]*	37	Promega
E. coli S17-1λ*pir*	TpR, SmR, recA, thi, pro, hsdR-M+RP4: 2-Tc:Mu: Km, Tn7, λpir	37	
E. coli JM109 pJOC100Km	Ap^R, Km^R; pJOC100-derivative containing miniUIB100-Km (IRL-Km^R-IRR-IS*Ppu12* without IRs)	37	(Christie-Oleza *et al.*, 2013)
S. coelicolor M145	SCP1- SCP2- derivative from A3(2) [wild type]	30	(Bentley *et al.*, 2002)
S. coelicolor M1152	M145 derivative Δ*act* Δ*red* Δ*cpk* Δ*cda rpoB*(C1298T)	30	(Gomez-Escribano & Bibb, 2011)
Streptomyces spp. 38640		30	KtedoGen collection
Actinoallomurus sp. 145689		30	KtedoGen collection
Pseudomonas putida KT2440	rmo- mod+	30	DMSZ
Pseudomonas putida BIRD-1	plant growth-promoting rhizobacteria (PGPR)	30	(Matilla *et al.*, 2011)

All bacterial strains were stored as liquid cultures or spore suspensions at - 20 °C (WCB) and - 80 °C (MCB) in 10 % (v/v) glycerol.

E. coli and *Pseudomonas* sp. liquid cultures were harvested after an overnight incubation at 37 °C or 30 °C at 200 rpm according to the strain optimal conditions. Actinomycetes strains were grown in 300 mL Erlenmeyer baffled flask containing 50 mL of culture at 30 °C at 200 rpm and collected after several days of growth (up to six days) in order to obtain biomass for DNA extraction. Spores of *Streptomyces* strains were collected from agar plates grown at 30 °C for five days.

All *E. coli* and *P. putida* strains were grown on Luria-Bertani liquid (10 g/L tryptone, 5 g/L yeast extract, 10 g/L sodium chloride) and solid (addition of 15 g/L agar) medium with the addition of different antibiotics and inducers according to the strains requirements.

Streptomyces spp. and *Actinoallomurus* sp. strains were grown on the following solid and liquid media:

- Oat Meal Agar (60 g/L oat meal, 20 g/L agar, 1 mL/L Trace Element Solution);
- SFM (20 g/L Soybean flour, 20 g/L D-mannitol, 20 g/L agar);
- GYM (4 g/L glucose, 4 g/L yeast extract, 10 g/L malt extract);
- 2xYT (16 g/L bactopeptone, 10 g/L yeast extract, 5 g/L NaCl), AF-N (1 g/L yeast extract, 10 g/L dextrose, 1.5 g/L MOPS, 0.5 g/L NaCl, 2 g/L L-Glutamine);
- AF-A (10 g/L dextrose, 4 g/L soybean meal, 1 g/L yeast extract, 0,5 g/L NaCl, 1.5 g/L MES);
- V6 (20 g/L dextrose, 5 g/L yeast extract, 5 g/L meat extract, 5 g/L peptone, 3 g/L casein hydrolizate, 1,5 g/L NaCl).

Specifically, *Streptomyces* strains were grown in the liquid media GYM, 2xYT and V6 at pH 7, while *Actinoallomurus* spp. 145689 was grown in AF-A at pH 5.5. The use of glass beads or springs was adopted in 50 mL and 300 mL Erlenmeyer flaks when baffled flasks were not available in order to obtain a better dispersion of the mycelium.

5.2.2 Vectors

Vectors used in the current study (Table 5.3) were maintained and replicated in various *E. coli* strains (e.g. DH5α, EPI300™-T1R or TOP10). They were extracted using GeneJET Plasmid Miniprep kit (Thermo Scientific™) and DNA stored at - 20 °C. For BAC vectors exceeding 10 kb in size, the BAXMAX™ DNA purification kit (Epicentre®) was used to recover intact BAC DNA vectors for further digestion, sequencing and expression experiments.

Table 5.3 Vectors used in this study

Vectors	Description and comments	References
pCC2FOS™	Metagenomic library creation, Chl^R	Epicentre®
pCC1BAC™	Original BAC vector, Chl^R	Epicentre®
pKD20	λ-RED (*gam*, *bet*, *exo*), *bla*, *araC*, *repA101*^{ts}	(Datsenko & Wanner, 2000)
pUZ8002	Conjugative helper plasmid, *tra*, *neo*, RP4	(Paget *et al.*, 1999)
pR9406	Conjugative helper plasmid, $Carb^R$, derivative of pUB307	(Jones *et al.*, 2013)
pSET152	*aac(3)IV*, *oriT*, *ΦC31 attP*, *int ΦC31*, *lacZα*, *rep*^{pUC}	(Bierman *et al.*, 1992)
pK18MobSac	*aphI* – Kanamycin cassette source	(Schafer *et al.*, 1994)
p34S-Gm	*aacC1*- Gentamicin cassette source	(Dennis & Zylstra, 1998)
pJB658	*trfA* gene source	(Aakvik *et al.*, 2009)
pGEM®-T Easy	Cloning vector	Promega
pBBR1MCS-2	Cloning vector	(Kovach *et al.*, 1995)
pJOC100Km	Cloning transposone vector; Ap^R, Km^R; IRL-Km^R-IRR-IS*Ppu12* without IRs;	(Christie-Oleza *et al.*, 2013)
pBCLI::GSTA	Ap^R, $Gent^R$; IRL-$Gent^R$-p^{sucC}-*trfA*-IRR-IS*Ppu12* without IRs	This study
pBCaBAC	Ap^R; *aac(3)IV*, *oriT*, *ΦC31 attP*, *int ΦC31*	This study
pBCkBAC	Km^R; *aphI, oriT*, *ΦC31 attP*, *int ΦC31*	This study

5.2.3 Primers for screening and cloning

Metagenomic libraries were screened using the primers described in Table 5.4. The amplification conditions were optimized on genomic DNA and metagenomic library pools in order to reduce background noise and identify positive clones containing NRPS and PKS genes. Two sets of primers were already available in the literature,

while the other two were previously designed in the Wellington group at the University of Warwick (Amos *et al.*, 2015).

Table 5.4 Primers for PCR screening of metagenomics libraries.

Gene targeted	Name	Sequence	Annealing T (°C)	Amplicon size (bp)	Reference
NRPS	NRPS_F NRPS_R	CGCGCGCATGTACTGGACNGGNGAYYT GGAGTGGCCGCCCARNYBRAARAA	63	480	(Amos *et al.*, 2015)
PKS	PKS_F PKS_R	GGCAACGCCTACCACATGCANGNNYT GGTCCGCGGGACGTARTCNARRTC	61.9	350	(Amos *et al.*, 2015)
NRPS	A3F A7R	GCSTACSYSATSTACACSTCSGG SASGTCVCCSGTSCGGTA	59.3	700	(Ayuso-Sacido & Genilloud, 2005, Reddy *et al.*, 2012, Owen *et al.*, 2013)
PKS	degKS2F degKS2R	GCIATGGAYCCICARCARMGIVT GTICCIGTICCRTGISCYTCIAC	57.1	760	(Schirmer *et al.*, 2005, Reddy *et al.*, 2012, Owen *et al.*, 2013)

Fosmids, BAC vectors and heterologous hosts were engineered using different primer sets designed in the current study (Table 5.5). The engineered vectors and the positive fosmids/ BACs identified in the metagenomic/ genomic libraries were sequenced using the primers reported in Table 5.6.

Table 5.5 Primers designed in this study for cloning experiments.

Targets and comments	Name	Sequence
pCC1BAC™/pSET152::*int*	int_cm_KO_R	ACCAGGCGTTTAAGGGCACCAATAACTGCCTTAAAAAAACTACGCCGCTACGTCTTCCG
pCC1BAC™/pSET152::*aac(3)IV*	apra_Cm_KO_F	GAGTTATCGAGATTTTCAGGAGCTAAGGAAGCTAAAATGGTGCAATACGAATGGCGAAAAGCCG
pCC1BAC™	Cm_out_F	AACCAGGCGTTTAAGGGCA
pCC1BAC™	Cm_out_R	AACTTTTGGCGAAAATGAGACG
pCC1BAC™	Cm_check_F	ACCGTAACACGCCACATCTT
pCC1BAC™	Cm_check_R	TTCTTGCCCGCCTGATGAAT
pBCaBAC	attP_F	AACCGCTTCTGACCTGGG
pBCaBAC	attP_R	TTCCCAGGTCAGAAGCGG
pBCaBAC	traJ_F	TTCGGTGATGCCACGATCC
pBCaBAC	ori_R	ATGTACTTCACCAGCTCCG
pBCaBAC	Apra_check_F	TCGAGAATGACCACTGCTGT
pBCaBAC	Apra_check_R	CGGATGCAGGAAGATCAACG
pBCaBAC	Int_check_F	GAAGTCGGCGAAGATTCTCG
pBCaBAC	Int_check_R	GGAAGGTGTTTGTGCGTCTT
pBCaBAC	Int_seq_F	ATGCTGACGTTGCACGTG
pBCaBAC	Int_seq_R	AAGGACTCTTACCGCTGCC
pBCaBAC/pK18Mob sacB	KanaIN_apraOUT_F	CGAGATTTTCAGGAGCTAAGGAAGCTAAAATGATTGAACAAGATGGATTGCACGCAGGT
pBCaBAC/pK18Mob sacB	KanaIN_apraOUT_R	CCAACGTCATCTCGTTCTCCGCTCATGAGCTCAGAAGAACTCGTCAAGAAGGCGATAGA
pBCkBAC	Kana_check_F	GAACAAGATGGATTGCACGC
pBCkBAC	Kana_check_R	AGAACTCGTCAAGAAGGCGAT
p34S-GM+KpnI/NdeI	p34_KpnI/NdeI_F	GGTACCCATATGGAGCTCGAATTGACATAAGCCTG
p34S-GM + SalI	p34_SalI_R	TTGTCACAACGCCGCGGCCAAGTCGAC
sucC promoter *P. putida* KT2440 + SalI	promSucC_SalI_F	GTCGACTTCGGCACCAGTGCCGAG
sucC promoter *P. putida* KT2440 + SalI	promSucC_SalI_R	GTCGACGCTTTTTACCGTCTTCGT
pJB658 + SalI	trfA_SalI_F	GTCGACATGAATCGGACGTTTGAC
pJB658 + SacI	trfA_SacI_R	GAGCTCAGGCCCTAGCGTTTGCAA

Table 5.6 Primers for Sanger sequencing.

Target	Name	Sequence	Reference
pCC1BAC™	pCC1_F	GGATGTGCTGCAAGGCGATTAAGTTGG	(Epicentre, 2013)
pCC1BAC™	pCC1_R	CTCGTATGTTGTGTGGAATTGTGAGC	(Epicentre, 2013)
pCC2FOS™	pCC2_F	GTACAACGACACCTAGAC	(Epicentre, 2012)
pCC2FOS™	pCC2_R	CAGGAAACAGCCTAGGAA	(Epicentre, 2012)
BACs/Fosmids	M13_F	GTAAAACGACGGCCAGT	
BACs/Fosmids	M13_R	CAGGAAACAGCTATGAC	

5.2.4 PKS and NRPS primers design and characterization

5.2.4.1 PKS and NRPS primers design

Degenerate PKS and NRPS primers were previously designed in the Wellington group at University of Warwick targeting ketosynthase and adenylation domains respectively. BLOCKMAKER and CODEHOP (Rose *et al.*, 2003) were used to design primer sets from nine adenylation domain (Table 5.7) and 18 type-II KS_{α} genes (Table 5.8) retrieved from GeneBank.

Table 5.7 Nonribosomal peptide synthases used for NRPS_F/R primer design (Amos *et al.*, 2015).

Accession	Description
gi|2894188|	PCZA363.3 [*Amycolatopsis orientalis*]
gi|4481933|	CDA peptide synthetase II [*Streptomyces coelicolor* A3(2)]
gi|4481934|	CDA peptide synthetase I [*Streptomyces coelicolor* A3(2)]
gi|45006|	Alpha-aminoadipyl-L-cysteinyl-D-valine synthetase [*Amycolatopsis lactamdurans*]
gi|987101|	Pipecolate incorporating enzyme [*Streptomyces rapamycinicus*]
gi|3798625|	GFK506 peptide synthetase [Streptomyces sp. MA6548]
gi|2052277|	Virginiamycin S synthetase [*Streptomyces virginiae*]
gi|2052249|	Pristinamycin I synthase 3 and 4 [*Streptomyces pristinaespiralis*]
gi|5051823|	Putative peptide synthetase [*Amycolatopsis orientalis*]

Table 5.8 Ketoacylsynthases used for PKS_F/R primer design (Amos *et al.*, 2015).

Accession	Description
gi|125235	KAS1_STRCO Putative polyketide beta-ketoacyl synthase 1 (WhiE ORF III)
gi|729871|	KAS1_STRHA putative polyketide beta-ketoacyl synthase 1 (KS) (polyketide condensing enzyme)
gi|729870|	KAS1_STRCN putative polyketide beta-ketoacyl synthase 1
gi|15823945|	3-oxoacyl-(acyl carrier protein) synthase I [*Streptomyces avermitilis*]
gi|11024335|	PKSA beta-ketoacylsynthase subunit alpha; PKSA-ORF1 [*Streptomyces collinus*]
gi|7209628|	Ketosynthase [*Streptomyces nogalater*]
gi|7209626|	Ketosynthase [*Streptomyces venezuelae*]
gi|2580442|	ORF 1 [*Actinomadura hibisca*]
gi|7433744|	Polyketide synthase *Actinomadura hibisca*
gi|5381247|	Polyketide synthase [*Actinomadura verrucosospora*]
gi|14486277|	B-ketoacyl-ACP synthase-like protein [*Streptomyces aureofaciens*]
gi|125237|	KAS1_STRVN granaticin polyketide putative beta-ketoacyl synthase 1
gi|510722|	jadomycin polyketide ketosynthase; JadA [*Streptomyces venezuelae* ATCC 10712]
gi|1076101	ketosynthase – *Streptomyces griseus*
gi|532245|	daunorubicin-doxorubicin polyketide synthase
gi|516109|	polyketide synthase [*Streptomyces*]
gi|7209618|	ketosynthase [*Streptomyces aureofaciens*]
gi|7209610|	ketosynthase [*Streptomyces capoamus*]

5.2.4.2 PKS and NRPS primers characterization

Primer pairs PKS_F/R and NRPS_F/R were initially tested against genomic DNA extracted form a variety of Actinomycetes strains using a PCR approach. In particular, PCR products obtained from the amplification of adenylation and ketosynthase domains of *S. coelicolor* M145 were cloned in the vector pGEM®-T Easy in *E. coli* JM109 and 96 products for each domain were sequenced by GATC Biotech, Germany. Sequences were identified using Blastn against *S. coelicolor* M145 genome sequence and the location on biosynthetic clusters was reported using the annotation obtained by antiSMASH v3.

5.2.5 Metagenomic libraries

5.2.5.1 Preparation and storage

Metagenomic libraries were prepared using the CopyControl™ Fosmid Library Production Kit with pCC2FOS™ Vector (in particular the MaxPlax™ Lambda Packaging Extract) following manufacturer's instructions (Epicentre, 2012). Briefly, DNA was extracted from soil using the lysis buffer containing CTAB and SDS described in Brady (2007), then eDNA was precipitated using 100 % isopropanol and cleaned from humic acids and short fragments using 1 % (w/v) agar gel electrophoresis overnight at 30V at 4 °C. HMW DNA was then recovered using dialysis and concentrated using Amicon® ultra centrifugal filter columns. The recovered clean DNA was then used to perform a ligation with pCC2FOS™ vector (Epicentre®) and phage packaging reaction to transfect *E. coli* EPI300™-T1R (Epicentre®) according to Epicentre (2012) protocol. Transfected *E. coli* colonies were selected on LB agar plates containing adequate antibiotic marker. After the overnight incubation at 37 °C, single colonies were individually picked in 96-well plates, grown overnight at 37 °C and stored at -20 °C after addition of glycerol (15% v/v final concentration).

5.2.5.2 PCR screening

Metagenomic libraries were screened by PCR using degenerated primers (Table 5.4) available from the literature and previously designed at University of Warwick (Amos *et al.*, 2015).

For single plate screening and identification of the single positive clones, colonies were incubated overnight in 96-well plate with LB + chloramphenicol 12.5 µg/mL + CopyControl™ Fosmid Autoinduction Solution 1X. To extract DNA from colonies a boiling method was used: for each single plate or row/column from a single plate 10µL were taken from each well and pooled together. Cells were pelleted at 13000 rpm for 5 min, resuspended in 100 µL of sterile water, vortex for 30 s, incubated at 100 °C for 5 min and used as template (1-2 µL per PCR reaction).

PCR reactions were set using 12.5 µL Master Mix 2X (Promega), 1.25 µL DMSO, 0.8 µM Primer Forward, 0.8 µM Primer Reverse, 1-2 µL DNA template and dH_2O up to

25 µL (final PCR reaction volume). For each primer set, *S. coelicolor* M145 genomic DNA was used as positive control and dH_2O as negative control. In all PCR runs, an initial denaturation at 95 °C for 5 min was performed, followed by 40 cycles of denaturation at 95 °C for 30 s, annealing at 54-63 °C (depending on the primer set) for 45 s and extension at 72 °C for 1 min 30 s. A final extension was run at 72 °C for 10 min before cooling to 4 °C. In order to optimise the PCR conditions for each primer sets, different annealing temperatures were tested using gradient PCR and the best temperature for each primer set was used for the screening (Table 5.4).

PCR products were verified by gel electrophoresis in 1 % (w/v) agarose gel. Orange F loading dye 6X (ABNOVA) was added to PCR products to a final concentration of 1X and loaded into the gel. Different DNA ladders were used according to the product size. Each run was performed at 100 V for approximately 50 min in Tris-Acetate EDTA (TAE) buffer 1 X.

Positive bands on gel or relative PCR products were purified with Qiagen Gel purification kit (Qiagen, UK) or Qiagen PCR product purification kit (Qiagen, UK) and sequenced by GATC Biotech, Germany.

5.2.5.3 Fosmids isolation, sequencing and analysis

GeneJET Plasmid Miniprep Kit (Thermo Scientific™) was used to purify fosmid DNA from 10 mL overnight culture of single clone in order to obtain high quality DNA for sequencing.

Fosmid DNA was prepared for Illumina® HiSeq sequencing using Nextera® EX (Illumina®) library kit preparation as per manufacture's instruction. The company performed quality filtering of the reads using FASTQ for quality trimming, *de novo* assembly trough SPAdes (Bankevich *et al.*, 2012) and annotation using Prokka (Seemann, 2014). Contigs containing either annotated PKS or NRPS genes were analysed using antiSMASH v3 for antibiotic cluster investigation (Weber *et al.*, 2015).

5.2.6 BAC vector

5.2.6.1 Engineering of the BAC vectors

The BAC vector pCC1BAC™ (Epicentre®) was engineered to replace the chloramphenicol resistance with a cassette containing the *ΦC31 attP* site, an integrase gene and the apramycin/ kanamycin resistance marker (Figure 5.2).
A PCR-targeting approach as described by Gust *et al.* (2004) was performed using the primers int_cm_KO_R/Apra_cm_KO_F (Table 5.5). The vector pCC1BAC™ DNA was transformed into electrocompetent *E. coli* BW25113/pKD20. The positive clones containing the BAC vector were then transformed with the *ΦC31 attP* site-integrase-apramycin resistance cassette amplified using high-fidelity DNA polymerase (Phusion DNA polymerase, NEB) according to manufacturer's instructions using pSET152 as DNA template and int_cm_KO_R/Apra_cm_KO_F primers (Table 5.5). Positive clones were selected on LB+Apramycin 50 mg/L and tested using primers sets targeting the apramycin resistance cassette (Apra_check_F/Apra_check_R), the integrase gene (int_check_F/int_check_R) and the chloramphenicol gene (Cm_out_F/Cm_out_R) separately (Table 5.5). The DNA insert was also sequenced using Sanger sequencing (GATC, Germany) with different primers sets (Table 5.5 - 1: Cm_out_F/int_seq_R = 715bp; 2: int_seq_F/int_check_F = 783bp; 3: int_check_R/attP_R = 800bp; 4: attP_F/ori_R = 953bp; 5: traJ_F/Apra_check_F = 762bp; 6: Apra_check_R/Cm_out_R = 852bp). The apramycin resistance gene was then replaced with a kanamycin resistance gene amplified using the primers KanaIN_apraOUT_F/KanaIN_ApraOUT_R and pK18MobSac vector as DNA template. The replacement was performed using the same technique previously described (Gust *et al.*, 2004) and *E. coli* BW25113/pKD20/pBCaBAC. Mutants were selected on LB agar + kanamycin 25 mg/L and checked by colony PCR using the following primer sets: Kana_check_F/Kana_check_R. The full integrated kanamycin gene was also sequenced using Sanger sequencing (GATC, Germany).

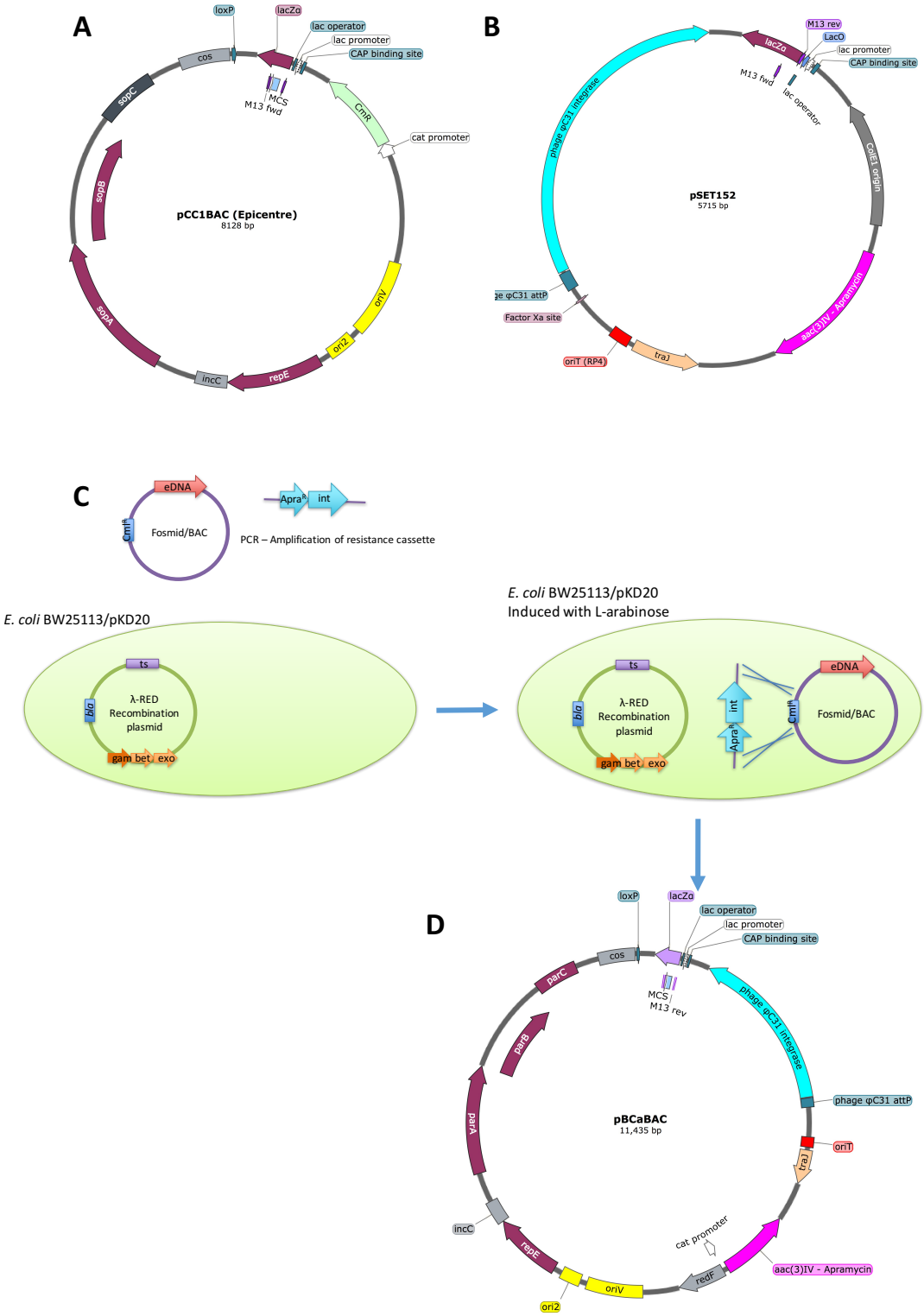


Figure 5.2 Schematic representation of the engineering process to modify the BAC vector. A) Map of the original pCC1BAC™ (Epicentre®); B) Map of the vector pSET152 from which the cassette containing the apramycin resistance gene, *ΦC31 attP* site and integrase gene was amplified; C) *E. coli* BW25113/pKD20 recombination system used to modify pCC1BAC™ with the PCR amplified cassette from pSET152 using homologous recombination of sequences flanking the chloramphenicol resistance gene to be replaced; D) Map of the engineered pBCaBAC.

5.2.6.2 Cloning of Actinomycetes HMW DNA in the new BAC vectors

Genomic DNA from *Streptomyces sp.* 38640 and *Actinoallomurus* sp. 145698 was extracted using an enzymatic lysis of the mycelium followed by phenol/chlorophorm treatment and isopropanol precipitation. In brief, 5 mL of mycelium were recovered and washed with 10 mL of 10.3 % (w/v) sucrose and resuspended in 5 mL of SET buffer (sodium EDTA 1mM, Tris-HCl 10 mM). Lysozyme to a final concentration of 20 mg/mL was added and incubation was performed at 37 °C 200 rpm for 30 h. Every 8-12 h 50 mg of lysozyme were added. After 24 h and 28 h of incubation 25 µg of ribonuclease A and 3 mg of proteinase K respectively were added to the mixture and incubation was continued at 37 °C. At the end of the lysis incubation time, sodium dodecyl sulphate to a final concentration of 1 % was added and incubation at 55 °C was performed overnight. After the SDS treatment, 2 mL of 5M NaCl were added to the sample prior purification with 4 mL of phenol:chloroform:isoamyl alcohol. Vortex mixing and centrifugation at 4000 rpm for 10 min were performed and the aqueous layer was recovered and treated a second time with phenol:chloroform:isoamyl alcohol solution and a third time with chloroform:isoamyl alcohol. The final aqueous phase recovered was mixed with an equal volume of ice-cold isopropanol until the formation of the DNA precipitate. DNA was pelleted and washed with ice-cold 70 % ethanol, then dried at 37 °C and resuspended in sterile deionized water.

An *in situ* lysis of the actinomycetes mycelium was also performed in low melting point agarose plugs to prevent shearing of the DNA and recovery of HMW genomic DNA. Fresh or frozen mycelium was washed with 10.3 % (w/v) sucrose, resuspended in TE25SUC (25 mM Tris-HCl pH8, 25 mM EDTA pH 8, 0.3 M sucrose) until an OD_{600nm} of 2, then it was mixed in equal volume with 1.5 % (w/v) low melting point agarose prepared in TE25SUC buffer. Plugs were prepared in the PFGE plug molds with 80 µL of the mixture per well. Plugs were then transferred in 10 mL of TE25SUC with the addition of 1 mg/mL of lysozyme and incubated at 37 °C 200 rpm for 2-3 h. Then plugs were transferred in 10 mL of NDS buffer (1 % SDS, 0.5 M EDTA pH 8, 10 mM Tris-HCl pH8) with 1 mg/mL of proteinase K and incubated at 50 °C for 40 h. After the proteinase treatment, plugs were transferred in 10 mL of TE25 buffer with 0.1 mM phenylmethylsulfonyl fluoride (PMSF) and incubated at 4 °C for 1 h. Plugs

were then washed three times in 10 mL TE25 buffer with 1 h incubation at 4 °C between each wash to allow the diffusion of the washing buffer inside the plugs. Plugs were stored in TE25 buffer at 4 °C.

Partial digestions of genomic DNA recovered from the direct method or the plug system were performed using different concentration of restriction enzyme (0.5 – 1 U BamHI) and incubation time (0-5-10-15-20-30-60 min). For each reaction 5-10 µg of gDNA or up to 3 plugs simultaneously were digested at 37 °C with BamHI (NEB) in CutSmart® Buffer (NEB; 50 mM potassium acetate, 20 mM tris-acetate, 10 mM magnesium acetate, 100 µg/mL BSA, pH 7.9). Digestion reactions were stopped with SDS 0.5 % (v/v final concentration) and checked on a 0.6 % (w/v) agarose gel run in 1 % (v/v) TAE buffer at 50 V at 4 °C for 24-27 h using λ-HindIII marker. Digested plugs were treated with GELase™ (Epicentre®) according to manufacturer's instructions.

5.2.6.3 Integration of the new BAC vectors in *Streptomyces* sp.

The integration of the newly engineered vectors pBCaBAC and pBCkBAC was performed in *S. coelicolor* M1152. *S. coelicolor* does not accept foreign methylated DNA, therefore both pBCaBAC and pBCkBAC were firstly transferred into a methylation-deficient *E. coli* ET12567 strain. The vector pBCaBAC was either electroporated in the electrocompetent *E. coli* ET12567/pUZ8002 or transferred by triparental mating into *E. coli* ET12567/pR9406. The vector pBCkBAC vector could only be transferred to *E. coli* ET12567/pR9406 by triparental mating as *E. coli* ET12567/pUZ8002 already contained the same resistance marker. For the triparental mating, *E. coli* ET12567 (recipient), *E. coli* TOP10 pR9406 (containing the self-transmissible helper plasmid) and *E. coli* EPI300™-T1R pBCaBAC or pBCkBAC (donors) were grown till exponential phase in fresh LB + adequate antibiotic selection at 37 °C 200 rpm. Cells were then recovered by centrifugation, washed twice with fresh LB without antibiotics and resuspended in 0.5 mL of LB. For each strain, 20 µL of cells were spotted onto the same location on LB agar plates without selection and incubated at 37 °C overnight. The next day, *E. coli* ET12567/pR9406 containing either pBCaBAC or pBCkBAC were selected on plates of LB + adequate antibiotic. Colony PCR was performed on ex-conjugants colonies to check the presence of the BAC

vectors. The methylation-deficient strains containing the two newly engineered vectors were then integrated into *S. coelicolor* M1152 by intergeneric conjugation as described by Kieser *et al.* (2000). The strains *E. coli* ET12567/pR9406/pBCaBAC, *E. coli* ET12567/pR9406/pBCkBAC and *E. coli* ET12567/pUZ8002/pBCaBAC were grown in fresh LB with antibiotics until an OD_{600nm} of 0.4-0.6, then cells were washed twice with fresh LB and resuspended in 0.5 mL of LB without antibiotics. While washing the *E. coli* cells, approximately 10^8 spores of *S. coelicolor* M1152 were germinated in 0.5 mL of 2xYT at 50 °C for 10 min. Heat-shocked spores and *E. coli* cells were then mixed in equal volumes and plated on SFM agar + 10 mM $MgCl_2$ and incubated at 30 °C for 16-20 h. An overlay of 1 mL of sterile water with nalidixic acid (25 μg/mL) and either apramycin (50 μg/mL) or kanamycin (200 μg/mL) BAC selection was distributed over the conjugation plates and the incubation at 30 °C was continued for 4-5 days or until ex-conjugant colonies appeared. The ex-conjugants were plated again on SMF agar plates + apramycin 50 μg/mL/ kanamycin 200 μg/mL + nalidixic acid 25 μg/mL to remove any possible residual *E. coli* cells, then they were checked by PCR for the presence of the integrated resistance genes. A conjugation between *S. coelicolor* M1152 and *E. coli* ET12567/pUZ8002/pSET152 or *E. coli* ET12567/pR9406/pSET152 was performed in parallel to the other conjugations as a control.

5.2.7 Heterologous host engineering

Pseudomonas putida KT2440 and BIRD-1 engineering was performed using a restriction enzyme guided cloning method to create a transposon vector with a cassette containing the *trfA* gene of interest under a strong promoter (*sucC* from *P. putida* KT2440) and a gentamicin marker cassette (Gm^R, *aacC1*). The random introduction of the *trfA* gene inside the bacterial genome will allow the replication of the engineered vectors (pBCkBAC and pBCaBAC) inside the hosts activating the origin of replication *oriV* present on the BACs. The *trfA* gene was cloned from pJB658 flanked by SalI and SacI restriction sites, *sucC* promoter was amplified from *P. putida* KT2440 genomic DNA with SalI restriction site at each end and the gentamycin marker from p34S-Gm vector with KpnI/NdeI and SalI restriction sites. All PCR products were amplified using Phusion DNA polymerase (NEB) with primers reported in Table 5.5. All PCR

amplicons were cloned individually into pGEM®-T Easy vector (Promega) following the manufacturer's instruction into *E. coli* JM109. Plasmid minipreps were prepared using GeneJET Plasmid Miniprep kit (Thermo Scientific™) and each gene was then cut using the restriction sites introduced in the primers sequences following the restriction enzymes manufacturer's instructions and cloned sequentially into the cloning vector pBBR1MCS-2 to form the Gm^R-p^{sucC}-*trfA* (GST) cassette flanked by NdeI and SacI restriction sites. The GST cassette was then transferred to the pJOC100Km using restriction enzyme digestion (NdeI and SacI) for both the plasmid containing the GST cassette (pBBR1MCS-2::GST) and the receiving one pJOC100Km. A large ligation reaction (100 µL) was set to ligate the digested pJOC100 and the GST cassette using T4 DNA ligase (NEB) following manufacturer's instructions. The resulting transposon system pBCLI::GSTA was electroporated into *E. coli* S17-1λ*pir* which was used for the conjugation with *P. putida* KT2440 and BIRD-1 according to Christie-Oleza *et al.* (2013). Ex-conjugant colonies were selected on gentamicin and controlled using PCR amplification of the complete GST cassette.

5.3 Results

5.3.1 Characterisation of novel degenerate primers for PKS and NRPS

5.3.1.1 *S. coelicolor* cloning A and KS domains

S. coelicolor M145 was used to characterise the hit rate of the primer sets NRPS_F/R and PKS_F/R as it was a well-known secondary metabolite producer and the genome sequence was available. The analysis of *S. coelicolor* A3(2) genome (GenBank accession number AL645882.2) using antiSMASH v3 (Weber *et al.*, 2015) showed the presence of four NRPS and eight PKS pathways. Within the NRPS clusters, a total of 20 adenylation domains were detected, with the majority of them belonging to the calcium-dependent antibiotic (CDA) cluster (Table 5.9).

Table 5.9 Number of adenylation domains present in *Streptomyces coelicolor* A3(2) NRPS clusters according to AntiSMASH and hit rate of primers for each cluster assuming unbiased and equal binding of primers to each adenylation domain.

Cluster	Type	No. of adenylation domains	BGC hit rate (%)
Coelichelin	Nrps	3	15
CDA	Nrps	11	55
Nogalamycin	Nrps	2	10
Coelibactin	T3pks-Nrps-Terpene	4	20
total		*20*	*100*

Within the eight PKS clusters, 13 ketosynthase domains were detected for type-I PKS and two chain length factors (KS_{β}) for type-II PKS. The cluster with the highest number of KS domains was coelimycin with six domains, followed by undecylprodigiosin with four (Table 5.10).

Table 5.10 Number of ketosynthase domains present in *Streptomyces coelicolor* A3(2) PKS clusters according to AntiSMASH and hit rate of primers for each cluster assuming unbiased and equal binding of primers to each ketosynthase domain according to KS type.

Cluster	Type	No. of KS domains	BGC KS-I hit rate (%)	BGC KS-II hit rate (%)
Leinamycin	Otherks-T1pks	2	15	
Herboxidiene	T3pks	nd	-	
Actinorhodin	T2pks	1(β)	-	50
Spore pigment	T2pks	1(β)	-	50
Undecylprodigiosin	T1pks	4	31	
Coelimycin	Butyrolactone-T1pks	6	46	
Arsenopolyketides	Otherks-T1pks	1	8	
Coelibactin	T3pks-Nrps-Terpene	nd	-	
total		13	100	100

The adenylation and ketosynthase domains were amplified from *S. coelicolor* M145 genomic DNA, cloned into pGEM®-T Easy vector in *E. coli* JM109 and sequenced. For NRPS primer amplicons, 76/ 96 clones containing amplicons from *S. coelicolor* genome were successfully sequenced. All the amplified adenylation domains matched

the calcium-dependent antibiotic cluster, which is the one with the highest number of adenylation domains. This cluster contains three NRPS genes (SCO3230, SCO3231 and SCO3232) with six, three and two adenylation domains respectively (Figure 5.3).

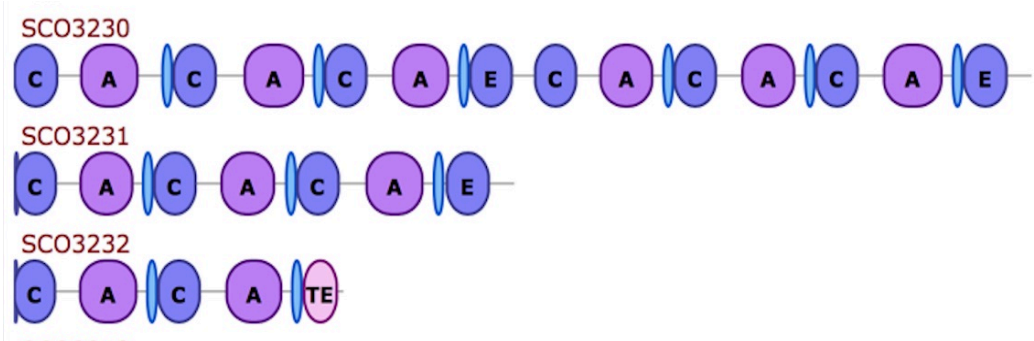


Figure 5.3 Domain organization details of the main NRPS genes (SCO3230, SCO3231 and SCO3232) in CDA cluster. A, adenylation domain; C, condensation domain; TE, thioesterase domain.

The NRPS_F/R primer set amplified multiple adenylation domains, with a higher hit rate for CDA peptide synthase I (SCO3230) with 45 % of sequences, then CDA peptide synthase II (SCO3231) with 33 % and CDA peptide synthase III (SCO3231) with 22 % (Table 5.11). In particular, 18 % of sequences matched the second A domain of CDAI and 11 % matched the fourth and the fifth A domain of CDAI. For CDAII all sequences matched only the first A domain and for CDAIII only the second A domain (Figure 5.4).

Table 5.11 NRPS_F/R primers amplicon distribution on each adenylation domain present in calcium-dependent antibiotic BGC.

Gene	Annotation	Total counts	Domain name	Domain counts
SCO3230	CDA peptide synthetase I	33	AI.1	3
			AI.2	13
			AI.3	0
			AI.4	8
			A1.5	8
			A1.6	1
SCO3231	CDA peptide synthetase II	24	A2.1	24
			A2.2	0
			A2.3	0
SCO3232	CDA peptide synthetase III	16	A3.1	0
			A3.2	16

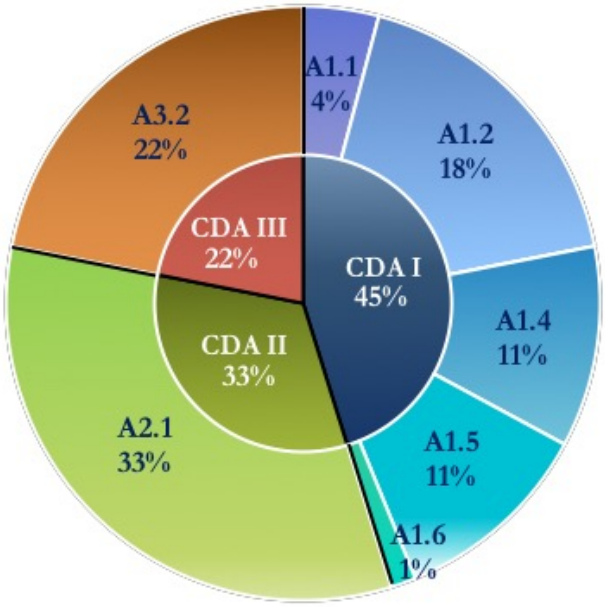


Figure 5.4 NRPS_F/R primers amplicon distribution on each adenylation domain present in calcium-dependent antibiotic BGC. Inner layer: hit rate on single peptide synthetase (CDA I, CDA II and CDA III); outer layer: hit rate per single adenylation domain in each NRPS gene (A1.1, A1.2, A1.3, A1.4, A1.5, A1.6, A2.1, A2.2, A2.3, A3.1 and A3.2).

For PKS primers amplicons, 84/ 96 clones were successfully sequenced. All ketosynthase domains matched the polyketide beta-ketoacyl synthase alpha (SCO5318) of type II PKS cluster for the production of spore pigments.

5.3.1.2 Comparison with other degenerate primers for A and KS domains

After testing the target specificity and the hit rate of primers on one genome, the primers NRPS_F/R and PKS_F/R were compared with other primers on diverse genomic DNA to test the ability to detect different A and KS domains.

Comparison of NRPS _F/R and PKS_F/R primers with two of the most commonly used degenerate primer sets for metagenomic library screening for novel natural products (A3F/A7R and degKS2_F/R) (Ayuso-Sacido & Genilloud, 2005, Owen *et al.*, 2013) showed that there are some differences in detecting known antibiotic clusters in actinomycetes isolated strains (Table 5.12). In particular, NRPS_F/R detected half of the hits amplified by A3F/A7R primers and PKS_F/R detected two additional hits than degKS2_F/R.

Table 5.12 Comparison of primer sets on genomic DNA of different actinomycetes. Examples of hypothetical targets (known antibiotic pathways present in the strains) are reported.

	Antibiotic pathways				PCR results				
Organism	**NRPS**	**PKSI**	**PKSII**	**Hybrid NRPS-PKS**	**NRPS_F/R**	**A3F/A7R**	**A3F/A7R**	**PKS_F/R**	**degKS2F/R**
Micromonospora fulvoviolaceus JCM 3258						+			
Streptomyces avermitilis MA-4680		Avermectin			+	+	+	+	+
Streptomyces coelicolor M1152						+		+	
Streptomyces coelicolor M1154						+		+	
Streptomyces coelicolor M145	CDA		Actinorhodin	Prodigiosin	+	+		+	+
Streptomyces flavogriseus								+	
Streptomyces griseus DSM 40660		Candicidin	Frederica-mycin						
Streptomyces hygroscopicus AM-3672		Herbimycin			+	+		+	+
Streptomyces hygroscopicus NRRL 3602				Geldana-mycin	+	+		+	+
Streptomyces hygroscopicus subsp. glebosus ATCC 14607									
Streptomyces lividans TK24									
Streptomyces parvulus		Borrelidin				+	+		+
Streptomyces rochei DSM 40231	Streptothricin	Lankamycin		Lankacidin	+	+	+	+	+
Streptomyces spectabilis					+	+		+	+
Streptomyces subrutilus						+			
Streptomyces violaceusniger KCC-S0850				Meridamycin	+	+	+	+	+

The creation of a metagenomic library from Cuban soil allowed testing the screening ability of the primers on captured eDNA. The screening of the Cuban metagenomic library with NRPS_F/R, A3F/A7R, PKS_F/R and degKS_F/R identified eight hits for NRPS and no hits for PKS clusters.

In particular, two clones were detected by both primer sets for adenylation domain (NRPS_F/R and A3F/A7R), two clones only with primers NRPS_F/R and two clones only with primers A3F/A7R. The hit rate of NRPS_F/R and A3F/A7R primer sets was

the same on metagenomic library with an additional effect on the number of diverse clones recovered (Table 5.13).

Table 5.13 Results of nucleotide sequences identity of the positive clones identified during the screening for NRPS and PKS genes of the metagenomic library created from Cuban soil using the blastn algorithm.

Clone	Primers set	Species	Annotation	% Identity	Accession No.	E value
ST1P6A4	NRPS2_F/ NRPS_R	*Delftia acidovorans* SPH-1	Amino acid adenylation domain protein	99	CP000884.1	0.0
ST1P6A4	A3F/A7R	*Delftia acidovorans* SPH-1	Amino acid adenylation domain protein	98	CP000884.1	0.0
ST1P6B6	NRPS2_F/ NRPS_R	*Delftia acidovorans* SPH-1	Amino acid adenylation domain protein	98	CP000884.1	0.0
ST1P6B6	A3F/A7R	*Delftia acidovorans* SPH-1	Amino acid adenylation domain protein	98	CP000884.1	0.0
ST1P9E10	NRPS2_F/ NRPS_R	*Saccharothrix espanaensis* DSM 44229	Non-ribosomal peptide synthetase	80	HE804045.1	3e-08
ST1P9D7	A3F/A7R	*Burkholderia gladioli* BSR3 chromosome 2	Arthrofactin synthetase/syringopeptin synthetase C-related non-ribosomal peptide synthetase module	85	CP002600.1	1e-07
ST1P19C8	NRPS2_F/ NRPS_R	*Stenotrophomonas maltophilia* R551-3	amino acid adenylation domain protein	97	CP001111.1	2e-174
ST1P29D1	A3F/A7R	*Streptomyces ansochromogenes*	nrps2 metabolite biosynthetic gene cluster	70	KF170330.1	4e-20

5.3.2 Creation of metagenomic libraries

Total eDNA was extracted from Cuban and Antarctic soils recovering an estimate of 141.8 ng/µL for Cuban soil and 124.5 ng/µL for Antarctic soil after concentration of the eDNA samples through the Amicon® ultra centrifugal filter columns. After the eDNA ligation into fosmid vectors, phages were packaged with fosmids (~ 50 Kb DNA per phage) and transfected into the *E. coli* EPI300™-T1R host recovering approximately 3,000 CFU for Cuban soil, while no colonies were obtained for Antarctic eDNA even after multiple transfection reactions. Potential loss of DNA might have occurred during the DNA purification steps affecting the ligation reaction efficiency.

For the Cuban metagenomic library, 3,000 clones with an approximate insert size of 35-40 kb DNA per fosmid vector captured an estimated size of 105 Mb, which corresponds to approximately a coverage of 24 genomes of 5 Mb each (see review: Land *et al.* (2015)).

5.3.3 Analysis of NRPS clusters recovered from the Cuban metagenomic library

From the Cuban metagenomic library screening six clones amplified NRPS adenylation domains. Only five (ST1P6A4, ST1P6B6, ST1P9D7, ST1P9E10, ST1P29D1) out of six fosmids were successfully sequenced using Illumina HiSeq® technology allowing the assembly of long contigs (>10 kb). For fosmid ST1P19C8 poor quality sequences were obtained from the first sequencing run and only very short contigs (<8-10 kb) were assembled from a second sequencing run. AntiSMASH v3 software identified antibiotic gene clusters for four out of five fosmids. Although ST1P29D1 clone was positive for the initial screening for NRPS adenylation domain (Table 5.11), an antimicrobial biosynthetic gene cluster was not detected by the antiSMASH algorithm. For the other four fosmids, clusters of an average of 29 kb were identified (Figure 5.5).

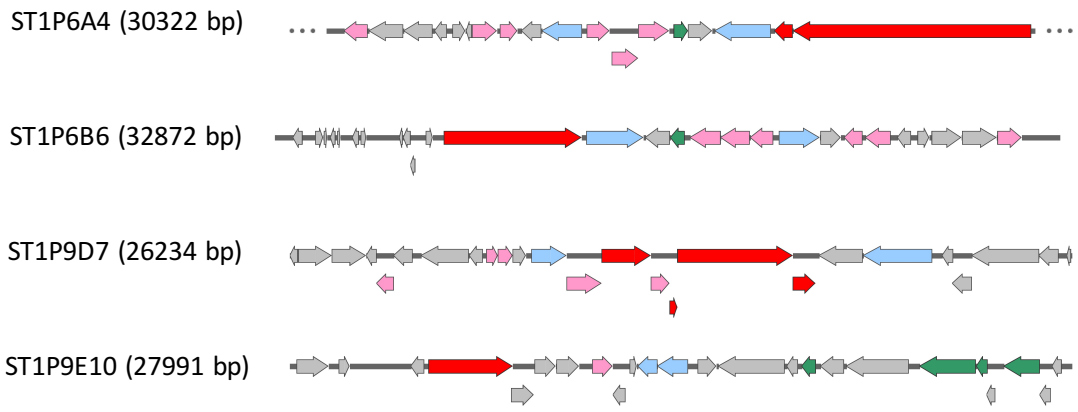


Figure 5.5 Organization of NRPS clusters recovered from fosmids of the Cuban metagenomic library according to antiSMASH. Red = core biosynthetic genes; rose = additional biosynthetic genes; green = regulatory genes; light blue = transport-related genes; grey = other genes.

5.3.3.1 ST1P6A4 cluster

The ST1P6A4 cluster presented 17 genes for a total length of 30322 bp. In particular, two genes were classified as core biosynthetic genes (*ctg4_26* and *ctg4_27*), two as transport-related genes (*ctg4_19* and *ctg4_25*), one as a regulatory gene (*ctg4_23*), six as additional biosynthetic genes (*ctg4_10*, *ctg4_16*, *ctg4_17*, *ctg4_20*, *ctg4_21* and *ctg4_22*) and the remaining seven genes as unknown (Figure 5.6).

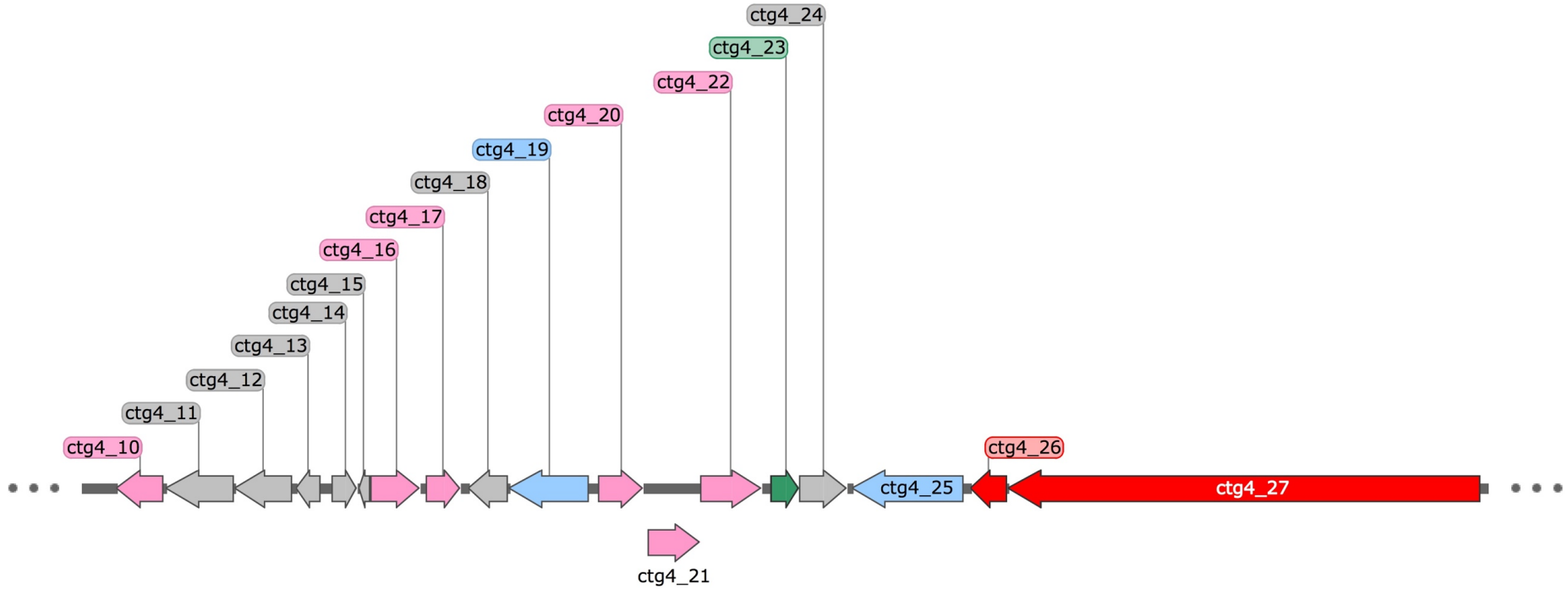


Figure 5.6 ST1P6A4 cluster. Red = core biosynthetic genes; rose = additional biosynthetic genes; green = regulatory genes; light blue = transport-related genes; grey = other genes.

Secondary metabolism clusters of orthologous groups (smCOGs) were identified for nine genes (Table 5.14). In particular, genes *ctg4_21* and *ctg4_25* grouped with genes related to siderophore biosynthesis and transport.

Table 5.14 Secondary metabolism Clusters of Orthologous Groups (smCOGs) of proteins for ST1P6A4 cluster.

Gene	smCOG hits	start	end	e-value	score
ctg4_10	SMCOG1072 dehydrogenase	3	322	5.30E-103	338.3
ctg4_16	SMCOG1143 sugar-binding lipoprotein	7	344	1.33E-10	38
ctg4_17	SMCOG1001 short-chain dehydrogenase/reductase SDR	5	245	1.10E-48	160.1
ctg4_19	SMCOG1288 ABC transporter related protein	11	547	2.10E-52	172.7
ctg4_20	SMCOG1066 alpha/beta hydrolase domain-containing protein	4	307	4.90E-85	279.5
ctg4_21	SMCOG1203 putative siderophore biosynthesis protein	27	366	6.80E-125	411.3
ctg4_22	SMCOG1080 lysine/ornithine N-monooxygenase	8	424	3.70E-168	553.9
ctg4_23	SMCOG1126 extracytoplasmic-function sigma-70 factor	4	171	1.20E-45	149.6
ctg4_25	SMCOG1051 TonB-dependent siderophore receptor	38	799	1.60E-214	708.5

Domain organization of the core biosynthetic genes showed the presence of a thioesterasae domain in *ctg4_26*, three adenylation, two condensation, one epimerization domains and two peptide carrier proteins in *ctg4_27* (Figure 5.7).

Figure 5.7 Domain organization of ST1P6A4 core biosynthetic genes (*ctg4_26* and *ctg4_27*). Purple = A (adenylation); light blue = PCP (peptide carrier protein); dark blue = C (condensation) and E (epimerization); rose = Te (thioesterase).

The analysis of the adenylation domains identified in the core biosynthesis genes predicted the recognition of three amino acids (serine-nrp-arginine). The second adenylation domain did not provide a specific amino acid as a consensus was not established amongst the predicted models used by the antiSMASH software (Figure 5.8).

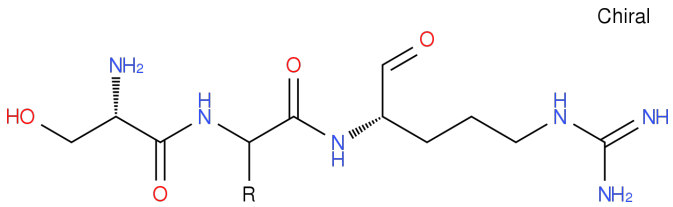


Figure 5.8 Prediction of core scaffold of ST1P6A4 based on assumed NRP collinearity without taking into account tailoring reactions. Consensus: Ser-Nrp-Arg, where nrp is a generic amino acid.

The analysis of gene cluster similarity with predicted BGC-like regions in genomes (Figure 5.9) showed that the majority of the genes were similar to genes present in different species of the *Delftia* genus, therefore suggesting that the DNA captured in clone ST1P6A4 might have been originally from this genus.

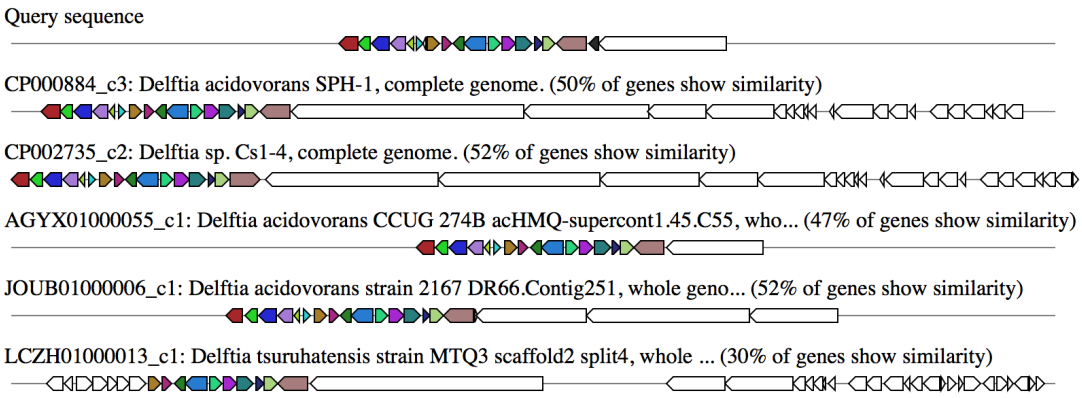


Figure 5.9 Homologous gene cluster analysis of ST1P6A4 (query) with BGC-like regions in genomes.

BLAST results against known clusters predicted that 28 % of the genes (in particular the core biosynthetic genes) showed similarity to delftibactin biosynthetic cluster (BGC0000984) on the MIBiG database (Medema *et al.*, 2015) (Figure 5.10). Other clusters such as taiwachelin, pyoverdine, serobactins and cupriachelin showed a similar or even higher percentage of gene similarity (28 % or 42 %). However, in these cases the genes that showed similarity were not related to the core biosynthetic genes but to regulatory genes, transport-related genes or tailoring reaction genes (Figure 5.10).

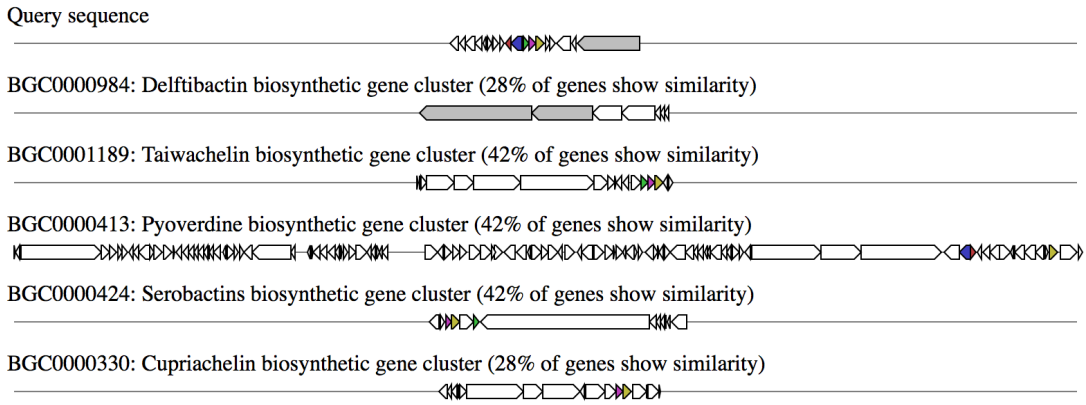


Figure 5.10 KnownClusterBlast results of ST1P6A4 cluster (query) against already known and characterized clusters deposited in the MIBiG database.

The most similar cluster to ST1P6A4 is related to the biosynthesis of the siderophore delftibactin (Figure 5.11) produced by *Delftia acidovorans* (Johnston et al., 2013).

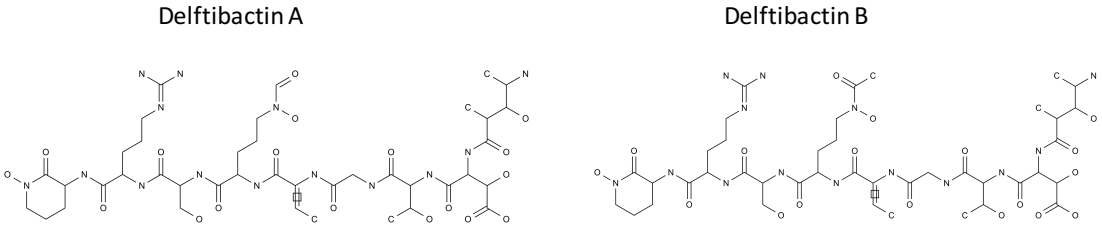


Figure 5.11 Delftibactin A and B structures. Compound structures of the products of Delftibactin biosynthetic gene cluster. Structure reported for the MIBiG entry BGC0000984 (Medema *et al.*, 2015).

5.3.3.2 ST1P6B6 cluster

The ST1P6B6 cluster was 32872 bp with a total of 27 genes. In particular, one core biosynthetic gene (*ctg1_12*), two transport-related genes (*ctg1_13* and *ctg1_19*), one regulatory gene (*ctg1_15*), six additional biosynthetic genes (*ctg1_16*, *ctg1_17*, *ctg1_18*, *ctg1_21*, *ctg1_22* and *ctg1_27*) and 17 unknown function genes were identified (Figure 5.12).

For ST1P6B6, 10 genes were grouped according to smCOGs (Table 5.15). As in the case of ST1P6A4, two genes (*ctg1_17* and *ctg1_13*) grouped with genes related to siderophore biosynthesis and transport.

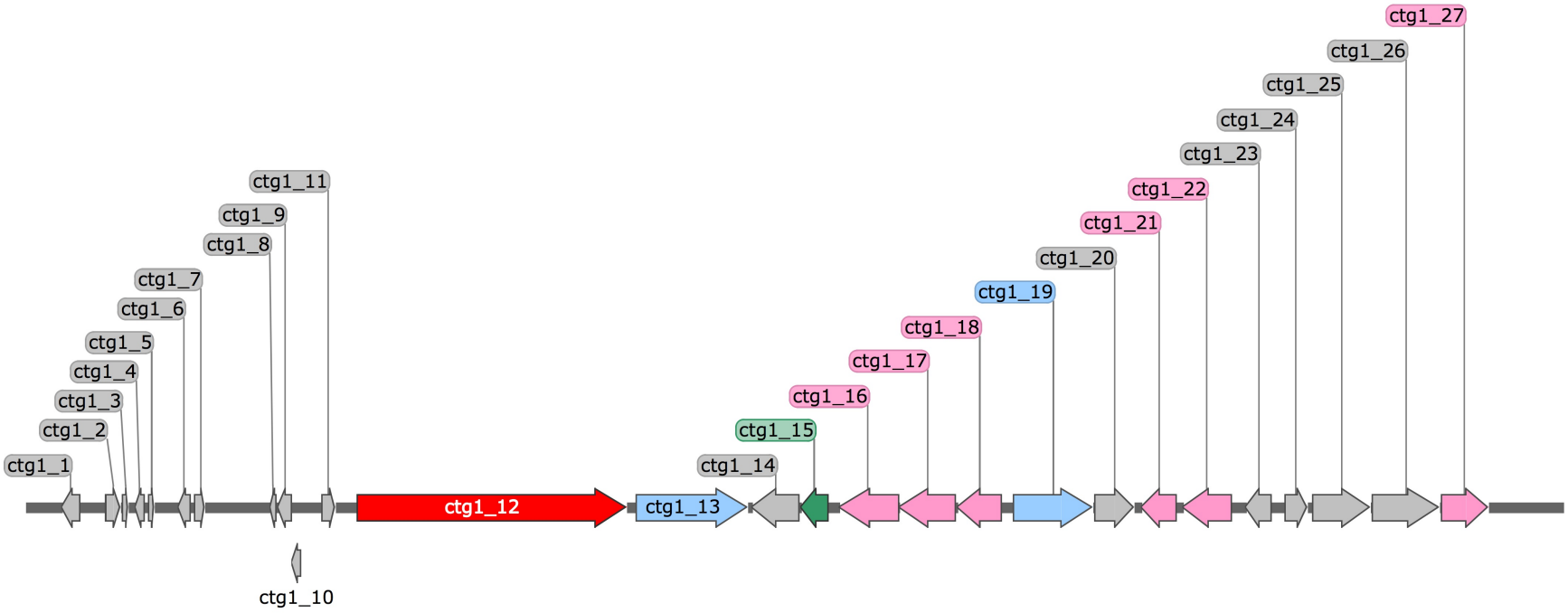


Figure 5.12 ST1P6B6 cluster. Red = core biosynthetic genes; rose = additional biosynthetic genes; green = regulatory genes; light blue = transport-related genes; grey = other genes.

Table 5.15 Secondary metabolism Clusters of Orthologous Groups (smCOGs) of proteins for ST1P6B6 cluster.

Gene	smCOG hits	start	end	e-value	score
ctg1_9	SMCOG1077 phage integrase family protein	3	72	1.00E-18	62.4
ctg1_13	SMCOG1051 TonB-dependent siderophore receptor	38	799	1.60E-214	708.5
ctg1_15	SMCOG1126 extracytoplasmic-function sigma-70 factor	4	171	1.20E-45	149.6
ctg1_16	SMCOG1080 lysine/ornithine N-monooxygenase	3	419	3.50E-168	553.9
ctg1_17	SMCOG1203 putative siderophore biosynthesis protein	38	377	7.50E-125	411.1
ctg1_18	SMCOG1066 alpha/beta hydrolase domain-containing protein	4	307	4.90E-85	279.5
ctg1_19	SMCOG1288 ABC transporter related protein	11	547	2.10E-52	172.7
ctg1_21	SMCOG1001 short-chain dehydrogenase/reductase SDR	5	245	1.10E-48	160.1
ctg1_22	SMCOG1143 sugar-binding lipoprotein	7	344	1.33E-10	38
ctg1_27	SMCOG1072 dehydrogenase	3	322	5.30E-103	338.3

Domain organization of the core biosynthetic gene *ctg1_12* showed the presence of one adenylation, one condensation, one epimerization and one thioesterase domain and two peptide carrier proteins (Figure 5.13).

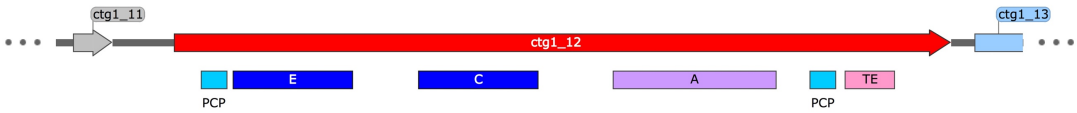


Figure 5.13 Domain organization of ST1P6B6 core biosynthetic gene (*ctg1_12*). Purple = A (adenylation); light blue = PCP (peptide carrier protein); dark blue = C (condensation) and E (epimerization); rose = Te (thioesterase).

The analysis of the adenylation domain specificity predicted the recognition of the consensus monomer arginine. Gene similarity analysis with known characterised clusters showed that 27 % of the ST1P6B6 cluster genes (including the core biosynthetic gene) were similar to the lipopetide siderophore taiwachelin biosynthetic gene cluster usually produced by *Cupriavidus taiwanensis* (Kreutzer & Nett, 2012). Lower gene similarity percentage were identified with serobactins (16 %), cupriachelin (11 %) and pyoverdine (11 %) biosynthetic gene clusters (Figure 5.14).

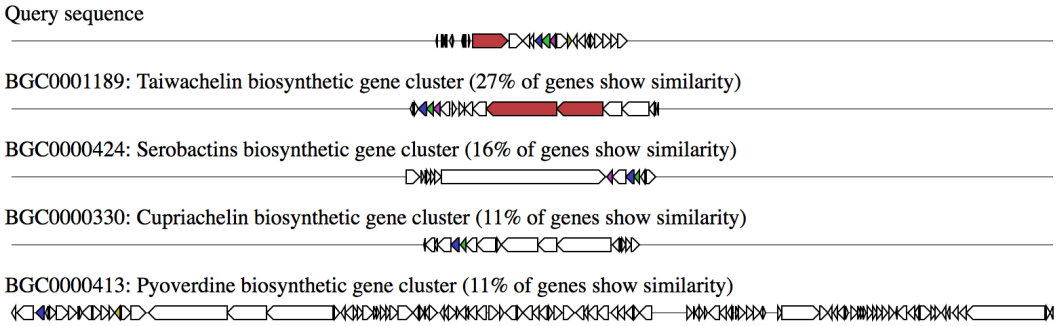


Figure 5.14 KnownClusterBlast results of ST1P6B6 cluster (query) against already known and characterized clusters deposited in the MIBiG database.

The analysis of cluster similarity with BGC-like regions in genomes showed 100 % similarity with *Delftia* sp., suggesting a possible relationship of the captured DNA with this genus (Figure 5.15).

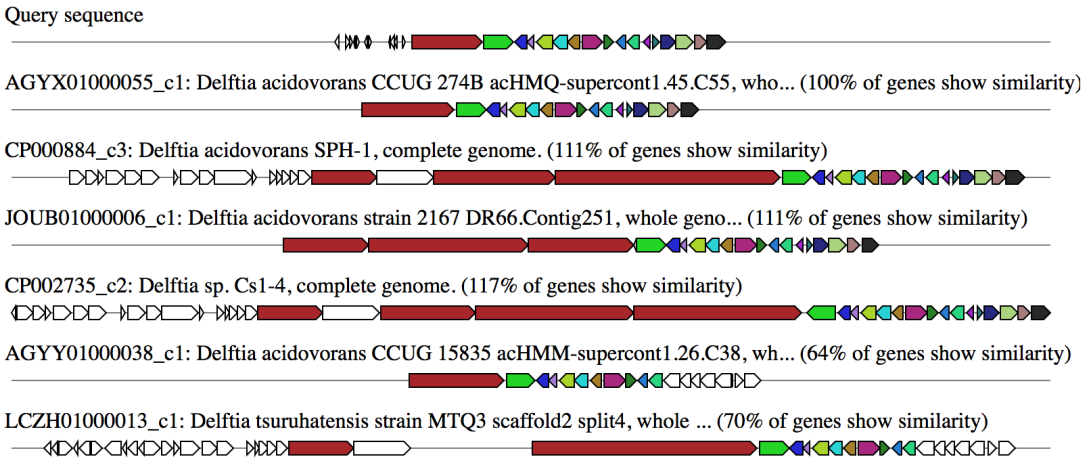


Figure 5.15 Homologous gene cluster analysis of ST1P6B6 (query) with BGC-like regions in genomes.

5.3.3.3 ST1P9D7 cluster

The ST1P9D7 cluster (26243 bp) showed a total of 25 genes: four core biosynthetic genes (*ctg2_14*, *ctg2_16*, *ctg2_17* and *ctg2_18*), two transport-related genes (*ctg2_12* and *ctg2_20*), five additional biosynthetic genes (*ctg2_5*, *ctg2_9*, *ctg2_10*, *ctg2_13* and *ctg2_15*) and fourteen other genes. No regulatory genes were detected in the cluster (Figure 5.16).

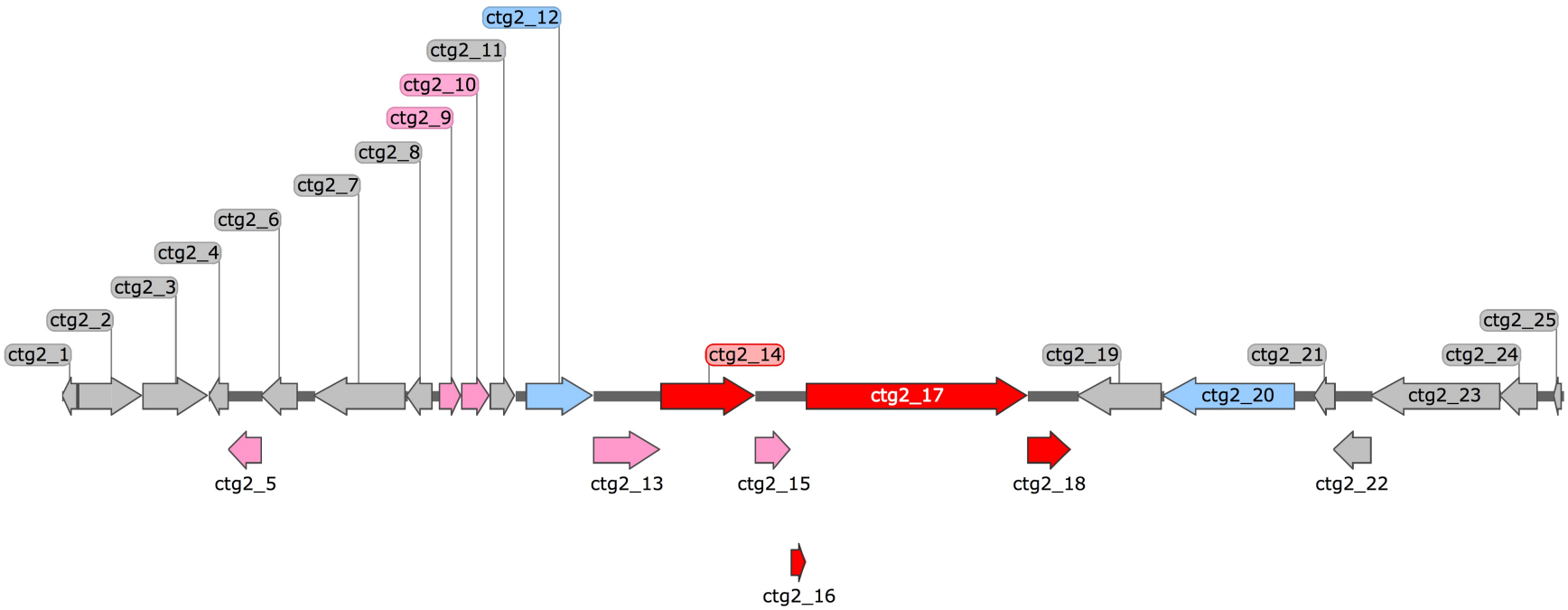


Figure 5.16 ST1P9D7 cluster. Red = core biosynthetic genes; rose = additional biosynthetic genes; light blue = transport-related genes; grey = other genes.

For ST1P9D7 gene cluster, 11 genes were grouped according to smCOGs (Table 5.16). In particular, gene *ctg2_20* grouped with siderophore transport related genes and genes *ctg2_13*, *ctg2_15* and *ctg2_16* grouped with genes related to the synthesis of isochorismate.

Table 5.16 Secondary metabolism Clusters of Orthologous Groups (smCOGs) of proteins for ST1P9D7 cluster.

Gene	smCOG hits	start	end	e-value	score
ctg2_2	SMCOG1094 ferredoxin	44	336	1.50E-70	231.9
ctg2_4	SMCOG1272 TPR repeat-containing protein	28	107	6.30E-07	23
ctg2_5	SMCOG1160 Flavin reductase domain protein FMN-binding	28	180	7.80E-40	130.4
ctg2_7	SMCOG1241 Lipoprotein dsbF	58	175	6.53E-08	30
ctg2_9	SMCOG1083 oxidoreductase	9	124	1.00E-35	117.8
ctg2_10	SMCOG1083 oxidoreductase	1	150	3.10E-13	44
ctg2_12	SMCOG1202 Major facilitator transporter	8	379	7.70E-33	107.8
ctg2_13	SMCOG1018 Isochorismate synthase	19	386	3.60E-101	333.8
ctg2_15	SMCOG1027 isochorismatase	0	208	1.60E-86	283.9
ctg2_16	SMCOG1027 isochorismatase	13	76	4.40E-15	49.8
ctg2_20	SMCOG1082 TonB-dependent siderophore receptor family	4	739	2.80E-42	139.7

Domain organization of the core biosynthetic gene *ctg2_14* showed the presence of one adenylation domain. *Ctg2_17* presented one condensation, one adenylation, one peptide carrier protein and one thioesterase domain. In *ctg2_18* a ketoreductase domain usually related to PKS was identified (Figure 5.17).

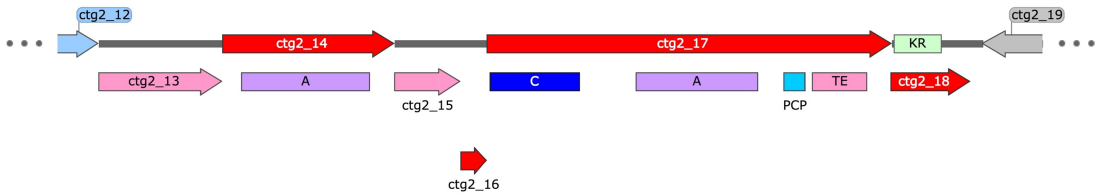


Figure 5.17 Domain organization of ST1P9D7 core biosynthetic genes (*ctg2_14*, *ctg2_17* and *ctg2_18*). Purple = A (adenylation); light blue = PCP (peptide carrier protein); dark blue = C (condensation); rose = Te (thioesterase); Light green = KR (ketoreductase).

The analysis of the adenylation domain specificity predicted the recognition of the 2,3-dihydroxybenzoate or salicylic acid (dhb|sal) from *ctg2_14* and serine (ser) from *ctg2_17*. Gene similarity analysis with known characterised clusters showed that the most similar known clusters are related to the siderophore griseobactin produced by *Streptomyces* sp. ATCC 700974 (Patzer & Braun, 2010) and fuscachelin biosynthetic clusters of the Actinobacteria *Thermobifica fusca YX* (23 % of genes showed similarity) (Figure 5.17). Other biosynthetic clusters such as bacillibactin, myxochelin, paenibactin, mirubactin, heterobactin and enterobactin also presented a lower percentage of gene (11-17 %) similar to the ST1P9D7 cluster genes (Figure 5.18).

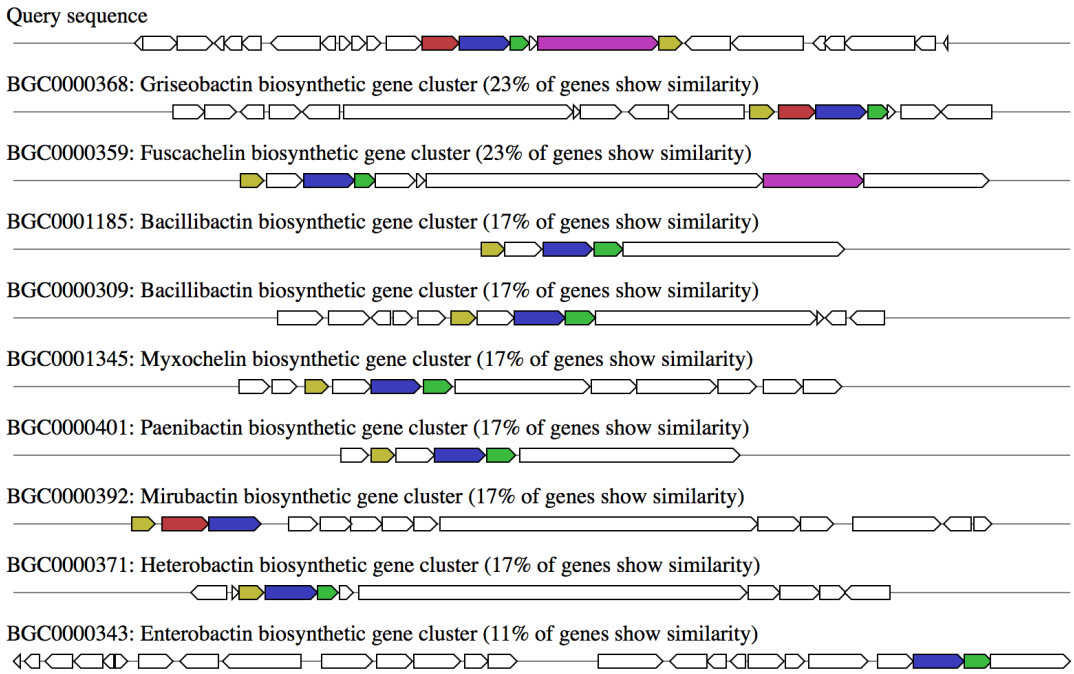


Figure 5.18 KnownClusterBlast results of ST1P9D7 cluster (query) against already known and characterized clusters deposited in the MIBiG database.

Analysis of similarity with BGC-like regions present in genomes showed that 54 % of genes were similar to genes belonging to different species of *Stenotrophomonas maltophilia* (Figure 5.19).

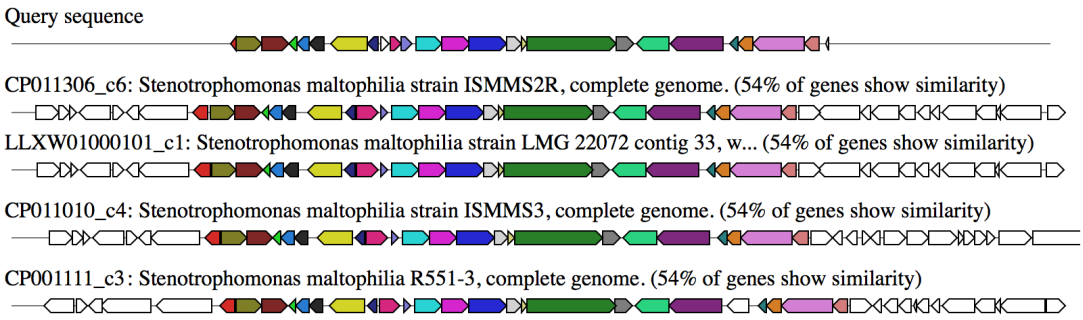


Figure 5.19 Homologous gene cluster analysis of ST1P9D7 (query) with BGC-like regions in genomes.

5.3.3.4 ST1P9E10 cluster

A total of 24 genes were identified for ST1P9E10 cluster (27991 bp. In particular, one core biosynthetic gene (*ctg1_4*), two transport-related genes (*ctg21_11* and *cyg1_12*), four regulatory genes (*ctg1_16*, *ctg1_19*, *ctg1_20* and *ctg1_22*), one additional biosynthetic gene (*ctg1_8*) and 16 unknown genes were identified (Figure 5.20). For the ST1P9E10 gene cluster, 11 genes were grouped according to smCOGs (Table 5.17).

Table 5.17 Secondary metabolism Clusters of Orthologous Groups (smCOGs) of proteins for ST1P9E10 cluster.

Gene	smCOG hits	start	end	e-value	score
ctg1_1	SMCOG1157 Hypothetical protein	107	350	1.50E-39	130.3
ctg1_8	SMCOG1007 Cytochrome P450	24	245	1.10E-59	196.6
ctg1_11	SMCOG1000 ABC transporter ATP-binding protein	9	227	7.00E-59	193.5
ctg1_12	SMCOG1029 RND family efflux transporter MFP subunit	17	343	6.10E-31	101.8
ctg1_14	SMCOG1061 methyl-accepting chemotaxis protein	92	737	5.10E-101	335.4
ctg1_16	SMCOG1201 Response regulator	3	144	4.50E-35	115.4
ctg1_18	SMCOG1061 methyl-accepting chemotaxis protein	75	694	1.17E-101	337.5
ctg1_19	SMCOG1003 Sensor histidine kinase	408	521	5.83E-07	25.6
ctg1_20	SMCOG1008 Response regulator	4	120	7.50E-33	108.2
ctg1_22	SMCOG1201 Response regulator	286	422	6.20E-24	78.9
ctg1_23	SMCOG1157 Hypothetical protein	0	112	2.60E-16	54.2

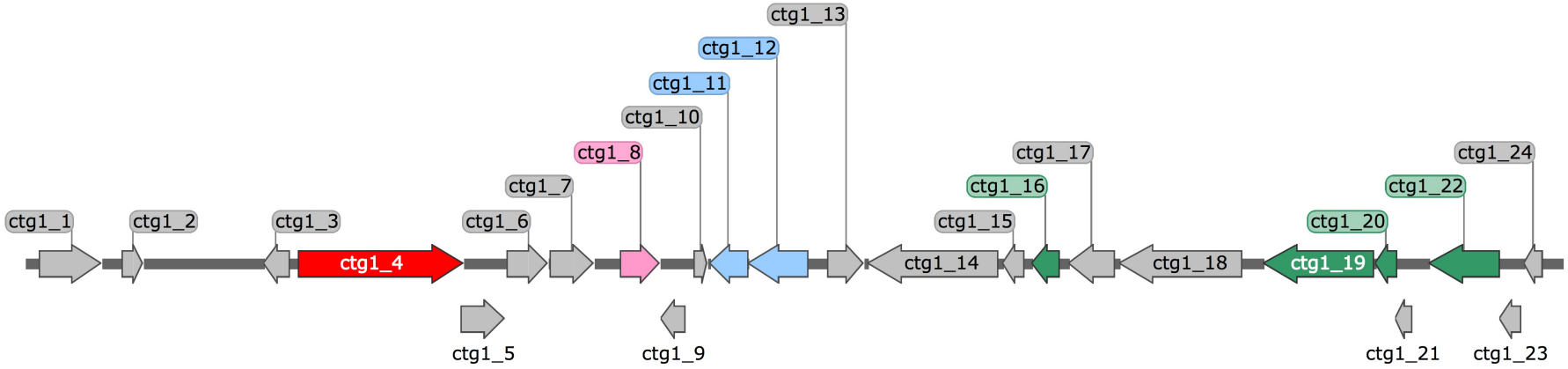


Figure 5.20 ST1P9E10 cluster. Red = core biosynthetic genes; rose = additional biosynthetic genes; light blue = transport-related genes; green = regulatory genes; grey = other genes.

Domain organization of the core biosynthetic genes *ctg1_4* showed the presence of one adenylation domain, one peptide carrier protein and one thioesterase domain (Figure 5.21). The analysis of the adenylation domain specificity predicted the recognition of leucine.

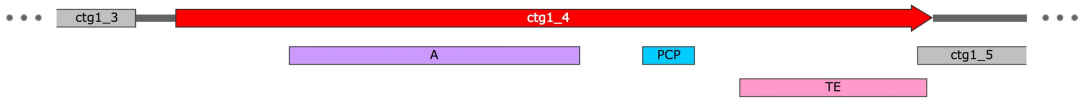


Figure 5.21 Domain organization of ST1P9E10 core biosynthetic gene (*ctg1_4*). Purple = A (adenylation); light blue = PCP (peptide carrier protein); rose = Te (thioesterase).

Gene similarity analysis with known characterised clusters did not show similarity of ST1P9E10 cluster to any of the biosynthetic cluster present in the MIBiG database. However, analysis of the homologous gene cluster with BGC-like regions in genomes presented a low similarity (18-13 %) to *Pseudomonas* species and even lower (7%) to *Burkholderia* species (Figure 5.22).

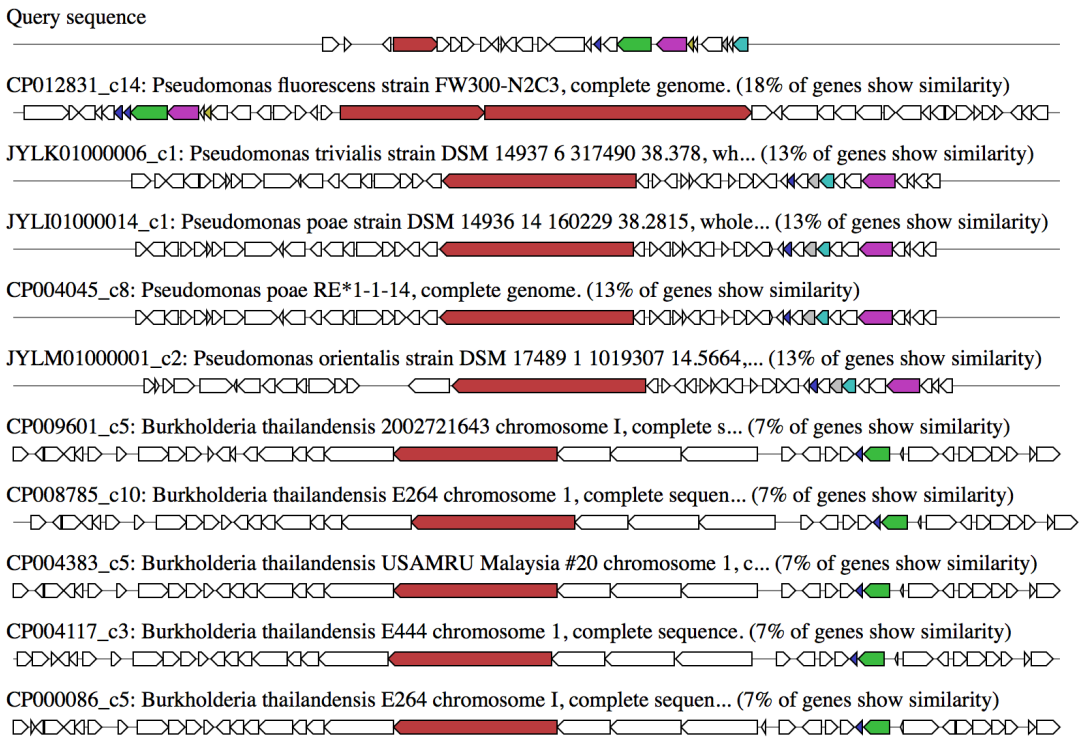


Figure 5.22 Homologous gene cluster analysis of ST1P9E10 (query) with BGC-like regions in genomes.

5.3.3.6 Summary of clusters analyses

All the clusters recovered showed similarity to siderophore biosynthetic clusters (Table 5.18) and to BGC-like regions in genomes of sequenced Gram-negative bacteria (Table 5.19). Clone ST1P9E10 was the only clone that did not show similarity to any known gene cluster.

Table 5.18 Clone cluster similarity to known clusters according to antiSMASH analysis. The most similar cluster was chosen according to the % of gene which showed similarity, in particular for similarity in the biosynthetic genes.

Clone	Known Cluster similarity	Organism	% of genes which showed similarity	MIBiG reference
ST1P6A4	Delftibactin	*Delftia acidovorans*	28	BGC0000984
ST1P6B6	Taiwachelin	*Cupriavidus taiwanensis*	27	BGC0001189
ST1P9D7	Griseobactin	*Streptomyces* sp. ATCC 700974	23	BGC0000368
ST1P9E10	nd	nd	nd	nd

Table 5.19 Cloned cluster similarity to BGC-like region of sequenced genomes.

Clone	BGC-like region in genome similarity	% of genes which showed similarity	Genome accession No.
ST1P6A4	*Delftia acidovorans*	50	CP000884_c3
ST1P6B6	*Delftia acidovorans*	100	AGYX01000055_c1
ST1P9D7	*Stenotrophomonas maltophilia*	54	CP011306_c6
ST1P9E10	*Pseudomonas fluorescens*	18	CP012831_c14

5.3.3.5 High-throughput modification of fosmids for transfer in heterologous hosts

In order to express the identified biosynthetic clusters in different heterologous hosts, the modification of the vector backbone of all fosmids containing sequenced biosynthetic clusters was necessary to integrate them in multiple hosts chromosomes.

The introduction of a *ΦC31 attP*: *ΦC31 integrase* cassette in the backbone of the vectors using a PCR-targeting approach (as described in Section 5.2.6.1) would have allowed the integration of the fosmids inside the chromosome of different heterologous hosts such as *S. coelicolor* for the expression of the captured clusters. However the introduction of the *aac(3)IV*: *ΦC31 attP*: *ΦC31 integrase* cassette (Figure 5.23) to replace the chloramphenicol resistance gene present in the fosmid (Figure 5.24) was not successful. The high-throughput option to modify all cloned vectors using the same strategy did not succeed. The unsuccessful modification of the vectors presented difficulties in transferring the clones to different heterologous hosts such as *S. coelicolor* or other superhosts usually used for secondary metabolites expression for further characterization of the clusters. These results showed the importance of engineering a new vector which would be able to transfer, replicate and/or integrate in different hosts, prior to the insertion of eDNA to create metagenomic libraries.

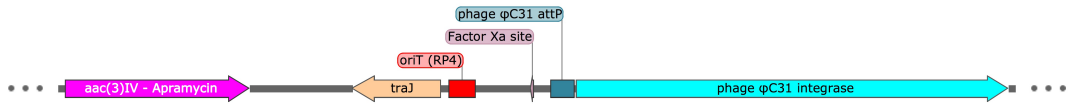


Figure 5.23 Cassette containing the apramycin marker gene, the integration site and the integrase gene used to transform the fosmid vectors and replace the chloramphenicol resistance gene.

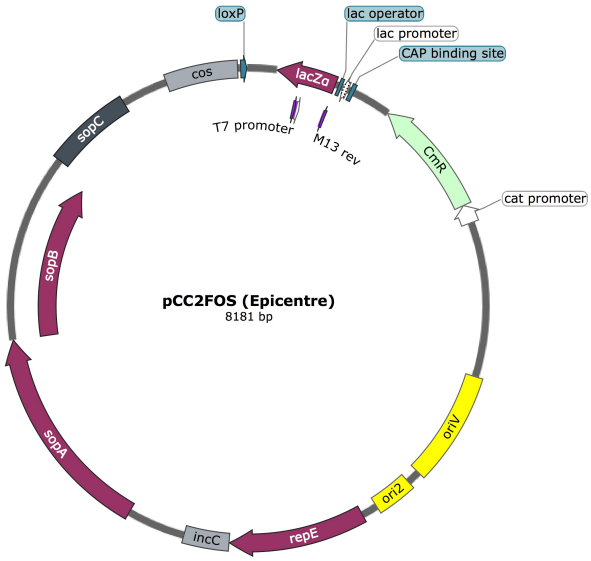


Figure 5.24 Map of the fosmid pCC2FOS™ used for the creation of the Cuban metagenomic library.

5.3.4 BAC vectors and heterologous host engineering

5.3.4.1 Engineering of BAC vectors

The BAC vector pCC1BAC™ (Epicentre®) (Figure 5.25 A) was modified in order to replace the chloramphenicol marker gene with a cassette containing the apramycin resistance gene *aac(3)IV*, the origin of transfer *oriT*, the *ΦC31 attP* site and the *ΦC31* integrase gene (Figure 5.22) cloned from the vector pSET152 (Bierman *et al.*, 1992). This modification would allow the integration of the vector in *Streptomyces* spp. heterologous hosts for the expression of captured eDNA inside the vector. The engineered BAC vector was named pBCaBAC (Figure 5.25 B).

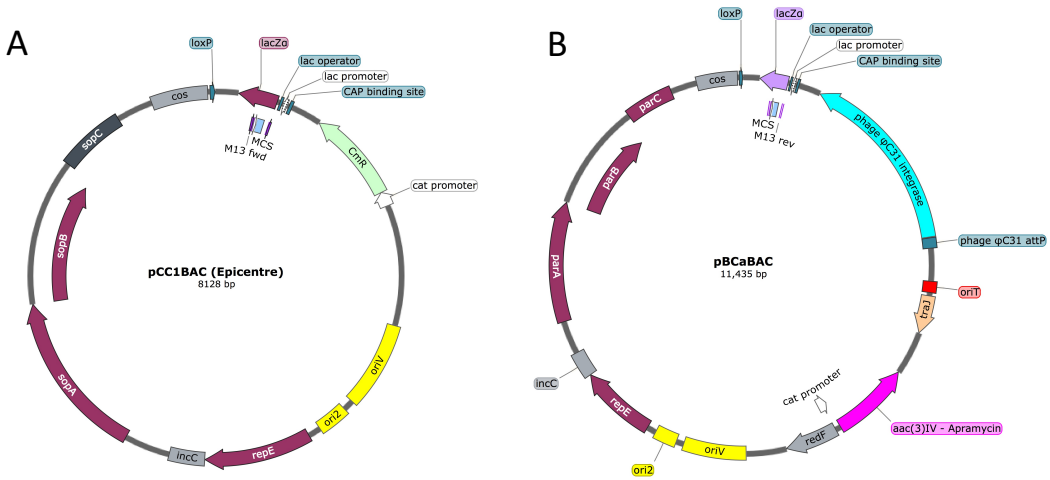


Figure 5.25 Maps of the original vector pCC1BAC™ (A) and the modified vector pBCaBAC (B).

The correct replacement of the chloramphenicol gene with the apramycin cassette was successfully checked by PCR using the primers described in Table 5.5 and reported in Figure 5.26 and Sanger sequencing.

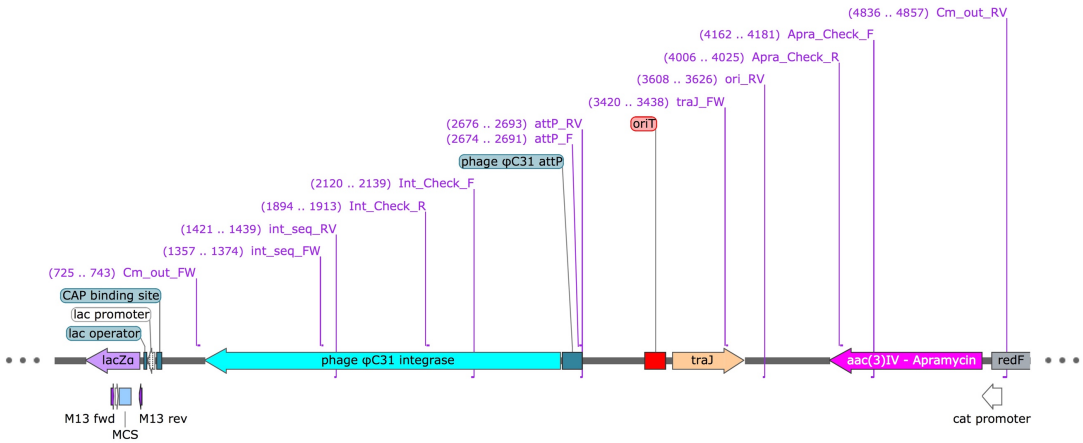


Figure 5.26 Maps of primers binding sites on the cassette used to replace the chloramphenicol gene in pBCaBAC vector.

Following the same PCR-targeting approach, the apramycin resistance gene was also successfully replaced with a kanamycin resistance gene (*nptII*, neomycin phosphotransferase) cloned by pK18MobsacB using primers described in Table 5.5. The full cassette (Figure 5.27) was successfully checked through sequencing. The new vector was named pBCkBAC (Figure 5.28).

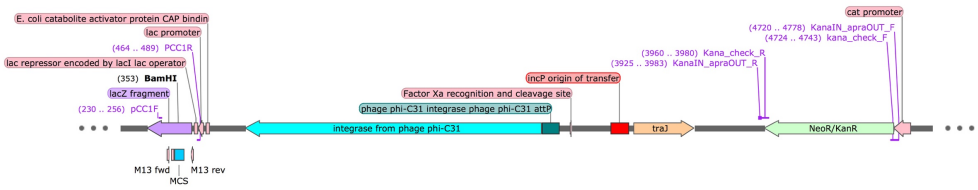


Figure 5.27 Maps of primers binding sites on the kanamycin cassette used to replace the apramycin resistance gene in pBCkBAC.

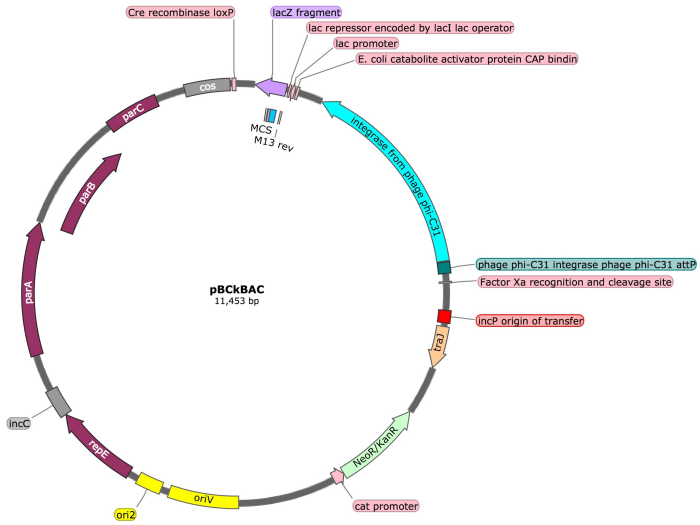


Figure 5.28 Maps of the modified vector pBCkBAC.

5.3.4.2 Integration of the engineered BAC vectors in *S. coelicolor* M1152

The empty vectors were successfully transferred to *E. coli* ET12567 strains either by electroporation or triparental mating as described in Section 5.2.6.3 using the vector pSET152 as control. The strains *E. coli* ET12567/pUZ8002/pSET152, *E. coli* ET12567/pUZ8002/pBCaBAC, *E. coli* ET12567/pR9406/pBCaBAC and *E. coli* ET12567/pR9406/pBCkBAC were then conjugated with *S. coelicolor* M1152. Ex-conjugants were obtained with selection at 50 µg/mL apramycin for pSET152 and pBCaBAC and 200 µg/mL kanamycin for pBCkBAC. Colony PCR showed the presence of the vectors inside the *S. coelicolor* ex-conjugants (Figures 5.29 and 5.30).

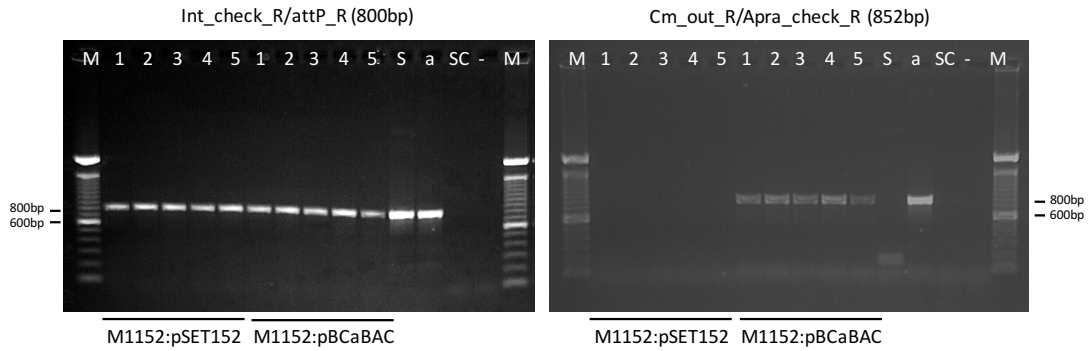


Figure 5.29 Colony PCR of ex-conjugants of *S. coelicolor* M1152:pSET152 and *S. coelicolor* M1152:pBCaBAC using primers Int_check_R/attp_R and Cm_out_R/Apra_check_R. M= 100 bp ladder (Invitrogen); 1-5= ex-conjugants; S = pSET152 vector; a = pBCaBAC vector; SC = *S. coelicolor* M1152; - = dH_2O.

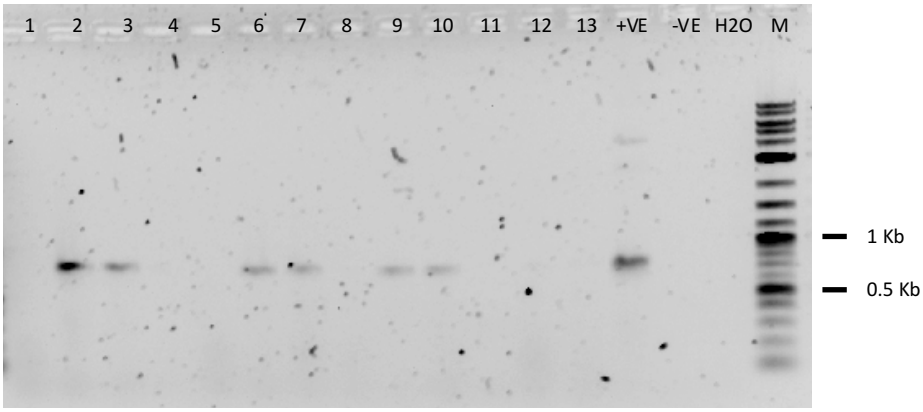


Figure 5.30 Colony PCR of ex-conjugants of *S. coelicolor* M1152:pBCkBAC using primers kana_checkF/R. M= 2-log marker (NEB); 1-13= ex-conjugants; +ve = pK18MobSacB vector; -ve = *S. coelicolor* M1152.

5.3.4.3 Control of the DNA insert size captured by the engineered BAC vectors

In order to check the size of the DNA that could be integrated in the modified BAC vector, genomic DNA was extracted from cell pellets of *Streptomyces sp.*38640 and *Actinoallomurus* sp. 145698 as described in Section 5.2.6.2 recovering 755 ng/µL for *Streptomyces sp.* 38640 and 220 ng/µL for *Actinoallomurus* sp. 145698. The plug system presented a less sheared DNA than the direct genomic DNA extraction from the cell pellet, but DNA yield could not be compared as the DNA was encapsulated in the plug for further digestion experiments. The partial digestion with BamHI on DNA in plugs did not show differences between incubation times, while longer digestions on genomic DNA recovered directly from the mycelium had an impact in the size of the fragments available (Figure 5.31). Transformation results of digested DNA ligated to linearized dephosphorylated pBCkBAC vector showed that ligation reactions performed with DNA digested for 5 min gave a better efficiency in the transformation step with the formation of more CFU and a higher percentage of clones with an insert (white colonies) (Tables 5.20 and 5.21). DNA recovered from digested plugs instead did not produce any CFU (Tables 5.20).

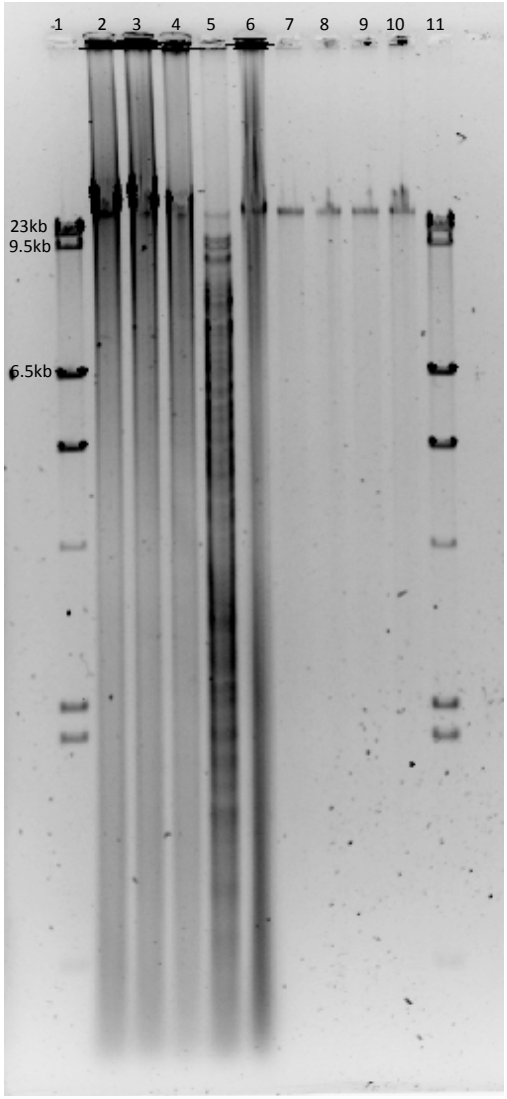


Figure 5.31 Gel electrophoresis on 0.7 % (w/v) agarose gel for 24 h at 4 °C at 50 V of digested gDNA and plugs of *Streptomyces sp.* 38640 with BamHI. 1,11= λ-HindIII marker; 2 = gDNA BamHI 1U 5 min; 3 = gDNA BamHI 1U 10 min; 4 = gDNA BamHI 1U 15 min; 5 = gDNA BamHI 60U 60 min; 6 = gDNA uncut; 7 = Plug gDNA BamHI 1U 20 min; 8 = Plug gDNA BamHI 1U 30 min; 9 = Plug gDNA BamHI 60U 60 min; 10 = Plug gDNA uncut.

Table 5.20 CFU counts of *E. coli* EPI300™-T1R transformed with pBCkBAC:*Streptomyces sp.*38640 DNA selected on LB+kanamycin/IPTG/X-Gal. G5 = gDNA BamHI 1U 5min; G10 = gDNA BamHI 1U 10min; G15 = gDNA BamHI 1U 15min; P20 = Plug OD=2 BamHI 1U 20min; P30 = Plug OD=2 BamHI 1U 30min.

Ligation	Control of colonies on plate				
	CFU counts (n.)			CFU counts (%)	
	Blue	White	Total	Blue	White
G5	177	237	414	43	57
G10	131	46	177	74	26
G15	164	37	201	82	18
P20	0	0	0	0	0
P30	0	0	0	0	0

Table 5.21 CFU counts of *E. coli* EPI300™-T1R transformed with pBCkBAC:*Actinoallomurus sp.* 145689 DNA selected on LB+kanamycin/IPTG/X-Gal. G5 = gDNA BamHI 1U 5min; G10 = gDNA BamHI 1U 10min.

Ligation	Control of colonies on plate				
	CFU counts (n.)			CFU counts (%)	
	Blue	White	Total	Blue	White
G5	122	26	148	82	18
G10	0	0	0	0	0

Gel electrophoresis of twelve positive clones fully digested with BamHI showed there was a variety in the size of DNA inserts inside the vector (Figure 5.32). Only 12 of the 21 randomly chosen white clones were successfully end sequenced using Sanger sequencing. DNA insert end sequences blasted against *Streptomyces sp.*38640 genome showed that the smallest insert was approximately 1.2 kb and the largest was 20.4 kb. The average size of DNA inserts was between 5-10 kb and three clones presented possible chimera inserts.

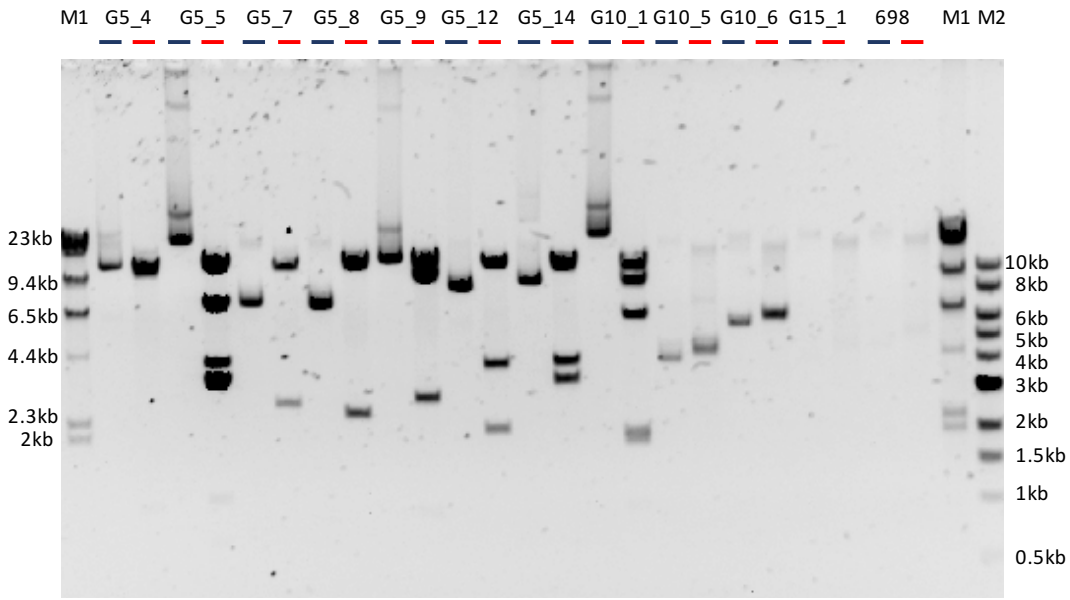


Figure 5.32 Gel electrophoresis on 0.7 % (w/v) agarose gel of undigested (blue) and BamHI completly digested (red) pBCkBAC vectors containing genomic DNA of *Streptomyces sp.* 38640 recovered from the ligation G5, G10 and G15 and genomic DNA of *Actinoallomurus sp.* 145689 (689). M1 = λ-HindIII marker; M2 = 1 kb marker (NEB).

5.3.3.4 Heterologous host (*P. putida*) engineering

For the expression of clusters captured in either the fosmid pCC2FOS™ (Epicentre®) used for the Cuban metagenomic library or the newly engineered pBCaBAC and pBCkBAC vectors, the heterologous host *P. putida* was lacking the *trfA* gene necessary for the activation of the replication system of the vectors. Therefore the engineering of the hosts required the integration of the *trfA* gene in the host chromosome. The construct containing a gentamycin cassette, *trfA* from *E. coli* and the promoter of *sucC* (p^{sucC}) from *P. putida* KT2440 was created amplifying each fragment using primers described in Table 5.6 (Figure 5.33). Each amplicon was successfully cloned into pGEM®-T Easy vectors.

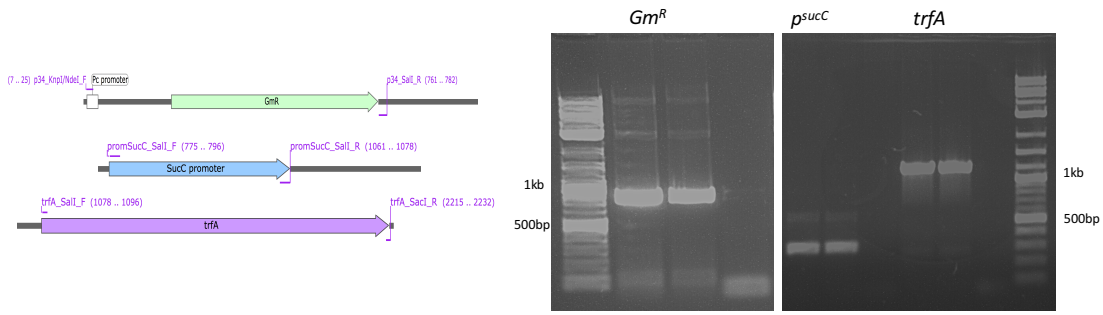


Figure 5.33 Genes amplification of gentamycin cassette (*Gm^R^*) from pS34-Gm vector, p^*sucC*^ from *P. putida* KT2440 genomic DNA and *trfA* gene from pJB658.

Each fragment was then digested from pGEM®-T Easy with the relevant restriction enzymes and cloned sequentially in pBBR1MCS-2 to prepare the GST cassette (Figure 5.34). The final construct was checked by PCR (Figure 5.34) and sequencing. Only a few clones containing the three fragments had p^{sucC} in the correct orientation (Figure 5.35).

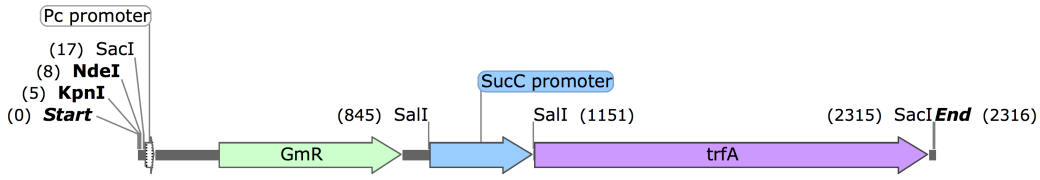


Figure 5.34 GST cassette cloned inside pBBR1MCS-2 and then pJOC100 transposon vector.

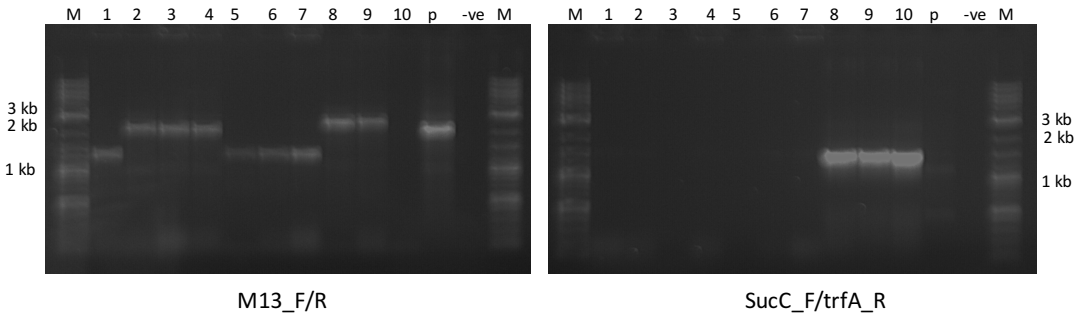


Figure 5.35 Control PCR of pBBR1MCS:GST cloning to detect clones with all three genes (M13_F/R) and correct orientation of the p^{sucC} in the construct (SucC_F/TrfA_R).

The SacI-GST-SacI construct was successfully cloned into pJOC100 vector forming the plasmid pBCLI::GSTA and transferred in *E. coli* S17-1λ*pir* via electroporation. pBCLI::GSTA was mobilised into *P. putida* KT2440 and *P. putida* BIRD-1 via

conjugation. Trans-conjugants were screened for the presence of pBCLI::GSTA by PCR (Figure 5.36).

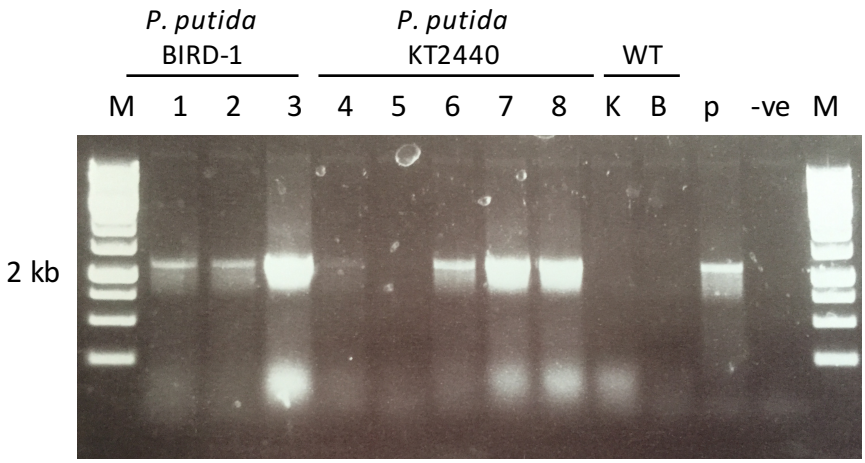


Figure 5.36 *P. putida* KT2440 and BIRD-1 ex-conjugants colony PCR with p34_KpnI/NdeI_F/TrfA_SacI_R primers. 1-3 = *P. putida* BIRD-1 pBCLI::GSTA ex-conjugants; 4-8 = *P. putida* KT2440 pBCLI::GSTA ex-conjugants; K = *P. putida* KT2440 WT; B = *P. putida* BIRD-1 WT; p = pBCLI::GSTA; -ve = dH_2O; M = 1 kb ladder (Invitrogen)

A genomic library constructed in the newly engineered pBCkBAC was conjugated in the engineered *P. putida* BIRD-1 proving the ability of the pBCkBAC vector to replicate in the new host. The use of the *P. putida* hosts for the expression of clusters recovered from the metagenomic library was not tested due to time constraints. Future experiments will use these hosts in parallel to the *Streptomyces* spp. hosts to improve the chances to express clusters recovered from eDNA.

5.4 Discussion

This study characterised two newly designed degenerate primer sets targeting adenylation and ketosynthase domains for NRPS and PKS, respectively. This allowed for the exploitation of functional metagenomic libraries for the discovery of genes involved in the synthesis of novel natural products. At the single-genome level, these new primers could detect the majority of the adenylation domains present in only one of many NRPS cluster of *S. coelicolor* and only one of the two ketosynthase alpha subunit of the PKS type-II clusters present in the strain. Although these results suggest

a low diversity hit rate for both primer sets, the comparison with already available primers targeting the same domains (Ayuso-Sacido & Genilloud, 2005, Owen *et al.*, 2013) showed that PKS primers had a comparable hit rate on multiple genomes to the already available primers on both type-I and type-II PKS despite the fact that they were mainly designed on type-II KSα domain. NRPS primers instead had a lower hit rate than the others, but still predicted the majority of the expected hits. Moreover, the use of the newly designed primers in the screening of the metagenomic library proved the importance of using different probes in parallel to enhance the detection of biosynthetic clusters from a given metagenomic library. Therefore, the application of multiple primer sets in PCR based screening of metagenomic libraries is necessary for efficient retrieval during high-throughput screening programmes for novel natural drug discovery. Another important aspect to consider during functional metagenomic screening is the ability to dereplicate the clusters recovered at an early stage. Owen *et al.* (2013) reported that the likelihood of an amplicon to have a truly functional relationship to the closest characterised cluster match is proportionally related to the identity observed between the amplicon and the identified domain of the cluster. In the current study, two out of four initial amplicons recovered from the library showed high identity (99 %, expected value e0.0) to *Delftia acidovorans* adenylation domains, which was later confirmed by analysis (antiSMASH) of the fully sequenced clusters from the same clones (ST1P6A4 and ST1P6B6). The current study identified four potentially novel clusters which might belong to Betaproteobacteria and Gammaproteobacteria. All clusters showed a low gene similarity to already known clusters suggesting the potential recovery of variants of known clusters or novel clusters. All recovered clusters matched various metallophore clusters belonging to different species, such as delftibactin produced by *Delftia acidovorans* that contributes to gold biomineralization in the environment (Johnston *et al.*, 2013) or siderophores e.g. taiwaichelin in *Cupriavidus taiwanensis* (Kreutzer & Nett, 2012), griseobactins in *Streptomyces sp.* ATCC 700974 (Patzer & Braun, 2010), serobactins in *Herbaspirillum seropedicae* Z67 (Rosconi *et al.*, 2013) and pyoverdine in *Pseudomonas aeruginosa* (Visca et al., 2007, Hannauer et al., 2012), which are involved in iron acquisition from the environment. It has been reported that siderophore NRPS clusters share the same NRPS biosynthetic system as antimicrobial NRPS compounds presenting a direct evolutionary link to them (Wang *et al.*, 2014). Therefore, the identification of these clusters showed the recognition of the correct

target in metagenomic library screening even though the related compounds do not present the activity of interest. However, the study of siderophore clusters remains of great importance for their ecological role complexing iron or other metals essential for the microorganisms' survival in the environment. Previous studies performed on the same Cuban soil showed that a high proportion of the microbial population was represented by Proteobacteria (Johnson-Rollings, 2012), which also resulted in a higher representation in the metagenomic library of the current study. The small size of the library and consequently the low number of represented genomes reduced the probability of identifying NRPS clusters and in particular of PKS clusters. A genome mining study of 2699 genomes by Wang *et al.* (2014) showed that NRPS clusters were more prevalent than PKS clusters in genomes, therefore in the current study the probability of identifying the less prevalent PKS cluster in nature were constrained by the size of the library itself. Nevertheless, PKS clusters were identified as well as NRPS clusters related to antimicrobial gene clusters in the screening of a larger Antarctic metagenomic library (Laskaris, 2009, Amos *et al.*, 2015). The screening with the same NRPS primers of a library approximately 10X bigger then the Cuban library used in the current study allowed the recovery of NRPS genes possibly related to potential antimicrobial clusters (Laskaris, 2009) (Figure 5.37).

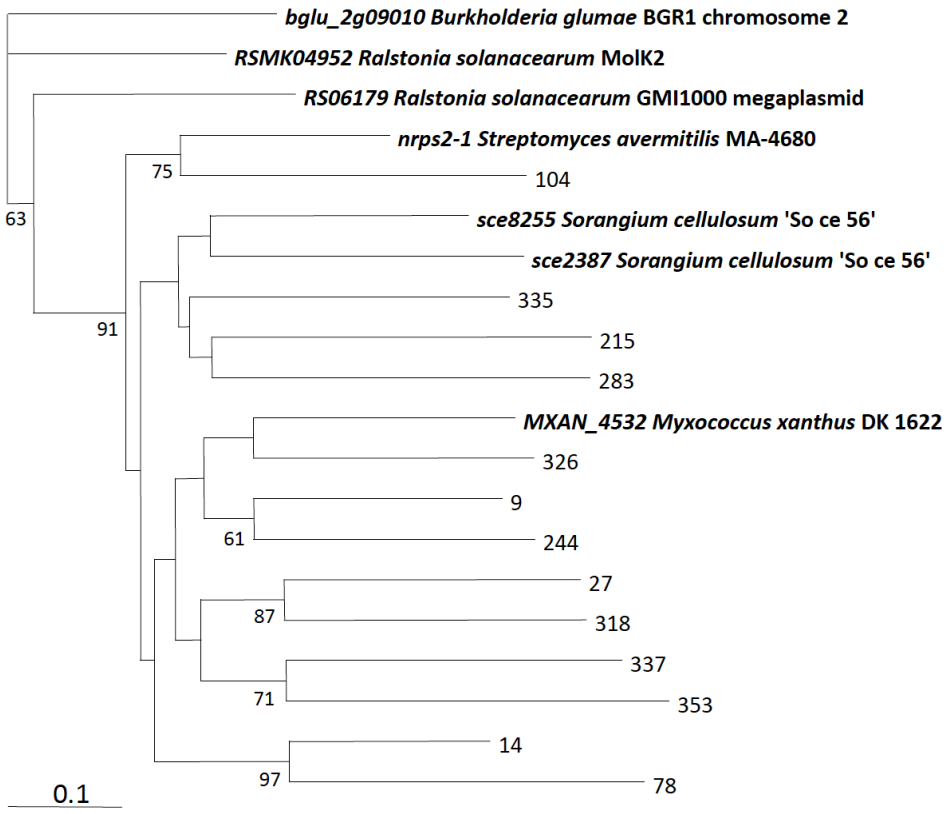


Figure 5.37 Neighbour-joining phylogenetic tree of NRPS sequences amplified from the Antarctic fosmid library with close homologous. Percentage bootstrap values of 10000 replicates are reported next to the branches. The scale indicates 10 % nucleotide dissimilarity (Laskaris, 2009).

The annotation and comparison of the clusters retrieved in the current study suggested that only partial clusters were captured in the fosmids vectors. TAR recombination could have been applied to potentially recover the rest of the clusters as already demonstrated by Feng *et al.* (2010), Kim *et al.* (2010) and Kallifidas & Brady (2012). However, the relatively small size of the metagenomic library of the current study (approximately 120 Mb) was prohibitive to following the aforementioned approach.

The co-extraction of inhibitory compounds with eDNA using a direct chemical lysis of soil required further steps of purification prior cloning of the eDNA inside fosmids vectors, therefore loss and possible further shearing of the DNA might have occurred, reducing the efficiency of packaging and transduction into the *E. coli* host to obtain a large library.

The partial clusters recovered in the current study were still of particular interest to attempt high-throughput transfer methods. The initial attempted modification of the fosmid backbone of clones containing insert DNA was unsuccessful, showing the difficulties of modifying already existing clones from a metagenomic library with a

common simple strategy. A novel BAC vector which could integrate in *Streptomyces* hosts and replicate in *Pseudomonas* engineered hosts without further manipulation of the vector backbone of metagenomic library clones proved necessary to facilitate expression and characterization of future recovered clones. Multiple modified BAC vectors and systems have been engineered to be used for library exploitation in the last two decades (Sosio *et al.*, 2000, Martinez *et al.*, 2004, Hain *et al.*, 2008, Aakvik *et al.*, 2009). For example, the pESAC vector, which can integrate in *Streptomyces* sp. and replicate in *E. coli*, has been commonly used to prepare genomic libraries from single strains (Jones *et al.*, 2013) or metagenomic library from soil eDNA (Berry *et al.*, 2003). Martinez *et al.* (2004) presented the engineering of a derivative of pBeloBAC11 (Shizuya *et al.*, 1992, Kim *et al.*, 1996) introducing in a cre-loxP site a *Streptomyces* sp. conjugative cassette (*oriT – aac(3)IV – attP - ΦC31 int*) and engineering the chromosome of *P. putida* KT2440 with an *attB ΦC31* for the integration of the BAC vector. Aakvik *et al.* (2009), instead, modified the commercially available pCC1FOS™ vector introducing the genes *parD* and *parE* to improve the stability of the vector and engineered *Pseudomonas fluorescens* and *Xanthomonas campestris* strains introducing through a suicide vector the *trfA* gene under an inducible promoter. Indeed, vectors based on pCC1FOS™ or pCC1BAC™ (Epicentre®) contains a modified broad-host range RK2 origin of replication (*oriV*) missing the replication initiation protein gene *trfA*. This function can be restored by the presence of this gene (*trfA*) in the hosts allowing the replication of the vectors in multiple copies. The current study, provided the construction of pCC1BAC derivative vectors containing an *oriT – aac(3)IV/nptII – attP - ΦC31 int* cassette which allows the integration in *Streptomyces* sp. hosts and potentially any other hosts containing an *attB(ΦC31)* site and the engineering of two additional heterologous hosts *P. putida* KT2440:*trfA* and *P. putida* BIRD-1:*trfA* for the expression of recovered clones from metagenomic libraries. This study provided a system which merged the advantages of the vectors and heterologous hosts described in the aforementioned studies. The preparation of good quality HMW DNA and good competent cells still remain the main problems for an efficient preparation of large insert metagenomic library in BAC vectors. The current cloning results of genomic DNA in the engineered BAC vectors resulted in small insert fragments as there were some difficulties in the recovery of partially digested HMW DNA. The transformation efficiency varied according to the

system used to recover HMW DNA and partial digestion incubation, suggesting a preference for non-plugs based HMW DNA recovery methods. However, the modified BAC vectors showed the ability to stably capture DNA, integrate in *S. coelicolor* and replicate in *P. putida* KT2440:*trfA* and *P. putida* BIRD-1:*trfA*. The ability to capture HMW DNA in BAC vectors able to transfer to different heterologous hosts with potentially high-throughput conjugation system as described by Martinez *et al.* (2004) will provide a useful tool for expression and characterization of potentially novel compounds.

In conclusion, the current study provided some solutions to tackle the major issues related to functional metagenomic libraries for novel drug discovery. It presented a full characterization of two primer sets targeting adenylation and ketosynthase domains of NRPS and PKS gene clusters which identified four NRPS siderophore clusters in a metagenome library from Cuban soil and developed a novel BAC-heterologous hosts system which will help in the characterization of the recovered clusters and for future libraries preparation and screening.

Chapter 6

General discussion

The use of metagenomic approaches was investigated for novel natural product discovery. In particular, new tools were developed to tackle some of the most challenging issues related to recovering secondary metabolite gene clusters through functional metagenomics.

An analysis of antibiotic biosynthetic diversity in soil was related to community diversity and established potential novel phylogenetic drivers of metabolic diversity (Chapter 4). Previous work has established a correlation between biosynthetic gene diversity and geographic location in addition to edaphic variables (Charlop-Powers *et al.*, 2014, Charlop-Powers *et al.*, 2015). The novelty of the current study was the focus on the inter-relationship between structure and function of microbial communities, thus identifying the taxonomic groups driving biosynthetic diversity. Surprisingly despite the considerable biogeographical range of the samples the microbial community structure was remarkably similar at the phylum level. This may be due to the poor discriminant ability of 16S rRNA gene in terms of predicting functional diversity. The 16S rRNA gene-based approach may provide information on structural differences in microbial communities. However, functional diversity within the prokaryotes is extensive and significant differences between species of the same genus can often been seen in soil bacteria such as the diversity in metabolism of phosphate by *Pseudomonas* species (Lidbury *et al.*, 2016). Understanding the relationship between community diversity and metabolic function is the cornerstone of microbial ecology but it is severely limited by our inability to culture a large proportion of the bacterial groups in any given soil (see review: Stewart (2012)). Studies focusing on the impact of biogeography have shown that limitations to dispersal such as physical barriers do not exist for microorganisms but environmental factors will act selectively on the metabolic traits within the population (Ramette & Tiedje, 2007). It has been possible to understand biogeographic patterns of specific functional groups such as nitrifiers because they have unique functional properties for examples studies on ammonia-oxidizing archaea (AOA) and bacteria (AOB) where AOA are the main ammonia oxidizer in acidic soils (Yao *et al.*, 2013). The target biosynthetic genes of

interest in the current study are widely distributed across many genera of bacteria within the prokaryotes whereas they appear to be less represented in archaeal lineages (Figure 6.1) (Tracanna *et al.*, 2017).

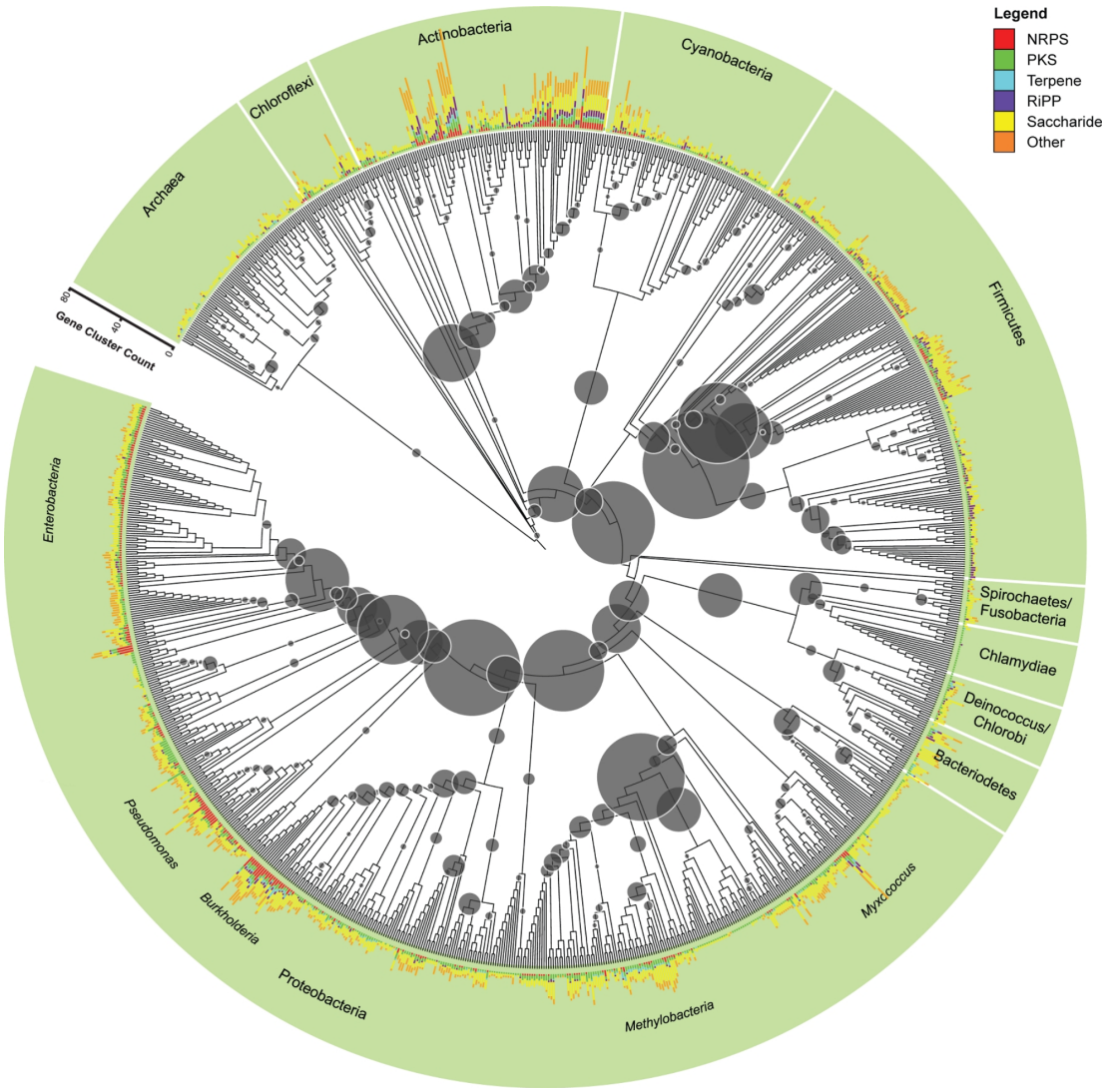


Figure 6.1 Taxonomic diversity of BGCs across bacteria and archaea. Gene cluster counts are indicated by bar plots, while within-taxon variation is reported by circles on the tree (image from Tracanna *et al.* (2017)).

Despite this, significant differences in metabolic diversity were detected within the range of samples studied indicating distinct biogeographic patterns in agreement with previous studies (Charlop-Powers *et al.*, 2015). Different phyla were driving the diversity of biosynthetic genes according to the location. These results indicated a selection by environmental conditions at a micro and macro scale and represented an interesting start for further investigation into the ecological role of these compounds

in the newly identified drivers. The presence of different secondary metabolites might represent an advantage in the adaptation of microorganisms attempting to colonise within pre-existing communities as a consequence of dispersal. The theory "*everything is everywhere*, but *the environment selects*" (Baas Becking, 1934) is therefore partially correct as not all microorganisms would be able to adapt, survive and be transported to different environments and might not be present everywhere. This is of particular relevance for extreme environments such as Antarctica which are rapidly disappearing because of global climate changes. For instance, the Mars Oasis (Antarctica) which was included in the study has been identified as a potential hotspot for drug discovery programmes but is now inaccessible (pers. comm. Dr. Kevin Newsham, British Antarctic Survey and Prof. David Pearce, Northumbria University, July 2016). The diversity of these extreme environments could be captured using metagenomic libraries and stored for exploitation not only for novel natural products but also for other enzymes of biotechnological interest. Metagenomic libraries will therefore constitute an invaluable resource representing a snapshot of the bacterial community and their functional genetic diversity, remaining available for exploitation for future generations.

The Cuban metagenomic library (Chapter 5) and the Antarctic library from previous work represented two examples of capturing unique community genetic potential for future exploitation. Various problems still need to be addressed to fully benefit from the diversity captured by libraries. For instance, the DNA preparation is usually one of the critical steps for an exhaustive representation of the community. DNA extraction optimization protocols were studied and it was clear that a combination of physical and chemical lysis was a compromise between efficiency of extraction and integrity of eDNA. Enzymatic *in situ* direct lysis of the soil community provided an alternative method, but further studies would be needed to confirm this approach on a wider range of soil types. Enrichment or amendments of soil represented a successful strategy to target groups of bacteria for specific enzymes of interest, even though it might potentially reduce the diversity present in the original environment (Cretoiu *et al.*, 2013, Johnson-Rollings *et al.*, 2014). The chitin amendment of Cuban soil (Chapter 3) had a dramatic effect on detection of NRPS genes which might have related to increased prevalence of non-targeted taxonomic groups such as Proteobacteria or Firmicutes well-known as chitin degraders (Cottrell *et al.*, 2000). Additional studies to test the "chitin effect" i.e. the effect of chitin amendment on

actinobacterial prevalence have provided insights into the significant role of Proteobacteria in soil chitin degradation (Jacquiod *et al.*, 2013). Proteobacteria are known to be fast-growing therefore the presence of additional nutritional sources might have promoted this group, out-competing the slower-growing Actinobacteria or Verrucomicrobia (Goodfellow & Williams, 1983, Sangwan *et al.*, 2005). In the Cuban soil a higher prevalence of Proteobacteria and Firmicutes than Actinobacteria was observed in the taxonomic study (Chapter 4) and was also confirmed by the diversity captured in the metagenomic library (Chapter 5). All the NRPS clusters recovered were associated with siderophores from Proteobacteria suggesting an environmental selection for groups containing metallophores to better adapt to low concentration of metals (Johnson-Rollings, 2012). Recovery of spore-forming bacteria was challenging for metagenomic library construction because spores require physical lysis which shears the DNA to the extent that it cannot be used for cloning large gene clusters (Chapter 2; Moré *et al.* (1994)). Attention has focused on exploitation of rare actinobacterial groups such as the genera *Actinoallomurus* (Pozzi *et al.*, 2011) and *Planobispora* (Tocchetti et al., 2015), but this involved isolation of representative strains as the small population size would prohibit recovery by metagenomic approaches which will always represent the most abundant groups. Methods to promote these taxa need to be further investigated in order to increase the chances to capture a wider variety of biosynthetic genes belonging also to the less characterised and lower abundant taxa. These methods which would mimic the natural environmental selection could then be considered as a potential method not only to select for the genes/compounds of interest but also to gain an insight into the ecological role played by these groups under different selection pressures. Understanding the ecological role of natural products is a difficult challenge. Recent studies showed how natural products play a crucial role in mutual symbiosis between specific bacteria and sponges (Fieseler *et al.*, 2007) or ants (Seipke *et al.*, 2013) or suppressive activity against crop pathogens in rhizosphere soil (Weller *et al.*, 2002, Yin *et al.*, 2013). Although studies have been done on microbial interaction and molecular signalling through co-culture experiments, little is still known about the role of secondary metabolites in the environment (see review: Shank & Kolter (2009), van der Meij *et al.* (2017);Traxler *et al.* (2012)). Moreover, genome mining has recently provided a variety of hypothetical natural product biosynthetic gene clusters in multiple microorganisms which have been difficult to express in laboratory conditions

(Laureti *et al.*, 2011, Duncan *et al.*, 2015). Biosynthetic gene clusters might be silent in laboratory conditions, but they could still be expressed in soil, clearly the precise regulation of clusters by environmental cues indicates an important role for these compounds. Interaction with other members of the microbial community or other environmental conditions which are difficult to replicate in the laboratory might be essential for the production of certain natural products (Whitt *et al.*, 2014). This is an important problem for metagenomic library exploitation as well as novel bacterial isolate bioprospecting. Indeed, the expression of novel natural product clusters can be problematic as they are usually highly regulated at different levels, not only by local activators or repressors but also by pleiotropic signals (see reviews: Bibb (2005), Liu *et al.* (2013), Licona-Cassani *et al.* (2015)).

Previous studies have shown the benefits of expressing metagenomic library clones in multiple hosts using systems able to transfer between hosts (Martinez *et al.*, 2004, Craig *et al.*, 2010). However, additional vectors and heterologous expression hosts are required to increase the rate of successful expression of genes from unknown sources. Novel BAC vectors and heterologous hosts were engineered with the aim of enabling the transfer of captured DNA to a variety of hosts using simple efficient natural systems such as conjugation for high-throughput screen in drug discovery programme (Chapter 5). Indeed, the expression in different organisms might enable the activation and expression of gene clusters which would be silent or not recognized in a different host.

There have been recent developments in recovering strains using cultivation-based approaches combined with physical segregation of individual bacterial cells thus reducing competitive exclusion and metabolic inhibition on agar plates (Zengler *et al.*, 2002, Nichols *et al.*, 2010, Stewart, 2012). In addition the use of more natural growth media have been used in an attempt to mimic soil (George *et al.*, 2011). This led to the development of novel isolation techniques such as alginate or gel beads to encapsulate single cells (Stormo & Crawford, 1992, Zengler *et al.*, 2002) or the iChip which segregated single cells in micro-chambers which were then incubated in the original soil (Nichols *et al.*, 2010). This last technique enabled the isolation of a novel unculturable β-proteobacteria species *Eleftheria terrae* which produces a novel antibiotic, teixobactin (Ling *et al.*, 2015). Culture dependent and independent techniques are usually applied separately to investigate and exploit the diversity

available in the environment. However, in order to obtain a more complete understanding of the microbial community structure and function, these techniques should be performed in sequence with meta-omics providing indicators for a focused isolation programme. Metagenomics has revealed Verrucomicrobia as a highly metabolically talented group recovered from diverse habitats including marine habitats (Freitas *et al.*, 2012) to human gut (Dubourg *et al.*, 2000) and rhizosphere (Bergmann *et al.*, 2011, Nunes da Rocha *et al.*, 2013). This group have rarely been isolated so little is known about their physiology (Janssen *et al.*, 2002, Sangwan *et al.*, 2005, Wagner & Horn, 2006). A perusal of the IMG database provided 111 verrucomicrobial genomes and 853 putative biosynthetic gene clusters related to this group (Table 6.1). The Talent Ratio (T_R) reported in Table 6.1 represents the potential number of secondary metabolite gene clusters per genome for each phylum which was calculated according to the following, Talent ratio (T_R) = BGC counts (total) / Genomes available.

Table 6.1 Counts of biosynthetic gene clusters (BGC) reported on IMG/ABC database according to phylum and evidence (experimentally or predicted only)*. BGC were identified from either genome sequences of isolate strains or metagenomes from various environments. Talent ratio (T_R) = BGC counts (total) / Genomes available. *(Data available on 18/05/2016)

Phylum	Genomes available	BGC counts (total)	BGC counts (experimentally proved)	BGC counts (predicted only)	Talent Ratio (T_R)
Acidobacteria	82	851	0	851	10.4
Actinobacteria	6054	235791	621	235146	38.9
Bacteroidetes	1960	22387	2	22385	11.4
Chloroflexi	250	1195	0	1195	4.8
Cyanobacteria	436	7729	60	7669	17.7
Firmicutes	14049	155530	106	155424	11.1
Planctomycetes	150	1295	0	1295	8.6
Proteobacteria	23850	344702	386	344515	14.5
Verrucomicrobia	111	853	0	853	7.7

The talent ratio (T_R) showed that the Actinobacteria phylum has the highest potential for secondary metabolites gene clusters, followed by Cyanobacteria and Proteobacteria. However, the number of genomes available for these groups is considerably higher than for Verrucomicrobia, therefore although Verrucomicrobia showed a low T_R index, this could be biased by the lack of information available. Bacteroidetes for instance showed a 10X less genomes available than Actinobacteria and the T_R is a quarter of the one showed by the Actinobacteria. The acquisition of

further genomes of the less represented phyla will enable a better understanding of the talent ratio of this group. Comparison of the predicted secondary metabolite gene clusters of these newly identified potential drivers of NRPS and PKS diversity in soil to the already known and experimentally characterised one would also provide a good index of the diversity entrapped in them promoting dereplication strategy for future drug discovery screening.

Current drug discovery programs still rely on the knowledge acquired from studying cultured actinobacterial strains. Genome mining approaches showed that many silent clusters still exist in this group and techniques have been developed to activate this gene expression (see reviews: Scherlach & Hertweck (2009), Choi *et al.* (2015)). A variety of regulatory mechanisms have been defined (quorum-sensing, nutritional cues and signalling compounds other than quorum-sensing) which are involved in the regulation of natural product expression in the environment (see reviews: (Papenfort & Bassler, 2016, van der Meij *et al.*, 2017).

Culture independent analysis of the soil microbial community will provide new directions for understanding the ecology of secondary metabolite production.

An example of this is the bioactive potential of the *Pseudomonas* species which have long been known as important plant growth promoters and inhibitors of soil-borne fungal pathogens producing compounds such as phenazine and 2,4-diacetylphloroglucinol (DAPG) (Chin-A-Woeng *et al.*, 2003, Lutz *et al.*, 2004). More recent culture independent studies provided evidence that *Pseudomonas* species in association with brown macroalgae *Saccharina latissima* were identified as (potentially) important antimicrobial members of the community protecting their macroalgal host (Wiese *et al.*, 2009).

6.1 Future work

This project has provided an insight into the metagenomic applications for novel natural product discovery. Assignation of BGCs clusters to the functional gene reads (Chapter 4) will be performed to dereplicate environments' potential and to investigate novel taxa potential towards specific classes of molecules. According to Owen *et al.* (2013), the BGCs identification through amplicons is a reliable approach to dereplicate environments in term of biosynthetic potential for specific class of

molecules and related activity. The association between the classes of molecules and novel taxa could help understanding the ecological role of these compounds in these novel potential natural product drivers enabling more targeted drug discovery approaches and more effective screening methods.

A comparison of the PKS and NRPS diversity related to the newly identified drivers (Verrucomicrobia and Bacteroidetes in particular) will be performed against multiple database to investigate our hypothesis that even though these taxa might be represented at low levels in the microbial community in soil, their biosynthetic diversity is clearly diverged from the main groups including Actinobacteria.

Metagenomic libraries will be prepared on identified hotspots for PKS and NRPS diversity not only related to the main natural product producers Actinobacteria, but also Verrucomicrobia and Bacteroidetes in order to increase the chances to capture novel gene clusters. Isolation of novel bacterial species will also be performed from Antarctic soil, in particular enrichment for Verrucomicrobia and Bacteroidetes. Exploitation of Bacteroidetes and Verrucomicrobia strains for novel natural products will be performed on isolate strains which will show activity against multiple gram-positive and multi-resistant gram-negative and chemical characterization will be attempted for the expressed compounds. Initial studies will also be performed on Bacteroidetes strains available in our culture collection to further characterise their potential for novel compounds and their regulation. Genetic manipulation will also be performed on these representative strains to developed optimal protocol for further studies.

Metaproteomic analyses of the microcosms reported in Chapter 3 will help in the study of microbial community shift in the presence of additional nutrients such as chitin and to explore the active protein fraction present in soil, in particular proteins potentially involved in antibiotic production.

Appendix

Table S2.1 NanoDrop™ results for DNA extraction methods (A-M) on Reading and Warwick soils.

Sample	Nucleic Acid Conc. (ng/µL)	260/280	260/230
Reading A1.1	177.5	1.43	0.30
Reading A1.2	178.6	1.43	0.27
Reading A1.3	217.0	1.41	0.36
Reading A2.1	61.3	1.66	0.14
Reading A2.2	68.9	1.62	0.15
Reading A2.3	76.7	1.57	0.18
Reading B.1	218.1	1.41	0.37
Reading B.2	207.1	1.41	0.36
Reading B.3	215.7	1.41	0.37
Reading C.1	55.0	1.69	0.13
Reading C.2	51.2	1.76	0.12
Reading C.3	56.6	1.68	0.13
Reading D.1	740.5	1.45	0.61
Reading D.2	830.7	1.43	0.63
Reading D.3	639.8	1.43	0.60
Reading E.1	46.3	1.70	0.12
Reading E.2	57.6	1.58	0.14
Reading E.3	55.1	1.60	0.14
Reading F.1	138.5	1.86	0.22
Reading F.2	137.0	1.87	0.20
Reading F.3	222.8	1.69	0.32
Reading G.1	14.8	3.03	0.04
Reading G.2	13.7	3.22	0.04
Reading G.3	12.2	4.50	0.04
Reading H.1	24.5	2.06	0.07
Reading H.2	23.0	2.14	0.06
Reading H.3	20.6	2.27	0.06
Reading I.1	16.8	2.65	0.05
Reading I.2	15.4	3.06	0.04
Reading I.3	15.0	3.42	0.04
Reading J.1	288.3	1.40	0.82
Reading J.2	251.8	1.37	0.92
Reading J.3	373.2	1.38	0.82
Reading K.1	215.6	1.35	0.85
Reading K.2	225.1	1.37	0.86

Continuation Table S2.1. NanoDrop™ results for DNA extraction methods (A-M) on Reading and Warwick soils.

Reading K.3	404.8	1.34	0.54
Reading L.1	372.1	1.37	0.74
Reading L.2	371.2	1.36	0.74
Reading L.3	420.5	1.36	0.71
Reading M.1	2598.1	1.39	0.89
Reading M.2	1458.0	1.38	0.76
Reading M.3	2486.7	1.39	0.89
Reading B100.1	898.3	1.35	0.11
Reading B100.2	5330.0	1.35	0.74
Reading B100.3	1186.4	1.35	0.71
Warwick A1.1	185.7	1.47	0.31
Warwick A1.2	191.9	1.46	0.32
Warwick A1.3	173.6	1.34	0.40
Warwick A2.1	54.1	1.68	0.12
Warwick A2.2	53.1	1.68	0.12
Warwick A2.3	94.3	1.55	0.21
Warwick B.1	330.5	1.41	0.48
Warwick B.2	238.4	1.45	0.39
Warwick B.3	267.5	1.45	0.42
Warwick C.1	73.8	1.65	0.16
Warwick C.2	78.6	1.64	0.17
Warwick C.3	59.8	1.75	0.14
Warwick D.1	195.2	1.46	0.36
Warwick D.2	197.3	1.46	0.37
Warwick D.3	163.5	1.46	0.32
Warwick E.1	154.5	1.76	0.19
Warwick E.2	127.2	1.84	0.17
Warwick E.3	132.4	1.86	0.21
Warwick F.1	172.6	1.84	0.21
Warwick F.2	164.8	1.87	0.24
Warwick F.3	148.8	1.91	0.21
Warwick G.1	27.6	2.01	0.08
Warwick G.2	22.4	2.23	0.06
Warwick G.3	20.0	2.30	0.06
Warwick H.1	18.9	2.73	0.05
Warwick H.2	21.7	2.31	0.06
Warwick H.3	18.5	2.72	0.05

Continuation Table S2.1. NanoDrop™ results for DNA extraction methods (A-M) on Reading and Warwick soils.

Warwick I.1	28.6	2.04	0.08
Warwick I.2	21.9	2.34	0.06
Warwick I.3	24.6	2.21	0.07
Warwick J.1	625.6	1.45	0.85
Warwick J.2	323.1	1.42	0.81
Warwick J.3	311.6	1.36	0.82
Warwick K.1	530.4	1.38	0.80
Warwick K.2	441.5	1.36	0.72
Warwick K.3	148.0	1.48	0.43
Warwick L.1	828.1	1.32	1.24
Warwick L.2	946.7	1.41	0.58
Warwick L.3	603.3	1.40	0.77
Warwick M.1	673.5	1.37	0.76
Warwick M.2	1033.1	1.37	0.75
Warwick M.3	793.4	1.38	0.75
Warwick B100.1	856.9	1.38	0.71
Warwick B100.2	473.0	1.38	0.72
Warwick B100.3	1722.3	1.36	0.60

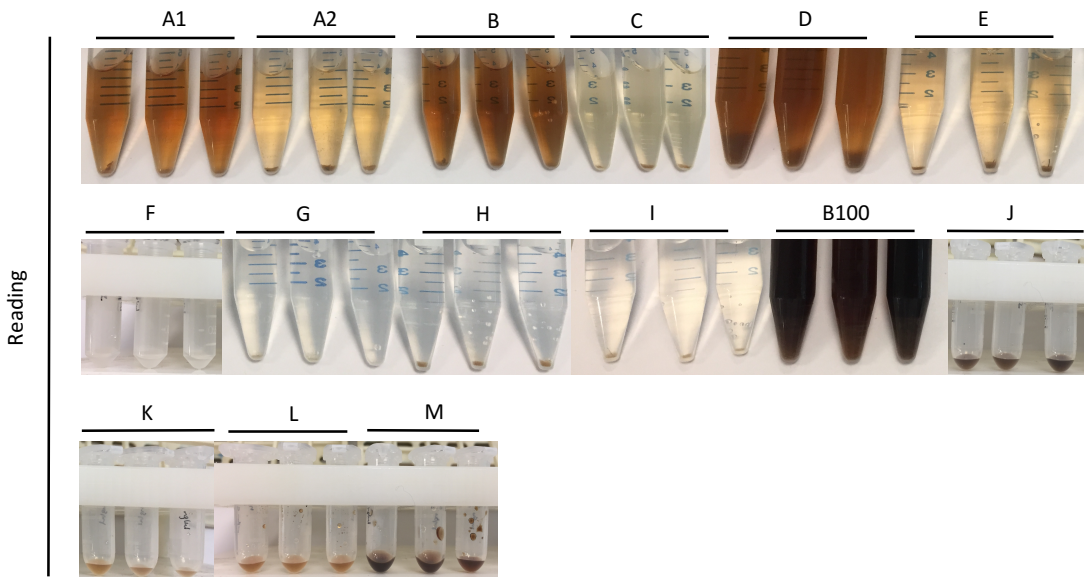


Figure S2.1 Reading soil eDNA samples extracted with methods A-M described in Chapter 2.

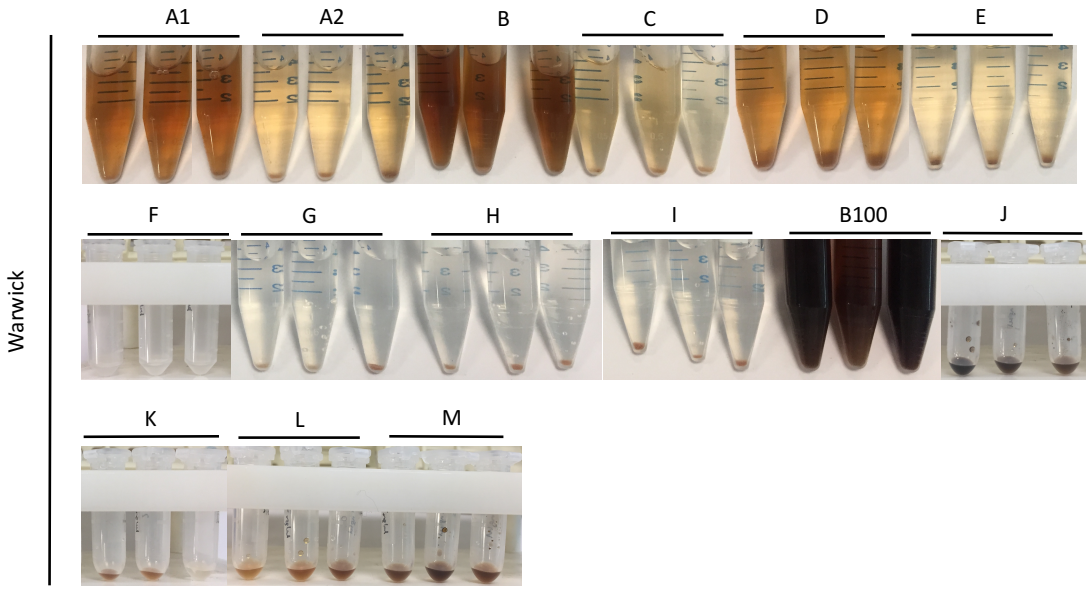


Figure S2.2 Warwick soil eDNA samples extracted with methods A-M described in Chapter 2.

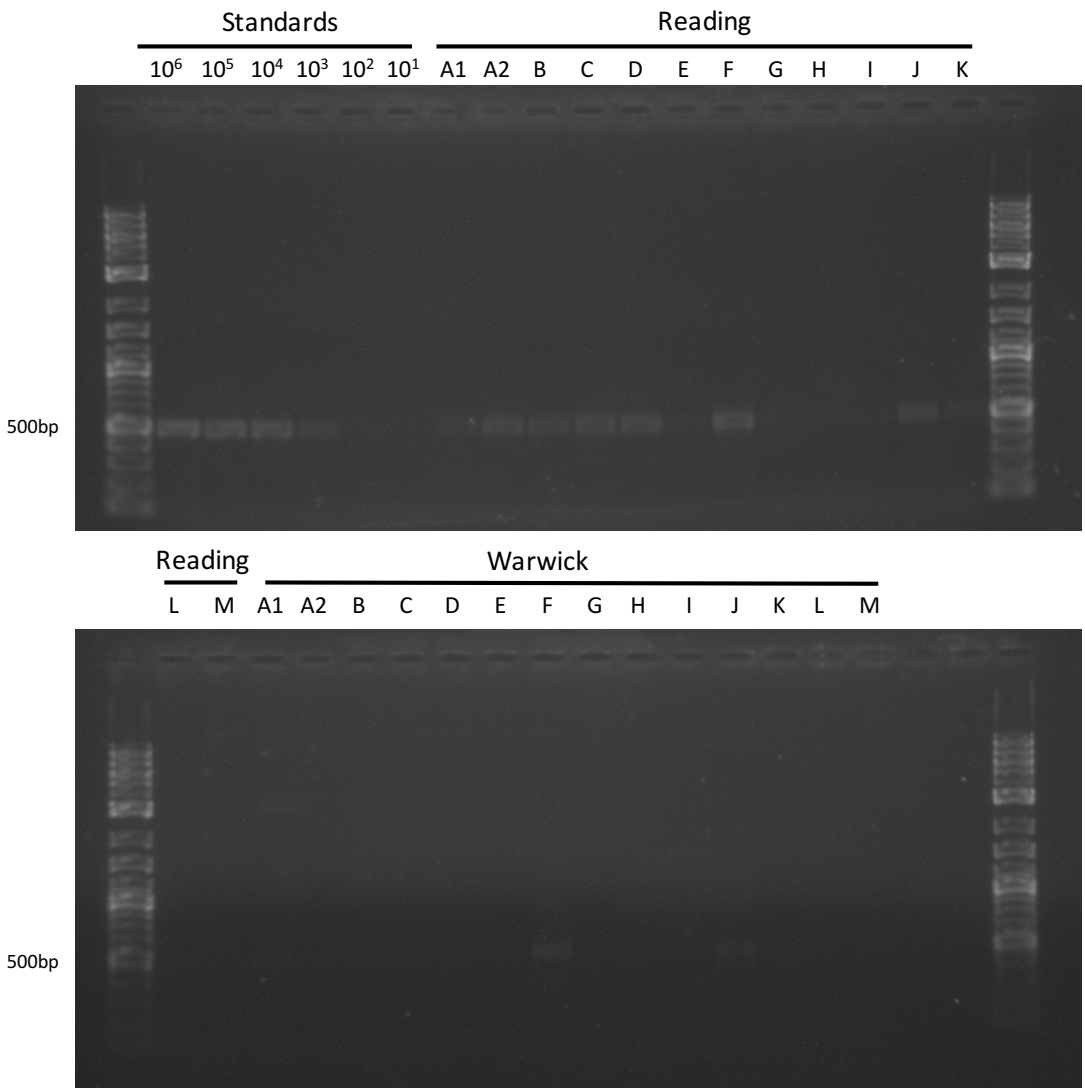


Figure S2.3 QPCR products for NRPS amplification of standards and eDNA samples extracted with methods A-M in Reading and Warwick soils.

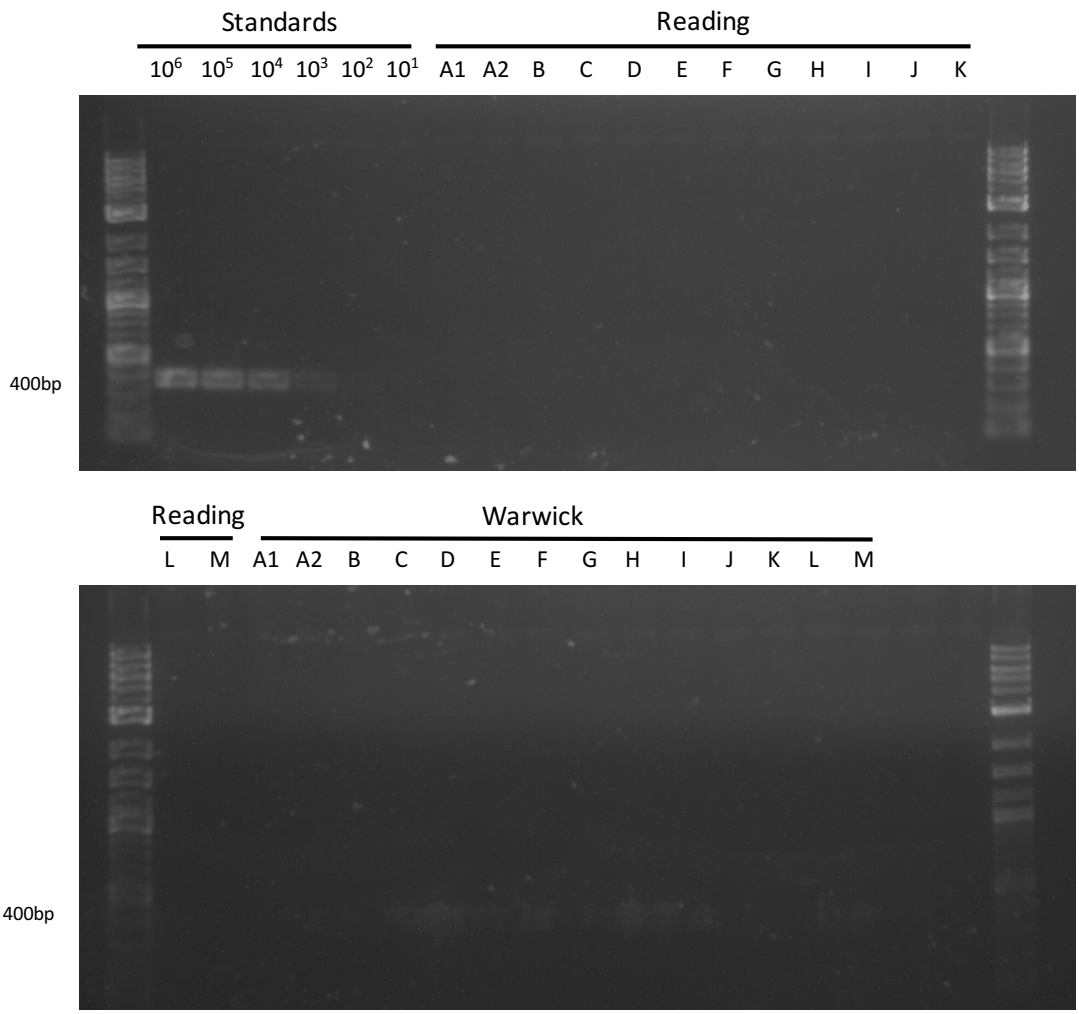


Figure S2.4 QPCR products for PKS amplification of standards and eDNA samples extracted with methods A-M in Reading and Warwick soils.

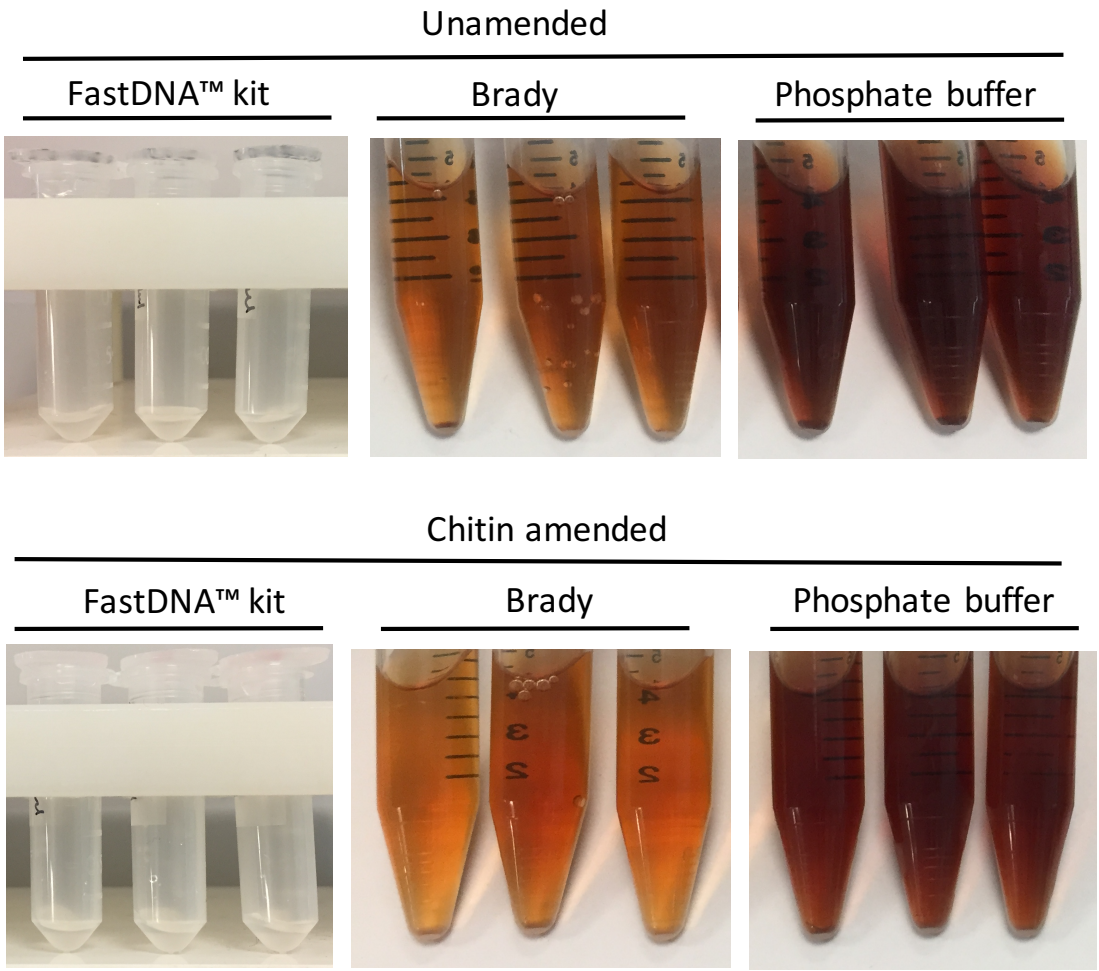


Figure S3.1 Cuban soil microcosms (chitin amended and unamended) eDNA samples extracted with methods F, A1 and D described in Chapter 2.

Table S4.1 Sequence counts for each sample for all targeted amplicon.

Counts/sample	16S rRNA gene	NRPS	PKS
S1	24288	95998	52239
S10	30206	failed	75988
S11	27478	21084	39822
S12	19819	22021	41516
S13	44074	5553	30244
S14	65959	38259	56486
S15	65750	31044	60355
S16	65933	27321	104394
S17	61958	40558	83
S18	59279	22	25334
S19	129754	398724	32878
S2	42937	64878	16
S20	105649	75290	38328
S21	84887	99925	21971
S22	69876	24008	55887
S23	222975	26700	43148
S24	114146	52437	46200
S25	90468	89102	26580
S26	61677	76766	37317
S27	62968	83784	63054
S28	79936	458	42728
S29	113245	423	38141
S3	26298	50549	28327
S30	176756	4274	54707
S31	132260	118870	12024
S32	148898	122156	8589
S33	110854	31134	820
S34	120526	2620	37851
S35	152972	4749	44096
S36	209039	1405	68300
S37	174621	37062	113981
S38	160953	153342	53902
S39	99009	115988	73447
S4	17457	46882	53747
S5	25563	45917	54
S6	30853	69226	51963
S7	32207	50096	53768
S8	80215	48652	106472
S9	80476	49860	49

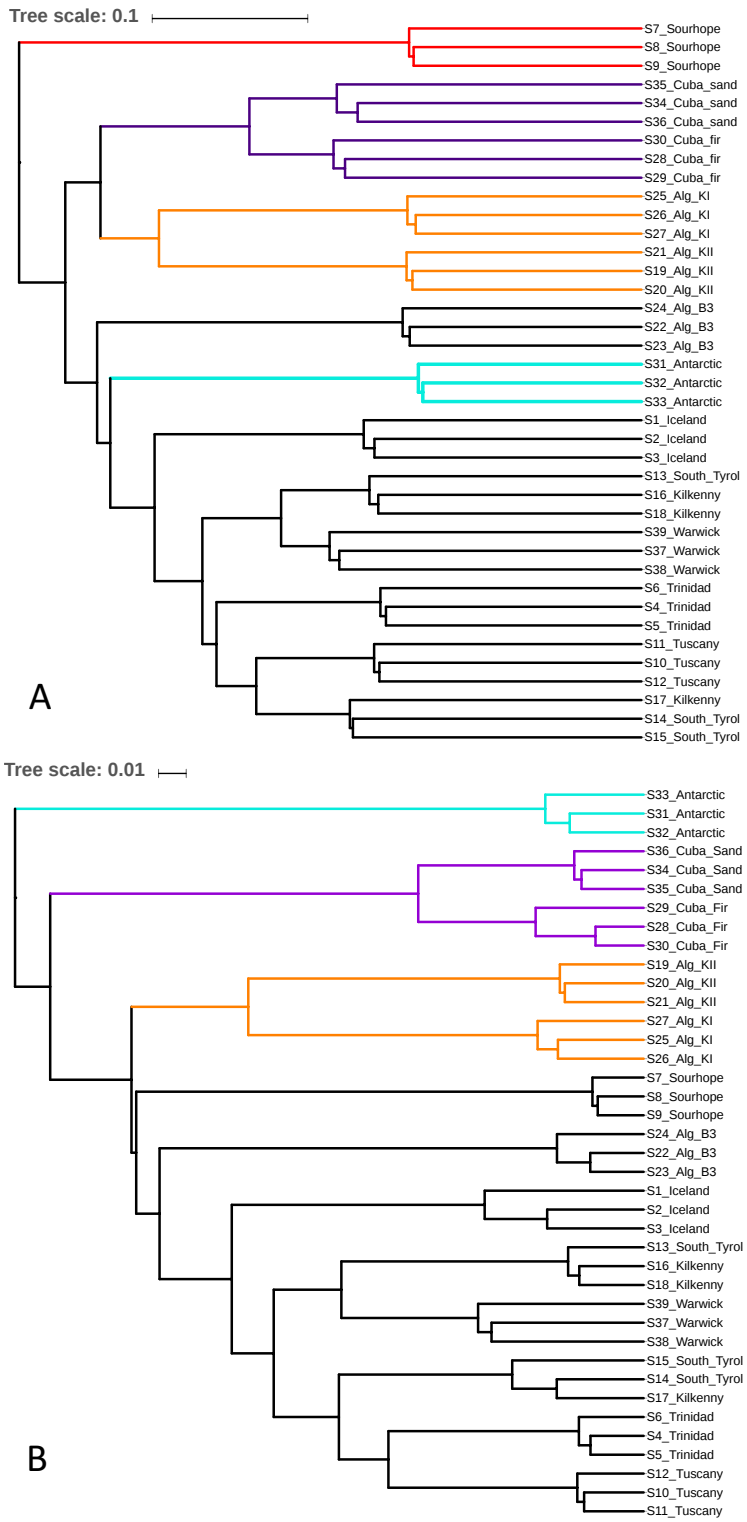


Figure S4.1 UPMGA trees of 16S rRNA genes of all samples based on unweighted unifrac (A) and weighted unifrac (B) metrics.

Bibliography

Aakvik T, Degnes KF, Dahlsrud R, Schmidt F, Dam R, Yu L, Volker U, Ellingsen TE & Valla S (2009) A plasmid RK2-based broad-host-range cloning vector useful for transfer of metagenomic libraries to a variety of bacterial species. *FEMS Microbiol Lett* **296**: 149-158.

Adrio JL & Demain AL (2014) Microbial enzymes: tools for biotechnological processes. *Biomolecules* **4**: 117-139.

Alanjary M, Kronmiller B, Adamek M, Blin K, Weber T, Huson D, Philmus B & Ziemert N (2017) The Antibiotic Resistant Target Seeker (ARTS), an exploration engine for antibiotic cluster prioritization and novel drug target discovery. *Nucleic Acids Res*.

Allen EE & Banfield JF (2005) Community genomics in microbial ecology and evolution. *Nat Rev Microbiol* **3**: 489-498.

Allen HK, Moe LA, Rodbumrer J, Gaarder A & Handelsman J (2008) Functional metagenomics reveals diverse [beta]-lactamases in a remote Alaskan soil. *ISME J* **3**: 243-251.

Allen NE, Hobbs JN & Alborn WE, Jr. (1987) Inhibition of peptidoglycan biosynthesis in gram-positive bacteria by LY146032. *Antimicrob Agents Chemother* **31**: 1093-1099.

Amos GCA, Borsetto C, Laskaris P, Krsek M, Berry AE, Newsham KK, Calvo-Bado L, Pearce DA, Vallin C & Wellington EMH (2015) Designing and implementing an assay for the detection of rare and divergent NRPS and PKS clones in European, Antarctic and Cuban Soils. *PLoS ONE* **10**: e0138327.

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. p.^pp.

Anon. (1952) Erythromycin. *British Medical Journal* **2**: 1085-1086.

Anon. (2014) 16S metagenomic sequencing library preparation. Preparing 16S ribosomal RNA gene amplicon for the Illumina MiSeq system. Vol. 2014 p.^pp. Illumina.

Aoi Y, Kinoshita T, Hata T, Ohta H, Obokata H & Tsuneda S (2009) Hollow-fiber membrane chamber as a device for in situ environmental cultivation. *Applied and Environmental Microbiology* **75**: 3826-3833.

Ayuso-Sacido A & Genilloud O (2005) New PCR primers for the screening of NRPS and PKS-I systems in Actinomycetes: detection and distribution of these biosynthetic gene sequences in major taxonomic groups. *Microbial Ecology* **49**: 10-24.

Baas Becking LGM (1934) *Geobiologie of inleiding tot de milieukunde*. W.P. Van Stockum & Zoon, Den Haag.

Babalola OO, Kirby BM, Le Roes-Hill M, Cook AE, Cary SC, Burton SG & Cowan DA (2009) Phylogenetic analysis of actinobacterial populations associated with Antarctic Dry Valley mineral soils. *Environ Microbiol* **11**: 566-576.

Baker BJ, Tyson GW, Webb RI, Flanagan J, Hugenholtz P, Allen EE & Banfield JF (2006) Lineages of acidophilic archaea revealed by community genomic analysis. *Science* **314**: 1933-1935.

Bakken LR & Lindahl V (1995) Recovery of bacterial cells from soil. *Nucleic Acids in the Environment: Methods and Applications,*(Van Elsas JD & J.T.; T, eds.), p.^pp. 9-27. Springer, Heidelberg.

Baltz RH (2010) *Streptomyces* and *Saccharopolyspora* hosts for heterologous expression of secondary metabolite gene clusters. *J Ind Microbiol Biotechnol* **37**: 759-772.

Banik JJ & Brady SF (2008) Cloning and characterization of new glycopeptide gene clusters found in an environmental DNA megalibrary. *Proc Natl Acad Sci U S A* **105**: 17273-17277.

Bankevich A, Nurk S, Antipov D*, et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**: 455-477.

Bardgett RD & van der Putten WH (2014) Belowground biodiversity and ecosystem functioning. *Nature* **515**: 505-511.

Bardgett RD, Freeman C & Ostle NJ (2008) Microbial contributions to climate change through carbon cycle feedbacks. *ISME J* **2**: 805-814.

Bardgett RD, Mommer L & De Vries FT (2014) Going underground: root traits as drivers of ecosystem processes. *Trends Ecol Evol* **29**: 692-699.

Bauer JD, King RW & Brady SF (2010) Utahmycins A and B, azaquinones produced by an environmental DNA clone. *Journal of natural products* **73**: 976-979.

Beier S & Bertilsson S (2013) Bacterial chitin degradation—mechanisms and ecophysiological strategies. *Frontiers in Microbiology* **4**: 149.

Beier S, Jones CM, Mohit V, Hallin S & Bertilsson S (2011) Global phylogeography of chitinase genes in aquatic metagenomes. *Applied and Environmental Microbiology* **77**: 1101-1106.

Bentley SD, Chater KF, Cerdeño-Tárraga AM*, et al.* (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**: 141-147.

Berdy J (1974) Recent developments of antibiotic research and classification of antibiotics according to chemical structure. *Adv Appl Microbiol* **18**: 309-406.

Berdy J (2012) Thoughts and facts about antibiotics: Where we are now and where we are heading. *J Antibiot* **65**: 385-395.

Berendsen RL, Pieterse CMJ & Bakker PAHM (2012) The rhizosphere microbiome and plant health. *Trends in Plant Science* **17**: 478-486.

Berg G & Smalla K (2009) Plant species and soil type cooperatively shape the structure and function of microbial communities in the rhizosphere. *FEMS Microbiology Ecology* **68**: 1-13.

Bergmann GT, Bates ST, Eilers KG, Lauber CL, Caporaso JG, Walters WA, Knight R & Fierer N (2011) The under-recognized dominance of Verrucomicrobia in soil bacterial communities. *Soil Biology and Biochemistry* **43**: 1450-1455.

Berry AE, Chiocchini C, Selby T, Sosio M & Wellington EM (2003) Isolation of high molecular weight DNA from soil for cloning into BAC vectors. *FEMS Microbiol Lett* **223**: 15-20.

Bibb MJ (2005) Regulation of secondary metabolism in streptomycetes. *Curr Opin Microbiol* **8**: 208-215.

Bierman M, Logan R, O'Brien K, Seno ET, Rao RN & Schoner BE (1992) Plasmid cloning vectors for the conjugal transfer of DNA from *Escherichia coli* to *Streptomyces* spp. *Gene* **116**: 43-49.

Bolger AM, Lohse M & Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120.

Borsetto C & Wellington EMH (2017) Bioprospecting soil metagenomes for antibiotics. *Bioprospecting: Success, Potential and Constraints,*(Paterson R & Lima N, eds.), p.^pp. 113-136. Springer International Publishing, Cham.

Brady SF (2007) Construction of soil environmental DNA cosmid libraries and screening for clones that produce biologically active small molecules. *Nat Protoc* **2**: 1297-1305.

Braid MD, Daniels LM & Kitts CL (2003) Removal of PCR inhibitors from soil DNA by chemical flocculation. *J Microbiol Methods* **52**: 389-393.

Brankatschk R, Bodenhausen N, Zeyer J & Burgmann H (2012) Simple absolute quantification method correcting for quantitative PCR efficiency variations for microbial community samples. *Appl Environ Microbiol* **78**: 4481-4489.

Branton D, Deamer DW, Marziali A*, et al.* (2008) The potential and challenges of nanopore sequencing. *Nat Biotechnol* **26**: 1146-1153.

Brewer TE, Handley KM, Carini P, Gilbert JA & Fierer N (2016) Genome reduction in an abundant and ubiquitous soil bacterium 'Candidatus Udaeobacter copiosus'. *Nature Microbiology* **2**: 16198.

Brigham RB & Pittenger RC (1956) *Streptomyces orientalis*, n. sp., the source of vancomycin. *Antibiot Chemother (Northfield)* **6**: 642-647.

Bru D, Martin-Laurent F & Philippot L (2008) Quantification of the detrimental effect of a single primer-template mismatch by real-time PCR using the 16S rRNA gene as an example. *Appl Environ Microbiol* **74**: 1660-1663.

Bulgarelli D, Rott M, Schlaeppi K*, et al.* (2012) Revealing structure and assembly cues for Arabidopsis root-inhabiting bacterial microbiota. *Nature* **488**: 91-95.

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.

Bustin SA, Benes V, Garson JA*, et al.* (2009) The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin Chem* **55**: 611-622.

Caporaso JG, Kuczynski J, Stombaugh J*, et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**: 335-336.

Carini P, Marsden PJ, Leff JW, Morgan EE, Strickland MS & Fierer N (2016) Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. *Nat Microbiol* **2**: 16242.

Challis GL, Ravel J & Townsend CA (2000) Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem Biol* **7**: 211-224.

Charlop-Powers Z, Banik JJ, Owen JG, Craig JW & Brady SF (2013) Selective enrichment of environmental DNA libraries for genes encoding nonribosomal peptides and polyketides by phosphopantetheine transferase-dependent complementation of siderophore biosynthesis. *ACS Chem Biol* **8**: 138-143.

Charlop-Powers Z, Owen JG, Reddy BV, Ternei MA & Brady SF (2014) Chemical-biogeographic survey of secondary metabolism in soil. *Proc Natl Acad Sci U S A* **111**: 3757-3762.

Charlop-Powers Z, Owen JG, Reddy BVB, Ternei MA, Guimarães DO, de Frias UA, Pupo MT, Seepe P, Feng Z & Brady SF (2015) Global biogeographic sampling of bacterial secondary metabolism. *eLife* **4**: e05048.

Chin-A-Woeng TFC, Bloemberg GV & Lugtenberg BJJ (2003) Phenazines and their role in biocontrol by *Pseudomonas* bacteria. *New Phytologist* **157**: 503-523.

Choi S-S, Kim H-J, Lee H-S, Kim P & Kim E-S (2015) Genome mining of rare actinomycetes and cryptic pathway awakening. *Process Biochemistry* **50**: 1184-1193.

Christie-Oleza JA, Brunet-Galmes I, Lalucat J, Nogales B & Bosch R (2013) MiniUIB, a novel minitransposon-based system for stable insertion of foreign DNA into the genomes of Gram-negative and Gram-positive bacteria. *Appl Environ Microbiol* **79**: 1629-1638.

Clardy J, Fischbach MA & Currie CR (2009) The natural history of antibiotics. *Current biology : CB* **19**: R437-441.

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR & Tiedje JM (2014) Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res* **42**: D633-642.

Collins J & Hohn B (1978) Cosmids: a type of plasmid gene-cloning vector that is packageable in vitro in bacteriophage lambda heads. *Proc Natl Acad Sci U S A* **75**: 4242-4246.

Cottrell MT, Wood DN, Yu L & Kirchman DL (2000) Selected chitinase genes in cultured and uncultured marine bacteria in the α- and γ-Subclasses of the Proteobacteria. *Applied and Environmental Microbiology* **66**: 1195-1201.

Courtois S, Frostegård Å, Göransson P, Depret G, Jeannin P & Simonet P (2001) Quantification of bacterial subgroups in soil: comparison of DNA extracted directly from soil or from cells previously released by density gradient centrifugation. *Environmental Microbiology* **3**: 431-439.

Cowan DA, Makhalanyane TP, Dennis PG & Hopkins DW (2014) Microbial ecology and biogeochemistry of continental Antarctic soils. *Frontiers in Microbiology* **5**: 154.

Craig JW, Chang FY & Brady SF (2009) Natural products from environmental DNA hosted in *Ralstonia metallidurans*. *ACS Chem Biol* **4**: 23-28.

Craig JW, Chang FY, Kim JH, Obiajulu SC & Brady SF (2010) Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria. *Appl Environ Microbiol* **76**: 1633-1641.

Cretoiu M, Berini F, Kielak A, Marinelli F & van Elsas J (2015) A novel salt-tolerant chitobiosidase discovered by genetic screening of a metagenomic library derived from chitin-amended agricultural soil. *Applied Microbiology and Biotechnology* 1-17.

Cretoiu MS, Korthals GW, Visser JH & van Elsas JD (2013) Chitin amendment increases soil suppressiveness toward plant pathogens and modulates the actinobacterial and oxalobacteraceal communities in an experimental agricultural field. *Appl Environ Microbiol* **79**: 5291-5301.

Daniel R (2005) The metagenomics of soil. *Nat Rev Microbiol* **3**: 470-478.

Datsenko KA & Wanner BL (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc Natl Acad Sci U S A* **97**: 6640-6645.

De Wit R & Bouvier T (2006) 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? *Environmental Microbiology* **8**: 755-758.

Debode F, Marien A, Janssen E, Bragard C & Beren G (2017) The influence of amplicon length on real-time PCR results. *Biotechnology, Agronomy, Society and Environment* **21**: 3-11.

Debode J, De Tender C, Soltaninejad S, Van Malderghem C, Haegeman A, Van der Linden I, Cottyn B, Heyndrickx M & Maes M (2016) Chitin mixed in potting soil alters lettuce growth, the survival of zoonotic bacteria on the leaves and associated rhizosphere microbiology. *Frontiers in Microbiology* **7**: 565.

Dejong CA, Chen GM, Li H, Johnston CW, Edwards MR, Rees PN, Skinnider MA, Webster AL & Magarvey NA (2016) Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. *Nat Chem Biol* **12**: 1007-1014.

Delmont TO, Eren AM, Maccario L, Prestat E, Esen OC, Pelletier E, Le Paslier D, Simonet P & Vogel TM (2015) Reconstructing rare soil microbial genomes using in situ enrichments and metagenomics. *Front Microbiol* **6**: 358.

DeLong EF (2005) Microbial community genomics in the ocean. *Nat Rev Microbiol* **3**: 459-469.

Dennis JJ & Zylstra GJ (1998) Plasposons: modular self-cloning minitransposon derivatives for rapid genetic analysis of gram-negative bacterial genomes. *Appl Environ Microbiol* **64**: 2710-2715.

Derakshani M, Lukow T & Liesack W (2001) Novel bacterial lineages at the (sub)division level as detected by signature nucleotide-targeted recovery of 16S rRNA genes from bulk soil and rice roots of flooded rice microcosms. *Appl Environ Microbiol* **67**: 623-631.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P & Andersen GL (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**: 5069-5072.

Donadio S, Monciardini P & Sosio M (2007) Polyketide synthases and nonribosomal peptide synthetases: the emerging view from bacterial genomics. *Nat Prod Rep* **24**: 1073-1109.

Donia MS, Cimermancic P, Schulze CJ, Wieland Brown LC, Martin J, Mitreva M, Clardy J, Linington RG & Fischbach MA (2014) A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell* **158**: 1402-1414.

Drews J (2000) Drug discovery: a historical perspective. *Science* **287**: 1960-1964.

Du L, Sánchez C, Chen M, Edwards DJ & Shen B (2000) The biosynthetic gene cluster for the antitumor drug bleomycin from *Streptomyces verticillus* ATCC15003 supporting functional interactions between nonribosomal peptide synthetases and a polyketide synthase. *Chemistry & Biology* **7**: 623-642.

Dubourg G, Lagier J-C, Armougom F, Robert C, Audoly G, Papazian L & Raoult D (2000) High-level colonisation of the human gut by Verrucomicrobia following

broad-spectrum antibiotic treatment. *International Journal of Antimicrobial Agents* **41**: 149-155.

Duggar BM (1948) Aureomycin: a product of the continuing search for new antibiotics. *Ann N Y Acad Sci* **1241**: 163-169.

Duncan KR, Crüsemann M, Lechner A*, et al.* (2015) Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from *Salinispora* species. *Chemistry & biology* **22**: 460-471.

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460-2461.

Edgar RC (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* **10**: 996-998.

Edgar RC & Flyvbjerg H (2015) Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics* **31**: 3476-3482.

Ehrlich J, Gottlieb D & et al. (1948) *Streptomyces venezuelae*, n. sp., the source of chloromycetin. *J Bacteriol* **56**: 467-477.

Elsaesser R & Paysan J (2004) Liquid gel amplification of complex plasmid libraries. *Biotechniques* **37**: 200, 202.

Epicentre (2012) CopyControl^{TM} Fosmid Library Production Kit with pCC1FOS^{TM} Vector. p.^pp.

Epicentre (2013) CopyControlTM BAC Cloning Kit with pCC1BACTM Vector. Vol. 2015 p.^pp.

Esmaeel Q, Pupin M, Kieu NP, Chataigné G, Béchet M, Deravel J, Krier F, Höfte M, Jacques P & Leclère V (2016) *Burkholderia* genome mining for nonribosomal peptide synthetases reveals a great potential for novel siderophores and lipopeptides synthesis. *MicrobiologyOpen* **5**: 512-526.

Esnard J, Potter TL & Zuckerman BM (1995) *Streptomyces costaricanus* sp. nov., isolated from nematode-suppressive soil. *Int J Syst Bacteriol* **45**: 775-779.

Fàbrega A, Sánchez-Céspedes J, Soto S & Vila J (2008) Quinolone resistance in the food chain. *International Journal of Antimicrobial Agents* **31**: 307-315.

Fægri A, Torsvik VL & GoksÖyr J (1977) Bacterial and fungal activities in soil: Separation of bacteria and fungi by a rapid fractionated centrifugation technique. *Soil Biology and Biochemistry* **9**: 105-112.

Feng Z, Kim JH & Brady SF (2010) Fluostatins produced by the heterologous expression of a TAR reassembled environmental DNA derived type II PKS gene cluster. *J Am Chem Soc* **132**: 11902-11903.

Feng Z, Chakraborty D, Dewell SB, Reddy BV & Brady SF (2012) Environmental DNA-encoded antibiotics fasamycins A and B inhibit FabF in type II fatty acid biosynthesis. *J Am Chem Soc* **134**: 2981-2987.

Fernandes P (2006) Antibacterial discovery and development--the failure of success? *Nat Biotechnol* **24**: 1497-1503.

Fierer N, Schimel JP & Holden PA (2003) Variations in microbial community composition through two soil depth profiles. *Soil Biology and Biochemistry* **35**: 167-176.

Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, Owens S, Gilbert JA, Wall DH & Caporaso JG (2012) Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc Natl Acad Sci U S A* **109**: 21390-21395.

Fieseler L, Hentschel U, Grozdanov L, Schirmer A, Wen G, Platzer M, Hrvatin S, Butzke D, Zimmermann K & Piel J (2007) Widespread occurrence and genomic context of unusually small polyketide synthase genes in microbial consortia associated with marine sponges. *Appl Environ Microbiol* **73**: 2144-2155.

Finlay BJ (2002) Global dispersal of free-living microbial eukaryote species. *Science* **296**: 1061-1063.

Finn RD, Attwood TK, Babbitt PC*, et al.* (2017) InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Research* **45**: D190-D199.

Fischbach MA & Walsh CT (2006) Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. *Chemical Reviews* **106**: 3468-3496.

Fleming A (1929) On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of *B. influenzae*. 1929. *Bulletin of the World Health Organization* **79**: 780-790.

Frearson J & Wyatt P (2010) Drug discovery in academia- the third way? *Expert opinion on drug discovery* **5**: 909-919.

Freitas S, Hatosy S, Fuhrman JA, Huse SM, Mark Welch DB, Sogin ML & Martiny AC (2012) Global distribution and diversity of marine Verrucomicrobia. *The ISME Journal* **6**: 1499-1505.

Frostegard A, Courtois S, Ramisse V, Clerc S, Bernillon D, Le Gall F, Jeannin P, Nesme X & Simonet P (1999) Quantification of bias related to the extraction of DNA directly from soils. *Appl Environ Microbiol* **65**: 5409-5420.

Gabor EM, Alkema WB & Janssen DB (2004) Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environ Microbiol* **6**: 879-886.

George IF, Hartmann M, Liles MR & Agathos SN (2011) Recovery of as-yet-uncultured soil acidobacteria on dilute solid media. *Applied and Environmental Microbiology* **77**: 8184-8188.

Gerwick WH & Moore BS (2012) Lessons from the past and charting the future of marine natural products drug discovery and chemical biology. *Chem Biol* **19**: 85-98.

Gillespie DE, Brady SF, Bettermann AD, Cianciotto NP, Liles MR, Rondon MR, Clardy J, Goodman RM & Handelsman J (2002) Isolation of antibiotics turbomycin a and B from a metagenomic library of soil microbial DNA. *Appl Environ Microbiol* **68**: 4301-4306.

Glenn TC (2011) Field guide to next-generation DNA sequencers. *Mol Ecol Resour* **11**: 759-769.

Gomez-Escribano JP & Bibb MJ (2011) Engineering *Streptomyces coelicolor* for heterologous expression of secondary metabolite gene clusters. *Microb Biotechnol* **4**: 207-215.

Gomez-Escribano JP, Song L, Fox DJ, Yeo V, Bibb MJ & Challis GL (2012) Structure and biosynthesis of the unusual polyketide alkaloid coelimycin P1, a metabolic product of the *cpk* gene cluster of *Streptomyces coelicolor* M145. Vol. 3 p.^pp. 2716-2720. Chemical Science.

Gooday GW (1990) The ecology of chitin degradation. *Advances in Microbial Ecology,*(Marshall KC, ed.) p.^pp. 387-430. Springer US, Boston, MA.

Goodfellow M & Williams ST (1983) Ecology of actinomycetes. *Annu Rev Microbiol* **37**: 189-216.

Graça AP, Calisto R & Lage OM (2016) *Planctomycetes* as novel source of bioactive molecules. *Frontiers in Microbiology* **7**: 1241.

Gullon S, Olano C, Abdelfattah MS, Brana AF, Rohr J, Mendez C & Salas JA (2006) Isolation, characterization, and heterologous expression of the biosynthesis gene cluster for the antitumor anthracycline steffimycin. *Appl Environ Microbiol* **72**: 4172-4183.

Gust B, Chandra G, Jakimowicz D, Yuqing T, Bruton CJ & Chater KF (2004) Lambda red-mediated genetic manipulation of antibiotic-producing *Streptomyces*. *Adv Appl Microbiol* **54**: 107-128.

Hadjithomas M, Chen IM, Chu K*, et al.* (2015) IMG-ABC: A knowledge base to fuel discovery of biosynthetic gene clusters and novel secondary metabolites. *MBio* **6**: e00932.

Hain T, Otten S, von Both U, Chatterjee SS, Technow U, Billion A, Ghai R, Mohamed W, Domann E & Chakraborty T (2008) Novel bacterial artificial chromosome vector pUvBBAC for use in studies of the functional genomics of *Listeria* spp. *Applied and Environmental Microbiology* **74**: 1892-1901.

Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews* **68**: 669-685.

Handelsman J, Rondon MR, Brady SF, Clardy J & Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* **5**: R245-249.

Hannauer M, Schafer M, Hoegy F, Gizzi P, Wehrung P, Mislin GL, Budzikiewicz H & Schalk IJ (2012) Biosynthesis of the pyoverdine siderophore of *Pseudomonas aeruginosa i*nvolves precursors with a myristic or a myristoleic acid chain. *FEBS Lett* **586**: 96-101.

Hanson CA, Fuhrman JA, Horner-Devine MC & Martiny JBH (2012) Beyond biogeographic patterns: processes shaping the microbial landscape. *Nat Rev Micro* **10**: 497-506.

Harbarth S, Theuretzbacher U, Hackett J & consortium D-A (2015) Antibiotic research and development: business as usual? *J Antimicrob Chemother* **70**: 1604-1607.

Herron PR & Wellington EM (1990) New method for extraction of streptomycete spores from soil and application to the study of lysogeny in sterile amended and nonsterile soil. *Appl Environ Microbiol* **56**: 1406-1412.

Heuer H, Krsek M, Baker P, Smalla K & Wellington EM (1997) Analysis of actinomycete communities by specific amplification of genes encoding 16S rRNA and gel-electrophoretic separation in denaturing gradients. *Appl Environ Microbiol* **63**: 3233-3241.

Hjort K, Bergström M, Adesina MF, Jansson JK, Smalla K & Sjöling S (2010) Chitinase genes revealed and compared in bacterial isolates, DNA extracts and a metagenomic library from a phytopathogen-suppressive soil. *FEMS Microbiology Ecology* **71**: 197-207.

Hoffmann A, Thimm T & Tebbe CC (1999) Fate of plasmid-bearing, luciferase marker gene tagged bacteria after feeding to the soil microarthropod *Onychiurus fimatus* (Collembola). *FEMS Microbiol Ecol* **30**: 125-135.

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 Microbiol* **54**: 703-711.

Horner-Devine MC, Carney KM & Bohannan BJM (2004) An ecological perspective on bacterial biodiversity. *Proceedings of the Royal Society B: Biological Sciences* **271**: 113-122.

Hrvatin S & Piel J (2007) Rapid isolation of rare clones from highly complex DNA libraries by PCR analysis of liquid gel pools. *J Microbiol Methods* **68**: 434-436.

Hunter S, Corbett M, Denise H*, et al.* (2014) EBI metagenomics--a new resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res* **42**: D600-606.

Ikeda H, Nonomiya T, Usami M, Ohta T & Omura S (1999) Organization of the biosynthetic gene cluster for the polyketide anthelmintic macrolide avermectin in *Streptomyces avermitilis*. *Proc Natl Acad Sci U S A* **96**: 9509-9514.

Ioannou PA, Amemiya CT, Garnes J, Kroisel PM, Shizuya H, Chen C, Batzer MA & de Jong PJ (1994) A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nat Genet* **6**: 84-89.

Jacquiod S, Franqueville L, Cécillon S, M Vogel T & Simonet P (2013) Soil bacterial community shifts after chitin enrichment: an integrative metagenomic approach. *PLoS One* **8**: e79699.

Janssen PH, Yates PS, Grinton BE, Taylor PM & Sait M (2002) Improved Culturability of Soil Bacteria and Isolation in Pure Culture of Novel Members of the Divisions Acidobacteria, Actinobacteria, Proteobacteria, and Verrucomicrobia. *Applied and Environmental Microbiology* **68**: 2391-2396.

Jenke-Kodama H, Sandmann A, Müller R & Dittmann E (2005) Evolutionary Implications of Bacterial Polyketide Synthases. *Molecular Biology and Evolution* **22**: 2027-2039.

Jevons MP (1961) "Celbenin" - resistant Staphylococci. *Bmj* **1**: 124-125.

Johnson-Rollings AS (2012) A polyphasic approach to the study of chitinolytic bacteria in soil. Thesis, University of Warwick.

Johnson-Rollings AS, Wright H, Masciandaro G, Macci C, Doni S, Calvo-Bado LA, Slade SE, Vallin Plou C & Wellington EMH (2014) Exploring the functional soil-microbe interface and exoenzymes through soil metaexoproteomics. *ISME J* **8**: 2148-2150.

Johnston CW, Wyatt MA, Li X, Ibrahim A, Shuster J, Southam G & Magarvey NA (2013) Gold biomineralization by a metallophore from a gold-associated microbe. *Nat Chem Biol* **9**: 241-243.

Johnstone DB (1947) Soil actinomycetes of Bikini Atoll, with special reference to an antibiotic-producing organism. *J Bacteriol* **54**: 25.

Johnstone DB & Waksman SA (1948) The production of streptomycin by *Streptomyces bikiniensis*. *Journal of Bacteriology* **55**: 317-326.

Jones AC, Gust B, Kulik A, Heide L, Buttner MJ & Bibb MJ (2013) Phage P1-Derived Artificial Chromosomes Facilitate Heterologous Expression of the FK506 Gene Cluster. *PLOS ONE* **8**: e69319.

Jones P, Binns D, Chang H-Y*, et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**: 1236-1240.

Jovetic S, Zhu Y, Marcone GL, Marinelli F & Tramper J (2010) β-Lactam and glycopeptide antibiotics: first and last line of defense? *Trends Biotechnol* **28**: 596-604.

Kallifidas D & Brady SF (2012) Reassembly of functionally intact environmental DNA-derived biosynthetic gene clusters. *Methods Enzymol* **517**: 225-239.

Kallifidas D, Kang HS & Brady SF (2012) Tetarimycin A, an MRSA-active antibiotic identified through induced expression of environmental DNA gene clusters. *J Am Chem Soc* **134**: 19552-19555.

Kaltenpoth M (2009) Actinobacteria as mutualists: general healthcare for insects? *Trends in microbiology* **17**: 529-535.

Kaltenpoth M, Goettler W, Dale C, Stubblefield JW, Herzner G, Roeser-Mueller K & Strohm E (2006) '*Candidatus Streptomyces philanthi*', an endosymbiotic streptomycete in the antennae of Philanthus digger wasps. *International Journal of Systematic and Evolutionary Microbiology* **56**: 1403-1411.

Kang HS & Brady SF (2014) Arixanthomycins A-C: Phylogeny-guided discovery of biologically active eDNA-derived pentangular polyphenols. *ACS Chem Biol* **9**: 1267-1272.

Kang S & Mills AL (2006) The effect of sample size in studies of soil microbial community structure. *J Microbiol Methods* **66**: 242-250.

Keller M & Zengler K (2004) Tapping into microbial diversity. *Nat Rev Micro* **2**: 141-150.

Kielak AM, Cretoiu MS, Semenov AV, Sørensen SJ & van Elsas JD (2013) Bacterial chitinolytic communities respond to chitin and pH alteration in soil. *Applied and Environmental Microbiology* **79**: 263-272.

Kieser T, Bibb MJ, Buttner MJ, Chater KF & Hopwood DA (2000) *Practical Streptomyces genetics*. John Innes Foundation, Norwich Research Park, Colney, Norwich NR4 7UH, UK.

Kim JH, Feng Z, Bauer JD, Kallifidas D, Calle PY & Brady SF (2010) Cloning large natural product gene clusters from the environment: piecing environmental DNA gene clusters back together with TAR. *Biopolymers* **93**: 833-844.

Kim U-J, Birren BW, Slepak T, Mancino V, Boysen C, Kang H-L, Simon MI & Shizuya H (1996) Construction and characterization of a human bacterial artificial chromosome library. *Genomics* **34**: 213-218.

Kim UJ, Shizuya H, de Jong PJ, Birren B & Simon MI (1992) Stable propagation of cosmid sized human DNA inserts in an F factor based vector. *Nucleic Acids Res* **20**: 1083-1085.

Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M & Glöckner FO (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research* **41**: e1-e1.

Kovach ME, Elzer PH, Hill DS, Robertson GT, Farris MA, Roop RM, 2nd & Peterson KM (1995) Four new derivatives of the broad-host-range cloning vector pBBR1MCS, carrying different antibiotic-resistance cassettes. *Gene* **166**: 175-176.

Kreutzer MF & Nett M (2012) Genomics-driven discovery of taiwachelin, a lipopeptide siderophore from *Cupriavidus taiwanensis*. *Organic & Biomolecular Chemistry* **10**: 9338-9343.

Krsek M & Wellington EMH (1999) Comparison of different methods for the isolation and purification of total community DNA from soil. *Journal of Microbiological Methods* **39**: 1-16.

Lam KN, Cheng J, Engel K, Neufeld JD & Charles TC (2015) Current and future resources for functional metagenomics. *Frontiers in Microbiology* **6**: 1196.

Land M, Hauser L, Jun S-R*, et al.* (2015) Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics* **15**: 141-161.

Laskaris P (2009) Evolution of the streptomycin and viomycin biosynthetic clusters and resistance genes. Thesis (Ph.D.) - University of Warwick, 2009. Thesis, S.l.

Laureti L, Song L, Huang S, Corre C, Leblond P, Challis GL & Aigle B (2011) Identification of a bioactive 51-membered macrolide complex by activation of a silent polyketide synthase in *Streptomyces ambofaciens*. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 6258-6263.

Lazzarini A, Cavaletti L, Toppo G & Marinelli F (2000) Rare genera of actinomycetes as potential producers of new antibiotics. *Antonie Van Leeuwenhoek* **78**: 399-405.

Le T-H, Sivachidambaram V, Yi X, Li X & Zhou Z (2014) Quantification of polyketide synthase genes in tropical urban soils using real-time PCR. *Journal of Microbiological Methods* **106**: 135-142.

Lee MH & Lee SW (2013) Bioprospecting potential of the soil metagenome: novel enzymes and bioactivities. *Genomics Inform* **11**: 114-120.

Letzel AC, Pidot SJ & Hertweck C (2013) A genomic approach to the cryptic secondary metabolome of the anaerobic world. *Nat Prod Rep* **30**: 392-428.

Levy SB & Marshall B (2004) Antibacterial resistance worldwide: causes, challenges and responses. *Nat Med* **10**: S122-129.

Li WJ, Zhang LP, Xu P, Cui XL, Lu ZT, Xu LH & Jiang CL (2002) *Streptomyces beijiangensis* sp. nov., a psychrotolerant actinomycete isolated from soil in China. *Int J Syst Evol Microbiol* **52**: 1695-1699.

Li Y, Tan W & Koopal LK (2013) The influence of humic acids on the activities of lysozyme and urease. *Functions of Natural Organic Matter in Changing Environment,*(Xu J, Wu J & He Y, eds.), p.^pp. 817-821. Springer Netherlands, Dordrecht.

Licona-Cassani C, Cruz-Morales P, Manteca A, Barona-Gomez F, Nielsen LK & Marcellin E (2015) Systems biology approaches to understand natural products biosynthesis. *Frontiers in Bioengineering and Biotechnology* **3**: 199.

Lidbury ID, Murphy AR, Scanlan DJ, Bending GD, Jones AM, Moore JD, Goodall A, Hammond JP & Wellington EM (2016) Comparative genomic, proteomic and exoproteomic analyses of three *Pseudomonas* strains reveals novel insights into the phosphorus scavenging capabilities of soil bacteria. *Environ Microbiol* **18**: 3535-3549.

Liles MR, Williamson LL, Rodbumrer J, Torsvik V, Goodman RM & Handelsman J (2008) Recovery, purification, and cloning of high-molecular-weight DNA from soil microorganisms. *Appl Environ Microbiol* **74**: 3302-3305.

Lindahl BD & Finlay RD (2006) Activities of chitinolytic enzymes during primary and secondary colonization of wood by basidiomycetous fungi. *New Phytologist* **169**: 389-397.

Lindahl V & Bakken LR (1995) Evaluation of methods for extraction of bacteria from soil. *FEMS Microbiology Ecology* **16**: 135-142.

Ling LL, Schneider T, Peoples AJ*, et al.* (2015) A new antibiotic kills pathogens without detectable resistance. *Nature* **517**: 455-459.

Linger JG, Adney WS & Darzins A (2010) Heterologous expression and extracellular secretion of cellulolytic enzymes by *Zymomonas mobilis*. *Appl Environ Microbiol* **76**: 6360-6369.

Liu G, Chater KF, Chandra G, Niu G & Tan H (2013) Molecular regulation of antibiotic biosynthesis in *streptomyces*. *Microbiol Mol Biol Rev* **77**: 112-143.

Logares R, Sunagawa S, Salazar G*, et al.* (2013) Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ Microbiol*.

Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J & Pallen MJ (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotech* **30**: 434-439.

Lombo F, Velasco A, Castro A, de la Calle F, Brana AF, Sanchez-Puelles JM, Mendez C & Salas JA (2006) Deciphering the biosynthesis pathway of the antitumor thiocoraline from a marine actinomycete and its expression in two streptomyces species. *Chembiochem* **7**: 366-376.

Lozupone C & Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71**: 8228-8235.

Lozupone C, Hamady M & Knight R (2006) UniFrac--an online tool for comparing microbial community diversity in a phylogenetic context. *BMC bioinformatics* **7**: 371.

Lozupone C, Lladser ME, Knights D, Stombaugh J & Knight R (2011) UniFrac: an effective distance metric for microbial community comparison. *ISME J* **5**: 169-172.

Luo C, Tsementzi D, Kyrpides N, Read T & Konstantinidis KT (2012) Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS One* **7**: e30087.

Lutz MP, Wenger S, Maurhofer M, Defago G & Duffy B (2004) Signaling between bacterial and fungal biocontrol agents in a strain mixture. *FEMS Microbiol Ecol* **48**: 447-455.

Mackey TK & Liang BA (2012) Integrating biodiversity management and indigenous biopiracy protection to promote environmental justice and global health. *American Journal of Public Health* **102**: 1091-1095.

Markowitz VM, Chen IM, Chu K*, et al.* (2014) IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Res* **42**: D568-573.

Marschner P, Yang CH, Lieberei R & Crowley DE (2001) Soil and plant specific effects on bacterial community composition in the rhizosphere. *Soil Biology and Biochemistry* **33**: 1437-1445.

Martinez A, Kolvek SJ, Yip CL, Hopke J, Brown KA, MacNeil IA & Osburne MS (2004) Genetically modified bacterial strains and novel bacterial artificial chromosome shuttle vectors for constructing environmental libraries and detecting heterologous natural products in multiple expression hosts. *Appl Environ Microbiol* **70**: 2452-2463.

Martiny JBH, Eisen JA, Penn K, Allison SD & Horner-Devine MC (2011) Drivers of bacterial β-diversity depend on spatial scale. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 7850-7854.

Martiny JBH, Bohannan BJM, Brown JH*, et al.* (2006) Microbial biogeography: putting microorganisms on the map. *Nat Rev Micro* **4**: 102-112.

Masella AP, Bartram AK, Truszkowski JM, Brown DG & Neufeld JD (2012) PANDAseq: paired-end assembler for illumina sequences. *BMC bioinformatics* **13**: 31.

Matilla MA, Pizarro-Tobias P, Roca A*, et al.* (2011) Complete genome of the plant growth-promoting rhizobacterium *Pseudomonas putida* BIRD-1. *J Bacteriol* **193**: 1290.

Mayfield CI, Williams ST, Ruddick SM & Hatfield HL (1972) Studies on the ecology of actinomycetes in soil IV. Observations on the form and growth of streptomycetes in soil. *Soil Biology and Biochemistry* **4**: 79-91.

McGarvey KM, Queitsch K & Fields S (2012) Wide variation in antibiotic resistance proteins identified by functional metagenomic screening of a soil DNA library. *Applied and Environmental Microbiology* **78**: 1708-1714.

Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E & Breitling R (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res* **39**: W339-346.

Medema MH & Kottmann R & Yilmaz P*, et al.* (2015) Minimum Information about a Biosynthetic Gene cluster. *Nat Chem Biol* **11**: 625-631.

Metcalfe AC, Krsek M, Gooday GW, Prosser JI & Wellington EMH (2002) Molecular analysis of a bacterial chitinolytic community in an upland pasture. *Applied and Environmental Microbiology* **68**: 5042-5050.

Metzker ML (2010) Sequencing technologies - the next generation. *Nat Rev Genet* **11**: 31-46.

Meyer F, Paarmann D, D'Souza M*, et al.* (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics* **9**: 386.

Meyers PR, Goodwin CM, Bennett JA, Aken BL, Price CE & van Rooyen JM (2004) *Streptomyces africanus* sp. nov., a novel streptomycete with blue aerial mycelium. *Int J Syst Evol Microbiol* **54**: 1531-1535.

Micallef ML, D'Agostino PM, Sharma D, Viswanathan R & Moffitt MC (2015) Genome mining for natural product biosynthetic gene clusters in the Subsection V cyanobacteria. *BMC Genomics* **16**: 669.

Mitchell A, Bucchini F, Cochrane G*, et al.* (2016) EBI metagenomics in 2016 - an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Research* **44**: D595-D603.

Moré MI, Herrick JB, Silva MC, Ghiorse WC & Madsen EL (1994) Quantitative cell lysis of indigenous microorganisms and rapid extraction of microbial DNA from sediment. *Applied and Environmental Microbiology* **60**: 1572-1580.

Morris NZ, Standage S & Wellington EMH (1999) Extraction of DNA from spores in soil. *Molecular Microbial Ecology Manual,*(van Elsas JD, Akkermans A & de Bruijn FJ, eds.), p.^pp. Kluwer, Dordrecht.

Moy TI, Conery AL, Larkins-Ford J, Wu G, Mazitschek R, Casadei G, Lewis K, Carpenter AE & Ausubel FM (2009) High-throughput screen for novel antimicrobials using a whole animal infection model. *ACS Chem Biol* **4**: 527-533.

Muyzer G, de Waal EC & Uitterlinden AG (1993) Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl Environ Microbiol* **59**: 695-700.

Nandi S, Maurer JJ, Hofacre C & Summers AO (2004) Gram-positive bacteria are a major reservoir of Class 1 antibiotic resistance integrons in poultry litter. *Proc Natl Acad Sci U S A* **101**: 7118-7122.

Nichols D, Cahoon N, Trakhtenberg EM, Pham L, Mehta A, Belanger A, Kanigan T, Lewis K & Epstein SS (2010) Use of ichip for high-throughput in situ cultivation of "uncultivable" microbial species. *Appl Environ Microbiol* **76**: 2445-2450.

Nunes da Rocha U, Plugge CM, George I, van Elsas JD & van Overbeek LS (2013) The rhizosphere selects for particular groups of Acidobacteria and Verrucomicrobia. *PLOS ONE* **8**: e82443.

Nybond S, Karp M & Tammela P (2013) Antimicrobial assay optimization and validation for HTS in 384-well format using a bioluminescent *E. coli* K-12 strain. *European Journal of Pharmaceutical Sciences* **49**: 782-789.

O'Neill J (2016) Tackling drug-resistant infections globally: final report and recommendations. p.^pp.

Ogram A, Sayler GS, Gustin D & Lewis RJ (1988) DNA adsorption to soils and sediments. *Environmental science & technology* **22**: 982-984.

Olander LP & Vitousek PM (2000) Regulation of soil phosphatase and chitinase activityby N and P availability. *Biogeochemistry* **49**: 175-191.

Owen JG, Reddy BV, Ternei MA, Charlop-Powers Z, Calle PY, Kim JH & Brady SF (2013) Mapping gene clusters within arrayed metagenomic libraries to expand the structural diversity of biomedically relevant natural products. *Proc Natl Acad Sci U S A* **110**: 11797-11802.

Paget E, Monrozier LJ & Simonet P (1992) Adsorption of DNA on clay minerals: protection against DNaseI and influence on gene transfer. *FEMS Microbiology Letters* **97**: 31-39.

Paget MS, Chamberlin L, Atrih A, Foster SJ & Buttner MJ (1999) Evidence that the extracytoplasmic function sigma factor sigmaE is required for normal cell wall structure in *Streptomyces coelicolor A3(2)*. *J Bacteriol* **181**: 204-211.

Papenfort K & Bassler BL (2016) Quorum sensing signal-response systems in Gram-negative bacteria. *Nat Rev Microbiol* **14**: 576-588.

Parsley LC, Linneman J, Goode AM, Becklund K, George I, Goodman RM, Lopanik NB & Liles MR (2011) Polyketide synthase pathways identified from a metagenomic library are derived from soil *Acidobacteria*. *FEMS Microbiol Ecol* **78**: 176-187.

Patzer SI & Braun V (2010) Gene cluster involved in the biosynthesis of griseobactin, a catechol-peptide siderophore of *Streptomyces sp*. ATCC 700974. *Journal of Bacteriology* **192**: 426-435.

Pearce DA, Newsham KK, Thorne MA, Calvo-Bado L, Krsek M, Laskaris P, Hodson A & Wellington EM (2012) Metagenomic analysis of a southern maritime antarctic soil. *Front Microbiol* **3**: 403.

Picard C, Ponsonnet C, Paget E, Nesme X & Simonet P (1992) Detection and enumeration of bacteria in soil by direct DNA extraction and polymerase chain reaction. *Applied and Environmental Microbiology* **58**: 2717-2722.

Pozzi R, Simone M, Mazzetti C, Maffioli S, Monciardini P, Cavaletti L, Bamonte R, Sosio M & Donadio S (2011) The genus *Actinoallomurus* and some of its metabolites. *J Antibiot* **64**: 133-139.

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP & Gu Y (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**: 341.

Rajamuthiah R, Fuchs BB, Jayamani E, Kim Y, Larkins-Ford J, Conery A, Ausubel FM & Mylonakis E (2014) Whole animal automated platform for drug discovery against multi-drug resistant *Staphylococcus aureus*. *PLoS One* **9**: e89189.

Ramette A & Tiedje JM (2007) Biogeography: an emerging cornerstone for understanding prokaryotic diversity, ecology, and evolution. *Microbial Ecology* **53**: 197-207.

Ranjan K, Paula FS, Mueller RC, Jesus EaC, Cenciani K, Bohannan BJ, Nüsslein K & Rodrigues JL (2015) Forest-to-pasture conversion increases the diversity of the phylum Verrucomicrobia in Amazon rainforest soils. *Front Microbiol* **6**: 779.

Reddy BV, Kallifidas D, Kim JH, Charlop-Powers Z, Feng Z & Brady SF (2012) Natural product biosynthetic gene diversity in geographically distinct soil microbiomes. *Appl Environ Microbiol* **78**: 3744-3752.

Reddy Boojala Vijay B, Milshteyn A, Charlop-Powers Z & Brady Sean F (2014) eSNaPD: a versatile, web-based bioinformatics platform for surveying and mining natural product biosynthetic diversity from metagenomes. *Chemistry & Biology* **21**: 1023-1033.

Robe P, Nalin R, Capellano C, Vogel TM & Simonet P (2003) Extraction of DNA from soil. *European Journal of Soil Biology* **39**: 183-190.

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. *Appl Environ Microbiol* **66**: 2541-2547.

Rosconi F, Davyt D, Martinez V, Martinez M, Abin-Carriquiry JA, Zane H, Butler A, de Souza EM & Fabiano E (2013) Identification and structural characterization of serobactins, a suite of lipopeptide siderophores produced by the grass endophyte *Herbaspirillum seropedicae*. *Environ Microbiol* **15**: 916-927.

Rose TM, Henikoff JG & Henikoff S (2003) CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. *Nucleic Acids Res* **31**: 3763-3766.

Rungin S, Indananda C, Suttiviriya P, Kruasuwan W, Jaemsaeng R & Thamchaipenet A (2012) Plant growth enhancing effects by a siderophore-producing

endophytic streptomycete isolated from a Thai jasmine rice plant (*Oryza sativa* L. cv. KDML105). *Antonie Van Leeuwenhoek* **102**: 463-472.

Rusch DB, Halpern AL, Sutton G*, et al.* (2007) The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS biology* **5**: e77.

Sangwan P, Kovac S, Davis KE, Sait M & Janssen PH (2005) Detection and cultivation of soil verrucomicrobia. *Appl Environ Microbiol* **71**: 8402-8410.

Schafer A, Tauch A, Jager W, Kalinowski J, Thierbach G & Puhler A (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.

Schatz A & Waksman SA (1944) Effect of streptomycin and other antibiotic substances upon *Mycobacterium tuberculosis* and related organisms. *Experimental Biology and Medicine* **57**: 244-248.

Scherlach K & Hertweck C (2009) Triggering cryptic natural product biosynthesis in microorganisms. *Organic & Biomolecular Chemistry* **7**: 1753-1760.

Schirmer A, Gadkari R, Reeves CD, Ibrahim F, DeLong EF & Hutchinson CR (2005) Metagenomic analysis reveals diverse polyketide synthase gene clusters in microorganisms associated with the marine sponge *Discodermia dissoluta*. *Applied and Environmental Microbiology* **71**: 4840-4849.

Schrader C, Schielke A, Ellerbroek L & Johne R (2012) PCR inhibitors - occurrence, properties and removal. *J Appl Microbiol* **113**: 1014-1026.

Seemann T (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*.

Seipke RF, Barke J, Heavens D, Yu DW & Hutchings MI (2013) Analysis of the bacterial communities associated with two ant–plant symbioses. *MicrobiologyOpen* **2**: 276-283.

Selama O, Amos GC, Djenane Z, Borsetto C, Laidi RF, Porter D, Nateche F, Wellington EM & Hacene H (2014) Screening for genes coding for putative antitumor compounds, antimicrobial and enzymatic activities from haloalkalitolerant and haloalkaliphilic bacteria strains of Algerian Sahara soils. *BioMed research international* **2014**: 317524.

Selenska S & Klingmuller W (1991) DNA recovery and direct detection of Tn5 sequences from soil. *Lett Appl Microbiol* **13**: 21-24.

Shank EA & Kolter R (2009) New developments in microbial interspecies signaling. *Current Opinion in Microbiology* **12**: 205-214.

Shen B (2003) Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Curr Opin Chem Biol* **7**: 285-295.

Sheridan C (2005) EPO neem patent revocation revives biopiracy debate. *Nat Biotech* **23**: 511-512.

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

Skinnider MA, Dejong CA, Rees PN, Johnston CW, Li H, Webster Andrew LH, Wyatt MA & Magarvey NA (2015) Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). *Nucleic Acids Research* **43**: 9645-9662.

Sosio M, Giusino F, Cappellano C, Bossi E, Puglia AM & Donadio S (2000) Artificial chromosomes for antibiotic-producing actinomycetes. *Nat Biotechnol* **18**: 343-345.

Sosio M, Kloosterman H, Bianchi A, de Vreugd P, Dijkhuizen L & Donadio S (2004) Organization of the teicoplanin gene cluster in *Actinoplanes teichomyceticus*. *Microbiology* **150**: 95-102.

Souza CP, Almeida BC, Colwell RR & Rivera ING (2011) The importance of chitin in the marine environment. *Marine Biotechnology* **13**: 823.

Srinivasan VB, Singh BB, Priyadarshi N, Chauhan NK & Rajamohan G (2014) Role of novel multidrug efflux pump involved in drug resistance in *Klebsiella pneumoniae*. *PLoS ONE* **9**: e96288.

Steffan RJ, Goksøyr J, Bej AK & Atlas RM (1988) Recovery of DNA from soils and sediments. *Applied and Environmental Microbiology* **54**: 2908-2915.

Stein JL, Marsh TL, Wu KY, Shizuya H & DeLong EF (1996) Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *Journal of Bacteriology* **178**: 591-599.

Stevenson BS, Eichorst SA, Wertz JT, Schmidt TM & Breznak JA (2004) New strategies for cultivation and detection of previously uncultured microbes. *Appl Environ Microbiol* **70**: 4748-4755.

Stewart EJ (2012) Growing unculturable bacteria. *J Bacteriol* **194**: 4151-4160.

Stormo KE & Crawford RL (1992) Preparation of encapsulated microbial cells for environmental applications. *Appl Environ Microbiol* **58**: 727-730.

Strieker M, Tanović A & Marahiel MA (2010) Nonribosomal peptide synthetases: structures and dynamics. *Curr Opin Struct Biol* **20**: 234-240.

Struvay C & Feller G (2012) Optimization to Low Temperature Activity in Psychrophilic Enzymes. *International journal of molecular sciences* **13**: 11643-11665.

Suzuki MT, Taylor LT & DeLong EF (2000) Quantitative analysis of small-subunit rRNA genes in mixed microbial populations via 5′-nuclease assays. *Applied and Environmental Microbiology* **66**: 4605-4614.

Tan GY, Robinson S, Lacey E & Goodfellow M (2006) *Amycolatopsis australiensis* sp. nov., an actinomycete isolated from arid soils. *Int J Syst Evol Microbiol* **56**: 2297-2301.

Tebbe CC & Vahjen W (1993) Interference of humic acids and DNA extracted directly from soil in detection and transformation of recombinant DNA from bacteria and a yeast. *Applied and Environmental Microbiology* **59**: 2657-2665.

Thaker MN, Waglechner N & Wright GD (2014) Antibiotic resistance–mediated isolation of scaffold-specific natural product producers. *Nat Protocols* **9**: 1469-1479.

Tocchetti A, Bordoni R, Gallo G*, et al.* (2015) A genomic, transcriptomic and proteomic look at the GE2270 producer *Planobispora rosea*, an uncommon Actinomycete. *PLoS One* **10**: e0133705.

Torsvik V, Goksoyr J & Daae FL (1990) High diversity in DNA of soil bacteria. *Appl Environ Microbiol* **56**: 782-787.

Torsvik V, Sørheim R & Goksøyr J (1996) Total bacterial diversity in soil and sediment communities—A review. *Journal of Industrial Microbiology & Biotechnology* **17**: 170-178.

Torsvik V, Daae FL, Sandaa RA & Ovreas L (1998) Novel techniques for analysing microbial diversity in natural and perturbed environments. *J Biotechnol* **64**: 53-62.

Tracanna V, de Jong A, Medema MH & Kuipers OP (2017) Mining prokaryotes for antimicrobial compounds: from diversity to function. *FEMS microbiology reviews*.

Traxler MF, Seyedsayamdost MR, Clardy J & Kolter R (2012) Interspecies modulation of bacterial development through iron competition and siderophore piracy. *Molecular Microbiology* **86**: 628-644.

Tseng M, Yang SF, Li WJ & Jiang CL (2006) *Amycolatopsis taiwanensis* sp. nov., from soil. *Int J Syst Evol Microbiol* **56**: 1811-1815.

Turpin PE, Maycroft KA, Rowlands CL & Wellington EM (1993) An ion-exchange based extraction method for the detection of salmonellas in soil. *J Appl Bacteriol* **74**: 181-190.

UN (1992) Convention on Biological Diversity. p.^pp.

van der Meij A, Worsley SF, Hutchings MI & van Wezel GP (2017) Chemical ecology of antibiotic production by actinomycetes. *FEMS microbiology reviews* **41**: 392-416.

van Dijk EL, Auger H, Jaszczyszyn Y & Thermes C (2014) Ten years of next-generation sequencing technology. *Trends in Genetics* **30**: 418-426.

Vester JK, Glaring MA & Stougaard P (2015) Improved cultivation and metagenomics as new tools for bioprospecting in cold environments. *Extremophiles* **19**: 17-29.

Vinje H, Liland KH, Almoy T & Snipen L (2015) Comparing K-mer based methods for improved classification of 16S sequences. *BMC bioinformatics* **16**: 205.

Vionis AP, Niemeyer F, Karagouni AD & Schrempf H (1996) Production and processing of a 59-kilodalton exochitinase during growth of *Streptomyces lividans* carrying pCHIO12 in soil microcosms amended with crab or fungal chitin. *Appl Environ Microbiol* **62**: 1774-1780.

Visca P, Imperi F & Lamont IL (2007) Pyoverdine siderophores: from biogenesis to biosignificance. *Trends in microbiology* **15**: 22-30.

Vogel TM, Simonet P, Jansson JK, Hirsch PR, Tiedje JM, van Elsas JD, Bailey MJ, Nalin R & Philippot L (2009) TerraGenome: a consortium for the sequencing of a soil metagenome. *Nat Rev Micro* **7**: 252-252.

Vos M, Wolf AB, Jennings SJ & Kowalchuk GA (2013) Micro-scale determinants of bacterial diversity in soil. *FEMS microbiology reviews* **37**: 936-954.

Wagner M & Horn M (2006) The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. *Current Opinion in Biotechnology* **17**: 241-249.

Wain J, Hoa NTT, Chinh NT*, et al.* (1997) Quinolone-Resistant Salmonella typhi in Viet Nam: Molecular Basis of Resistance and Clinical Response to Treatment. *Clinical Infectious Diseases* **25**: 1404-1410.

Waksman SA, Reilly HC & Harris DA (1948) *Streptomyces griseus* (Krainsky) Waksman and Henrici. *J Bacteriol* **56**: 259-269.

Wall DH, Bardgett RD & Kelly E (2010) Biodiversity in the dark. *Nature Geosci* **3**: 297-298.

Wang G-Y-S, Graziani E, Waters B, Pan W, Li X, McDermott J, Meurer G, Saxena G, Andersen RJ & Davies J (2000) Novel natural products from soil DNA libraries in a *Streptomycete* host. *Organic Letters* **2**: 2401-2404.

Wang H, Fewer DP, Holm L, Rouhiainen L & Sivonen K (2014) Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. *Proceedings of the National Academy of Sciences of the United States of America* **111**: 9259-9264.

Wang J, Soisson SM, Young K*, et al.* (2006) Platensimycin is a selective FabF inhibitor with potent antibiotic properties. *Nature* **441**: 358-361.

Wang Y, Yang Q & Wang Z (2014) The evolution of nanopore sequencing. *Front Genet* **5**: 449.

Watve MG, Tickoo R, Jog MM & Bhole BD (2001) How many antibiotics are produced by the genus *Streptomyces*? *Archives of Microbiology* **176**: 386-390.

Weber T, Blin K, Duddela S*, et al.* (2015) antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Research* **43**: W237-W243.

Weissman KJ (2014) The structural biology of biosynthetic megaenzymes. *Nat Chem Biol* **11**: 660-670.

Weller DM, Raaijmakers JM, Gardener BB & Thomashow LS (2002) Microbial populations responsible for specific soil suppressiveness to plant pathogens. *Annu Rev Phytopathol* **40**: 309-348.

Whitt J, Shipley SM, Newman DJ & Zuck KM (2014) Tetramic acid analogues produced by coculture of *Saccharopolyspora erythraea* with *Fusarium pallidoroseum*. *Journal of Natural Products* **77**: 173-177.

WHO (2014) *Antimicrobial resistance: global report on surveillance 2014*.

Wiese J, Thiel V, Nagel K, Staufenberger T & Imhoff JF (2009) Diversity of antibiotic-active bacteria associated with the brown alga *Laminaria saccharina* from the Baltic Sea. *Mar Biotechnol (NY)* **11**: 287-300.

Wright GD & Poinar H (2012) Antibiotic resistance is ancient: implications for drug discovery. *Trends in microbiology* **20**: 157-159.

Yao H, Campbell CD, Chapman SJ, Freitag TE, Nicol GW & Singh BK (2013) Multi-factorial drivers of ammonia oxidizer communities: evidence from a national soil survey. *Environmental Microbiology* **15**: 2545-2556.

Yin C, Hulbert SH, Schroeder KL, Mavrodi O, Mavrodi D, Dhingra A, Schillinger WF & Paulitz TC (2013) Role of bacterial communities in the natural suppression of *Rhizoctonia solani* bare patch disease of wheat (*Triticum aestivum L.*). *Applied and Environmental Microbiology* **79**: 7428-7438.

Yong D, Toleman MA, Giske CG, Cho HS, Sundman K, Lee K & Walsh TR (2009) Characterization of a new metallo-beta-lactamase gene, *bla(NDM-1)*, and a novel erythromycin esterase gene carried on a unique genetic structure in *Klebsiella pneumoniae* sequence type 14 from India. *Antimicrob Agents Chemother* **53**: 5046-5054.

Young K, Jayasuriya H, Ondeyka JG*, et al.* (2006) Discovery of FabH/FabF Inhibitors from natural products. *Antimicrobial Agents and Chemotherapy* **50**: 519-526.

Zengler K, Toledo G, Rappe M, Elkins J, Mathur EJ, Short JM & Keller M (2002) Cultivating the uncultured. *Proc Natl Acad Sci U S A* **99**: 15681-15686.

Zhang H, Wang Y, Wu J, Skalina K & Pfeifer BA (2010) Complete biosynthesis of erythromycin A and designed analogs using *E. coli* as a heterologous host. *Chemistry & Biology* **17**: 1232-1240.

Zhou J, Bruns MA & Tiedje JM (1996) DNA recovery from soils of diverse composition. *Applied and Environmental Microbiology* **62**: 316-322.

Ziemert N, Podell S, Penn K, Badger JH, Allen E & Jensen PR (2012) The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS One* **7**: e34064.