An Investigation into the Biosynthesis of Proximicins

Pollyanna E. J. Moreland

Ph.D. Thesis

Molecular Microbiology, School of Biology

February 2018

Abstract

The proximicins are a family of three compounds – A-C – produced by two marine Actinomycete *Verrucosispora* strains – *V.* maris AB18-032 and *V.* sp. str. 37 - and are characterised by the presence of 2,4-disubstituted furan rings. Proximicins demonstrate cell-arresting and antimicrobial ability, making them interesting leads for clinical drug development. Proximicin research has been largely overshadowed by other *Verrucosispora* strain secondary metabolites (SM), and despite the publication of the *V. maris* AB18-032 draft, the enzymatic machinery responsible for their production has not been established. It has been noted in related research into a pyrrole-containing homolog – congocidine –due to the structural similarity exhibited, proximicins likely utilise a similar biosynthetic route.

The initial aim of this research was to confirm the presumed pathway to proximicin biosynthesis. Following the sequencing, assembly and annotation of the second proximicin producer, *Verrucosispora* sp. str. MG37, and genome mining of *V. maris* AB18-032, no common clusters mimicked that of congocidine, casting doubt on the previously assumed analogous biosynthetic routes. A putative proximicin biosynthesis (*ppb*) cluster was identified, containing non-ribosomal peptide synthetase (NRPS) enzymes, exhibiting some homology with congocidine. NRPS-systems represent a network of interacting proteins, which act as a SM assembly line: crucially, adenylation (A)- domain enzymes act as the 'gate-keeper', determining which precursors are included into the elongating peptide. To elucidate the route to proximicins, activity characterisation of the four A-domains present in *ppb* cluster was attempted. The A-domain Ppb120 was shown to possess novel activity, demonstrating a high promiscuity towards heterocycle containing precursors, in addition to the absence of an apparent essential domain. This discovery refutes previous work outlining the core residues which dictate A-domain activity, while also presenting a facile route to novel heterocycle-containing compounds. Despite extensive work, A-domains *ppb195* and *ppb210*, were ineffectively purified in the active form – informing future work into A-domains activity characterisation. Finally, the *ppb220* A-domain which lies at the border of *ppb*, was inactive suggesting over-estimation of the cluster margins. To confirm *ppb220* redundancy and confirm *ppb* boundaries, CRISPR/Cas gene editing studies were done. The gene responsible for the orange pigment of *Verrucosispora* strains was initially targeted and successfully deleted, and *ppb* studies commenced.

The research here refutes the previously presumed route to proximicin biosynthesis; the *ppb* cluster instead comprises enzymes exhibiting unique activity and structure. The findings represent the foundations for allowing exploitation of chemistry exhibited within the proximicin family. The novelty exhibited can be utilised in the search for antimicrobial clinical leads, by allowing the production of compounds containing previously inaccessible heterocycle chemistry.

Acknowledgments

First and foremost, I would like to thank my supervisor Dr. Jem Stach. I am extremely grateful for his contributions of time, ideas and support to make my Ph.D. experience productive and exciting. The opportunity to work and learn from someone with such expertise in the field of molecular biology, as well as enthusiasm for the subject, has been invaluable.

The technical support – Ros Brown, Miriam Earnshaw and Dr. Matthew Peak – have contributed immensely to my personal and professional time at Newcastle. They have been a source of great advice and guidance throughout my Ph.D., as well as friendships and support. I would also like to thank Dr. Jon Marles-Wright for help with protein crystallization, and my secondary supervisor, Dr. Michael Hall for synthetic chemistry guidance.

Members of the Sylvie Garneau-Tsodikova Group, University of Kentucky, I thank for the immense help with adenylation domain purification, as well as radioactive assays. I also gratefully acknowledging my funding source – BBSRC- that made my Ph.D. work possible.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

A-domain	Adenylation domain
A#	Adenylation domain motif of number #
ACP	Acyl-carrier protein
BP	Base pairs
C-domain	Condensation domain
Co-A	Co-enzyme A
CRISPR	Clustered regularly interspaced short palindromic repeats
CS	Carboxyaminoimidazole synthase
CY	Cyclodehydratase
DH	Dehydrogenase
DSB	Double stranded break
El	Elution
FMN	Flavin mononucleotide
FT	Flow through
gDNA	Genomic DNA
HDR	Homology dependent repair
hr	Hour
Kb	Kilo bases
Min.	Minute
MP	Mate-pair
MT	Methyltransferase
NGS	Next generation sequencing
NHEJ	Non-homologous end joining
NP	Natural product
NRP	Non-ribosomal peptide
NRPS	Non-ribosomal peptide synthetase
PAM	Protospacer adjacent motif
PCR	Polymerase chain reaction
PE	Paired End
PKS	Polyketide synthetase
ppb	Putative proximicin biosynthetic cluster/gene
PP_i	Pyrophosphate
QC	Quality control
R	Read
SBC	Secondary biosynthetic cluster
sgRNA	Synthetic guide RNA
SM	Secondary metabolite
SMC	Secondary metabolite cluster
Spp.	Species
SR	Single read
Str.	Strain
T-domain	Thiolation domain

Te-domain	Thioesterase domain
trRNA	Tracer RNA
W	Wash
WGS	Whole genome sequencing

Chapter 1. Introduction to Research

1.1 Introduction

The unprecedented, large scale emergence of antimicrobial resistance, and depletion of antimicrobial leads, means we currently balance on the precipice of the post-antibiotic era (Willyard, 2017). The appearance of multi-drug resistant microbes continues to increase at an alarming rate, yet world-wide prevention programs continue to be of low-priority (WHO, 2017); thus, it is vital that we are united in this war. It has been demonstrated that some of the most effective antibiotics are those identified from microorganisms themselves; exploiting the millennia of evolution driven by microbe-on-microbe warfare. In the recent decades, misdirected efforts preoccupied with financial initiatives has led to a complete lack of novel antibiotic identification (Tegos et al., 2012). I propose that if we are to stall this impending catastrophe, we return to the previously fruitful pursuit of antibiotics of microbial origin, utilizing cheaply available scientific advances to result in a financially attractive research avenue, to propel the identification of a whole collection of novel antibiotic compounds.

1.2 A Brief History of Drug Discovery

The discovery and wide scale exploitation of antimicrobial compounds has been the singular most influential medicinal discovery of the twentieth century; revolutionizing healthcare, allowing treatment of infections that were once largely fatal and safeguarding procedures once unthinkable. The development of penicillin represented a tide-change in how life threatening infections were treated; due to the subsequent synthesis of other microbial agents, vaccines and antiseptics, victory against pathogens was declared (Sigerist, 1971). Convinced the arsenal of antimicrobial agents was well stocked, by the 1980's research emphasis was poised to shift to other clinically relevant problems such as cancer, diabetes and heart disease, effectively closing the book on infectious diseases. However, after the introduction in the early 1940's of the proclaimed 'miracle drug', penicillin-resistant *Staphylococcus* spp. were widespread by 1944 (Livermore, 2004). This was followed

by progressive ineffectiveness exhibited by every compound deployed in the fight against infectious disease; pathogenic microorganisms demonstrate a defiantly impressive ability to adapt and develop resistance. The optimistic mindset towards microbial control turned to skepticism as the chain of outbreaks and epidemics of new, re-emerging and resistant infections continued to appear. Giving rise to the term ‘super bugs’, these organisms dominated the scientific literature, commanding the attention of academics, governments, public health officials and the general public, alike. Focus shifted to quickly identifying new compounds to circumvent the requirement of those obsolete; the ‘Golden Era’ of antibiotic drug discovery ensued.

The Golden Age of antimicrobial drug discovery was characterized by large scale screening of fermentation broths and extracts from microorganisms, with the aim of simply inhibiting the growth of bacterial organisms of interest, with little regard for the mechanism by which it exerts this action. This worked efficiently for a period, resulting in the discovery of the most common natural product antimicrobials (Fig. 1); however, issues ensued regarding high rediscovery rates and in the 1960s methods of de-replication were developed. Traditional screening methods were modified to limit hits to detect inhibitors of a specific pathway, this lay the basis of what we consider today as target based screening. These focused pathway-screens, resulted in the discovery of clinically useful classes such as carbapenems and fosfomycin, shown to target the bacterial cell wall (Silver, 2012). It was suggested towards the late 1970’s that all the low hanging fruit had been picked (Silver, 2012), and efforts were directed to increasingly intelligent screening designs, utilizing the growing ability to clone genes and manipulate bacterial strains, converting previously rudimentary approaches to enzyme specific target-directed screens (Kraus, 2008). As a result, a bottle neck toward chemical novelty began to form. Interest was awarded to compounds with identifiable targets with established amenability to downstream biochemical and physical analysis; this meant the perceived attractiveness of compounds with novel cellular targets was small, and they were hence given low priority. The diversity of natural compounds was limited further with steps being taken away from expensive crude microbial fermentation studies, with focus shifting to screening large libraries of compounds, especially those of combinatorial chemical origin (Demain, 1999; Kingston & Newman 2002). This concept of synthetically tailoring a small group of scaffolds to yield

antimicrobials without resistance, as a drug discovery approach, was largely fruitless. However, the emergence of resistance and the propulsion of research specifically aimed at its avoidance only extended the futile implementation of this approach. Specific cellular targets would be selected by the relative ease a microbe could develop the ability to circumvent its necessity, and then large scale high throughput screens would be employed to find compounds which specifically inhibit this target. Much of the chemical diversity tested were of synthetic origin, groups of modified molecules with the aim of increasing target selectivity, potency and lesser off-target effect. This success of this approach was limited and marked the beginning of what is considered ‘the discovery void’ (Fig. 1) (Fischbach & Walsh, 2009), which is typically greatly understated; some of the most recently registered representatives of novel drug classes were first reported in the latter half of the twentieth century but disregarded as poor leads, hence, to consider these as recent novelties, is tenuous at best (Silver, 2009). Despite creative, rational and technologically cutting edge screening approaches, innovation has failed to result in sufficient biologically active compounds. The current situation has been only exacerbated by the resilience of microbes, quickly acquiring resistance to the few compounds remaining in our repertoire; the full extent this threat poses to the societies of today is being modestly related to the mases, the crisis point looming being fundamentally dismissed (WHO, 2017).

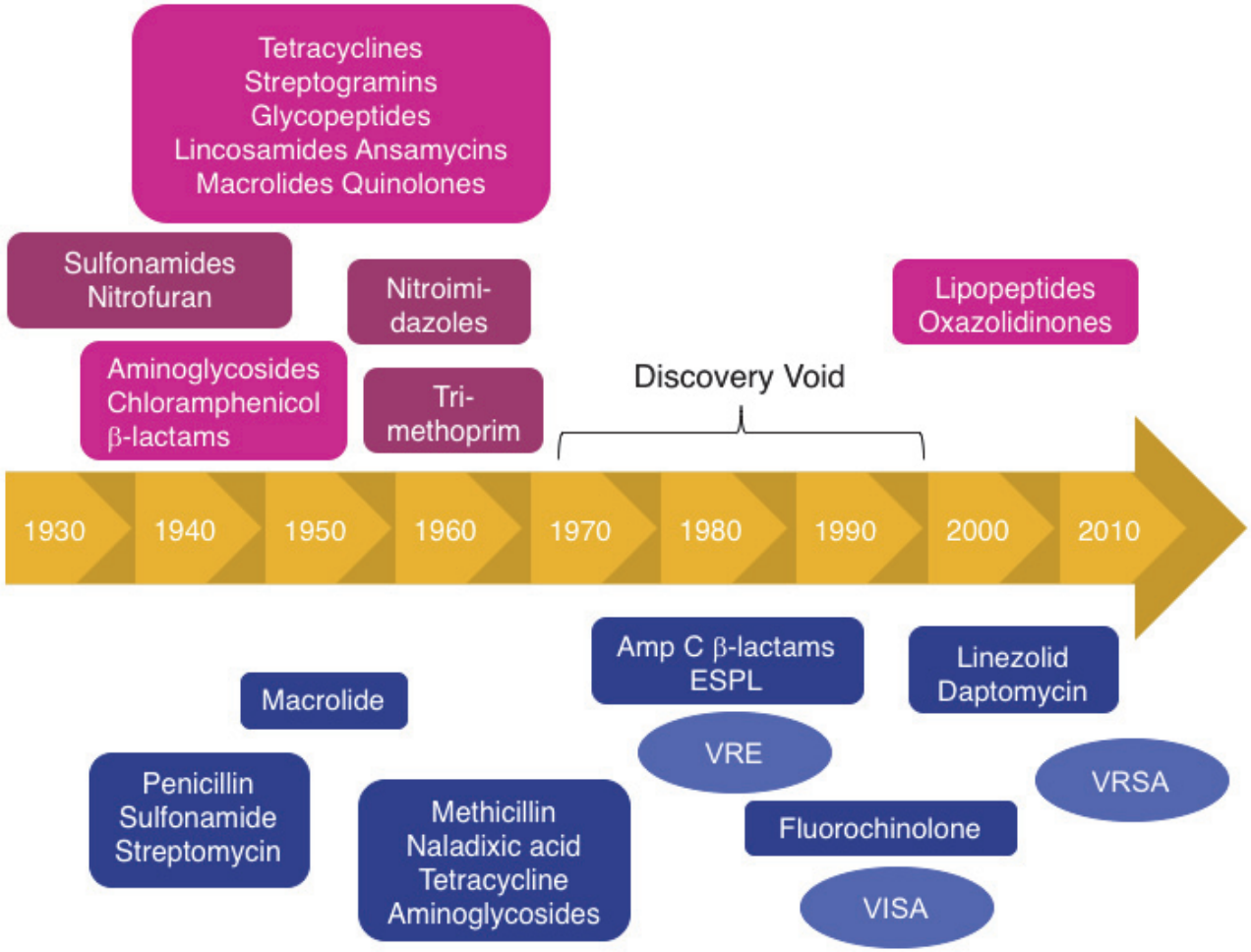


Figure 1. History of antibiotic drug discovery and the occurrence of resistance. An overview of the evolutionary arms race between our ability to produce novel compounds and microbes acquiring immunity. Synthetic antibiotics are in dark pink, with naturally arising compounds in pink. Dark blue denotes resistance to the specified antibiotic, and blue ovals denote the appearance of clinically important multi-resistant pathogens. Where the boxes span across the timelines denote the initial report of compounds or drug resistant microbes. VRE = Vancomycin Resistant Enterococci, VISA = Vancomycin intermediate *Staphylococcus aureus* and VRSA = Vancomycin resistant *S. aureus*

1.3 Drug Resistance: the emerging threat

It is estimated in the United Kingdom in the year of 2017, more people will die as a result of microbial infection causing sepsis, than the two leading cancers combined (Antibiotic Research UK, 2017). And yet, as cancer research continues to steadily decrease mortality associated with the disease, drug discovery pipelines remain eerily empty. When identifying emerging threats to public health, four main classes are typically highlighted (Tegos et al., 2012; WHO,2017):

1. Methicillin resistant *Staphylococcus aureus* (MRSA)
2. Multi-drug resistant (MDR) Gram-negative bacteria
3. MDR and extensively drug-resistant *Mycobacterium tuberculosis* (TB)
4. *Candida* species

The major four classes are such as they result in the most deaths annually, due to typically affecting people with already compromised health, including the young, old, malnourished and immune-comprised (Tegos et al., 2012). The exact extent of mortality contributed to all resistant pathogens is undisclosed, Fig. 2 shows the numbers for 2015 combined from various sources reporting on singular infections. Each group is associated with specific characteristics facilitating their persistent pathogenicity, a niche they occupy in which they thrive; hindering their successful eradication. MRSA represents the only resistance level which has been steadily declining since the 1980's, however, it remains such a threat as it increases the likelihood of vancomycin resistant *S. aureus,* increasing the difficulty associated with treating the infection (Chavers et al., 2003; Nordmann et al., 2007). Alternatively, the opportunistic health-care associated infections typical of resistant Gram-Negative bacteria are less prevalent, but are truly untreatable (McGowan, 2006; Dijkshoorn et al., 2007; Nordmann et al., 2007; Baldry, 2010; McKenna, 2011). These organisms have evolved the ability to modify the cellular targets of Gram-Negative specific antibiotics, intensified by the intrinsic difficulty associated with developing novel compounds to treat these pathogens attributed to their outer membrane preventing hydrophilic compound entry (Lee et al., 2000). Multi-Drug Resistant *Mycobacterium tuberculosis* (MDR-TB) poses an enormous threat to developing countries where healthcare access is low, as infections requires extensive treatment which is typically not administered leading to the associated high fatality rate (Dye, 2009;

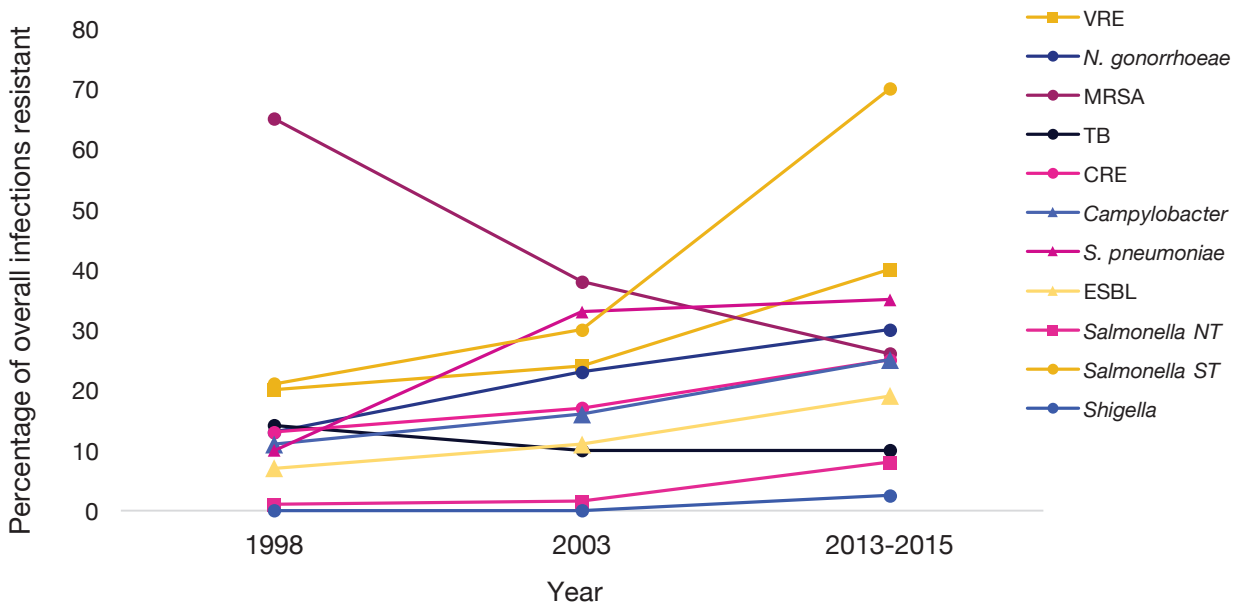


Figure 2. Increasing prevalence of common antibiotic-resistant pathogens. A graph of combined data (WHO, 2017; McGowan, 2006; Dijkshoorn et al., 2007; Nordmann et al., 2007; Baldry, 2010; McKenna, 2011). All but MRSA is steadily increasing. VRE = Vancomycin resistant Enterococci; MRSA = Methicillin resistant S. *aureus;* TB = Resistant *Mycobacterium tuberculosis*; CRE = Carbapenem-resistant Enterobacteriaceae; ESBL = Extended spectrum beta-lactamases incidence; *Salmonella* NT = Non-typhoidal; *Salmonella* ST = Sequence Type

Jassal & Bishai, 2009). Finally, resistant fungal infections are responsible for the most deaths associated of any catheter-related infection with the incredibility high mortality rates of ~15-49% (Crump & Collignon, 2000; Gudlaugsson et al., 2003). Each of these clinically prevalent pathogens pose the possibility of drastically increasing the risk associated with everyday procedures, such as hip replacements and appendectomies, or making them completely impossible, hurling our healthcare into an era distressingly similar to that before wide scale utilization of antimicrobials.

These four groups stand as a conservative estimate of the groups which pose a major threat; a total of 18 were outlined by the US government, in a 2013 report with at least the hazard level: serious, requiring 'prompt and sustained action' (CDC, 2017). It is becoming increasingly apparent that these infections are not only concerning the old, young or already unwell, resistant pathogen are spreading too amongst the healthy; one recently publicized example being MDR *Nesisseria gonorrhoeae*. It is responsible for the second most reported infection across developed countries, with current reports suggesting that as much as 30% of cases are due to pathogens resistant to typical treatment approaches (CDC, 2017). It is

hence not difficult to imagine a time in the immediate future when it will be completely untreatable, as such, *N. gonorrhoeae* has been identified as a global health threat with intensive programs being put in place, poised to address the heavy burden resistant infections will exert on healthcare worldwide (CDC, 2017). Despite the evident threat these pathogenic microbes pose, along with the enormous costs associated with their treatment, they have garnered little re-direction of funds from other research areas; a situation which must be resolved.

The World Health Organization has recently proposed a multifaceted approach of resolution, summarized as (i) novel compound discovery, and (ii) curtailing resistance occurrence (Willyard, 2017). The latter has been extensively studied since its first report in the first half of the twentieth century. Put simply, drug resistance is the natural evolution of pathogens in response to a large selection pressure being

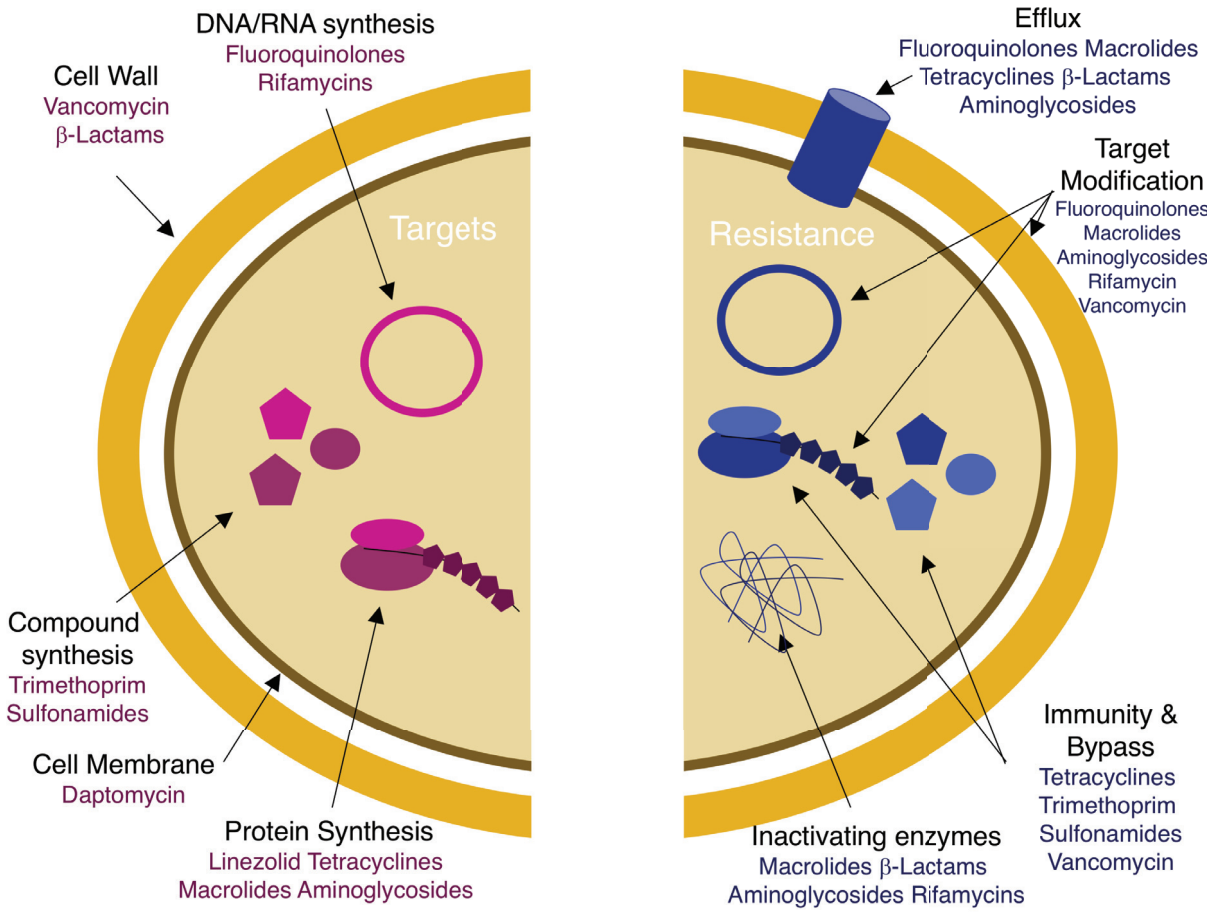


Figure 3. Cellular targets of antibiotics and mechanisms utilised by pathogens for resistance. Antibiotics work by attacking an essential cellular target or process within a pathogenic microorganism, such as DNA synthesis or the cell membrane. Microbes evolve ways to circumvent their effect, primarily through excreting the antibiotic through efflux pumps, but other ways such as modifying the target so that the antibiotic can longer take effect on it, also exist.

exerted, exacerbated by the vast scale implementation, and irresponsible use, of antibiotics. How it arises can be typically categorized into either endogenously in the pathogen by mutation or selection, or exogenously, driven by horizontal gene transmission (HGT) (Bryan, 1989; D'Costa et al., 2006; Martinez et al., 2007). Just as the potential chemical scaffolds for antibiotics seems endless, so does the collective genomic repertoire of possible mechanisms within microbes to render them obsolete. This ability is often termed the antibiotic 'resistome' (D'Costa et al., 2006), and consists of mechanisms outlined in Fig. 3; regardless of cellular target, chemical modification or breakdown of antibiotics, target protection, efflux, or specific changes of the cellular target can prevent their activity (Woodford & Ellington, 2007) (Fig. 3). Crucially, the evolutionary event leading to endogenous resistance mechanism only need occur once, in a singular strain and it can then be propelled by HGT to any other bacterial strain. It is hence no surprise that exogenous resistance is responsible for some of the most clinically important resistance, inconspicuous strains act as reservoirs of resistance genes, acquired over years of challenges from other microbes (Courvalin, 2008; D'Costa et al., 2007; Martinez, 2009). The prompt appearance of MDR microbes, evolving an array of efficient resistance mechanisms, demonstrates that the prospects for long term avoidance of resistance is poor and hence, the need for the large scale discovery of novel chemistry antimicrobials has never been greater. In this current predicament, all recent advances in scientific capability need to be redirected and repurposed with this aim, yielding not only novel compounds, but the refinement of discovery pipelines for optimal exploitation.

1.4 Current Discovery Strategies: an evolving necessity

The onset of the genomic era and the impending gene editing revolution may signify an end to the discovery hiatus, by accelerating the identification of novel compounds from microorganisms. Microbes possess an unparalleled ability to produce a wealth of structurally diverse metabolites which maintain a range of biological activity lending themselves to a wide variety of applications, crucially, in the treatment of infectious disease (Fowler et al., 2006; Demain & Sanchez, 2009; Louie et al., 2011). Compounds of microbial origin, and their semisynthetic derivatives, provide the basis for much of the healthcare of today, and contain

chemical scaffolds far more complex than combinatorial approaches could feasibly yield (Demain & Sanchez, 2009). Hence, it is no surprise that recent research has returned to microorganisms for help tackling the current resistance crisis. This has been propelled by the onset of the genomic era over the last decade, resulting in the wide scale implementation of sequencing technologies; in doing this, one fact became increasingly apparent: microorganisms have far greater potential to produce specialized metabolites, than what was revealed by traditional bioactive screens (Omura et al., 2001; Bentley et al., 2002; Backhage et al., 2011). Whole genome sequencing (WGS) and improved bioinformatical abilities has led to the discovery of many silent or cryptic biosynthetic gene clusters, the products of which, were never realized during prior screens. Demonstrating the extent of this, the soil bacterium *Streptomyces coelicolor* A3(2), an organism which has been extensively studied for more than five decades and regarded as a model antibiotic producing Actinobacteria, was shown to contain the biosynthetic gene clusters (BGCs) for 22 distinct specialized metabolites compared to the previously identified 6 (Challis, 2011). This was a trend not limited to the actinobacteria but also genomes of fungi such as *Aspergillus*, and Gram-Negative bacteria *Pseudomonas* and *Clostridium*, which were also demonstrated to be promising reservoirs of novel secondary metabolite gene clusters (SMC's) (Keller et al., 2005; Donadio et al., 2007; Letzel et al., 2013; Wilson et al., 2014). This has stimulated the development of a range of experimental pipelines for identifying these clusters (Fig. 4), and predicting their products, giving rise to the new field of genomic-driven natural product discovery.

It is tempting to envision a time in which next generation sequencing (NGS)-based approaches would allow the identification of the full complement of putative SMC's within the DNA of any given microorganism. The genomic revolution, leading to the full comprehension of what microbes have to offer, has started a renaissance in antibiotic identification (Zerikly & Challis, 2009; Weber et al., 2015). When novel product discovery began to dwindle towards the end of the 20th century, so did interest in the industry; big pharma redirected funds to areas of research more financially fruitful. This desertion of drug discovery is further exacerbated by the unattractive long term prospects of any identified novel drugs; they are only used

Figure 4. Typical pipeline for genome mining approaches to novel compound discovery. Research into identifying novel biologically active compounds via genome mining typically occurs in the four steps outlined resulting in the discovery of not only novel compounds but also their pathways which aids large scale production.

short term, unlike some of the cancer and HIV drugs and importantly, in the aim to avoid resistance they are not used in such high quantities, primarily as a last resort. This makes the prospect of large scale financial investment less than alluring, and so steps are required to lower the financial burden associated with novel compound generation. The advent, and continual refinement, of next generation sequencing approaches offers an unparalleled potential resolution; they possess the ability to rapidly and cheaply identify multiple potentially profitable gene clusters and lend themselves to high throughput approaches (Metzker, 2010; Gomez-Escribano et al., 2012; Izawa et al., 2013; Schorn et al., 2013). Unlike costs and efforts associated with whole genome sequencing (WGS) at its conception, drastic improvements have greatly accelerated the rate and reduced cost of genomic acquisition, making it commonplace for draft genomes of organisms of interest to be generated and utilized as an investigative tool (Medema et al., 2011; Blin et al., 2013; Blin et al., 2014). The result is that the SMCs, and large-scale production, of novel compounds can be done in a matter of weeks with relatively low associated costs. This has resulted in well-established frameworks for novel compound exploration utilizing genome mining approaches (Fig. 4). Bioinformatical programs, despite lagging behind NGS technologies which continue to surge ahead, possess the ability to identify common biosynthetic cluster types present in bacteria which may contain novel compound chemistry, and hence, potential antimicrobial leads. One area of particular interest, owed to their highly conserved domain structure permitting easy identification and exploitation, are compounds produced via non-ribosomal peptide synthases; these systems are responsible for the biosynthesis of clinically important compounds such as daptomycin, and NGS has allowed the unveiling of the true breadth and potential these systems encompass (Felnagle et al., 2008).

1.5 Importance of Non-Ribosomal Peptide Synthetase Enzymes

Non-ribosomal peptide synthesis represents a complex ribosomal-independent mechanism utilised by microorganisms to produce a large array of chemically and structurally diverse and clinically important secondary metabolites. They represent a network of interacting proteins, which act as an enzymatic assembly line: accepting precursor molecules, catalysing peptide bond formation before shuttling intermediates down the Non-Ribosomal Peptide Synthases (NRPS) system for the next addition. The ability NRPS systems possess to introduce novel chemistry, in a highly predictable and exploitable manner, dictated by their characteristic domain arrangements, make them an alluring target for novel antibiotic compound discovery.

Non-ribosomal peptide synthetase (NRPS) systems are not a recent discovery; Alexander Fleming's detection of penicillin in the late 1920's unwittingly initiated the scrutiny of these large enzyme complexes. Non-ribosomal peptides (NRPs) comprise a group of some of our most important antibacterial, antifungal and anticancer drugs, ensuring their continual garnering of significant interest in areas of drug development. Research has been steadily building on these complex systems for decades; commencing with Mach et al., (1963) in the demonstration that a mechanism independent of the ribosome was responsible for biosynthesis of the cyclic decapeptide, tyrocidine. This work was further confirmed by investigation into gramicidin S and ediene, showing that addition of ribosome inhibitors or RNase enzymes did not abolish their production (Berg et al., 1965; Yukioka et al., 1965; Fujikawa et a., 1966; Borowska et al., 1966; Spaeren et al., 1967; Tomino et al., 1967). As the ability to manipulate DNA progressed, so did the insight into these systems. Gevers et al., (1968) executed an array of elegant experiments, culminating in the exposure of mechanisms responsible; initially utilising partially purified enzymes known to be involved in gramicidin biosynthesis, it was revealed that the ATP-dependent reaction catalysed, occurs in two steps. The first step involves the release of PP_i and the second, the release of AMP, resulting in an amino acid being covalently tethered to the NRPS machinery (Gevers et al., 1968; Kleinkauf et al., 1969). This evidence established the presence of an amino acid 'carrier', furthered by the finding that it occurs as a progressive synthesis of enzyme-linked peptides,

one amino acid added at a time (Froshov et al., 1970; Lipmann et al., 1970). Analysis of the chemical stability of the covalent linkages present, suggested a thioester bond between the NRPS enzyme and amino acid (Gevers et al., 1969). This work was propelled by that being done simultaneously on fatty acid biosynthesis, in which it was established that intermediates were thioesterfied to a 4'-phosphopantetheinyl (4'-Ppant) cofactor (Bremer, 1968). Once confirmed the involvement of 4'-PPant (Gilhuus-Moe et al., 1970; Kleinkauf et al., 1970; Lee et al., 1973), it was initially hypothesised that this was done on a single molecule of 4'-Ppant through a series of thiol exchange reactions. This research led to the initial 'thiotemplate mechanism' for NRPS enzymology, proposed by various groups concurrently (Lipmann et al., 1971; Laland & Zimmer, 1973; Kurahashi, 1974); this system was modified slightly, with the advent of NGS, which revealed that each module of catalytic domains involved the activation of an amino acid, contained its own 4'-Ppant cofactor (Felnagle et al., 1998). This subtle revision led to the 'multiple carrier model' we use today (Schlumbohm et al., 1991; Stein et al., 1994; Stein et al., 1996); each amino acid is activated as an aminoacyl-AMP intermediate and subsequently tethered to their individual 4'-Ppant cofactor as an aminoacylthioester. The transthiolation reactions proposed in the initial mechanism, were made redundant, as biosynthesis proceeds via head-to-tail condensations of the amino acids by transpeptidation reactions. The characteristic modularity of NRPS systems was noted very early in the research, with the observation of ~70-75 kDa of protein existed per amino acid activated (Lee et al., 1973; Lipmann et al., 1973). It is typically noted that there are a single set of enzymatic domains or modules associated with each incorporated amino acid. These repeating modules catalyse the actions of three core domains: the adenylation, thiolation and condensation domains which together comprise a minimal NRPS module. They are typically arranged chronologically, dictating the peptide sequence; ending at the C-terminal end with a thioesterase domain which catalyses the termination of NRP synthesis, marked by the release of the peptide from the enzyme complex. As with any naturally occurring scheme, the multiple carrier model does not encompass the enzymology utilised in all NRPS systems; for the purpose of clarity, however, it is the only one discussed here.

The repeating steps, and distinctive domain nature of NRPS systems, conveys a simplicity in determining the product they are responsible for producing; by understanding the roles of enzymatic machineries present, it allows the route to potentially novel compounds to be established and exploited.

1.5.1 Adenylation domain (A-domain)

Arguably the most important component of the NRPS system, the A-domain acts as the 'gate-keeper' to peptide elongation, determining the chemistry of the next building block added (Haese et al., 1994; Diechmann et al., 1995). They do this by catalysing a two-step, ATP dependent reaction which involves the activation of the carboxylate group of a specific substrate as an aminoacyl-AMP intermediate and subsequent transfer onto the 4'-Ppant arm of the neighbouring thiolation domain (Fig. 6a). Raised as early as 1967, Plant et al., (1967) noted the functional similarities exhibited by NRPS A-domains, with Acyl-CoA synthetases and firefly luciferases. All members of these groups, although catalysing distinct reactions utilise a two-step reaction: first activating a carboxylate substrate by reacting with ATP for form the acyladenylate and inorganic PP_i which provides the energy for the second partial reaction which differs depending on subfamily. This diverse group of adenylating enzymes has since been termed the ANL superfamily, of which NRPS A-domains have recently been classified (Gulick, 2009).

A-domains represent the first level of substrate selectively, and are responsible for the huge chemical diversity exhibited by NRPS as they are not limited to proteogenic amino acids, but have been shown to have activity towards hundreds of substrates (Von Dohren, 1990). Because of this, extensive research efforts have been dedicated to understanding how these enzymes elicit this selectivity, with aim of modifying and exploiting it for novel compound production (Strieker et al., 2010; Thrilway et al., 2012). Traditional work, supported by recent crystal structures of A-domains, either alone or co-crystallised with a known substrate, has led to some fundamental findings. Key residues involved in substrate specificity and catalytic activity have been revealed culminating in the bioinformatical ability to establish an A-domains substrate pool with the amino acid sequence (Challis et al., 2000; Schwarzer et al., 2003; Rausch et al., 2005; Tanovic et al., 2008). The basis of much of what we know is built on innovative works done by Marahiel et al., (1997) and

Shachelhaus et al., (1999) (Fig. 5a), in which ten core sequence motifs were identified (A1-A10) shown to be important for structural, catalytic and substrate selectively roles. The latter of these was furthered by the identification of 10 substrate selectivity conferring motifs, shown to occupy vital positions within the substrate binding pocket located between A4 and A5, whose chemistry dictates the potential substrate pool (Shachelhaus et al., 1999) (Fig. 5b). This information has been supported by numerous gene editing studies in the model A-domain – the phenylalanine activating GrsA involved in gramicidin synthesis, to confirm the role of each of these domains in which substrate selectively has been relaxed or altered entirely (Shachelhaus et al., 1999; Challis et al., 2000). Some of this work is shown in Fig. 5c, in which specificity towards phenylalanine of GrsA was altered to allow the activation of other precursors by mutating single residues (shown in pink) present within the binding pocket defined by Shachelhaus et al., (1999). Further, extensive specificity conferring residues mutational studies have shown some residues are more resistant to change than others; for example, Asp_{235} and Lys_{517} are always invariant, and have been shown to stabilise the α-amino acid group and to interact with the α-carboxylate, present on the substrate, respectively (Shachelhaus et al., 1999). Other positions are very tolerant of variation, leading to relaxation or complete changes in specificity (Shachelhaus et al., 1999). Garnering information regarding the components which determines an A-domains activity is vital in the continual expansion of the pool of potential compounds, possessing biological activity, which can be produced. Understanding the mechanism for catalytic activity exhibited by adenylation domains has permitted the development of many bioinformatical tools which identify substrates by the residues which line the binding pocket (Medema et al., 2011; Rottig et al., 2011). These applications allow the identification of potentially novel BGCs, propelling the detection of novel compounds for medicinal exploitation.

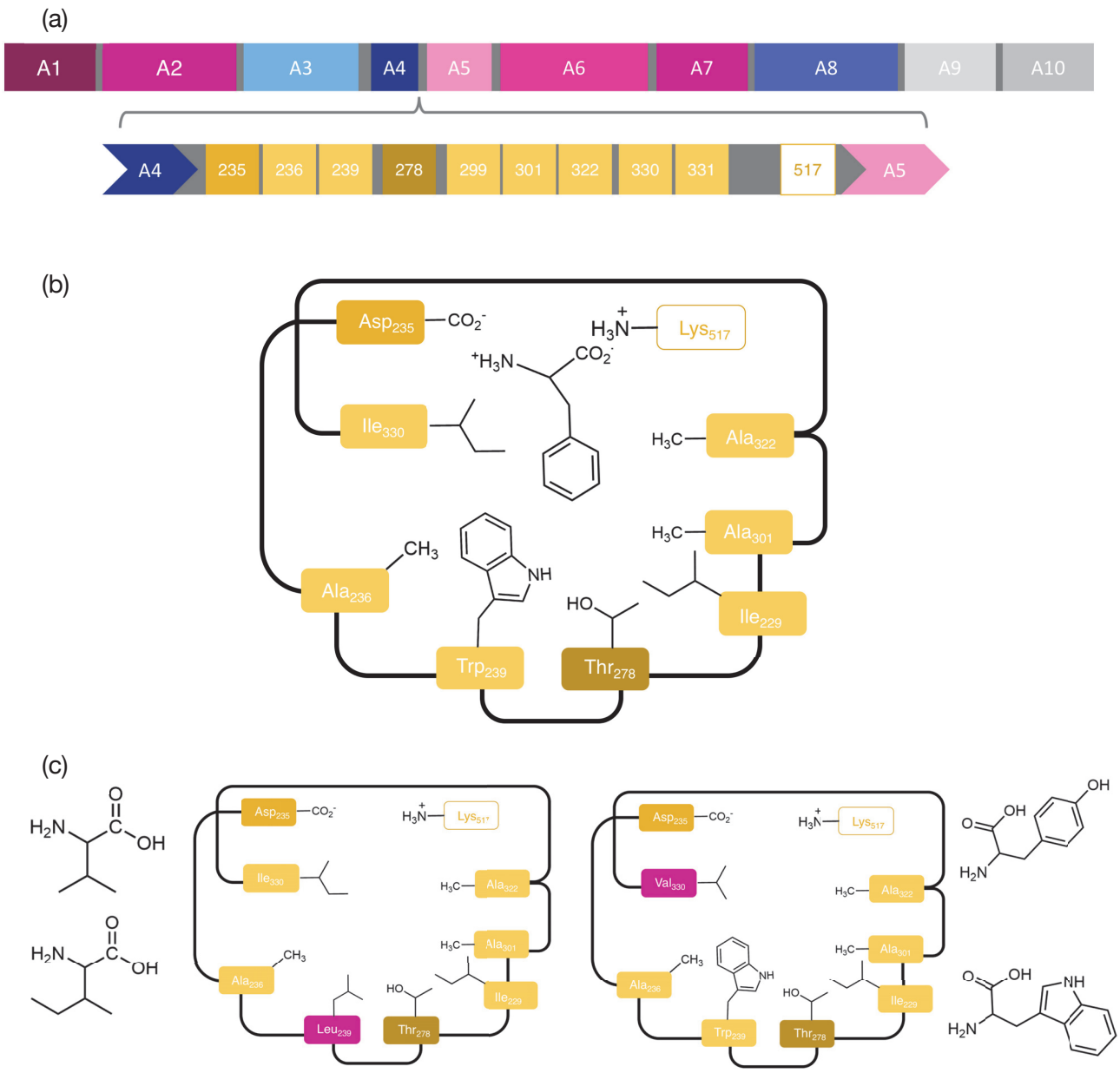


Figure 5. Important Adenylation domain active site residues, and research into their manipulation. (a) A typical NRPS-adenylating domain contains 10 well conserved motifs outlined by Marahiel et al., (1997) demonstrated to be important in structure (pink), catalytic (grey) activity and substrate selection (blue); located between domain A4 and A5 sits the active site of the A-domain which is responsible for substrate binding. Shachelhaus et al., (1999) outlines these 9 core specificity conferring residues shown in yellow. **(b)** the GrsA adenylation domain binding pocket with the phenylalanine substrate present outlined by Shachelhaus et al., (1999). Different shade of yellow denotes the chemical nature of the largest proportion of residues noted at this position: White – Basic amino acids; light yellow: hydrophobic; medium yellow: acidic and dark yellow: hydrophobic/polar. **(c)** the mutations to single amino acids known to sit within the GrsA binding pocket, and the alteration from phenylalanine substrate selectivity to other substrates located to the side of the mutated binding pocket, outlined by Shachelhaus et al., (1999). Binding pockets shown in the style of Challis et al., (2000).

1.5.2 Thiolation domain (T-domain)

T-domains, sometimes referred to as PCP domains, work in conjunction with A-domains once the aminoacyl-AMP intermediate has been formed (Fig. 6b). T-domains and acyl carrier proteins (ACP) involved in fatty acid and polyketide biosynthesis belong to the same superfamily, and share a conserved four-helix bundle structure Weber et al., 2000; Samel et al., 2007). Similar to ACPs, T-domains are required to be post-translationally modified with the addition of a 4'-Ppant group on a conserved seryl residue located at the centre of the T-domain by a 4'-Ppant transferase; this results in the change into its holo-form, and activation of the T-domain (Quadri et al., 1998). The T-domain has been shown to be the location of the aminoacylthioester formation noted by initial researchers in the early 1970's (Lipmann et al., 1971; Laland & Zimmer, 1973; Kurahashi, 1974), as the newly acquired 4'-Ppant arm attacks the carboxyl group of the aminoacyl-AMP intermediate yielding the aminoacylthioester (fig. 6b)

T-domains play a central role in NRP synthesis, as they are responsible for a variety of reactions with many partners, which must be conducted in an exquisitely timed sequence. Understanding how T-domain enzymes achieved this was the focus of much research (Koglin et al., 2006), and it was determined that they can maintain three different conformations and transitioning between these requires significant repositioning of the 4'-Ppant arm; this gave the first detailed evidence of the 4'-Ppant 'swinging arm' motion required for shuttling peptide intermediates down the NRPS assembly line and established a structural basis for the differential interactions exhibited.

1.5.3 Condensation domain (C-domain)

As the name suggests, C-domains are responsible for a condensation reaction, and are typically situated between consecutive pairs of A and T domains – precisely positioned for catalysing peptide bond formation between two tethered substrates. This results in the linkage of the upstream and downstream building blocks which surround them. The currently accepted model of C-domain activity suggests they possess two distinct substrate binding sites: one binding the preceding and the other, the proceeding amino acid, denoted the accepter and donor site, respectively (Fig. 6c). As there are only two crystal structures solved of C-domains (Keating et al.,

2002; Samel et al., 2007), neither with the substrate bound, the mechanism of substrate recognition remains elusive; however, it has been determined that, similarly to A-domains, they contain a core motif (HHxxxxDG) that is essential for catalysis (Stachelhaus et al., 1998; Keating et al., 2002). The mechanism they utilise is thought to be similar to that of ribosome-catalysed peptide bond formation, in which ionic interactions stabilise catalytic intermediates, as opposed to acid-base catalysis as previously thought (Bieling et al., 2006; Samel et al., 2007).

1.5.4 Thioesterase Domains (Te domain)

The Te-domain works to release the now fully synthetized tethered peptide, and hence, is typically located as part of the final C-domain. This enzyme can catalyse either (i) the hydrolysis of the peptide (Fig. 6d (a)), or (ii) intramolecular cyclization either by amide or ester bond formation (Fig. 6d (b)), both of which result in the release of the peptide from the assembly line. They do this by catalysing the formation of an ester-linked intermediate the terminal carboxyl group located on the peptide with a conserved seryl residue within the Te domain, resulting in the cleavage of the peptide from the final 4'-Ppant arm. Te domains represent another barrier of substrate specificity as it has been shown that selectivity is limited to the NRP residues which bind at or near the site of the conserved seryl residue. Further, domains which catalyse cyclisation reactions have a larger substrate pool, with some selectively being based on the size of the resultant cyclic peptide and the secondary structure it forms. (Trauger et al., 2000; Trauger et al., 2001; Kohli et al., 2002; Tseng et al., 2002). As with A-domains, the specificity conferring nature and the ability to induce increased substrate flexibility has been investigated, generating cyclic peptide libraries resulting in novel derivatives of daptomycin and cryptophycin (Kohli et al., 2002; Beck et al., 2005; Kopp et al., 2006).

In addition to these core domains comprising the minimal requirement for NRPS activity, other strategically embedded auxiliary domains can be present, serving to further increase the structural diversity and functional groups present. Normally located between core motifs A8 and A9, these can include cyclization, methyltransferase, monooxygenases, oxidases and reductase domains (Labby et al., 2015). Further, it has recently been seen that these additional domains can sit within A-domains; initially thought to be inactivated by this interruption, it is

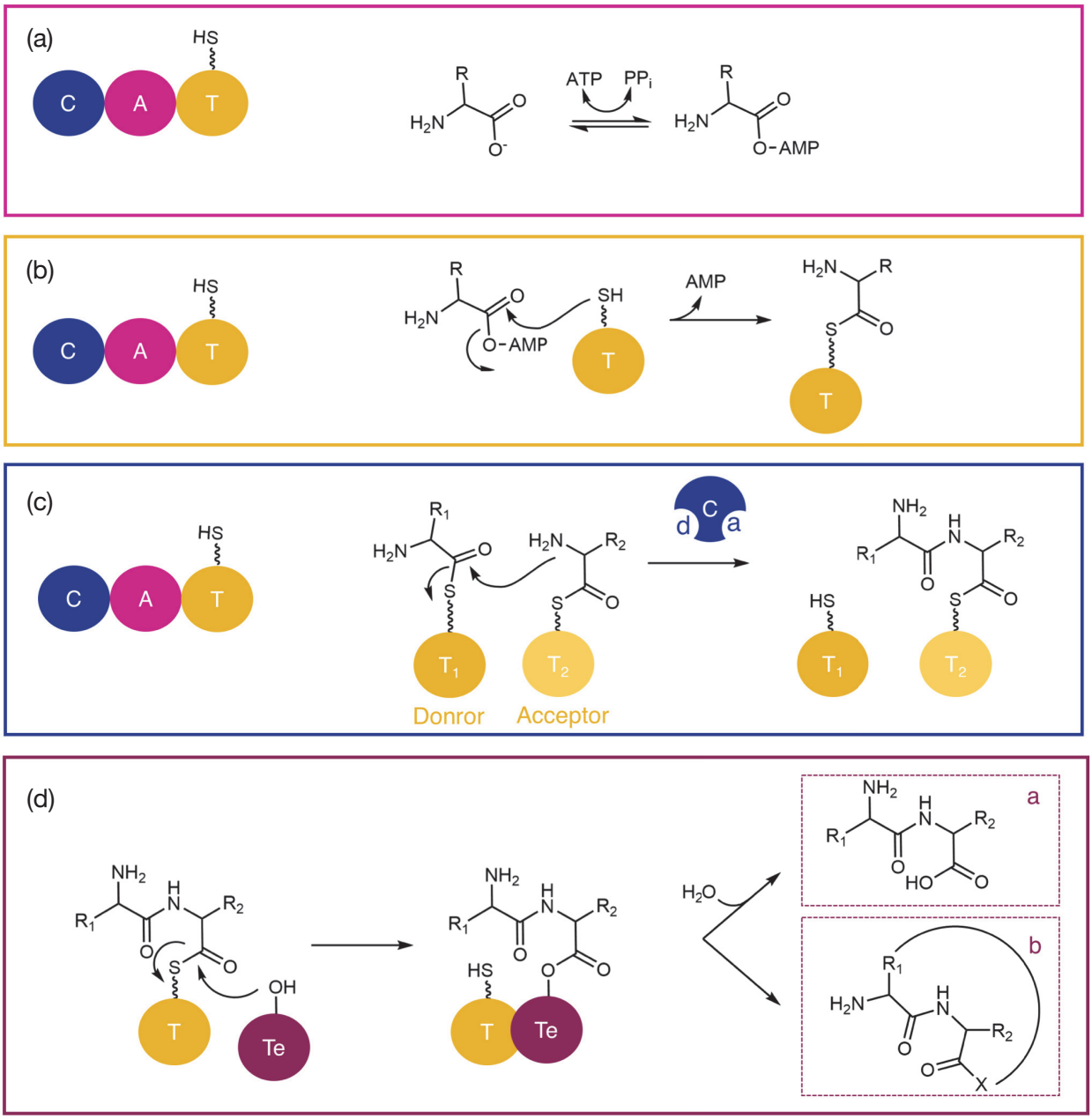


Figure 6. Schematic representation of the reactions catalysed by each NRPS module. The domain involved in each reaction is that of the same colour box which surrounds the reaction. **(a)** The reversible A-domain adenylation reaction in which catalyses aminoacyl-AMP formation. **(b**) transfer of the activated substrate to the T-domain releasing AMP, and the formation of the aminoacyl thioester. **(c)** The condensation domain then forms the peptide bond between two aminoacyl thioester substrates. In the schematic, T_1 and T_2 denote the T domains from neighbouring NRPS modules. And on the condensation enzyme, the 'a' and 'd' denote the donor and acceptor sites, respectively. **(d)** release of the substrate by the Te domain, first by aminoacylester formation on the Te domain then either by hydrolysis of the peptide (a) or cyclization (b). The X represents either a nitrogen or oxygen. The 4'Ppant co-factor is represented by the SH bonded to each T domain. Adapted from Felnagle et al., (2008).

increasingly becoming clear that these A-domains maintain activity, and take an active role in the biosynthesis of many natural products such as thiocoraline and cereulide (Magarvey et al., 2006; Zolova et al., 2014). The incorporation of an auxiliary domain within an A-domain has been demonstrated to allow successive adenylating then auxiliary action, further increasing the already colossal potential product structures (Labby et al., 2015). NRPS systems possess, acquired by millennia of evolution, a largely unexploited reservoir of compounds exhibiting chemical and structural diversity. This untapped potential could harbour the answer to stalling the ever impending post-antibiotic era; research efforts need to be redirected towards these systems with aim of (i) novel compound production, and (ii) low cost yet efficient and profitable discovery pipelines to entice interest back into the largely deserted antibiotic drug discovery industry.

1.6 Introduction to Proximicins

As previously discussed, the necessity for antimicrobial and anticancer with novel chemistry and mode of action has never been greater: proximicins represent a group of compounds which meet this demand. Proximicins are a family of three molecules (A-C) (Fig. 7), each consisting of two distinctive 2,4-disubstituted furan groups differing in what appears to be an additional tryptophan or tyrosine group being present on B and C, respectively. Proximicin A was first discovered in the

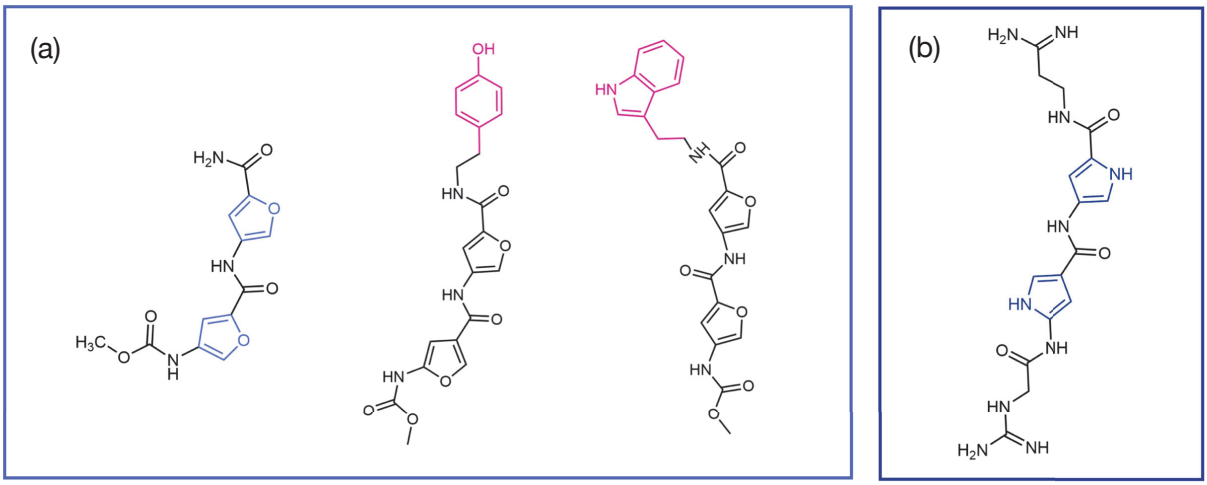


Figure 7. The Proximicin family of compounds, and the similar congocidine molecule. (a) Proximicins A-C (left to right) showing the core 2,4-disubstituted furan group highlighted in light blue and the unique groups to B and C highlighted in pink. **(b)** Congocidine molecule with core pyrrole groups highlighted in dark blue.

marine Actinomycete species, *V. maris* AB18-032 followed by that of A-C in *Verrucosispora sp.* str. MG37 (Fiedler et al., 2008). The research into proximicins is limited, however it has been shown that proximicins have the ability to arrest the cell cycle, although the mechanism of this remains elusive (Schneider et al., 2008). Schneider et al., (2008) demonstrated that they cause an upregulation of p53 and cyclin kinase inhibitor p21 - regulatory proteins involved in the transition from the G1 to the S phase of the cell cycle, suggesting a related cellular target, although this remains unverified. Other than their likely production via an NRPS system, little is also known about the by which proximicins are synthesised in the producing organism; interest has been largely directed towards other potential drug leads produced by *Verrucosispora* strains, such as the abyssomicin antimicrobial family (Fiedler et al., 2008). It has been suggested in passing by Lautru et al., (2012), that proximicins are likely biosynthesised via an NRPS assembly line mimicking that of the pyrrole containing analogue produced by *Streptomyces netropsis* (Finlay et al., 1954) congocidine (Fig. 7). It was noted that a simple replacement of the reductive amination step with a ketoreduction, could repurpose the congocidine pathway for proximicin biosynthesis (Lautru et al., 2012). However, despite the continual garnering of research interest directed to abyssomicin (Nicolaou et al., 2006; Keller et al. 2007a; Keller et al., 2007b; Gottardi et al., 2011; Goodfellow et al., 2012; Savic et al., 2013; Vieweg et al., 2014, Hashimoto et al., 2015), and congocidine (Juguet et al., 2009; Lipfert et al., 2010; Lautru et al., 2012; Al-Mestarihi et al., 2015), enabling the discovery and exploitation of their biosynthetic routes, information concerning proximicin biosynthesis remains largely elusive.

Proximicins present an exciting research opportunity for novel compound identification, for two principal reasons: (i) heterocycle containing compounds have previously been shown to possess anticancer and antimicrobial properties (Dmiitrienko et al., 2007; Wetzler et al., 2007; Arora et al., 2012; Mazimba et al., 2015), and it has been demonstrated that furans specifically alter the cellular target (Schneider et al., 2008); leading on to, (ii) 2,4-disubstituted furans preparation methods are rare, and typically low yielding and expensive (Lautru et al., 2012), hence, a cheap and rapid route to 2,4-disubstituted furan production would be advantageous to novel compound elucidation.

1.6.1 Proximicins as a scaffold of novel cell arresting drugs

In the quest for novel antibacterial compounds, the presence or addition of many chemical groups has been assessed for the activities they contribute, underlining the principle of combinatorial chemistry approaches to novel compound discovery. It was identified that compounds containing heterocycle groups possessed DNA binding activity; two compounds widely studied of note here are congocidine and distamycin produced by *S. netropsis* (Finlay et al., 1954) and *S. distallicus* (Arcamone et al., 1964), respectively. These compounds have a similar structure to that of the proximicin family, distinct by the presence of a N-methyl-pyrrole heterocycle, opposed to the furan of proximicins. These pyrrole analogues possess the ability to bind to the minor groove of DNA, representing the first compounds to demonstrate AT-selective DNA binding. This ability has made them good targets for the production of synthetic derivatives capable of binding to a specific sequence for use as antitumor agents (Kopka et al., 1985a; Kopka et al., 1985b; Schneider et al., 2008; Ganji et al., 2016). This interest has led to the refinement of the route responsible for congocidine biosynthesis in *S. netropsis*, outlining the enzymes responsible for incorporation of the activity-conferring N-methyl-pyrrole ring (Juguet et al., 2009; Al-Mestarihi et al.,2015). How this occurs was of specific interest as no adenylation domain has previously been shown to have specificity towards precursors containing this chemistry. Due to the high level of homology exhibited by proximicins and congocidine, it was suggested that their synthesis routes route be analogous. This led to the enzymatic machinery responsible for proximicin biosynthesis in *Verrucosispora* strains being considered established, and focus quickly progressed to other attractive metabolite investigation (Lautru et al., 2012). The differing mode of actions exhibited conferring the antitumor activity by congocidine and proximicin - DNA binding and interrupting the cell cycle independent of DNA binding respectively – was investigated, leading to an array of hybrid molecules being synthesised (Schneider et al., 2008; Wolter et al., 2009). Derivatives varied by the nature of terminal modifications, and the central furan/pyrrole groups were not altered. These apparently minor alterations resulted in an array of altered activities including improved antitumor activity compared to original compounds (Wolter et al., 2009). A change in mode of action was also shown, with furan containing homologs of congocidine exhibiting the ability to bind to the minor groove of DNA – a capability not displayed by proximicins (Wolter et al.,

2009). Interestingly, furan containing derivatives of congocidine, maintain higher levels of antitumor activity (Schneider et al., 2008; Wolter et al., 2009); this is likely due to the increased polarity of the molecules, and additional hydrogen-bond acceptors conferred by the oxygens of the furan group. The novel cell arrest activity demonstrated, and the presence of the novel, activity conferring chemistry, make it difficult to believe that the proximicin family of molecules have been overlooked to such an extent.

1.6.2 A simple source of novel chemistry

Heterocycles are an abundant structural element of natural products of almost every class and they typically contribute significantly to their biological activity (Pozharskii et al., 2011; Taylor et al., 2014). Despite nature providing many routes to their incorporation into biological molecules, reproducing this in a lab setting is extremely difficult. This is true for proximicins, resulting in the distinctive 2,4-disubstituted furan group of proximicins which makes them an alluring drug discovery target, also being responsible for the lack of research interest surrounding them. Excluding their initial discovery, all proximicin research has relied entirely on chemical synthesis of the 2,4-disubstituted group (Schneider et al., 2008; Wolter et al., 2009); as the biosynthetic pathway responsible in the producing organism has not been outlined, preventing its exploitation. The route to total synthesis of proximicin is extremely time consuming and financially expensive, typically resulting in low yields and over substitution of the furan ring (Wolter et al., 2009); this hinders the ability to investigate them, lowering their perceived attractiveness to large scale research investment. An inexpensive route amenable to scaling-up, resulting in 2,4-disubstiuted furan production would negate this problem; one approach would be to exploit the natural route to furan production in *Verrucosispora* strains, enabling their industrial scale recombinant production. To allow this, the biosynthetic route would have to be elucidated; a complex task considering only a singular *Verrucosispora* strain has a genome draft published (Roh et al., 2011), no biosynthetic clusters have been identified and, current work is minimal. Despite these barriers, the potential which 2,4-disubstituted furan exploitation poses is huge, not only for the construction of libraries of pyrrole/furan hybrid for novel DNA binding agents, but the ability to easily access 2,4-disubstituted furan chemistry, would allow its exploitation in small molecule libraries. The modular nature of the NRPS system

responsible for proximicin synthesis lends itself to this application. Analysis of the proximicin family of molecules, shows the repeating 2,4-disubstituted furan group, suggesting that it is formed from iterative activation of two furan containing precursors. The machinery which produces these compounds, and the unique enzymes responsible for their incorporation into the peptide backbone, provide extremely attractive targets to exploit for novel compound production. If these enzymes could be identified and characterised, it would represent an important milestone on the way to using these enzymes as a chemo-enzymatic tool in novel natural product synthesis, similar to previous work done (Friedrich et al., 2015; Schmidt et al., 2016). Effectively unlocking the potential this chemistry contains, allowing the production of the much needed novel antimicrobial, anticancer and antifungal compounds.

Proximicins present such an exciting lead for novel compound production as both a family of compounds and a potential source of novel chemistry for introduction into other families. To facilitate this, the gene cluster responsible for proximicin production was identified in both producers *V. maris* AB18-032 and *V. sp.* str. MG37; done by the sequencing, assembly and annotation of the *V. sp.* str. MG37 genome. Large scale genome comparison studies followed, allowing the identification of a common secondary metabolite gene cluster encoding NRPS enzyme machinery exhibiting the predicted similarities to that of congocidine. Large scale recombinant protein production and A-domain activity assays revealed the A-domain responsible for 2,4-disubstitued furan incorporation and allowed the outlining of the putative proximicin biosynthetic route. Preliminary studies utilising CRISPR/Cas technologies for gene editing were done in the proximicin producing organism, resulting in a phenotypical variant, paving the way for future work into proximicin cluster deletion and re-complementation studies.

1.7. References

Al-Mestarihi, A.H., Villamizar, G., Fernández, J., Zolova, O.E., Lombó, F. and Garneau-Tsodikova, S., 2014. Adenylation and S-methylation of cysteine by the bifunctional enzyme TioN in thiocoraline biosynthesis. *Journal of the American Chemical Society*, *136*(49), pp.17350-17354.

Antibiotic Research UK (2017) http://www.antibioticresearch.org.uk/about-antibiotic-resistance/
accessed 30/08/2017

Arora, P., Arora, V., Lamba, H.S. and Wadhwa, D., 2012. Importance of heterocyclic chemistry: A Review. *International Journal of Pharmaceutical Sciences and Research*, *3*(9), p.2947.

Baldry, S., 2010. Attack of the clones. *Nature Reviews Microbiology*, *8*(6), pp.390-391.

Baltz, R.H., 2011. Function of MbtH homologs in nonribosomal peptide biosynthesis and applications in secondary metabolite discovery. *Journal of industrial microbiology & biotechnology*, *38*(11), p.1747.

Baltz, R.H., 2014. MbtH homology codes to identify gifted microbes for genome mining. *Journal of industrial microbiology & biotechnology*, *41*(2), pp.357-369.

Beck, Z.Q., Aldrich, C.C., Magarvey, N.A., Georg, G.I. and Sherman, D.H., 2005. Chemoenzymatic synthesis of cryptophycin/arenastatin natural products. *Biochemistry*, *44*(41), pp.13457-13466

Bentley, S.D., Chater, K.F., Cerdeno-Tarraga, A.M., Challis, G.L., Thomson, N.R., James, K.D., Harris, D.E., Quail, M.A., Kieser, H., Harper, D. and Bateman, A., 2002. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3 (2). *Nature*, *417*(6885), pp.141-147.

Berg, T.L., Frøholm, L.O. and Laland, S.G., 1965. The biosynthesis of gramicidin S in a cell-free system. *Biochemical Journal*, *96*(1), p.43.
9

Bieling, P., Beringer, M., Adio, S. and Rodnina, M.V., 2006. Peptide bond formation does not involve acid-base catalysis by ribosomal residues. *Nature structural & molecular biology*, *13*(5), pp.423-428.

Blin, K., Medema, M.H., Kazempour, D., Fischbach, M.A., Breitling, R., Takano, E. and Weber, T., 2013. antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. *Nucleic acids research*, *41*(W1), pp.W204-W212.

Blin, K., Kazempour, D., Wohlleben, W. and Weber, T., 2014. Improved lanthipeptide detection and prediction for antiSMASH. *PLoS One*, *9*(2), p.e89420.

Borowska, Z.K. and Tatum, E.L., 1966. Biosynthesis of edeine by Bacillus brevis Vm4 in vivo and in vitro. *Biochimica et Biophysica Acta (BBA)-Nucleic Acids and Protein Synthesis*, *114*(1), pp.206-209.

Brakhage, A.A. and Schroeckh, V., 2011. Fungal secondary metabolites–strategies to activate silent gene clusters. *Fungal Genetics and Biology*, *48*(1), pp.15-22.

Bremer, D.J., 1968. *Cellular compartmentalization and control of fatty acid metabolism* (Vol. 12). Academic Press.

Bryan, L.E., 1989. Two forms of antimicrobial resistance: bacterial persistence and positive function resistance. *Journal of Antimicrobial Chemotherapy*, *23*(6), pp.817-820.

CDC (2017) https://www.cdc.gov/drugresistance/threat-report-2013/pdf/ar-threats-2013-508.pdf#page=18 Accessed 30/08/2017

Challis, G.L., Ravel, J. and Townsend, C.A., 2000. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chemistry & biology*, *7*(3), pp.211-224.

Challis, G.L., 2014. Exploitation of the Streptomyces coelicolor A3 (2) genome sequence for discovery of new natural products and biosynthetic pathways. *Journal of industrial microbiology & biotechnology*, *41*(2), pp.219-232.

Chavers, L.S., Moser, S.A., Benjamin, W.H., Banks, S.E., Steinhauer, J.R., Smith, A.M., Johnson, C.N., Funkhouser, E., Chavers, L.P., Stamm, A.M. and Waites, K.B., 2003. Vancomycin-resistant enterococci: 15 years and counting. *Journal of Hospital Infection*, *53*(3), pp.159-171.

Courvalin, P., 2008. Predictable and unpredictable evolution of antibiotic resistance. *Journal of internal medicine*, *264*(1), pp.4-16.

Crump, J.A. and Collignon, P.J., 2000. Intravascular catheter-associated infections. *European Journal of Clinical Microbiology & Infectious Diseases*, *19*(1), pp.1-8.

D'costa, V.M., McGrann, K.M., Hughes, D.W. and Wright, G.D., 2006. Sampling the antibiotic resistome. *Science*, *311*(5759), pp.374-377.

D'Costa, V.M., Griffiths, E. and Wright, G.D., 2007. Expanding the soil antibiotic resistome: exploring environmental diversity. *Current opinion in microbiology*, *10*(5), pp.481-489.

Demain, A.L., 1999. Pharmaceutically active secondary metabolites of microorganisms. *Applied microbiology and biotechnology*, *52*(4), pp.455-463.

Demain, A.L. and Sanchez, S., 2009. Microbial drug discovery: 80 years of progress. *Journal of Antibiotics*, *62*(1), p.5.

Dieckmann, R., Lee, Y.O., van Liempt, H., von Döhren, H. and Kleinkauf, H., 1995. Expression of an active adenylate-forming domain of peptide synthetases corresponding to acyl-CoA-synthetases. *FEBS letters*, *357*(2), pp.212-216.

Dijkshoorn, L., Nemec, A. and Seifert, H., 2007. An increasing threat in hospitals: multidrug-resistant Acinetobacter baumannii. *Nature Reviews Microbiology*, *5*(12), pp.939-951.

Dmitrienko, G.I., 2007. Heterocyclic Antitumor Antibiotics. Topics in Heterocyclic Chemistry, 02 Edited by Moses Lee (Hope College, Holland, MI). Series Edited by RR Gupta. Springer: Berlin, Heidelberg, New York. 2006. xiv+ 252 pp. $149.00. ISBN 3-540-30982-9.

Donadio, S., Monciardini, P. and Sosio, M., 2007. Polyketide synthases and nonribosomal peptide synthetases: the emerging view from bacterial genomics. *Natural product reports*, *24*(5), pp.1073-1109.

Drake, E.J., Cao, J., Qu, J., Shah, M.B., Straubinger, R.M. and Gulick, A.M., 2007. The 1.8 Å crystal structure of PA2412, an MbtH-like protein from the pyoverdine cluster of Pseudomonas aeruginosa. *Journal of Biological Chemistry*, *282*(28), pp.20425-20434.

Dye, C., 2009. Doomsday postponed? Preventing and reversing epidemics of drug-resistant tuberculosis. *Nature reviews. Microbiology*, *7*(1), p.81.

Felnagle, E.A., Jackson, E.E., Chan, Y.A., Podevels, A.M., Berti, A.D., McMahon, M.D. and Thomas, M.G., 2008. Nonribosomal peptide synthetases involved in the production of medically relevant natural products. *Molecular pharmaceutics*, *5*(2), pp.191-211.

Fiedler, H.P., Bruntner, C., Riedlinger, J., Bull, A.T., Knutsen, G., Goodfellow, M., Jones, A., Maldonado, L., Pathom-Aree, W., Beil, W. and Schneider, K., 2008. Proximicin A, B and C, novel aminofuran antibiotic and anticancer compounds isolated from marine strains of the actinomycete Verrucosispora. *Journal of Antibiotics*, *61*(3), p.158.

Finlay, A.C., Hochstein, F.A., Sobin, B.A. and Murphy, F.X., 1951. Netropsin, a new antibiotic produced by a Streptomyces. *Journal of the American Chemical Society*, *73*(1), pp.341-343.

Fischbach, M.A. and Walsh, C.T., 2009. Antibiotics for emerging pathogens. *Science*, *325*(5944), pp.1089-1093.

Friedrich, S. and Hahn, F., 2015. Opportunities for enzyme catalysis in natural product chemistry. *Tetrahedron*, *71*(10), pp.1473-1508.

Frøshov, Ø., Zimmer, T.L. and Laland, S.G., 1970. The nature of the enzyme bound intermediates in gramicidin s biosynthesis. *FEBS letters*, *7*(1), pp.68-71.

Fowler Jr, V.G., Boucher, H.W., Corey, G.R., Abrutyn, E., Karchmer, A.W., Rupp, M.E., Levine, D.P., Chambers, H.F., Tally, F.P., Vigliani, G.A. and Cabell, C.H., 2006. Daptomycin versus standard therapy for bacteremia and endocarditis caused by Staphylococcus aureus. *New England Journal of Medicine*, *355*(7), pp.653-665.

FUJIKAWA, K., SUZUKI, T. and KURAHASHI, K., 1966. Incorporation of L-leucine-C14 into tyrocidine by a cell-free preparation of Bacillus brevis Dubos strain. *The Journal of Biochemistry*, *60*(2), pp.216-218.

Ganji, M., Kim, S.H., van der Torre, J., Abbondanzieri, E. and Dekker, C., 2016. Intercalation-based single-molecule fluorescence assay to study DNA supercoil dynamics. *Nano letters*, *16*(7), pp.4699-4707.

Gevers, W., Kleinkauf, H. and Lipmann, F., 1968. The activation of amino acids for biosynthesis of gramicidin S. *Proceedings of the National Academy of Sciences*, *60*(1), pp.269-276.

Gevers, W., Kleinkauf, H. and Lipmann, F., 1969. Peptidyl transfers in gramicidin S biosynthesis from enzyme-bound thioester intermediates. *Proceedings of the National Academy of Sciences*, *63*(4), pp.1335-1342.

Gomez-Escribano, J.P., Song, L., Bibb, M.J. and Challis, G.L., 2012. Posttranslational β-methylation and macrolactamidination in the biosynthesis of the bottromycin complex of ribosomal peptide antibiotics. *Chemical Science*, *3*(12), pp.3522-3525.

Goodfellow, M., Stach, J.E., Brown, R., Bonda, A.N.V., Jones, A.L., Mexson, J., Fiedler, H.P., Zucchi, T.D. and Bull, A.T., 2012. Verrucosispora maris sp. nov., a novel deep-sea actinomycete isolated from a marine sediment which produces abyssomicins. *Antonie Van Leeuwenhoek*, *101*(1), pp.185-193.

Gottardi, E.M., Krawczyk, J.M., von Suchodoletz, H., Schadt, S., Mühlenweg, A., Uguru, G.C., Pelzer, S., Fiedler, H.P., Bibb, M.J., Stach, J.E. and Süssmuth, R.D., 2011. Abyssomicin Biosynthesis: Formation of an Unusual Polyketide, Antibiotic-Feeding Studies and Genetic Analysis. *ChemBioChem*, *12*(9), pp.1401-1410.

Gudlaugsson, O., Gillespie, S., Lee, K., Berg, J.V., Hu, J., Messer, S., Herwaldt, L., Pfaller, M. and Diekema, D., 2003. Attributable mortality of nosocomial candidemia, revisited. *Clinical Infectious Diseases*, *37*(9), pp.1172-1177.

Gilhuus-Moe, C.C., Kristensen, T., Bredesen, J.E., Zimmer, T.L. and Laland, S.G., 1970. The presence and possible role of phosphopantothenic acid in gramicidin S synthetase. *FEBS letters*, *7*(3), pp.287-290.

Gulick, A.M., 2009. Conformational dynamics in the Acyl-CoA synthetases, adenylation domains of non-ribosomal peptide synthetases, and firefly luciferase. *ACS chemical biology*, *4*(10), pp.811-827.

Haese, A., Pieper, R., von Ostrowski, T. and Zocher, R., 1994. Bacterial expression of catalytically active fragments of the multifunctional enzyme enniatin synthetase. *Journal of molecular biology*, *243*(1), pp.116-122.

Hashimoto, T., Hashimoto, J., Teruya, K., Hirano, T., Shin-ya, K., Ikeda, H., Liu, H.W., Nishiyama, M. and Kuzuyama, T., 2015. Biosynthesis of versipelostatin: identification of an enzyme-catalyzed [4+ 2]-cycloaddition required for macrocyclization of spirotetronate-containing polyketides. *Journal of the American Chemical Society*, *137*(2), pp.572-575.

Imker, H.J., Krahn, D., Clerc, J., Kaiser, M. and Walsh, C.T., 2010. N-acylation during glidobactin biosynthesis by the tridomain nonribosomal peptide synthetase module GlbF. *Chemistry & biology*, *17*(10), pp.1077-1083.

Izawa, M., Kawasaki, T. and Hayakawa, Y., 2013. Cloning and heterologous expression of the thioviridamide biosynthesis gene cluster from Streptomyces olivoviridis. *Applied and environmental microbiology*, *79*(22), pp.7110-7113.

Jassal, M. and Bishai, W.R., 2009. Extensively drug-resistant tuberculosis. *The Lancet infectious diseases*, *9*(1), pp.19-30.

Juguet, M., Lautru, S., Francou, F.X., Nezbedová, Š., Leblond, P., Gondry, M. and Pernodet, J.L., 2009. An iterative nonribosomal peptide synthetase assembles the pyrrole-amide antibiotic congocidine in Streptomyces ambofaciens. *Chemistry & biology*, *16*(4), pp.421-431.

Keating, T.A., Marshall, C.G., Walsh, C.T. and Keating, A.E., 2002. The structure of VibH represents nonribosomal peptide synthetase condensation, cyclization and epimerization domains. *Nature Structural & Molecular Biology*, *9*(7), p.522.

Keller, N.P., Turner, G. and Bennett, J.W., 2005. Fungal secondary metabolism--from biochemistry to genomics. *Nature reviews. Microbiology*, *3*(12), p.937.

Keller, S.(a), Nicholson, G., Drahl, C., Sorensen, E., Fiedler, H.P. and Süssmuth, R.D., 2007. Abyssomicins G and H and atrop-abyssomicin C from the marine Verrucosispora strain AB-18-032. *The Journal of antibiotics*, *60*(6), pp.391-394.

Keller, S. (b), Schadt, H.S., Ortel, I. and Süssmuth, R.D., 2007. Action of atrop-Abyssomicin C as an Inhibitor of 4-Amino-4-deoxychorismate Synthase PabB. *Angewandte Chemie International Edition*, *46*(43), pp.8284-8286.

Kingston, D.G. and Newman, D.J., 2002. Mother nature's combinatorial libraries; their influence on the synthesis of drugs. *Current opinion in drug discovery & development*, *5*(2), pp.304-316.

Kleinkauf, H., Gevers, W. and Lipmann, F., 1969. Interrelation between activation and polymerization in gramicidin S biosynthesis. *Proceedings of the National Academy of Sciences*, *62*(1), pp.226-233.

Kleinkauf, H., Gevers, W., Roskoski, R. and Lipmann, F., 1970. Enzyme-bound phosphopantetheine in tyrocidine biosynthesis. *Biochemical and biophysical research communications*, *41*(5), pp.1218-1222.

Koglin, A., Mofid, M.R., Löhr, F., Schäfer, B., Rogov, V.V., Blum, M.M., Mittag, T., Marahiel, M.A., Bernhard, F. and Dötsch, V., 2006. Conformational switches modulate protein interactions in peptide antibiotic synthetases. *science*, *312*(5771), pp.273-276.

Kohli, R.M., Takagi, J. and Walsh, C.T., 2002. The thioesterase domain from a nonribosomal peptide synthetase as a cyclization catalyst for integrin binding peptides. *Proceedings of the National Academy of Sciences*, *99*(3), pp.1247-1252.

Kopka, M.L. (a), Yoon, C., Goodsell, D., Pjura, P. and Dickerson, R.E., 1985. Binding of an antitumor drug to DNA: Netropsin and CGCGAATT-BrC-GCG. *Journal of molecular biology*, *183*(4), pp.553-563.

Kopka, M.L. (b), Yoon, C., Goodsell, D., Pjura, P. and Dickerson, R.E., 1985. The molecular origin of DNA-drug specificity in netropsin and distamycin. *Proceedings of the National Academy of Sciences*, *82*(5), pp.1376-1380.

Kopp, F., Grünewald, J., Mahlert, C. and Marahiel, M.A., 2006. Chemoenzymatic design of acidic lipopeptide hybrids: new insights into the structure– activity relationship of daptomycin and A54145. *Biochemistry*, *45*(35), pp.10474-10481.

Kraus, C.N., 2008. Low hanging fruit in infectious disease drug development. *Current opinion in microbiology*, *11*(5), pp.434-438.

Kurahashi, K., 1974. Biosynthesis of small peptides. *Annual review of biochemistry*, *43*(1), pp.445-459.

Labby, K.J., Watsula, S.G. and Garneau-Tsodikova, S., 2015. Interrupted adenylation domains: unique bifunctional enzymes involved in nonribosomal peptide biosynthesis. *Natural product reports*, *32*(5), pp.641-653.

Laland, S. and Zimmer, T.L., 1973. The protein thiotemplate mechanism of synthesis for the peptide antibiotics produced by Bacillus brevis. *Essays in biochemistry*, *9*, pp.31-57.

Lautru, S., Oves-Costales, D., Pernodet, J.L. and Challis, G.L., 2007. MbtH-like protein-mediated cross-talk between non-ribosomal peptide antibiotic and siderophore biosynthetic pathways in Streptomyces coelicolor M145. *Microbiology*, *153*(5), pp.1405-1412.

Lautru, S., Song, L., Demange, L., Lombès, T., Galons, H., Challis, G.L. and Pernodet, J.L., 2012. A Sweet Origin for the Key Congocidine Precursor 4-Acetamidopyrrole-2-carboxylate. *Angewandte Chemie International Edition*, *51*(30), pp.7454-7458.

Lee, S.G., Roskoski Jr, R., Bauer, K. and Lipmann, F., 1973. Purification of the polyenzymes responsible for tyrocidine synthesis and their dissociation into subunits. *Biochemistry*, *12*(3), pp.398-405.

Letzel, A.C., Pidot, S.J. and Hertweck, C., 2013. A genomic approach to the cryptic secondary metabolome of the anaerobic world. *Natural product reports*, *30*(3), pp.392-428.

Lipfert, J., Klijnhout, S. and Dekker, N.H., 2010. Torsional sensing of small-molecule binding using magnetic tweezers. *Nucleic acids research*, *38*(20), pp.7122-7132.

Lipmann, F., Roskoski Jr, R., Kleinkauf, H. and Gevers, W., 1970. Isolation of enzyme-bound peptide intermediates in tyrocidine biosynthesis. *Biochemistry*, *9*(25), pp.4846-4851.

Lipmann, F., 1971. Attempts to map a process evolution of peptide biosynthesis. *Science*, *173*(4000), pp.875-884.

Lipmann, F., 1973. Nonribosomal polypeptide synthesis on polyenzyme templates. *Accounts of Chemical Research*, *6*(11), pp.361-367.

Livermore, D.M., 2000. Antibiotic resistance in staphylococci. *International Journal of Antimicrobial Agents*, *16*, pp.3-10.

Louie, T.J., Miller, M.A., Mullane, K.M., Weiss, K., Lentnek, A., Golan, Y., Gorbach, S., Sears, P. and Shue, Y.K., 2011. Fidaxomicin versus vancomycin for Clostridium difficile infection. *New England Journal of Medicine*, *364*(5), pp.422-431.

Marahiel, M.A., Stachelhaus, T. and Mootz, H.D., 1997. Modular peptide synthetases involved in nonribosomal peptide synthesis. *Chemical reviews*, *97*(7), pp.2651-2674.

Mach, B., Reich, E. and Tatum, E.L., 1963. Separation of the biosynthesis of the antibiotic polypeptide tyrocidine from protein biosynthesis. *Proceedings of the National Academy of Sciences*, *50*(1), pp.175-181.

Magarvey, N.A., Ehling-Schulz, M. and Walsh, C.T., 2006. Characterization of the cereulide NRPS α-hydroxy acid specifying modules: activation of α-keto acids and chiral reduction on the assembly line. *Journal of the American Chemical Society*, *128*(33), pp.10698-10699.

Martínez, J.L., Baquero, F. and Andersson, D.I., 2007. Predicting antibiotic resistance. *Nature reviews. Microbiology*, *5*(12), p.958.

Martinez, J.L., 2009. The role of natural environments in the evolution of resistance traits in pathogenic bacteria. *Proceedings of the Royal Society of London B: Biological Sciences*, *276*(1667), pp.2521-2530.

Mazimba, O., 2015. Antimicrobial activities of heterocycles derived from thienylchalcones. *Journal of King Saud University-Science*, *27*(1), pp.42-48.

McGowan, J.E., 2006. Resistance in nonfermenting gram-negative bacteria: multidrug resistance to the maximum. *The American journal of medicine*, *119*(6), pp.S29-S36.

McKenna, M., 2011. The enemy within. *Scientific American*, *304*(4), pp.46-53.

Medema, M.H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M.A., Weber, T., Takano, E. and Breitling, R., 2011. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic acids research*, *39*(suppl_2), pp.W339-W346.

Metzker, M.L., 2010. Sequencing technologies--the next generation. *Nature reviews. Genetics*, *11*(1), p.31.

Nicolaou, K.C. and Harrison, S.T., 2006. Total synthesis of abyssomicin C and atrop-abyssomicin C. *Angewandte Chemie International Edition*, *45*(20), pp.3256-3260.

Nordmann, P., Naas, T., Fortineau, N. and Poirel, L., 2007. Superbugs in the coming new decade; multidrug resistance and prospects for treatment of Staphylococcus aureus, Enterococcus spp. and Pseudomonas aeruginosa in 2010. *Current opinion in microbiology*, *10*(5), pp.436-440.

Ōmura, S., Ikeda, H., Ishikawa, J., Hanamoto, A., Takahashi, C., Shinose, M., Takahashi, Y., Horikawa, H., Nakazawa, H., Osonoe, T. and Kikuchi, H., 2001. Genome sequence of an industrial microorganism Streptomyces avermitilis: deducing the ability of producing secondary metabolites. *Proceedings of the National Academy of Sciences*, *98*(21), pp.12215-12220.

Plant, P.J., White, E.H. and McElroy, W.D., 1968. The decarboxylation of luciferin in firefly bioluminescence. *Biochemical and biophysical research communications*, *31*(1), pp.98-103.

Pozharskii, A.F., Soldatenkov, A.T. and Katritzky, A.R., 2011. *Heterocycles in life and society: an introduction to heterocyclic chemistry, biochemistry and applications*. John Wiley & Sons.

Quadri, L.E., Weinreb, P.H., Lei, M., Nakano, M.M., Zuber, P. and Walsh, C.T., 1998. Characterization of Sfp, a Bacillus subtilis phosphopantetheinyl transferase for peptidyl carrier protein domains in peptide synthetases. *Biochemistry*, *37*(6), pp.1585-1595.

Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W. and Huson, D.H., 2005. Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic acids research*, *33*(18), pp.5799-5808.

Roh, H., Uguru, G.C., Ko, H.J., Kim, S., Kim, B.Y., Goodfellow, M., Bull, A.T., Kim, K.H., Bibb, M.J., Choi, I.G. and Stach, J.E., 2011. Genome Sequence of the Abyssomicin and Proximicin-Producing Marine Actinomycete Verrucosispora maris AB-18-032. *Journal of bacteriology*, pp.JB-05041.

Röttig, M., Medema, M.H., Blin, K., Weber, T., Rausch, C. and Kohlbacher, O., 2011. NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. *Nucleic acids research*, *39*(suppl_2), pp.W362-W367.

Samel, S.A., Schoenafinger, G., Knappe, T.A., Marahiel, M.A. and Essen, L.O., 2007. Structural and functional insights into a peptide bond-forming bidomain from a nonribosomal peptide synthetase. *Structure*, *15*(7), pp.781-792.

Savic, V., 2013. Abyssomicins: Isolation, Properties, and Synthesis. *Studies in Natural Products Chemistry*, *40*, p.133.

Schorn, M., Zettler, J., Noel, J.P., Dorrestein, P.C., Moore, B.S. and Kaysser, L., 2013. Genetic basis for the biosynthesis of the pharmaceutically important class of epoxyketone proteasome inhibitors. *ACS chemical biology*, *9*(1), pp.301-309.

Schlumbohm, W., Stein, T., Ullrich, C., Vater, J., Krause, M., Marahiel, M.A., Kruft, V. and Wittmann-Liebold, B., 1991. An active serine is involved in covalent substrate amino acid binding at each reaction center of gramicidin S synthetase. *Journal of Biological Chemistry*, *266*(34), pp.23135-23141.

Schmidt, N.G., Eger, E. and Kroutil, W., 2016. Building Bridges: Biocatalytic C–C-Bond Formation toward Multifunctional Products. *ACS catalysis*, *6*(7), pp.4286-4311.

Schneider, K., Keller, S., Wolter, F.E., Röglin, L., Beil, W., Seitz, O., Nicholson, G., Bruntner, C., Riedlinger, J., Fiedler, H.P. and Süssmuth, R.D., 2008. Proximicins A, B, and C—antitumor furan analogues of netropsin from the marine actinomycete Verrucosispora induce upregulation of p53 and the cyclin kinase inhibitor p21. *Angewandte Chemie International Edition*, *47*(17), pp.3258-3261.

Schwarzer, D., Finking, R. and Marahiel, M.A., 2003. Nonribosomal peptides: from genes to products. *Natural product reports*, *20*(3), pp.275-287.

Sigerist, H.E., 1958. *The great doctors: A biographical history of medicine*. Doubleday.

Silver, L.L., 2012. Rational approaches to antibacterial discovery: pre-genomic directed and phenotypic screening. In *Antibiotic Discovery and Development* (pp. 33-75). Springer US.

Spæren, U., Frøholm, L.O. and Laland, S.G., 1967. Further studies on the biosynthesis of gramicidin S and proteins in a cell-free system from Bacillus brevis. *Biochemical Journal*, *102*(2), p.586.

Stachelhaus, T., Mootz, H.D. and Marahiel, M.A., 1999. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chemistry & biology*, *6*(8), pp.493-505.

Stein, T., Vater, J., Kruft, V., Wittmann-Liebold, B., Franke, P., Panico, M., Mc Dowell, R. and Morris, H.R., 1994. Detection of 4'-phosphopantetheine at the thioester binding site for L-valine of gramicidinS synthetase 2. *FEBS letters*, *340*(1-2), pp.39-44.

Stein, T., Vater, J., Kruft, V., Otto, A., Wittmann-Liebold, B., Franke, P., Panico, M., McDowell, R. and Morris, H.R., 1996. The multiple carrier model of nonribosomal peptide biosynthesis at modular multienzymatic templates. *Journal of Biological chemistry*, *271*(26), pp.15428-15435.

Strieker, M., Tanović, A. and Marahiel, M.A., 2010. Nonribosomal peptide synthetases: structures and dynamics. *Current opinion in structural biology*, *20*(2), pp.234-240.

Tanovic, A., Samel, S.A., Essen, L.O. and Marahiel, M.A., 2008. Crystal structure of the termination module of a nonribosomal peptide synthetase. *Science*, *321*(5889), pp.659-663.

Tatham, E., sundaram Chavadi, S., Mohandas, P., Edupuganti, U.R., Angala, S.K., Chatterjee, D. and Quadri, L.E., 2012. Production of mycobacterial cell wall glycopeptidolipids requires a member of the MbtH-like protein family. *BMC microbiology*, *12*(1), p.118.

Taylor, R.D., MacCoss, M. and Lawson, A.D., 2014. Rings in drugs: miniperspective. *Journal of medicinal chemistry*, *57*(14), pp.5845-5859.

Tegos, G., Tegos, G. and Mylonakis, E., 2012. *Antimicrobial drug discovery: emerging strategies*. Cabi.

Thirlway, J., Lewis, R., Nunns, L., Al Nakeeb, M., Styles, M., Struck, A.W., Smith, C.P. and Micklefield, J., 2012. Introduction of a non-natural amino acid into a nonribosomal peptide antibiotic by modification of adenylation domain specificity. *Angewandte Chemie International Edition*, *51*(29), pp.7181-7184.

Tomino, S., Yamada, M., Itoh, H. and Kurahashi, K., 1967. Cell-free synthesis of gramicidin S. *Biochemistry*, *6*(8), pp.2552-2560.

Trauger, J.W., Kohil, R.M., Mootz, H.D., Marahiel, M.A. and Walsh, C.T., 2000. Peptide cyclization catalysed by the thioesterase domain of tyrocidine synthetase. *Nature*, *407*(6801), p.215.

Trauger, J.W., Kohli, R.M. and Walsh, C.T., 2001. Cyclization of backbone-substituted peptides catalyzed by the thioesterase domain from the tyrocidine nonribosomal peptide synthetase. *Biochemistry*, *40*(24), pp.7092-7098.

Tseng, C.C., Bruner, S.D., Kohli, R.M., Marahiel, M.A., Walsh, C.T. and Sieber, S.A., 2002. Characterization of the surfactin synthetase C-terminal thioesterase domain as a cyclic depsipeptide synthase. *Biochemistry*, *41*(45), pp.13350-13359.

Vieweg, L., Reichau, S., Schobert, R., Leadlay, P.F. and Süssmuth, R.D., 2014. Recent advances in the field of bioactive tetronates. *Natural product reports*, *31*(11), pp.1554-1584.

Von Döhren, H., 1990. Compilation of peptide structures-A biogenetic approach. *Biochemistry of peptide antibiotics*, pp.411-507.

Weber, T., Baumgartner, R., Renner, C., Marahiel, M.A. and Holak, T.A., 2000. Solution structure of PCP, a prototype for the peptidyl carrier domains of modular peptide synthetases. *Structure*, *8*(4), pp.407-418.

Weber, T., Charusanti, P., Musiol-Kroll, E.M., Jiang, X., Tong, Y., Kim, H.U. and Lee, S.Y., 2015. Metabolic engineering of antibiotic factories: new tools for antibiotic production in actinomycetes. *Trends in biotechnology*, *33*(1), pp.15-26.

Wetzler, M., 2007. *Advances in the synthesis of DNA-binding polyamide ligands, their biophysical characterization, and effects on bacterial sporulation*. University of California, Berkeley.

Willyard, C., 2017. The drug-resistant bacteria that pose the greatest health threats. *Nature News*, *543*(7643), p.15.

Wilson, M.C., Mori, T., Rückert, C., Uria, A.R., Helf, M.J., Takada, K., Gernert, C., Steffens, U.A., Heycke, N., Schmitt, S. and Rinke, C., 2014. An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature*, *506*(7486), p.58.

WHO (2017) http://www.who.int/antimicrobial-resistance/en/ accessed 30/08/2017

Wolter, F.E., Schneider, K., Davies, B.P., Socher, E.R., Nicholson, G., Seitz, O. and Süssmuth, R.D., 2009. Total Synthesis of Proximicin A– C and Synthesis of New Furan-Based DNA Binding Agents. *Organic letters*, *11*(13), pp.2804-2807.

Woodford, N. and Ellington, M.J., 2007. The emergence of antibiotic resistance by mutation. *Clinical Microbiology and Infection*, *13*(1), pp.5-18.

Yukioka, M., Tsukamoto, Y., Saito, Y., Tsuji, T., Otani, S. and Otani, S., 1965. Biosynthesis of gramicidin S by a cell-free system of Bacillus brevis. *Biochemical and biophysical research communications*, *19*(2), pp.204-208.

Zerikly, M. and Challis, G.L., 2009. Strategies for the discovery of new natural products by genome mining. *ChemBioChem*, *10*(4), pp.625-633.

Zhang, W., Heemstra Jr, J.R., Walsh, C.T. and Imker, H.J., 2010. Activation of the pacidamycin PacL adenylation domain by MbtH-like proteins. *Biochemistry*, *49*(46), pp.9946-9947.

Zolova, O.E. and Garneau-Tsodikova, S., 2014. KtzJ-dependent serine activation and O-methylation by KtzH for kutznerides biosynthesis. *Journal of Antibiotics*, *67*(1), p.59.

Chapter 2. Genome sequencing, assembly, annotation of *Verrucosispora sp.* str. MG37, and identification of putative proximicin cluster

2.1 Introduction

2.1.1 Overview of Proximicins

Proximicins are a family of heterocycle containing compounds, shown to possess the ability to arrest the cell cycle, in addition to slight antimicrobial activity (Wolter et al., 2009). They are produced by two species of marine Actinomycete *Verrucosispora – maris* AB18-032 and *sp.* str. MG37– first discovered in sediments isolated from the Sea of Japan and Raune Fjord, Norway, respectively (Fiedler et al., 2008). Fiedler et al., (2008) used HPLC-diode array to monitor extracts from the isolated *Verrucosispora* species, which led to the identification of the secondary metabolites (SM) – specifically proximicins – from the two producers. Proximicins show high structural similarity with characterised pyrrolamidone antibiotics – congocidine and distamycin – produced by *S. netropsis* (Finlay et al., 1954) and *S. distallicus* (Arcamone et al., 1964), respectively. These γ-peptide antibiotics have a characteristic structural element – an N-methyl-pyrrole ring, conferring antitumor activity caused by their ability to selectively bind to the AT-rich sequences found in the minor groove of DNA (Bailly et al., 1998). Distinct from the pyrrole-amides, proximicins have been shown to have a different mode of anti-tumour action not reliant on DNA binding (Fiedler et al., 2008). The ability to produce biologically active compounds produced by *Verrucosispora* strains pose the possibility of novel clinically relevant molecule characterisation. *V. maris* AB18-032 was the subject of much research, resulting in its genome being sequenced, assembled and annotated (Roh et al., 2011) – due to its ability to produce the abyssomicin family of compounds. Abyssomicin's are polycyclic polyketide natural products with impressive biological activity and present a substantial potential clinical lead (Nicolaou & Harrison, 2006); this family of compounds are not produced by *V. sp.* str. MG37 *(*Fiedler et al., 2008). Due to the combination of structural novelty and activity of the abyssomicin family, proximicins have largely been overlooked. What is known is that there are three proximicin members: A, B and C (Fig. 7); all contain

two 2,4-disubstitued furan groups, with B and C having what appears to be an additional tyrosine or tryptophan moiety present, respectively. Only B exhibited moderate growth inhibition of Gram-Positive bacteria, whereas C showed only slight inhibition against *Brevibaccillus brevis* DSM30 (Fiedler et al., 2008). All Gram-Negative and fungi tested by Fiedler et al., (2008) where shown to be insensitive to all proximicins; however, substantial inhibition of tumour cell line growth was demonstrated. Despite their novel chemistry and activity, little is known about the enzymatic machinery responsible for proximicin biosynthesis. To harness the novel chemistry and activity exhibited by this family, whole genome sequencing (WGS) and annotation studies must be done to: (i) identify a putative proximicin biosynthesis (*ppb*) cluster present in both producers; (ii) identify the route to proximicin synthesis and, (iii) determine the genetic basis for the discrepancies in proximicin production between the two *Verrucosispora* species.

2.1.2 Pre-genomic era SBC identification

Secondary metabolites isolated from microbes have had an unparalleled impact on modern-day medicine. These small molecule natural products, which include important clinical drugs, such as: antibiotics (Kardos & Demain et al., 2011); cholesterol-lowering compounds (Endo, 2010); antifungals, and immunosuppressive cyclosporins (Britton & Palacios, 1982), have revolutionised the treatment and prevention of illness and infection. Despite the plethora of compounds identified and in current use, microbial secondary metabolites continue to be an important source of molecules for drug discovery. Many initial insights into SM biosynthetic pathways came during the early 1950's, and were obtained by organic chemists using isotopic tracers (Bentley, 1999). The research methods quickly shifted to investigate the molecular biology with the advent of DNA cloning and sequencing tools; this culminated in many landmark papers, which described the molecular cloning of entire secondary metabolite clusters (SMC) from Actinomycete bacteria (Malpartida & Hopwood, 1984; Cortes et al., 1990; Donadio et al., 1991). After preliminary identification of genes involved in SM biosynthesis, the hallmark trait of gene clustering was quickly observed in both fungi and bacteria; this tendency of genes to cluster on a chromosomal locus greatly accelerated the elucidation of enzymatic steps involved in the biosynthesis of compounds. In the pre-genomic era, identification of SMC was a laborious, and time consuming task. It was done

traditionally by either complementation of blocked mutants by cosmid libraries or insertional mutagenesis followed by plasmid rescue (Mayorga & Timberlake, 1990; Hendrickson et al., 1999); both of these approaches rely on extensive screening, which is simply not a feasible response to increasing incidence of bacterial multiple-resistance (Yang et al., 1996). Other approaches, such as the genome scanning method described by Zazopoulos et al., (2003) in which the enediyne antitumor antimitotic pathways were characterised offered other approaches; however, the introduction of low cost, rapid sequencing platforms and the genomic revolution it triggered completely transformed SMC elucidation.

The introduction of DNA sequencing, bypassed the necessity for these cumbersome SMC identification methods, by revealing the inventory of all the SMC present in the producing organism. This paradigm shift led to some renaissance of natural product research in academia and industry. This was reinforced when the first genome sequence of the model organism *Streptom*yces *coelicolor* A3(2) (Bentley et al., 2002) and the avermectin producing strain *S. avermitilis* (Omura et al., 2001; Ikeda et al., 2003) were both found to possess more SMC than previously estimated. This was especially remarkable as these strains both served as model organisms and industrial production strains, and so it was widely presumed that all was already known about the SM's they are capable of producing. As sequencing ability and applicability increased, high numbers of cryptic or silent secondary biosynthetic cluster (SBC) – especially in the Actinomycetes (Weber et al., 2015), became a common feature. Table 1 shows novel compound discoveries solely from gene mining approaches. This has been further propelled by the rapid improvement in NGS technologies in the last decade, dramatically lowering the cost of DNA sequencing, putting the power of bacterial WGS in the hand of individual laboratories.

Table 1. Examples of novel compounds discovered by mining genomes for specific biosynthetic genes. Here are specific examples of how genome mining is becoming commonplace for novel compound identification. Included are examples where identification of the biosynthetic gene clusters preceding identification of novel compound. Highlighted in pink of compounds identified from model organism S. coelicolor, which had been studies for decades.

Compound	Cluster type	Means of identification	Reference
Asperfuranone	Type I polyketide	Searching for PKS genes in WGS	Chiang et al., (2009)
Stambomycins		Searching for PKS genes in *S. ambofaciens*	Laureti et al., (2011)
Aureusimines	NRPS	Searching for NRPS genes in *S. aureus*	Wyall et al., (2010)
Coelichelin		**Searching for NRPS genes in *S. coelicolor***	**Lautru et al., (2005)**
Poeamide		Searching for NRPS genes in *Pseudomonas poae*	Zachow et al., (2015)
Orfamide		Searching for NRPS genes in *P. fluorescens*	Gross et al., (2007)
Thanapeptin		Searching for NRPS genes	Van Der Voort et al. (2015)
Venezuelin	Lantipeptides	Searching for genes with specific domains	Goto et al., (2010)
Streptocollin		Searching for genes with specific domains	Iftime et al., (2015)
Cembrane	Terpenoids	BLAST homology searches	Merguro et al., (2013)
Kolavelools		BLAST homology searches	Nakano et al., (2015)
Stellatic acid		BLAST homology searches	Matsuda et al., (2015)
Sesquiterpene (+)-epi-isozizaene		**Searching for SMC in *S. coelicolor***	**Lin et al., (2006)**

2.1.3 Advent and impact of Whole Genome Sequencing

The information of biochemical and hereditary properties conferred by the DNA sequence of an organism underpins much of today's scientific knowledge; this is due to the advent of relatively cheap and fast DNA sequencing technologies. The primary break-through came with the introduction of Sanger Sequencing in 1977, and served as the blue print for today's technologies. Sanger Sequencing utilises the classical chain-termination method; put simply, the DNA code is read by identifying the incorporation of a modified fluorescent ddNTPs, which has blocked the DNA polymerase extension of the DNA strand. This led on to the development and implementation of increasingly automated approaches, leading to the first crop of commercial DNA sequencing machines. The advent of Sanger Sequencing was undoubtedly revolutionary; however, limitations arose, specifically related to the time consuming and tedious cloning steps necessary. The year of 2006 marked a turning point in sequencing abilities; an explosion of new methods, techniques and protocols were revealed with the arrival of Next Generation Sequencing (NGS) technologies which quickly overcame previously encountered issues. Typical approaches of NGS are: sequencing by synthesis (Illumina, Pacific Biosciences, Ion Torrent); Nanopore (Oxford Nanopore); Pyrophosphorolysis (Base4), and Sequencing by ligation (SOLiD, Complete Genomics). There are many excellent reviews (Levy & Myers, 2016) which explain each sequencing methodology in detail, and so this will not be discussed here. However, it should be noted that NGS platforms can be divided by three axes of sequencing approach: (i) single molecule detection per reaction, well or sensor (Pacific Biosciences) or detection of clonally amplified DNA (Illumina, Ion Torrent); (ii) the use of optical detection to make the sequencing base calls (Illumina, Pacific Biosciences, Ion Torrent and Oxford Nanopore), and (iii) the use of polymerase or ligation process to drive a sequencing by synthesis reaction (Illumina, Ion Torrent, Pacific Biosciences and Roche 454). Each approach has similarities and distinctions relative to the others depending on the chemistry and detection method utilised; these lead to a spectrum of capabilities and applications. The differences in each platform has resulted in much research into their comparison under similar conditions. Comparisons are typically done using two specifications: the number of reads produced in a single run, and the length of those reads; with other metrics such as sample preparation time and cost, being much harder to compare between platforms. Table 2 Shows the currently marketed

NGS compared by these specifications. The application of the sequencing data dictates which approach is best suited; for example, sequencing by ligation (SOLiD) struggles to resolve areas of palindromic repeats and would be a poor choice for bacterial genomes which are known to contain these complex regions (Utturkar et al., 2014). Further, sequence by synthesis methods, such as Illumina, require large amounts of DNA template and so are not well suited for metagenomic analysis. Recently, deviation of sequencing approaches can be seen as platforms become specialised and tailored, to a given purpose. Despite this diversification, all share one major limitation: due to a reduction in read length from ~1 to 0.3 Kb, the sheer volume of reads produced per run has increased dramatically. This imposes a progressively high demand on the bioinformatical capabilities responsible for processing the raw data into clinically actionable knowledge. The computational ability to analyse the data generated from NGS platforms is advancing; however, it remains largely over eclipsed by the ability of NGS machines. To challenge this bottle-neck, just as sequencing approaches are becoming increasingly purpose-designed, it is vital to test and compare bioinformatical approaches in an application-specific manner to produce refined pipelines.

Table 2. Comparison of currently available NGS platforms. Method of sequencing, typical read length, accuracy and typical applications. Data taken from Mardis, (2017). Highlighted in blue is the MiSeq technique which was chosen for work here due to a compromise on price, accuracy and speed.

	Method	Sequencing by	Read length (bp)	Accuracy of single read	Reads per run	Time per run	Cost per 1 million bases ($)	Applications
Pacific Biosciences	Single-molecule real-time	Single molecule	10,000-15,000	87%	50,000	30 min – 4 hrs	0.06	Small scale, specific research questions / resolving high GC regions
Life Technologies	Ion Torrent	Synthesis	< 400	98%	> 80 million	2 hours	1	Bacterial genomes / Large sequencing projects
	SOLiD	Ligation	80-100	99.9%	1.2-1.4 billion	1-2 weeks	0.13	Metagenomics
Illumina	**MiSeq**	**Fluorescence**	**50-600**	**99.9%**	**1-25 million**	**1-11 days**	**0.05-0.15**	**Bacterial genomes / Large sequencing projects**
Roche	Pyro 454	Synthesis	700	99.9%	1 million	24 hours	10	Metagenomics
	Sanger	Chain termination	400-900	99.9	N/A	20 min – 3 hrs	2400	Small scale, specific research questions / resolving high GC regions

2.1.4 Next generation bacterial sequencing and assembly

The introduction of second-generation sequencing technologies has resolved the traditional Sanger sequencing paradigm. However, the intrinsic complexity of genomes, even that of small microbes, means it remains a challenge to generate complete assemblies without laborious laboratory work, such as PCR gap closing (Finotello et al., 2011; Quail et al., 2012). The incidence of long repeat regions and short read length characteristic of bacterial genomes and NGS platforms respectively, often lead to assemblies being left as unfinished, fragmented drafts (Ferrarini et al., 2013; Koren et al., 2013). Attempts to resolve these issues have been made, principally with the introduction of so-called 'third generation' sequencing – such as PacBio – which have significantly longer read length; however, these too have inherent shortcomings such as low accuracy (~15% error rate) (Eid et al., 2009). This has led to production of hybrid approaches combining reads from two or more sequencing platforms to allow optimal sequence assembly; these reads are combined using hybrid programs such as ALLPATHS-LG (Ribeiro et al., 2012), PacBio correct reads (PBcR) pipeline (Bashir et al., 2012) and SPAdes (Prjibelski et al., 2014). For example, the authors of ALLPATHS-LC (Ribeiro et al., 2012), proposed a unique methodology which is incorporated into ALLPATHS-LG to use Illumina paired-end reads of two libraries, in addition to PacBio long reads. It was demonstrated by Ribeiro et al., (2012) that this pipeline was able to produce nearly perfect bacterial assembly; despite this, few bacterial genomes have been assembled in this way likely owing to the cost associated with a such an extensive sequencing approach (Shibata et al., 2013; Ku et al., 2012).

To circumvent costly multiple platform implementation, work is also focusing on improving bioinformatical capabilities and specifically the intelligent treatment of reads to allow optimal data extraction. Review of previous assembly pipelines and the quality of the resulting assemblies, allows the effectiveness of each program in a bacterial genome application to be evaluated, hence, enlightening future work. To aid in pipeline selection here, recent work which utilised the Illumina MiSeq platform specifically was reviewed and is summarised in Fig.8. In line with the typical NGS output, raw data from Illumina Miseq is two collections of reads – the paired-end (PE) and mate-pair (MP) library; the two are not compatible with each other and require different pre-processing treatments. Pre-processing typically involves quality

control of reads and trimming of adapter content; many programs are available for PE processing, with less emphasis been put on MP read processing. This is due to MP's being harder to utilise, and has resulted in previous research largely disregarding the library, or incorporating it at the very end of assembly. The wealth of information held within the MP reads has recently been realised, and a tide-change in pipeline has resulted in additional effort into producing programs which can utilise the data. Taking advantage of this sequence information, many programs – such as SPAdes 3.5 (Bankevich et al., 2012) – are incorporating MP libraries much earlier in the assembly process. All assemblers differ in intensiveness and assembly speed but typically all use de Bruijn graph algorithms. Put simply, de Bruijn graphs are directed graphs representing overlaps between sequences; to be generated, the amount of overlap between reads must be dictated – this is set by the k-mer length, to result in the production of a vertex by reconstructing a string from the set of k-mers. Recently, advances have been made into producing algorithms which uses a set of pairs of k-mers (k-bimers), utilising the data present in the MP library. As mentioned, the potential of this approach is arguably the most poorly explored stage of assembly; this was due to programmers aiming to have a fixed distance between reads – unrealistic due to the variation in the read length characteristic of NGS machines. This made bridging the gap between theoretical idea to practical implementation difficult; this was resolved with the introduction of SPAdes which uses k-bimer adjustment which reveals exact distances for the vast majority of the adjusted k-bimers. It has been demonstrated that this type of approach, utilising the MP libraries can help close gaps, resolve palindromic regions, and increase contig length threefold (Greninger et al., 2017). The final task is locating the target SMC which is typically done by either directly searching (TBLASTN search against a database generated from the assembled genome scaffolds) or using a SMC searching program (Prokka (Seeman, 2014), RAST (Aziz et al., 2008), AntiSMASH (Medema et al., 2012)) to identify genes which encode backbone enzymes that synthesise the specific class of compounds which correspond to the target SM. As shown in Fig. 8 there are many programs which use different algorithmic approaches for bacterial genome assembly; if the potential of NGS platforms is to avoid bottleneck, the applicability of these programs in non-specialised settings must be scrutinised.

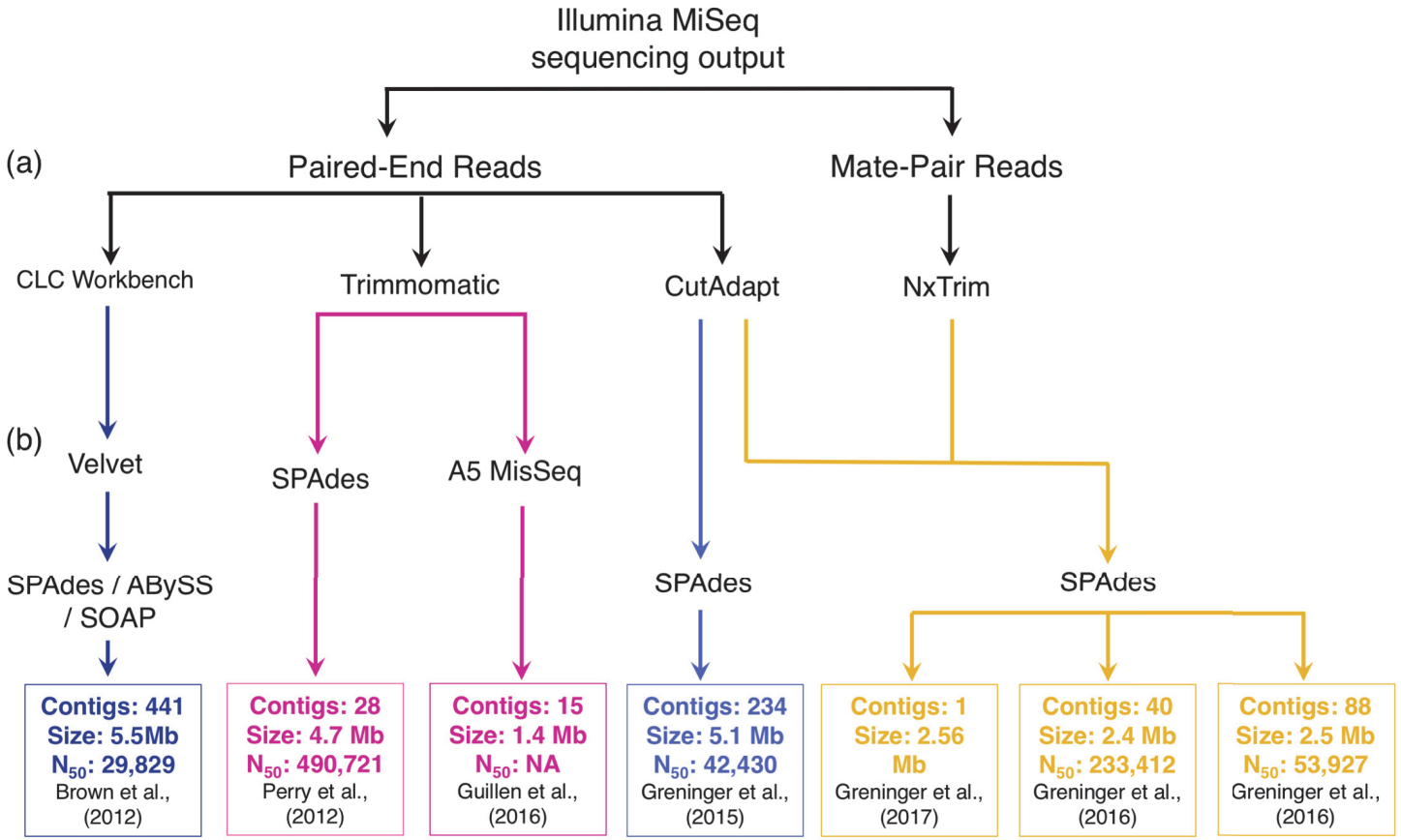


Figure 8. Previously utilised assembly pipelines by prior researchers, and the resultant assembly. (a) the libraries utilised, **(b)** the programs used and analysis of the resulting assembly. For each assembly method, the number of contigs the genome was assembled into, the size of the genome and the N_{50} was recorded. It can be clearly seen that incorporating both the PE and MP library data at an early stage, gives a superior assembly.

Here, using the *Verrucosispora sp.* str. MG37 genome as an example, we test and compare many leading bioinformatical programs to describe the most efficient pipeline for microbial genome assembly and annotation, major issues encountered and how they were resolved. The principal objective was to identify the *ppb* cluster and outline the biosynthetic route resulting in proximicin, and different SMC searching programs were used and compared; in addition to this, the applicability of NGS genome assembly and annotation programs to be utilised in non-centralised, individual laboratory was assessed.

Chapter 2. Sequencing, assembly & annotation of *Verrucosispora sp.* str. MG37, and identification of the biosynthetic gene cluster responsible for proximicin biosynthesis

2.2 Materials and Methods

2.2.1 Media and Reagents

GYM Streptomyces medium was used for routine growth of *Verrucosispora* strains: 4.0 g Glucose, 4.0 g Yeast extract, 10.0 g Malt extract, 2.0 g $CaCO_3$, 12 g Agar (optional) and 1L distilled water. The pH was adjusted to 7.2 prior to addition of agar, and $CaCO_3$ excluded if medium was used, followed by autoclaving. All DNA fragments were visualised by running on 2% agarose gels stained with ethidium bromide, unless otherwise stated. Explanation of output terms is explained in Appendix B.

2.2.2 I*dentification of putative proximicin biosynthetic (ppb) gene cluster in Verrucosispora species and initial investigation into adenylation domains.*

Identifying likely genes to be present in the ppb cluster

The component chemistry of proximicins A, B and C was analyzed, common structures identified, and differences highlighted. Possible biosynthetic routes were hypothesized and genes which would be required predicted – this information was used to identify gene cluster characteristics that should be prioritized. This was done manually, with support from previous di-substituted heterocycle biosynthesis research (c.f. congocidine biosynthesis).

Identifying the ppb cluster

Mining of the published *V. maris* AB18-032 genome using a secondary metabolite gene cluster analyzer – antiSMASH (Blin et al., 2013) – was done to reveal NRPS-containing biosynthetic gene clusters. Clusters with the characteristics previously identified to be likely present for proximicin production were prioritized, with those likely to produce a compound of similar size to proximicin. Prediction of adenylation

domain substrates was done using Maryland NRPS analyzer (Bachmann & Ravel, 2009) and the function of non-NRPS proteins by BLAST (Altschu et al., 1990) and HHPred (Söding et al., 2005). Finally, 3D protein prediction was done using iTASSER (Zhang, 2008) and PyMOL (DeLano, 2002).

Initial bioinformatic analysis of ppb cluster adenylation domains

Adenylation domains were initially analysed using Bioedit (Hall, 1991) to identify core domains required for adenylation domain function. This was done by sequence alignment with 11 described functional A-domains. Core consensus sequences were highlighted, and their presence / absence / modifications from consensus, were noted and potential enzymatic outcomes from these deviations hypothesised.

Analyzing adenylation domain activity

Programs were used to initially identify potential substrates of the adenylation domains identified in the *ppb* cluster; these programs include: Maryland PKS/NRPS analysis (Bachmann et al., 2009); NRPSpredictor2 (Röttig et al., 2011) and Prism (Tunchag et al., 2011; Baspinar et al., 2014); this was also done by comparing known a selection of 10 known adenylation domains to identify present/absent core sequence domains.

Route to proximicin biosynthesis outlined

From the proposed functions of the genes within the putative proximicin cluster, various biosynthetic routes were outlined. This was done manually, using the similarities the genes in *ppb* have to give potential routes to a 2,4-disubstituted heterocycle containing compound.

2.2.3 Genomic sequencing of Verrucosispora sp. str. MG37

Genomic DNA extraction

Genomic DNA was extracted from *Verrucosispora* sp. str. MG37 using the cetyltrimethyl ammonium bromide (CTAB) method. 30 mL cultures were grown up in GYM media for 16 hours and harvested at 16,000 X g and re-suspended in 5 mL TE25S buffer (25 mM Tris-HCl pH. 8; 25 mM EDTA pH. 8 and 0.3 M sucrose) lysozyme added to final concentration 2 mg/mL and incubated for 1 hr at 37°C. Proteinase K and SDS added to final concentration 0.18 mg/mL and 0.5 %

respectively, and incubated for 1 hr with occasional inversion at 55°C. NaCl was added to final concentration of 0.8 M and mixed; CTAB/NaCl (10 %CTAB in 0.7 M NaCl) was added and the mixture was incubated at 55°C for 10 mins followed by cooling to 37°C. Chloroform/isoamyl alcohol added and mixed by inversion for 30 min then centrifuged at 13,500 X g at 20°C for 15 mins. The supernatant was decanted and 0.6 volumes of isopropanol added and mixed; after 3 min DNA was spooled, rinsed in ethanol and air dried before being dissolved in 1 mL TE buffer (1mM Tris-HCl; 1 mM EDTA pH. 8) at 55°C. Successful gDNA extraction was confirmed by agarose gel electrophoresis.

Checking presence of ppb genes

HPLC-purified primers were designed to amplify *ppb* genes chosen to span the entire cluster identified in *V. maris* AB18-032: *ppb* genes *ppb090* (Primer #: 1 & 2); *ppb110* (Primer #: 3 & 4); *ppb120* (Primer #: 5 & 6); *ppb155* (Primer #: 7 & 8) and *ppb180* (Primer #: 9 & 10). The extracted gDNA was used in a PCR mixture (25 µL) contained 2XMyfi™ High-Fidelity polymerase master-mix (Bioline), primers (0.5 µM each), template DNA (25ng) and DMSO (3%). The PCR was performed using a specified MyFi™ Thermal Cycling Procedure: 1 cycle, 95°C for 60 s; 30 cycles, 95°C for 15 s, 71°C for 15 s and 72°C for 3:00 min; and 1 cycle 72°C for 15 min. PCR was checked using gel electrophoresis, 4 µL of reaction mix was ran on a 1% agarose gel. Products were E-Gel purified and quantified using Nanodrop.

Genome sequencing of Verrucosispora sp. str MG37

For shotgun sequencing (Paired End) libraries, DNA was sheared into 100 - 1500 bp range in a microTUBE using a focused-ultrasonicator (Covaris, USA) and quantified on a Qubit instrument (Thermofisher, USA). Sheared DNA was end-repaired, A-tailed, adapter ligated and amplified with a Kapa Hyper library construction kit (Kapa Biosystem, USA). For Mate-Pair libraries, DNA was tagmented, circularized, fragmented, enriched and adapter ligated using a Nextera Mate Pair sample preparation kit (Illumina, USA). The circularization adaptor for mate-pair library was 5'-CTGTCTCTTATACACATCT-3'. Adapters for sequencing both paired end and mate pair libraries were: read 1: 5'-AGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNATCTCGTATGCCGTCTTCTGCTTG-3' where (NNNNNN = 6 nt index) and 5'-

AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATC AT-3'. Index sequences for the paired end library was GCCAAT, and ACAGTG for the Mate Pair library. The libraries were pooled and quantified by qPCR prior to sequencing on a MiSeq DNA instrument; one MiSeq V3 flowcell for 311 cycles and a MiSeq 600-cycle sequencing kit, was used to generate sequences from each end of the fragments. Fastq files were generated and demultiplexed with bcl2fastq v1.8.4 conversion software (Illumina, USA). All the above steps were done by the High-Throughput Sequencing and Genotyping Unit at the Roy J. Carver Biotechnology Centre, University of Illinois.

2.2.4 *Verrucosispora* sp. str MG37 genome assembly and annotation

A bioinformatics pipeline was employed based on previous work in *de novo* microbial genome assembly and annotation, and is outlined in Fig. 9.

Quality check of raw sequencing data

Both Mate-Pair and Paired-End read libraries were checked using FastQC (Andrews, 2010), this program analyses the sequencing to give an output file detailing: total sequence length, flagged poor quality sequences, average sequence length, GC percentage, at what point in the reads the phred score drops below 20, adapter and Kmer content.

Trimming of Paired-End library

Trimmomatic (Bolger et al., 2014) and Cutadapt (Martin, 2011) were used to trim the paired end library. Trimmomatic is a command line tool used to trim and crop Illumina data, as well as remove sequencing adapters. Paired end mode was used to maintain correspondence of read pairs and to utilize the additional information contained in paired reads to better find adaptors and DNA fragments introduced by the library preparation process. Within the Trimmomatic program there is an array of steps (Appendix C). Four output files were created: Forward Reads which remained Paired; Reverse Reads that remained Paired; Forward Reads which are Unpaired and Reverse Reads which are Unpaired. The unpaired reads are due to one out of the pair being removed. Reads which failed to meet the specified criteria, were deleted; no library was made with them. Cutadapt is a command line program which finds and removes adapter sequences, primers, poly-A tails and

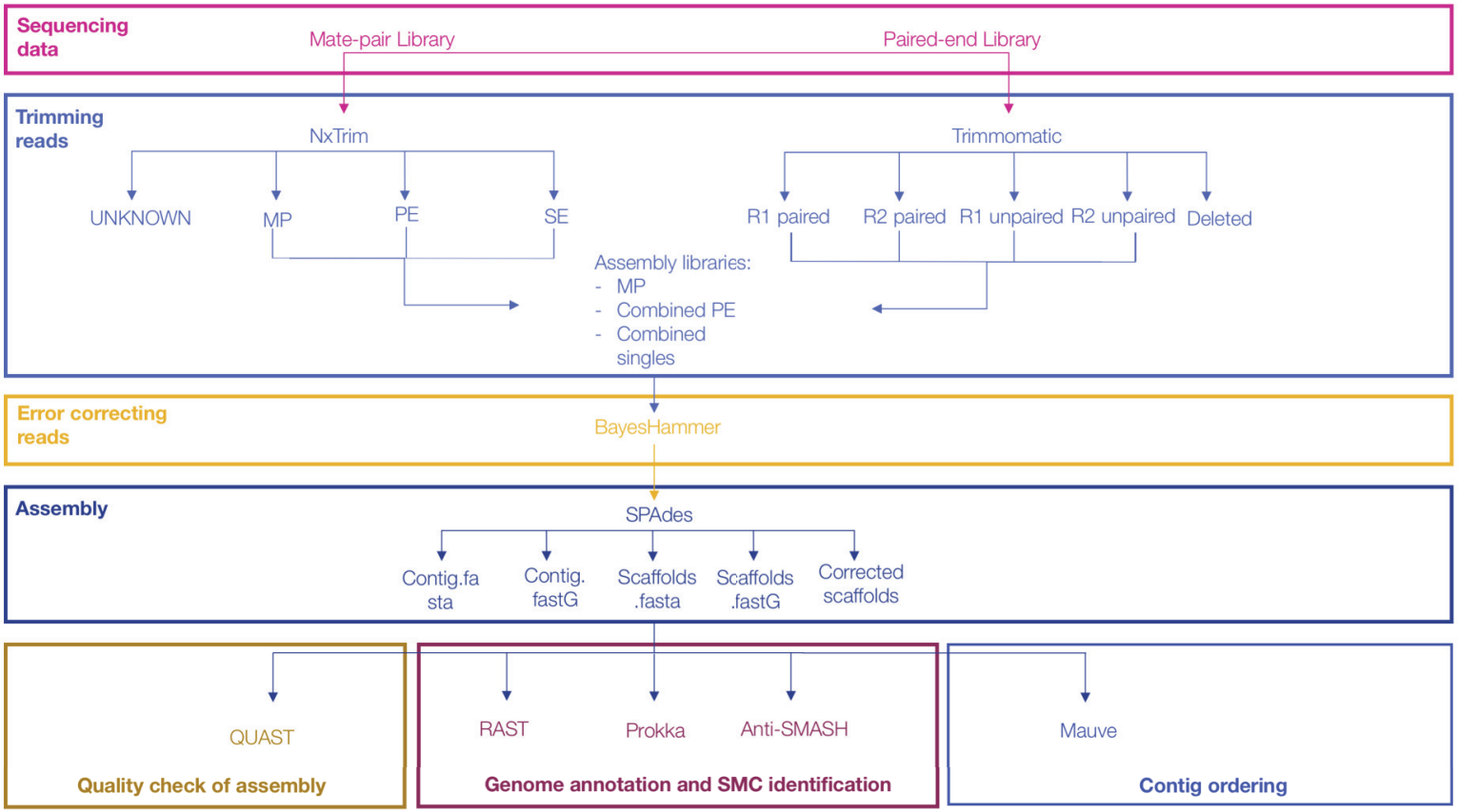


Figure 9. The bioinformatic pipeline employed in the research described here. The process is split into five main stages, separated by boxes. Each stage shows the program used and the library outputs generated and how they were processed and progressed onto the next stage of assembly. MP = Mate Pair reads, PE = Paired End reads, R = Read. The specific use of each program is noted in the text.

other unwanted sequences. Cutadapt searches for multiple adapters in a single run, either discarding or trimming reads where adapters are found. Low quality ends of reads are trimmed. The programs used within Trimmomatic are outlined in Appendix C.

Trimming of Mate-Pair library

NxTrim (O'Connel et al., 2014) is an optimized trimming program designed specifically for Mate-Pair reads. It can interpret each potential scenario resulting from the fragmentation of the 6-8 Kb fragments produced during Illumina Mate-pair sequencing, and splits these into four libraries: Mate-pair; UNKNOWN; Paired-end and a set of single reads. This allows the use of not just the typical Mate-pair reads (as is the case with other trimming software), but also reads which are Paired-end and singletons. This tool allows more information to be retained from the trimming steps to aid assembly. The four output files are the standard Mate-Pair, Unknown, Paired Ends and a set of single reads, and how each is created is outlined in Fig. 14.

Quality check after trimming steps

FastQC was again used, to determine the quality of the reads after the trimming steps. Primarily to ensure all adapter sequences have been removed and all sequences with a low quality (Phred score >20) has been removed, this should increase the average quality of reads but lower the number of reads.

Assembly of reads

SPAdes 3.5.0 (Bankevich at al., 2012), an iterative short read genome assembly program, was chosen as the assembly tool, this has the error correcting program BayesHammer (Nikolenko et al., 2013) within it. BayesHammer and SPAdes, although possible to run together, were run separately (Appendix D). As the input data was a multi-cell data set with read length for each pair >250, K-mer lengths were altered to take advantage of the long read length. SPAdes works iteratively to calculate the optimal read length from a set specified; for this genome assembly K-mer lengths: 21, 33, 55, 77 and 127 were chosen. Increasing the K-mer length increases the chances of mismatches and so SPAdes was run in the 'careful' mode to reduce the number of mismatches in the final contig. For both, Paired end, Mate-pair and single reads are the input files, accounting for orientation in the command

line. The output files are five directories: reads corrected by BayesHammer; resulting contigs; resulting contigs in FASTG format; resulting scaffolds and resulting scaffolds in FASTG format. The workflow of SPAdes is summarized in Appendix D.

Assessment of assembly quality

QUAST (Gurevich et al., 2013) is a command line program that evaluates the assemblies of genomes by comparing it against a known and annotated genome. Both the assembled *Verrucosispora* sp. str. MG37 and *V. maris* AB18-032 genomes were input as FASTA files. The output is an output report in a .html file with interactive plots on quality. All inputs were left as default.

Contig ordering

MAUVE (Darling et al., 2004) uses highly similar sub-sequences to employ an anchor alignment approach, which only requires modest computational resources. To avoid problems encountered with highly repetitive regions and contig ordering, Mauve employs multiple maximal unique matches (MUMs) of some minimal length k as alignment anchors. These 'MUM's' are exactly matching subsequences shared by two or more genomes that occur only a single time in each genome and are bounded either side by mismatched nucleotides. Mauve then uses a recursive anchoring strategy that progressively reduces k, searching for smaller anchors in the remaining unmatched regions. Input files were *Verrucosispora* sp. str, MG37 genome scaffolds and the *V. maris* AB18-032 genome. The output is an interactive graphical platform, allowing manual rearrangement of contigs. The rearranged contigs were output as a .fasta file.

Genome annotation and identification of secondary metabolite gene clusters

Four programs – RAST (Aziz et al., 2004), Prokka (Seeman, 2014), antiSMASH (Blin et al., 2013) and Prism (Tunchag et al., 2011; Baspinar et al., 2014); - were used to allow comparison of results:

RAST identifies protein encoding, rRNA and tRNA genes, attempts to assign function, and to reconstruct the metabolic network of the given genome. RAST uses the data and procedures established with the SEED framework to allow high quality gene calling annotation (Overbeek et al., 2005). The SEED project is an open source

resource which focus is to develop curated genomic data. The annotation is done via subsystems – functional roles that an annotator has decided should be thought of as related, e.g. functional roles or a class of proteins. This allows high quality gene calling and functional annotation. The annotated genome can be downloaded in GenBank or Fasta format, or browse the genome in the comparative environment of SEED-viewer.

Prokka uses external feature prediction tools to identify the coordinates of genomic features within contigs. Within Prokka five tools are used which predicts different features: Prodigal (coding sequence); RNAmmer (ribosomal RNA genes); Aragorn (transfer RNA genes); SignalP (signal leader peptides) and infernal (non-coding RNA). Protein coding regions are identified in two stages: Prodigal identifies the region and coordinates of the candidate gene, however, instead of using the traditional method of comparing this sequence to a large database of known sequences and finding the most closely matched, Prodigal uses a hierarchal method to start comparison with a small, trustworthy database, moving onto domain specific and then curated models of protein families. The antiSMASH program is a genome mining tool specifically designed to identify antibiotic and secondary metabolite gene clusters.

Prism is a computational resource for the identification of biosynthetic gene cluster, along with the predication of genetically encoded NRP's; it implements novel algorithms which render it uniquely capable of making comprehensive genome guided chemical structure predictions.

Comparative genomics

The draft genomes of *Verrucosispora* sp. str. MG37 and the whole genome of *V. maris* AB18-032 were aligned using ACT (Carver et al., 2005) software after running a pairwise blastn search. Mauve (Darling et al., 2004); was used to reorder the contigs of *Verrucosispora* sp. str. MG37 using *V. maris* AB18-032 as the reference genome. Regions of difference were manually identified. The *ppb* cluster was examined to identify any genetic differences which could account for differential production of proximicins observed between the strains. Genome illustrations were done using Circos (Krzywinski et al., 2009). Species identification based on genome

sequence was done as defined by the average nucleotide identity between *Verrucosispora* sp. str. MG37 and *Verrucosispora maris* AB18-032 (Goris et al., 2007) and by Genome-to-Genome distance calculation (Auch et al., 2010)

2.2.5 Confirming the ppb cluster

Identifying potential proximicin biosynthetic cluster

Of the clusters present in both *Verrucosispora* sp. str. MG37 and *V. maris* AB18-032, any clusters not common to both were excluded, then any which did not contain NRPS-adenylating proteins. The remaining clusters were ranked depending on their similarity to the predicted proximicin clusters previously outlined. This was done to confirm that the previously determined *ppb* is the most likely candidate in both producers.

Chapter 2. Sequencing, assembly & annotation of *Verrucosispora* sp. str. MG37 and identification of the biosynthetic gene cluster responsible for proximicin biosynthesis

2.3 Results

2.3.1 Putative proximicin cluster in V. maris AB18-032

Identifying genes likely to be present in the ppb cluster

Proximicins all contain a two 2,4-disubstituted furans, with B and C containing a tyramide and tryptamine (likely to be of tyrosine or tryptophan origin), respectively. They are likely synthesized by an NRPS route (amide bond between non-proteogenic amino acids), and it is possible that the two furan groups are added iteratively by a single A-domain (c.f. congocidine biosynthesis); also, if an A-domain does activate furan containing precursors, it would have non-canonical active site residues – as furans are not α-amino acids. If this is the similar to the biosynthetic route to congocidine, genes likely to be included in the cluster are: at least two NRPS adenylating enzymes, at least one having a non-canonical active site sequence (depending on whether the furan group is added iteratively or by two separate A-domains). This would be accompanied by a tryptophan and/or tyrosine activating A-domain, as well as the cognate T-domains and a TE domain to release the peptide. If this is the route utilized, it is likely the cluster will have similarities to congocidine biosynthetic cluster, as it is responsible for the production of a similarly structured compound. An alternative hypothesis is that a linear molecule could be initially activated and condensed, with cyclization to form the furan moieties occurring post peptide formation; this would negate the requirement for furan precursor synthesis. In this case, it would be a linear residue accepting A-domain, still accompanied by the tryptophan and/or tyrosine activating A-domain(s), but with the presence of a singular or multiple cyclizing enzymes. It is also likely that an MbtH-like protein will be present, as these proteins are often required for correct functioning of NRPS enzymes.

Genome mining of Verrucosispora maris AB18-032

Genome mining of *V. maris* AB18-032 showed many biosynthetic gene clusters (19 in total), five of which are NRPS systems (Table 3). Different prediction programs gave different outputs for the boundaries of the clusters, for example antiSMASH suggested a single cluster, whereas PRISM suggested two individual smaller clusters. This highlights the necessity for hand curation. Consequently, one cluster was identified as the putative proximicin biosynthetic cluster. This cluster was one that was predicted to be part of a larger cluster by antiSMASH, however, annotation of this cluster (see below) showed similarities to the gene clusters responsible for congocidine (Al-Mestarihi et al., 2015) and pacidamycins (Rackham et al., 2010, Zhang et al., 2010) biosynthesis – specifically all having dissociated atypical NRPS module systems – and so its prediction as one cluster (containing multiple NRPS genes) is likely due to its proximity to a second NRPS gene cluster (PRISM identified them as separate gene clusters). One similarity between the *ppb* cluster and the congocidine gene cluster is the presence of an adenylation domain with an unusual amino acid substitution in the residues comprising the substrate-recognition motif – Ser_{235} to Asp_{235} in Ppb120 and Cgc18, respectively. Ser_{235} is invariant in adenylation domains that select and activate α-amino acids making it highly likely that both activate a non-proteogenic amino acid; at the time of analysis, Cgc18 was predicted to activate a heterocyclic γ-amino acid (4-aminopyrrole-2-carboxylate) a substrate differing from the predicted proximicin precursor by the presence of a pyrrole, rather than furan, ring. This further supported the identification of ppb as the proximicin biosynthetic gene cluster. The components of the putative proximicin gene cluster and their predicted function are given in Table 4 and the genetic organization is shown in Figure 10.

Table 3. Major natural product gene clusters of *Verrucosispora* AB18-032

Cluster Type*	Genome location (bases)		Most similar known cluster biosynthetic gene cluster
	start	end	
Terpene	191982	212893	Sioxanthin
Terpene	278534	299589	Phosphonoglycans
NRPS-fatty acid-Arylpolyene	489000	581354	Kedarcidin
T1pks-fatty acid	2554963	2631233	Abyssomicin
T1pks	2679427	2778878	Streptazone E
T2pks	2948612	3007502	Xantholipin
NRPS-T1pks	3021143	3109012	Sporolide
NRPS	3194386	3247461	Arsenopolyketides
Bacteriocin-Lantipeptide	3439443	3502185	Desferrioxamine B
Terpene-Thiopeptide-Lantipeptide	3630549	3691043	n/a
NRPS	**3894106**	**3992199**	**Pacidamycin**
Lantipeptide-saccharide	4044111	4086687	SapB
T2pks	4255327	4299256	Spore pigment
T1pks-NRPS-fatty acid	4459178	4574561	Maduropeptin
Terpene-Bacteriocin	4750901	4778203	Lymphostin
Terpene	4913335	4934267	n/a
T3pks	6212354	6253403	Alkyl-O-Dihydrogeranyl-Methoxyhydroquinones

* As defined by Cimermancic et al. 2014, NRPS = Non-ribosomal peptide synthetase, T1pks = Type 1 Polyketide synthase, T2pks = Type 2 Polyketide synthase. Putative proximicin biosynthetic gene cluster (*ppb*) is shown in blue.

Table 4. Genes present in the putative proximicin biosynthetic cluster. NRPS and related proteins are shown in blue; regulatory genes in pink and other genes of interest in light blue.

Gene	Product size (aa)	Identity	Function	Present in both strains
ppb055	1951	99%	Short chain alcohol dehydrogenase	Y
ppb060	184	99%	Transcriptional regulator	Y
ppb065	323	100%	Hydrolase/ reductase	N
ppb070	211	64%	Integral membrane protein	Y
ppb075	337	99%	Membrane protein	Y
ppb080	319	98%	Transcriptional regulator	Y
ppb085	942	99%	DNA binding transcriptional activator of the SARP family	Y
ppb090	206	100%	Cyclodehydratase	Y
ppb095	340	99%	Transcriptional regulator	Y
ppb100	342	99%	DNA binding protein	Y
ppb105	442	100%	Transcriptional regulator	Y
ppb110	596	99%	NRPS module	Y
ppb115	342	100%	DAHP synthase	Y
ppb120	764	99%	NRPS module	Y
ppb125	72	100%	MbtH protein	Y
ppb130	240	99%	Hydrolase	Y
ppb135	286	100%	Orotidine-5'-phosphate decarboxylase	Y
ppb140	1160	99%	NRPS modules	Y
ppb145	123	99%	SAM-dependent methyl-transferase	Y
ppb150	437	100%	Permeases of the major facilitator family	Y
ppb155	353	99%	Carbomoyl-phosphate synthase large subunit	Y
ppb160	422	99%	Pyridoxal phosphate-dependent aminotransferase	Y
ppb165	247	99%	Thioesterase in NRPSynthesis	Y
ppb170	316	99%	Nonheme iron dioxygenase	Y
ppb175	344	99%	Threonin aldolase	Y
ppb180	757	99%	Cysteine synthase	Y
ppb185	299	99%	Protein involved in propanediol utilization and related proteins	Y
ppb190	512	99%	Argininosuccinate lyase	Y
ppb195	938	99%	NRPS module	Y
ppb200	258	99%	SAM-dependent methyl transferase	Y
ppb205	409	99%	Cytochrome P450	Y
ppb210	836	99%	NRPS module	Y
ppb215	305	99%	Transcriptional regulator	Y
ppb220	622	99%	NRPS module	Y

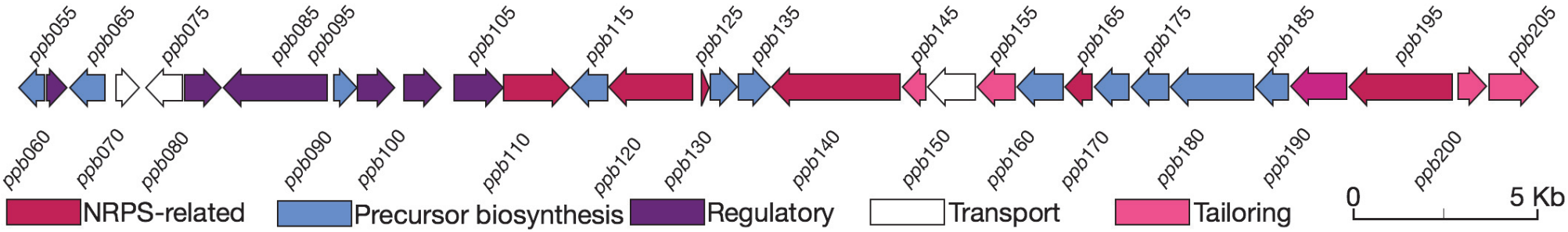


Figure 10. Genetic organization of the putative proximicin biosynthetic gene cluster (*ppb*). Gene names correspond to those published for the *V. maris* AB18-032 genome.

Confirmation of ppb cluster presence in V. sp. str. MG37 by PCR

If the *ppb* cluster was correct, it would also need to be present in the genome of *Verrucosispora* sp. str. MG37. This was successfully confirmed using *V. maris* AB18-032 as the control (Fig. 11). Genes *ppb* genes *ppb*090; *ppb*110; *ppb*120; *ppb*115 and *ppb*180 with the respective expected base pair fragment sizes: 528bp; 993bp; 774bp; 1008bp and 1008bp were shown to be present in both proximicin producers.

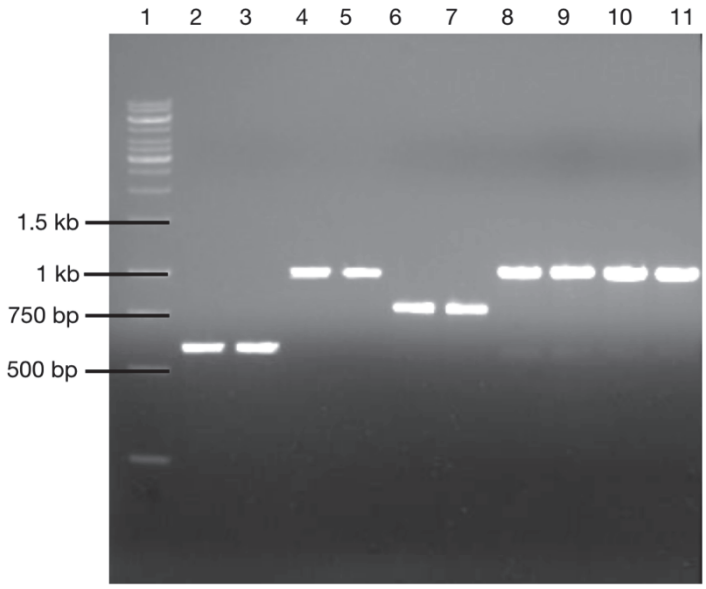


Figure 11. PCR amplicons of *ppb* cluster spanning genes in *Verrucosispora* strains. 1% agarose gel showing PCR products spanning of *ppb* genes first in *V.* sp. st. MG37 and then *V. maris* AB18-032 **1.** Generuler 1Kb **2 & 3** *ppb090* **4 & 5** *ppb110* **6 & 7** *ppb120* **8 & 9** *ppb155* **10 & 11** *ppb190* **11.**

2.3.2 Genome sequencing, assembly and annotation of Verrucosispora sp. str. MG37

Genome sequencing

To recover the whole *ppb* cluster from *Verrucosispora* sp. str. MG37, a whole genome sequence approach was used. This approach also allowed for comparative genomics; other biosynthetic gene clusters present in both strains of *Verrucosispora* were identified (see below) in case the *ppb* cluster was later experimentally rejected.

Trimming libraries and quality control of sequence reads

For the paired end library, both Trimmomatic and TrimAdapt programs were used to trim adapter sequences and any reads of low quality. The FastQC report before (raw) and after processing are shown in Table 5. Although both programs use the paired end information (using both reads) to identify where the sequences fall below a quality threshold, and to remove adaptor sequences, TrimAdapt left almost 60,000 more reads available to use in assembly. This is likely due to Trimmomatic only keeping sequences with both a Phred score >20 and a length of >304 bp, where TrimAdapt keeps any sequence of Phred score >20, with a less stringent requirement for length of read. Despite the difference in read number between programs, both very similar outcomes according to FastQC – maintaining a mean quality score >30, and removing 100% of the adapter content. Because of this, the TrimAdapt reads were used for assembly. Nxtrim was successfully used to trim the adapter sequences from the mate pair library and remove any low-quality reads.

Genome assembly of Verrucosispora sp. str. MG37

The genome was successfully assembled into 6 major scaffolds, with the largest contig spanning half the genome (3.01Mb), followed by two slightly smaller contigs at 2.1Mb and 1.2Mb. The quality of assembly was assessed using QUAST, it showed that 6.75Mb (>99%) of DNA could be mapped to the *V. maris* AB18-032 genome, with a low number of miss-assemblies (<200). The majority of these miss-assemblies were matches to entirely different species – and so were considered contamination from either the genome extraction or sequencing process. *Verrucosispora* sp. str. MG37 assembled contigs were successfully reordered according to the reference genome *V. maris* AB18-032 using MAUVE.

Table 5. Analysis of reads before (raw) and after trimming treatments. Two programs: Trimmomatic (blue) and TrimAdapt (pink) were compared to determine which is more suited for microbial genome assembly.

	Raw Read R1	Raw Read R2	After Trimmomatic R1	After Trimmomatic R2	After TrimAdapt R1	After TrimAdapt R2
Total sequence	4951654	4951654	4892495	4892495	4951654	4951654
Sequence flagged as poor quality	0	0	0	0	0	0
Sequence length	310	310	36 – 310	36 - 310	1 - 310	2 - 310
% GC	69	70	70	70	70	71
Mean quality score	36	32	36	35	36	30
Over represented sequences	None	None	None	None	None	None
Base pair in read Phred >20	289	244	>304	>304	>304	229
Adapter content (% of read)	13	9	0	0	0	0

Annotation of Verrucosispora sp. str. MG37 genome

Bacterial genome annotation and secondary metabolite cluster finding programs were used for annotation and mining of the *Verrucosispora* sp. str. MG37 genome. These included: Prokka, RAST, antiSMASH and Prism. After manual rearrangement of the large contigs, the two *Verrucosispora* strains had very similar genomic arrangement, as expected with two such closely related organisms. Indeed, *Verrucosispora* sp. str. MG37 very likely belongs to the species *maris* AB18-032, as the average nucleotide identity (ANI) between the two strains is > 97% (suggested cut-off for placing two strains in the same species is 95%). Furthermore, the genome-to-genome comparison, that is considered more reliable than ANI, gave a > 80% probability that both strains belong to the same species. The secondary metabolite biosynthetic clusters identified in *Verrucosispora* sp. str. MG37 were then compared to those in *V. maris* AB18-032, and corresponding clusters matched (Table 6), in total 21 common clusters were found, with 4 additional undescribed clusters being present in *Verrucosispora* sp. str. MG37. Importantly, there were four NRPS-containing clusters common to both proximicin producers, including the *ppb* cluster identified in *V. maris* AB18-032. Although largely similar in the secondary metabolite clusters present in both, some cluster differences between the two proximicin producers are of note: the abyssomicin cluster present in *V. maris* AB18-032 is identified as not being present in *Verrucosispora* sp. str. MG37 – close inspection of the cluster shows a very low retained similarity with the abyssomicin in *V. maris* AB18-032. This supported previous metabolite studies noting that *Verrucosispora* sp. str. MG37 doesn't produce this compound; and gave confidence the genome assembly was of high quality. Whole genome sequences and their annotated features are given in Fig. 12 and 13 and genome alignment of the two strains in Fig. 13. Alignment of the two genomes suggests a reason for strain differences: there is a large (ca. 700 kb) genome inversion region (red links in center circle Fig. 14), this inversion and subsequent repair is likely responsible for the loss of the abyssomicin gene cluster from *Verrucosispora* sp. str. MG37 (center of inversion).

Table 6. Major natural product gene clusters of *Verrucosispora* sp. str. MG37
Putative proximicin biosynthetic gene cluster (*ppb*) is shown in pink. Gene cluster present in *Verrucosispora* sp. str. MG37 but absent from *Verrucosispora maris* AB18-032 are shown in blue.

Cluster Type*	Genome location (bases)		Present in *V. maris* AB18-032	Most similar known biosynthetic gene cluster
	start	end		
Terpene	444471	465415	Y	Sioxanthin
Terpene	512879	533901	Y	Phosphonoglycans
NRPS-fatty acid-Arylpolyene	736987	852216	Y	Kedarcidin
NRPS	**1376505**	**1421245**	**N**	**Gentamicin**
Lantipeptide	**2578981**	**2599115**	**N**	**Kanamycin**
T1pks	2735950	2839089	Y	Streptazone E
T2pks	2998270	3057162	Y	Xantholipin
Thiopeptide-Lantipeptide-Terpene	**3268391**	**3325749**	**N**	**n/a**
NRPS	3731421	3784604	Y	Arsenopolyketides
NRPS-T1pks	3874671	3980228	Y	Sporolide
NRPS	**3946979**	**4067954**	**Y**	**Pacidamycin**
Lantipeptide-Bacteriocin	3457644	3534582	Y	Desferrioxamine B
Lantipeptide-saccharide	4120838	4163906	Y	SapB
T2pks	4322755	4365279	Y	Spore pigment
Lantipeptide	**4681496**	**4707311**	**N**	**n/a**
T1pks-NRPS-fatty acid	4550380	4664939	Y	Maduropeptin
Terpene-Bacteriocin	4998528	5025833	Y	Lymphostin
Terpene	5182313	5203245	Y	n/a
T3pks	6537542	6578591	Y	Alkyl-O-Dihydrogeranyl-Methoxyhydroquinones

* As defined by Cimermancic et al. 2014, NRPS = Non-ribosomal peptide synthetase, T1pks = Type 1 Polyketide synthase, T2pks = Type 2 Polyketide synthase.

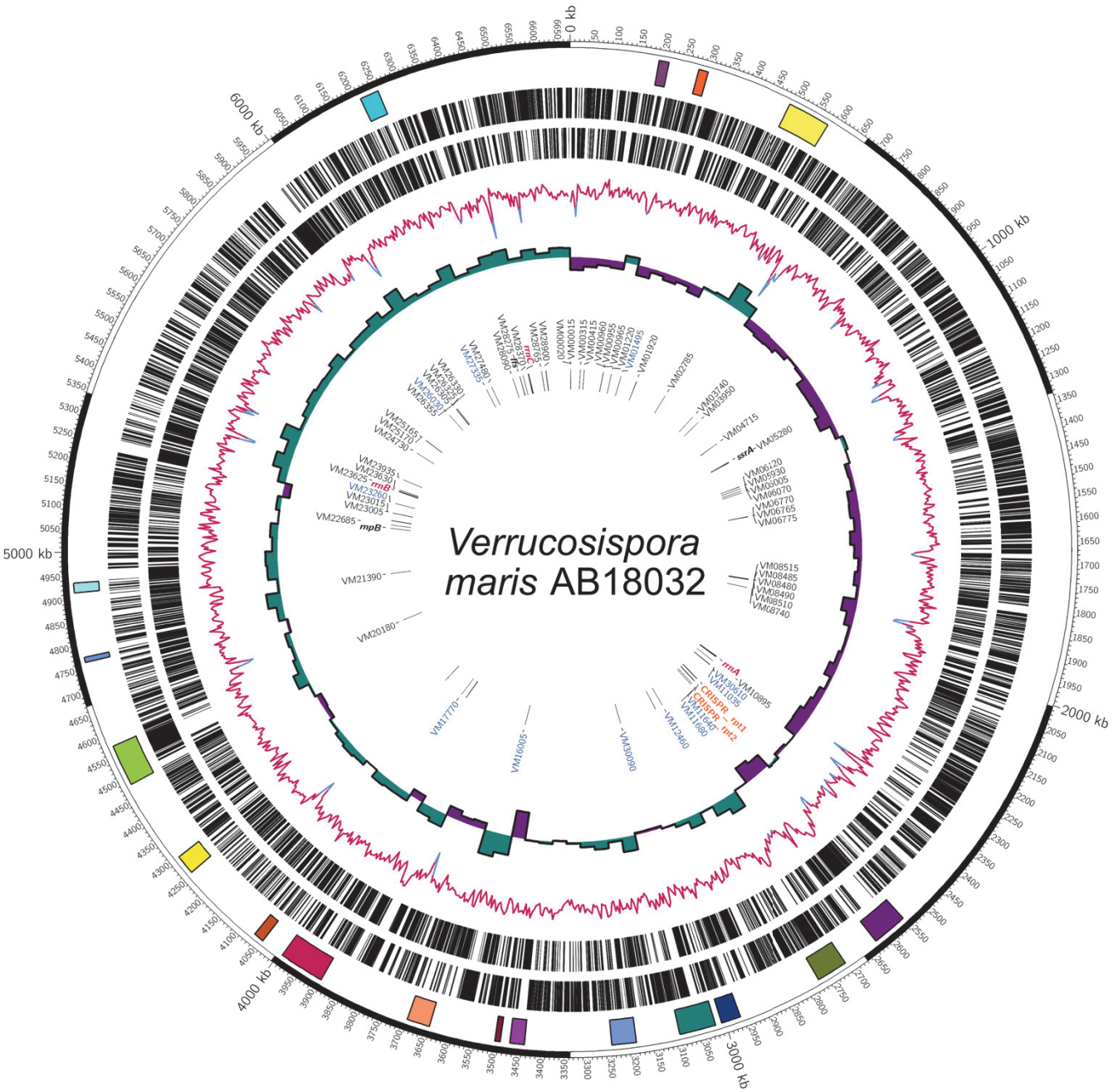


Figure 12. *V. maris* AB18-032 genome. Rings (outer to inner) represent: a) scale (kb), b) size and location of NP clusters as predicted by antiSMASH, c) RAST predicted genes positive strand, d) predicted genes negative strand, e) %G+C (red ≥ 66%, blue ≤ 66%; max = 76%, min = 57%), f) GC skew (green ≥ 0.287, purple ≤ 0.287; max = 5.5, min = -6.12), g) RNAs: black = tRNAs, bold italics = other none-coding small RNAs, red italics = ribosomal RNA operons, orange = CRISPR arrays and blue = regulatory RNAs.

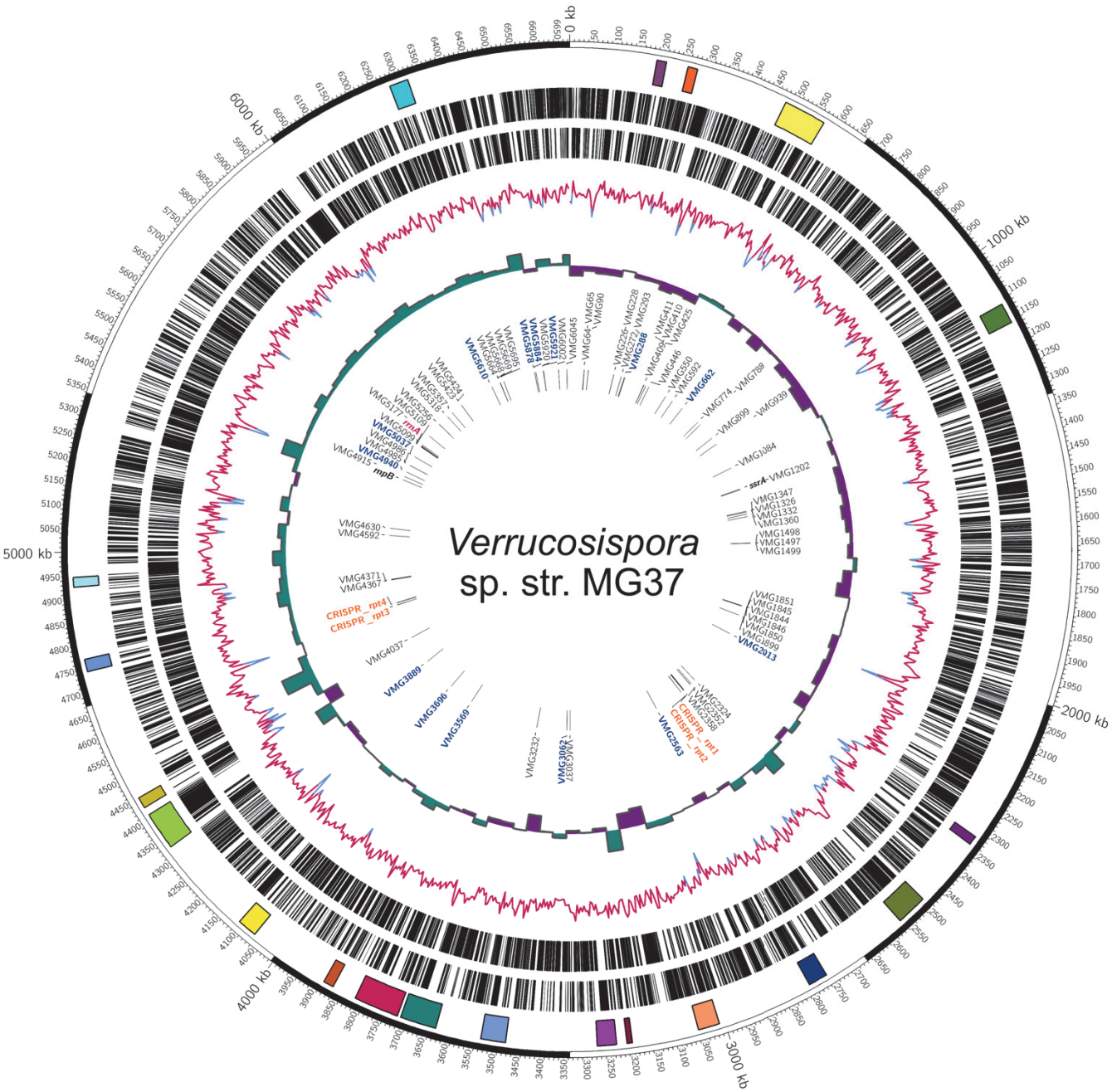


Figure 13. *Verrucosispora* sp. str. MG37 genome. Rings (outer to inner) represent: a) scale (kb), b) size and location of NP clusters as predicted by antiSMASH, c) RAST predicted genes positive strand, d) predicted genes negative strand, e) %G+C (red ≥ 68%, blue ≤ 68%; max = 76%, min = 60%), f) GC skew (green ≥ 0.202, purple ≤ 0.202; max = 5.5, min = -6.12), g) RNAs: black = tRNAs, bold italics = other none-coding small RNAs, red italics = ribosomal RNA operons, orange = CRISPR arrays and blue = regulatory RNAs.

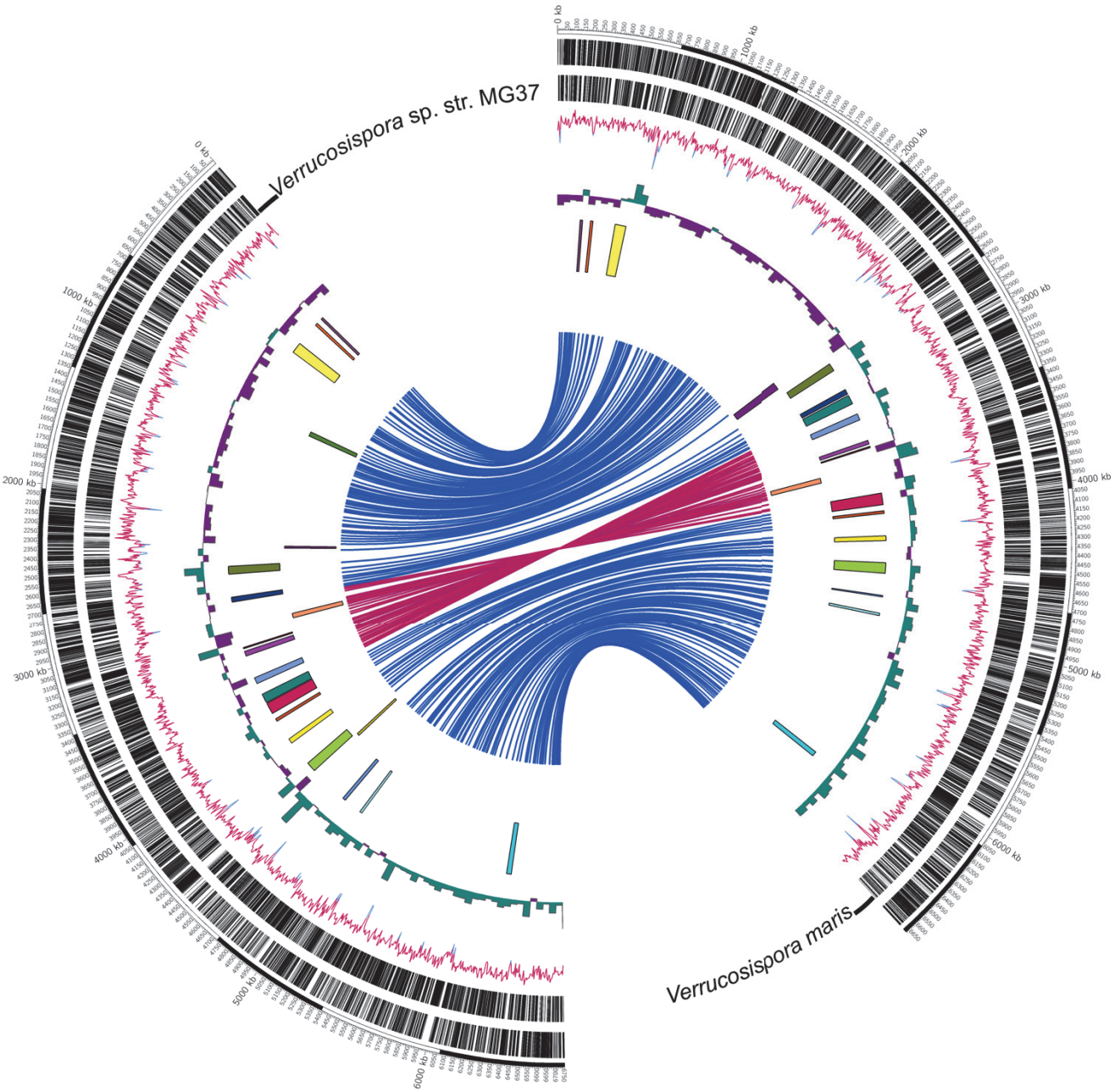


Figure 14. Genomic comparison of *Verrucosispora* sp. str. MG37 and *V. maris* AB18-032. Rings (outer to inner) represent: a) scale (kb), b) RAST predicted genes positive strand, d) predicted genes negative strand, c) %G+C, d) GC skew, e) shared NP clusters, f) NP non-shared NP clusters. The links in the centre of the figure show regions of DNA with >97% similarity over 10 kb, red lines show regions that are inverted relative to one another.

2.3.3 Analysis of ppb cluster in Verrucosispora sp. str. MG37

Confirmation of ppb cluster in Verrucosispora sp. str. MG37

To ensure the correct cluster is identified, all *Verrucosispora* sp. str. MG37 clusters were analyzed independently of the *ppb* cluster in *V. maris* AB18-032*,* with the hope that it would result in the same cluster being identified in both producers. All SBC were analyzed to determine their predicted products and it was found that the most likely candidate cluster present in *Verrucosispora* sp. str. MG37 was homologous with the *ppb* cluster identified in *V. maris* AB18-032. There were no other common SBCs identified in *Verrucosispora* sp. str. MG37 that were thought to produce a molecule similar in chemistry to proximicin, defined by constraints previously outlined to likely be present in *ppb*; this further confirmed our hypothesis that the *ppb* identified in *V. maris* AB18-032 is correct.

Alignment of ppb cluster in both strains

ACT was used to align both genomes to highlight regions of similarity and looking specifically at the *ppb* region in both producers. If *ppb* is the biosynthetic cluster responsible for proximicin production, then it should be present in its entirety in both strains. This was effectively shown using ACT, showing large regions of similarity with the only difference apparent being the lack of a reductase gene (*ppb*065) in *Verrucosispora* sp. str. MG37 which is present in *V. maris* AB18-032. No other secondary biosynthetic clusters identified in *Verrucosispora* sp. str. MG37 were determined to be likely candidates for proximicin biosynthesis, (using constraints previously outlined above)

Analysis of ppb genes

As the *ppb* cluster has been identified as being present in both proximicin producers, and it appears to be the only likely candidate for proximicin production, the genes present were analyzed for likely function. From this six NRPS-like proteins were present, all with >99% amino acid identity with *V. maris* AB18-032; in addition, there were a selection of regulatory proteins as well as a cyclodehydratase. Also present is an MbtH-like protein, which most likely has a role in adenylation activity.

Analysis of A-domains present in the putative proximicin biosynthetic cluster

Of the six NRPS-like proteins present in *ppb*, four contain adenylation domains; the other two contain only condensation and thiolation domains. The A-domains were examined for potential substrates they activate, and these results given from the NRPS-predictor programs were further elaborated by looking at the core motifs manually. Ppb120 was shown to be very atypical in a specific core residue – a change serine to aspartate at the Asp_{235} position. It also had a very minimally conserved A3 and A5 domains, 50% and 33%, respectively. Ppb210 was predicted to activate either tryptophan or tyrosine consistently by all predication programs, and was shown to contain all core motifs indicative of an active A-domain. No substrates could be predicted for Ppb220 and Ppb195; Ppb195 contains all conserved core residues which would imply activity, Ppb220 however, has a completely absent A1 domain, suggesting that the protein would be inactive.

Table 7. Summary of adenylation domains found in *ppb* cluster. Domains present, residues which line the binding pocket and predictions show. Also, the percentage each residue conforms to known boundaries as determined by Stachelhaus et al., (1999). Blue represents known regions to be involved in substrate binding, and any very poorly conserved regions (<40%) are highlighted in bold. The unique serine residue mutation observed in *ppb120*, is highlighted in pink.

	Ppb120	**Ppb195**	**Ppb210**	**Ppb220**
Domains present **	A-T	A-T-Te	A-T	A
Residues in binding pocket	**S**YGXIVXX	DIWQCTAD	DAWTVAAV	DILTFALM
Maryland NRPS analyzer (Bachmann & Ravel, 2009)	No-prediction	No-prediction	Phe	No-prediction
NRPS predictor2 (Röttig et al., 2011)	No-Prediction	Gln, Arg	Phe, Trp	Gly, Ala
A1	100%	83%	**33%**	**0%**
A2	50%	66%	75%	100%
A3	**61%**	**83%**	**100%**	**92%**
A4	**75%**	**100%**	**100%**	**100%**
A5	**33%**	**83%**	**100%**	**83%**
A6	81%	**36%**	72%	90%
A7	**100%**	**33%**	**100%**	**83%**
A8	75%	85%	100%	100%
A9	85%	100%	100%	80%
A10	100%	83%	83%	83%

**A-Adenylation T = Thiolation Te = Thioesterase

Chapter 2. Genome sequencing, assembly, annotation of *Verrucosispora sp.* str. MG37, and identification of putative proximicin cluster

2.4 Discussion

2.4.1 Overview of Findings

Whole genome sequencing and assembly of the second proximicin producer - *Verrucosispora sp.* str. MG37 – was reported here for the first time, using a next generation sequencing (NGS) platform. By using comparative bioinformatical analysis of both proximicin producing *Verrucosispora* species, a putative proximicin biosynthetic cluster was identified, and was shown to present and complete in both, with largely >99% identity. Extensive bioinformatical analysis of the cluster allowed the outlining of two opposing potential proximicin biosynthetic routes, detailing the incorporation of the 2,4-disubstitued furan group.

2.4.2 Assembly and annotation of Verrucosispora sp. str. MG37

In the research described here, Illumina MiSeq was chosen to sequence the genome of *V. sp. str. MG37*. This platform was selected for many reasons, principally, the high speed and coverage, and low relative cost associated with Illumina sequencing approaches in comparison to other NGS platforms. Illumina MiSeq harnesses sequencing by synthesis (SBS) technology – the reversible terminator method, with fluorescently labelled nucleotides to detect single bases as they are incorporated into the growing DNA strands - to allow the delivery of highly accurate and robust performance. As with the typical cyclic array sequencing workflow, the library was prepared by adding sequencing adapters and the libraries transferred to the flow cells and sequenced. This allows the output of 2 X 300 bp read lengths, 25 million reads per run and a run time of 4-55 hours to run a maximum of 15 Gb. The output format of the data is FASTQ, which was developed from the original FASTA format to allow the incorporation of Phred-scaled base quality scores to facilitate the assessment of sequence quality. Phred score aids in the determination of accurate, quality based assemblies and, despite increasing computational complexity to the sequencing output, should be retained. This vast amount of resulting data from NGS

platforms, and the continual adaptation of bioinformatical programs to process it, make deciding on an assembly pipeline difficult. Here I briefly discuss issues encountered in read treatments, assembly and SMC identification, and programs used to minimise their influence; this will hopefully give insight for future researchers attempting assembly and annotation of bacterial genomes.

Quality assessment & treatment of raw reads

The output given from the Illumina MiSeq sequencing technology is four read libraries: Mate-Pair (MP); Single Read (SE); Paired End (PE), and UNKNOWN; how each of these are generated from the sequencing reaction is shown in Fig. 15. As each library contains different information – determined from where the ligated adaptor is located – they each have to be processed differently, in order to most efficiently use the information contained. The first step in genome assembly involves checking the quality of the reads produced and pre-processing of reads – to make

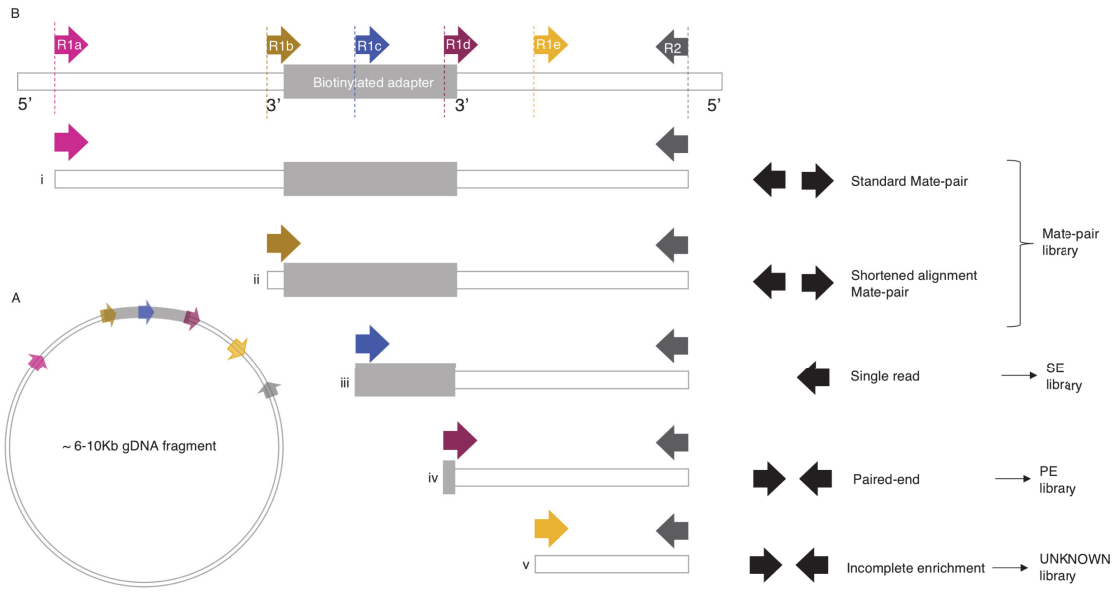


Figure 15. How each read library is produced during Illumina MiSeq sequencing depending on where reads are taken. (A) shows each read on circular DNA produced by adapter arms attached to each end of the DNA fragment then annealed. **(B)** once linearized, reads are taken and they are sorted depending on where the adapter is in the read. Looking just at the 5' read: If the adapter is not encountered at all, it is a standard MP (i); if the start of the read is adapter then native DNA, it is a shortened alignment MP, and pooled with the MP's (ii). If the read contains only adapter content, then the 3' read is a single read (iii). If the read contains adapter content, then native DNA it is a paired end read (iv) and if the read contains no adapter content, due to incomplete enrichment of adapters, it is put into the UNKNOWN library.

them compatible with one another. Several tools have been developed for quality control (QC) of raw NGS data; these include FastQC (Andrews, 2010) and PRINSEQ (Schmieder et al., 2011). The first of these – FastQC – was used here; it is a java application that generates many useful data diagnostics, such as the Phred score distribution along the reads, GC content and sequence duplication level – this information enables what future steps are required prior to alignment. Standard pre-processing steps include adapter removal and trimming reads identified by QC programs. Removal of adaptor sequences is important to prevent alignment errors and an increased number of unaligned of reads, yet preserving the maximum of useful sequencing data. Most trimming tools available provide this generic functions, along with their own custom feature; for example, CutAdapt (Martin et al., 2011) specifically allows detection of adaptor sequences anywhere in the read, whereas Trimmomatic is tailored to process PE reads. Both of these programs were analysed to identify which was best to use for Illumina MiSeq data processing, with FastQC was used after each program to identify which gave the highest quality reads after trimming steps. Although both programs successfully trimmed reads of adapter sequences, TrimAdapt left more reads for subsequent assembly while still removing adapter and low quality sequences. Trimmomatic removed an additional ~60,000 sequences, leaving a large amount of 'orphaned' reads; for this reason, it was excluded from the computational pipeline, and replaced by the more conservative trimmer – CutAdapt, in an effort to retain the largest potential amount of information available.

Assembly of Verrucosispora sp. str. MG37

Much research has focused on determining how genome assemblers compare, for example the original GAGE assembly comparison (Genome Assembly Gold-standard Evaluation) (Salzberg et al., 2012) which led to GAGE-B (Genome Assembly Gold-standard Evaluation for Bacteria). These large scale program evaluations were deigned to answer key questions, such as which program generates the best assemblies? and what depth of coverage and software parameters should be used to produce the optimal assemblies? This research demonstrated that the relatively new genome assembly programme – SPAdes (Bankevich et al., 2012) – was the best for microbial genomes, and hence, utilised here (Utturkar et al., 2014). One key advantage of SPAdes is the use of paired de

Bruijn graphs which allow the integration of read-pair information at a much earlier stage of assembly than most other assemblers which employ approach. This results in an improvement in contig length by incorporating MP information into the graph structure itself, in comparison to other programs which typically ignoring available MP information until post-processing steps. The *Verrucosispora* sp. str. MG37 genome was successfully assembled into 6 major scaffolds, with the largest three spanning >98% of *V. maris* AB18-032 genome. The availability of this assembled closely related genome allows it to be utilised as a high quality reference, and so laborious de novo assembly was not necessary. This also meant we were able to bypass many issues regarding assessing the quality of our assembly, as this could be judged largely by the amount which matched with the reference genome. However, to validate the assembly pipeline employed here, results were compared to that of other, similarly assembled genomes (Fig. 8). Where miss-assemblies were found, further analysis proved that these were not of *Verrucosispora* strain origin, and so are likely an artefact of contamination during the sequencing process. These miss-assemblies were disregarded and not included in further steps involving genome annotation, and *ppb* cluster identification.

Identification of Secondary Metabolite Clusters (SMC)

Despite the chemical diversity of small molecule natural products, the biosynthetic principles that govern their production are highly conserved. Enzyme families are typically associated with specific production of SM production classes; this information can be used to mine genomes for the presence of SMC. There are two principal strategies implemented by bioinformatical tools: (i) rule based approaches in which specific enzymatic domains are searched for, and then predefined rules are used to associate the presence of such hits with defined classes of SM. Such rule based strategies can easily detect clusters with well-defined domains, such as NRPS or polyketide synthetase (PKS) systems, and is implemented in such programs as antiSMASH (Medema et al., 2011), PRISM (Skinnider et al., 2015), and NP.searcher (Li et al., 2009). One major issue with this search tool is that it requires rules, and hence, cannot uncover SBC's with novel enzymology. This limitation was circumvented by the introduction of (ii) rule independent approaches, used in programs such as ClusterFinder (Cimermancic et al., 2014) and EvoMining (Weber et al., 2016), which utilises machine based learning or automated phylogenics analyses

to make predictions. As we predicted that proximicin is produced by a NRPS, we were confident that rule based programs would successfully identify the *ppb* cluster; three programs were used: antiSMASH (Medema et al., 2011), RAST (Aziz et al., 2008) and Prism (Skinnider et al., (2015). AntiSMASH (antibiotics and Secondary Metabolite Analysis Shell) can detect 44 different classes of SMC's, especially clusters containing modular enzymes such as PKS and NRPS genes. It also provides detailed annotation of the domain structure of NRPS, substrate prediction, genome-scale metabolic modelling and comparative genomic tools, making it the most comprehensive software for mining genomes. AntiSMASH is hence referred to as the 'gold-standard' of annotation tools, and was used here to allow the comparison of capabilities of less exploited programs – Prism and RAST. RAST – rapid annotation subsystem technology - identified the fewest number of SMC. RAST utilises the SEED project of annotation integration (Overbeek et al., 2013), to identify genes by comparison to the SEED collection. Continual addition of new subsystems into the SEED database will likely lead to the increase in unannotated regions of the *Verrucosispora* sp. str. MG37 genome. Prism made some important distinctions in comparison to antiSMASH. Prism is a web based tool specifically designed to mine for and analyse PKS and NRPS pathways; it includes the identification of similar known pathways. And, unlike antiSMASH – although not utilised in this research - is closely connected with the metabolomics platform iSNAP (Ibrahim et al., 2012), which can exploit liquid chromatography/tandem mass spectrometry (LC-MS/MS) data to identify peaks based on the predicted products. Similar to antiSMASH, PRISM identified the *ppb* cluster; however, interestingly it was a smaller cluster containing only six NRPS-proteins consisting of four A-domains, in comparison to the large SMC identified by antiSMASH. PRISM identified the same cluster boundaries as what was predicted by hand curation of the antiSMASH report; a small SBC with only four A-domains. AntiSMASH predicted that the *ppb* cluster was part of a larger NRPS cluster, including four other NRPS genes which lie upstream, responsible for producing a much larger SM. The connecting A-domain-containing protein Ppb220 – is likely inactive, supported by research discussed later, as many of its domains are incomplete; antiSMASH, using the rule-based approach, was unable to discern that although domains were present, they did not retain enough homology in core residues to allow activity. This is demonstrated in Table 7, were the presence of core residues was curated by

hand, and given as a percentage. As PRISM is a NRPS-specific annotator it was able to resolve this ambiguity; this demonstrates that even the 'gold standard' of annotation programs still requires support, either by hand curation or a specified program – if available.

Using the SBC annotation reports from all three discussed programs, identified clusters for both *V. maris* AB18-032 and *Verrucosispora* sp. str. MG37 were compared (Table 6). There were some important differences to note between the two *Verrucosispora* species: primarily, only *V. maris* AB18-032 contained the abyssomicin cluster which supports previous metabolite work suggesting *Verrucosispora* sp. str. MG37 does not produce this compound. Clusters which had characteristics previously identified to be required for proximicin biosynthesis were identified; primarily we expected to find (i) an NRPS-system cluster which (ii) contains an A-domain with a novel active site, capable of accepting a 2,4-disubstituted containing precursor, (iii) which may have likeness to that seen in congocidine biosynthesis. Clusters were systematically excluded or prioritised depending on their likeness to these three criteria. Only one common cluster with the required components was found; this was given the denotation of *ppb* – putative proximicin cluster.

2.4.3 Analysis of genes present in the ppb cluster

The *ppb* cluster contains a total 32 genes, composed of the expected regulatory, transportation and proximicin biosynthesis proteins (Fig. 10). Six were identified as NRPS-related modules, four of which contain a singular adenylation domain – *ppb120*, *ppb195*, *ppb210* and *ppb220*. To confirm that it is this cluster that is responsible for proximicin biosynthesis, the proteins it encodes, and in particular the structure of the adenylation domains present were further investigated.

Adenylation domains determine the chemistry of the amino acid incorporated next along the NRPS assembly line, and hence, are responsible for the structure of the final SM produced. Through crystal structure analysis, much has been learnt about the formation and configuration of these proteins; they are made up of three domains: a large N-terminus, a smaller C-terminus and an ATP binding site situated between. By investigating the sequence of an A-domain it can be determined (i)

whether it is likely that the enzyme is active and (ii) the potential precursor it is likely to activate. This is important for resolving the proximicin biosynthetic pathway, as by identifying the activity of the A-domains, we can determine chemistry, and the order, in which substrates are incorporated. There are many programs dedicated to A-domain characterisation which is typically a two-fold endeavour: (i) A-domain proteins are initially identified by well conserved domains – denoted as A1-A10 which are involved in both structural and catalytic roles established by Marahiel et al., (1997) summarised in Table 8; (ii) A-domain substrate selectivity is then analysed by a 10 residue code located at specific positions in the protein located between A4 and A5, resulting in them being present in the active site – and hence involved in substrate coordination and catalysis (Stachelhaus et al., 1999) (Table 8, pink). Adenylation domain core motifs are mentioned here only to aid understanding of the bioinformatical programs implemented; the origin and specific roles of each, as well as their organization in *ppb* genes, will be evaluated extensively in a later chapter. Programs PRISM, NRPSpredictor2 and PKS/NRPS Analysis Web-site exploit this well characterised, conserved structure-function relationship for substrate recruitment to allow prediction of potential substrate pools. The computational basis of these predictions are reliant on the prerequisite that the substrate of a given A-domain can be deciphered by specificity conferring code. PKS/NRPS Analysis Web-site delivers predictions based on BLAST analysis against the signatures determined by Challis et al., (2000) with more recent tools such as PRISM and NRPSpredictor2 introducing the use of profile Hidden Markov Models (HMM) and machine learning, respectively. As expected by the complexity of the algorithms they utilise, these programs were able to match substrates to three out of the four of the A-domains tested in this study. The less extensive analysis used in Maryland PKS/NRPS Analysis Web-site was reflected in its ability to only predict a singular A-domain substrate profile – for Ppb210 – which has an extremely conventional A-domain structure (Table 7). Other approaches for A-domain prediction are available, for example, there have been many successful reports of using structural models consisting of crystal or homology models, as well as docking analysis with putative substrates. This was not utilised here, despite being shown to contribute greatly to predicting substrate specificity, due in part to the lack of automated tools meaning it remains very computer intensive. But primarily due to the key A-domain of interest – Ppb120 or whichever is shown to be responsible for incorporating the furan-

containing precursor – being novel, and so there would be no homology or crystal structures available, preventing comparison affording useful information.

Table 8. Core domains present in NRPS-adenylation domains. Core domains present in A-domains first outlined by Marahiel et al., (1997), shown to have roles in structure, substrate binding and catalytic activity. Consensus sequence of each domain shown to be present. The substrate specificity conferring residues located between motifs A4 and A5 for model A-domain GrsA outlined by Stachelhaus et al., (1999) shown in pink. X denotes any amino acid residue.

Core	Consensus sequence	Role
A1	L(T/S)YxEL	Structural; at N-terminus of domain
A2	LKAGxAYL(V/L)P(L/I)D	Structural; properly aligns Gly_{78}
A3	LAYxxYTSG(S/T)TGxPKG	Substrate binding; acts as a loop
A4	*DxS	Substrate binding; aromatic residue terminates an A-helix that forms side of the acyl-binding pocket.
	D_{235} A_{236} W_{239} T_{278} I_{299} A_{301} A_{322} I_{330} C_{331} K_{517} **	
A5	NxYGPTE	Structural and substrate binding; invariant glutamic acid coordinates Mg^{2+} ion
A6	GELxIGx(V/L)ARGYL	Structural; stabilizes distorted beta sheets in the N-terminus domain
A7	Y(R/K)TGDL	Substrate binding and catalytic; aspartic residue is 100% conserved
A8	GRxDxxxKxxGxxxELxxxE	Structural and substrate binding
A9	(L/V)Px*M(L/V/I)P	Catalytic
A10	NGK(V/L)DR	Catalytic

* Aromatic residue ** GrsA numbering

Identifying the substrates of each present A-domain in the *ppb* cluster was the initial approach utilised in attempting to identifying a potential biosynthetic route to proximicin. As discussed, it was predicted that at least one of the adenylation domains present in the *ppb* cluster would have a unique structure, as it is required to activate the non α-amino acid group of a 2,4-disubstitued furan. It was also hypothesised that it would possess similarity with the A-domain responsible for pyrrole incorporation in congocidine biosynthesis – which at the time of the research was Cgc18. Cgc18 has been shown to be unique – it possesses a mutation at the core residue Asp_{235}, known to be responsible for the alignment of the amino group of the substrate in the active site. The idea that this substitution made the enzyme a good candidate for a protein which would be able to activate a heterocycle precursor was shared by Juguet et al. (2009). However, the congocidine biosynthesis pathway has since been revised and it has been shown that it is actually another protein – Cgc13* - which is responsible for pyrrole activation. Although discussed at length in future chapters, it should be noted that the Asp_{235} mutation exhibited in Cgc18, although proven to not be indicative of a heterocycle precursor specifically, is suggestive of a non α-amino acid being accepted. One such similar A-domain was identified – Ppb120 – which conformed to both of these initial predictions; it was unique in its predicted structure, no precursor predictor programs could determine its substrate and it possess the same Asp_{235} mutation. As Ppb120 was the singular unique A-domain present in the *ppb cluster*, we suggest that it works in an iterative manner to load precursors (see below). This repetitive activation, loading and trans-adenylation activity of A-domains been shown in other NRPS synthesis pathways, for example in yersiniabactin biosynthesis, a single adenylation domains loads three T domains (Keating et al., 2000) and also, more specifically in the revised congocidine pathway, a pyrrole molecule is added iteratively by a single A-domain to two different T-domains.

The adenylation domain of Ppb210 contains all the required domains for activity, and is predicated to activate a tyrosine or tryptophan like molecule; this prediction was shared by all utilised A-domain substrate prediction programs utilised, with a high likelihood. The A-domain of Ppb195 is also likely to be active as it contains all domains known to be required for activity, however, a likely substrate could not be predicted. The final A-domain present in the enzymes of the *ppb* cluster, located on

the boundary of the cluster - *Ppb220* - is most likely completely inactive due to many of its crucial activity core domains being mutated or completely absent. This was supported by in the inability of any of the NRPS-predictor programs to predict potential substrates. The inactivity of Ppb220 makes sense as it is the last adenylation domain in the gene cluster; as noted, the proximicin gene cluster is interpreted by some secondary metabolite clusters finding programs as being part of another cluster which lies further along the 3' end of the DNA. If experimental work supports the bioinformatical analysis, and Ppb220 is in fact inactive, it could mark the end of the proximicin cluster, and may suggests that previously a larger compound was made but inactivity of Ppb220 lead to disconnection of the two halves of the cluster, resulting in proximicin production. There are many characteristics of Ppb220 which make it extremely unlikely that it would be active, many of the core domains vary a lot from the known core consensus sequences, or are entirely absent. From the A-domains identified and characterised, along with other genes present in *ppb* cluster, two potentially conflicting routes for proximicin biosynthesis were outlined.

2.4.4 Discrepancies in proximicin production

It has been demonstrated that members of the proximicin family are produced by two *Verrucosispora* species, however they differ in which they produce; *Verrucosispora* sp. str. MG37 can produce all three – A, B and C- and *V. maris* AB18-032 has only been shown to produce A. It was previously thought that some mutational event led to this discrepancy exhibited in production. A diversion event pushed by differing conditions – potentially limited carbon sources – leading to *V. maris* AB18-032 only producing proximicin A most likely because it is a smaller molecule and doesn't require the addition of the tryptophan or tyrosine moiety. A selection event which was not imposed on *Verrucosispora* sp. str. MG37, and so it continued on producing all three. Or perhaps, as *V. maris* AB18-032 has the ability to produce abyssomicin compounds – another family of antimicrobials – proximicins were made surplus. Alternatively, *Verrucosispora* sp. str. MG37 may have been under selection pressure by competing bacteria – an evolutionary arms race – leading to it retaining the ability to produce proximicins B and C. Abyssomicin biosynthesis involves the biosynthesis of a three-carbon unit derived from fructose-6-phosphate, this is also a key intermediate in the biosynthesis of aromatic amino

acids. Thus, during secondary metabolism there may be competition for precursor molecules required for the biosynthesis of both abyssomicin and proximicins B and C, but not for proximicin A. The absence of the abyssomicin gene cluster in *Verrucosispora* sp. str. MG37 removes this competition resulting in the biosynthesis of all proximicins. This hypothesis could be tested by deleting the abyssomicin gene cluster in *V. maris* AB18-032 and observing whether or not proximicins B and C are produced. Whatever the driving force, the differences in production presented potential insight into the divergence of these two closely related species. However, analysis of the two producers and the *ppb* cluster – shows that there is no genetic basis which would result in *V. maris* AB18-032 only being able to produce proximicin A. The *ppb* cluster has over >99% identity between the two producers, and the only difference is the absence of Ppb065 – a hydrolase/reductase – in *Verrucosispora* sp. str. MG37. which would not account for the apparent differences in production of proximicins. Both proximicin producers are genetically capable of producing the entire family of proximicins, but only one does; this means the genes must be under some type of regulatory control preventing expression of certain compounds. Regulatory control of gene expression, and hence metabolite production, is extremely complex; and so further analysis of proteins governing proximicin production is required. It is extremely difficult to determine the regulatory control of a biosynthetic cluster, as proteins controlling production may be located distantly on the genome to the SBC they control; and typically, many proteins work in parallel in a complex homeostasis. One simpler approach would be growing *V. maris* AB18-032 in different media and determine whether it – when exposed to differing stresses - begins to produce other proximicins. Alternatively, by alleviating potential stresses - such as limited nutrient resources or microbial competition leading to the high production of other costly compounds – its metabolic profile may change. *V. maris* AB18-032 and *Verrucosispora sp. str. MG37* are both marine organisms but were isolated from different locations and at different depths, the Sea of Japan at 289 M and Raune Fjord, Norwary at 250 M, respectively. Fiedler et al., (2008) when describing the differing production ability of the two species uses a singular growth media for metabolite studies: SSG media. Although this is known to be a good media for growth of Streptomyces, it is not discussed to what extent this media is similar to that of which the bacteria have been extracted. It would be interesting to determine whether, if put in an environment similar to that of

Verrucosispora sp. str. MG37's isolation, *V. maris* AB18-032 would produce all three proximicins. Or whether, by adding excess tryptophan or tyrosine – the component differences between proximicin A, B and C – would induce production, meaning amino acid concentration is the limiting step to proximicins production in *V. maris* AB18-032. If this was successful, a transcriptomics approach could be used to determine genes with differing regulation in the new conditions resulting in proximicin production – and hence, gain insight into the regulatory control of the proximicins.

2.4.5 Conflicting biosynthetic route proposals

Preliminary review of the *ppb* gene cluster revealed that it resembles that of congocidine, and hence, it was initially presumed that a similar biosynthetic route was shared by both enzyme clusters. The biosynthesis of the 2,4-disubstitute pyrrole that forms the core dipeptide of congocidine has been determined (Fig. 16a); it is derived from fructose-6-phosphate (Lautru et al., 2012). The authors also predicted that a similar route would yield the 2,4-disubstituted furan of the proximicins; the difference being that a key reductive amination reaction in the formation of the pyrrole, would be replaced by a ketoreduction to form the furan (Fig. 16b). However, searches of the *Verrucosispora* genomes using the key congocidine biosynthetic genes involved in the formation of the pyrrole (Cgc8, 9, 10 and 11) failed to identify any homologues. This indicates that the furan moiety of the proximicins has a distinct biosynthetic origin. Also, on further analysis of the *ppb* gene cluster, it was shown that it contained many genes which would not be required in the initially outlined biosynthetic route; raising the questions – why are they present? What is their role? Of these proteins, one of particular interest is Ppb090, which shares high sequence similarity with cyclodehydratase enzymes responsible for cyclization of linear peptides into heterocycle containing compounds, specifically seen in secondary metabolite activation. This led to the proposal of a second potential biosynthetic route leading to proximicin biosynthesis, in which an unusual linear molecule is accepted by an A-domain and cyclised while on the NRPS assembly line by the cyclodehydratases *ppb090*. The support for both these potential routes is outlined below.

4-aminofuran-2-carboxylic acid route

In the first route predicted route a 2,4-disubstituted furan (4-aminofuran-2-carboxylic or a derivative thereof) is selected and activated by Ppb120 (its activity dependent on the MbtH protein Ppb125), Ppb110 loads the adenylated 2,4-disubstitued furan on to its cognate T domain, and also loads the T domain of Ppb140 (a C-T domain protein). Formation of the core proximicin structure (dipeptide of 4-aminofuran-2-carboxylic) is achieved through peptide bond formation using the two furan aminoacyls as substrates. Ppb165 then catalyzes release of the dipeptide from the T domain of Ppb110. Tailoring steps would include Ppb155 (carboxyaminoimidazole synthase-like protein) and Ppb145 (a methyltransferase) forming the N-terminal methyl carbamate. This route would yield proximicin A (Fig. 16b). Biosynthesis of proximicins B and C would be very similar except that Ppb120 would load itself and Ppb140 (a C-T-C domain protein) (Fig. 16b). The domain organization of Ppb140 is very unusual in having two condensation domains and no adenylation domains, but it makes sense from the perspective of the proposed proximicin biosynthesis pathway: the first condensation domain would form the peptide bond as for Ppb110 above; the second would then form a peptide bond between peptidyl-S-T of Ppb140 and the aminoacyl-S-T of Ppb210, ultimately yielding proximicin B if tyrosine is selected and loaded by Ppb210, or proximicin C if tryptophan is selected. Ppb165 would catalyze release of the tripeptide from Ppb140 (premature release of the dipeptide would lead to proximicin A biosynthesis). The N-terminal modifications would occur as above. The C-terminal tyramine in the case of proximicin B and of tryptamine of proximicin C are predicted to come from the decarboxylation of tyrosine and tryptophan, respectively. The most likely candidate enzyme for this decarboxylation reaction is Ppb135 (decarboxylase), but Ppb205 (cytochrome P450) and Ppb130 (hydroloase) could feasibly play a role in decarboxylation (Fig. 16b).

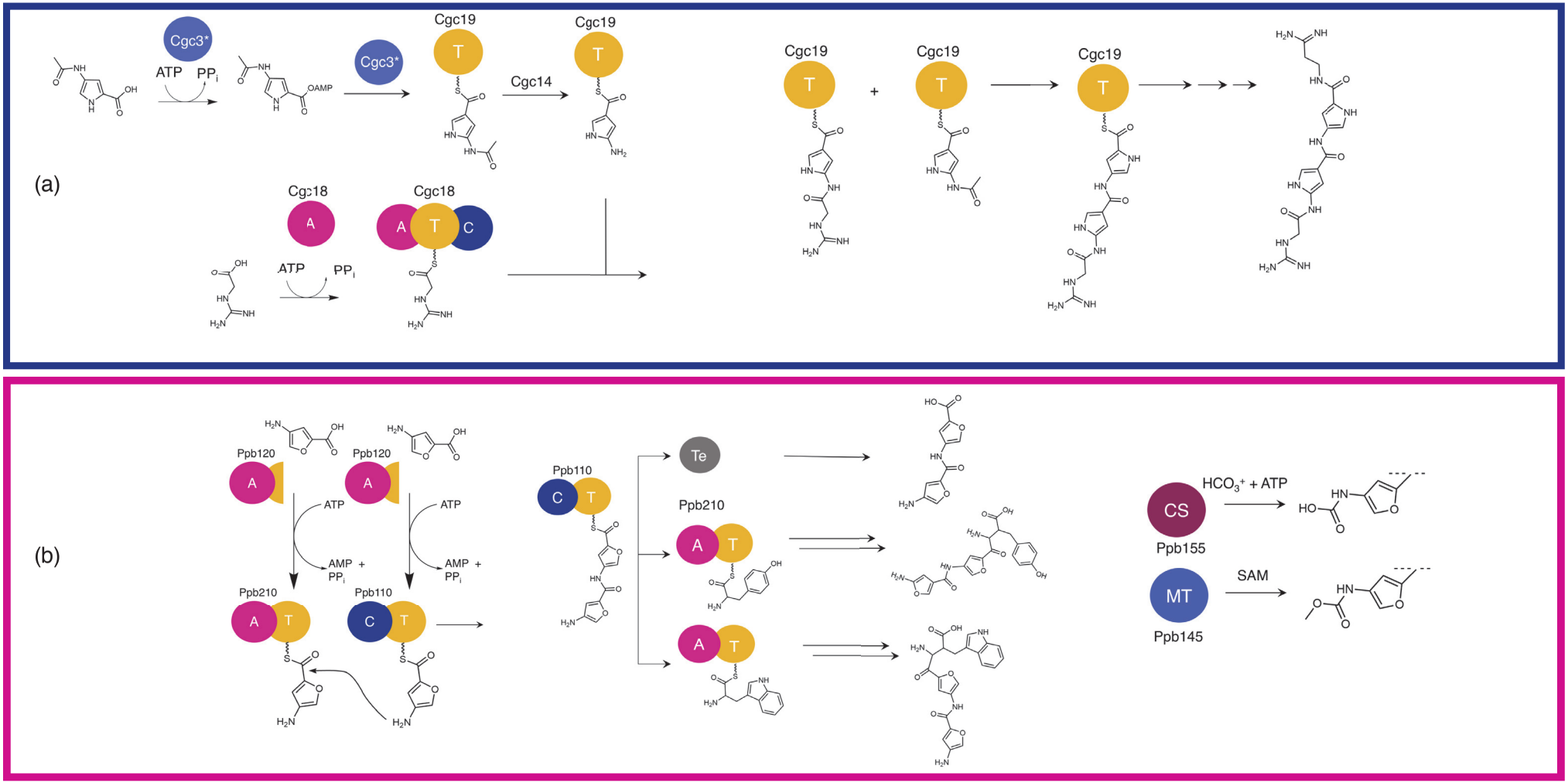


Figure 16. Congocidine and proximicin biosynthetic routes. (a) summary of the route to congocidine biosynthesis Al-Mestarihi et al., (2015) showing the non-NRPS adenylation domain iteratively activating pyrrole precursor, with Cgc18 activating a GA molecule. **(b)** potential route to proximicin biosynthesis via a similar mechanism: a furan containing precursor is added in an iterative manner by Ppb120 producing the core peptide of proximicin A (top), with the tryptophan or tyrosine moiety being added by Ppb210 for proximicin B (middle) and C (bottom). Tailoring steps being catalyzed by Ppb155 (dark pink) and Ppb145 (blue).

Cyclohydratase-mediated formation of heterocycles

Adenylation domains present in the *ppb* cluster support the route outlined above; however, other genes present in *ppb* cluster introduce potential ambiguity. The cyclodehydratases present within the cluster make it unclear at what point the di-substituted heterocycle is formed: pre- or post-amide bond formation. If it occurred post amide bond formation, it would require a heterocyclase to catalyse this reaction. Peptide-based biologically active natural products produced by bacteria which contain heterocyclic rings are common (Arnison et al., 2013); the presence of thiazolines and oxazolines in a variety of approved and potential drugs, has made them the focus of much research (Jin et al., 2011). Of particular note here, are three homologous systems – the Sag, Mcb and Tru pathways – responsible for the production of the antimicrobials streptolysin, microcin B17 and macrocyclic cyanobactins, respectively. These are of interest as each pathway involves the cyclisation of residues – normally the amino acid cysteine, serine or threonine – on the precursor molecule to produce the biologically active final compound. This post-transcriptional modification of inactive compounds to active toxins – suggests that it is a method utilised by microorganisms to prevent self- damage. This class of ribosomally synthesized and post- translationally modified peptides (RiPPS) is rapidly growing, as research continues into biologically active molecules. Peptide residues which are cyclised share some commonalities in their genetic organisation and enzyme requirements (Fig. 17a), such as (i) precursor peptide to be modified, which contains a conserved N-terminal leader and protease, micro-cyclisation domain and the peptide to be processed; (ii) cyclodehydratase enzyme; (iii) dehydrogenase enzyme and, (iv) a docking protein. This collection of enzymes typically appears next to one another in the genome and are co-transcribed to produce a trimeric synthetase complex, and through the action of this complex, oxazole and thiazole heterocycles are incorporated into the peptide scaffold. Using the biosynthesis of microcins as an example, it has been shown that heterocycles are installed through a two-step process: a cyclodehydration to generate an azoline heterocycle and a subsequent Flavin mononucleotide (FMN)-dependent dehydrogenation to give the aromatic azole (Fig.17b) (Li et al., 1996; Lee et al., 2008; McIntosh et al., 2010). It has been shown that these enzymes work in a systematic manner to cyclise the precursor molecule, and that the cyclodehydratases enzymes exhibit a high level promiscuity, meaning that chemical versatility exists in relation to

which can be processed. Further work has shown regions of invariantly conserved residues are present within both the cyclodehydratase and dehydrogenase which are indicative of activity. Multiple sequence alignment of these enzymes families previously done shows two almost invariant CXXC motifs in the cyclodehydratases thought to be involved in the structural coordination of a Zn^{2+} molecule – shown to play only a structural role (Zamble et al., 2000) - as well as regions containing nearly invariant tyrosines and arginine residues thought to be involved in FMN binding in dehydrogenases. Adding complexity to this reaction, ATP hydrolysis is required (Milne et al., 1998; McIntosh & Schmidt, 2010), yet in some clusters the cyclodehydratases contain no bioinformatically identifiable nucleotide binding sites (a notable exception which will be discussed later is TruD), and it is only the docking protein which strangely contains any resemblance of a nucleotide binding site (Milne et al., 1998). However, these residues are not conserved across all docking proteins and so ATP is most likely involved in some other way. Although the requirement for ATP hydrolysis with respect to azoline formation remains largely undescribed, it has been suggested that it may have a role as a dynamic regulator of the trimeric synthetase complex or potentially a role in the activation of the peptide backbone during cyclisation (Li et al., 1996 & Schmidt, 2010). Recently, Dunbar et al., (2012) demonstrated that ATP can be used directly to phosphorylate the peptide backbone during heterocycle formation, and that interestingly, the docking protein component of the trimeric synthetase complex is also able to perform the ATP-dependent cyclodehydration reaction when other related proteins are not present. This could give insight into how the furan moiety is introduced into the proximicin biosynthetic pathway.

It is logical to assume that a homologous system could be involved in proximicin biosynthesis – a precursor molecule is produced via the NRPS system, which is then subjected to processing to incorporate the furan groups. For example, the cyclised groups resulting in oxadole-groups present in microcin - if the nitrogen was replaced by carbon it would result in a furan group with the same 2,4-disubstitution pattern exhibited in proximicins. Interestingly, within the *ppb* cluster, components of a trimeric synthetase system are present: a cyclodehydratase - *ppb090*, and a dehydrogenase – *ppb055*, both of which have moderate, to high similarities to systems reported to utilise this trimeric synthetase complex. No homologs to the

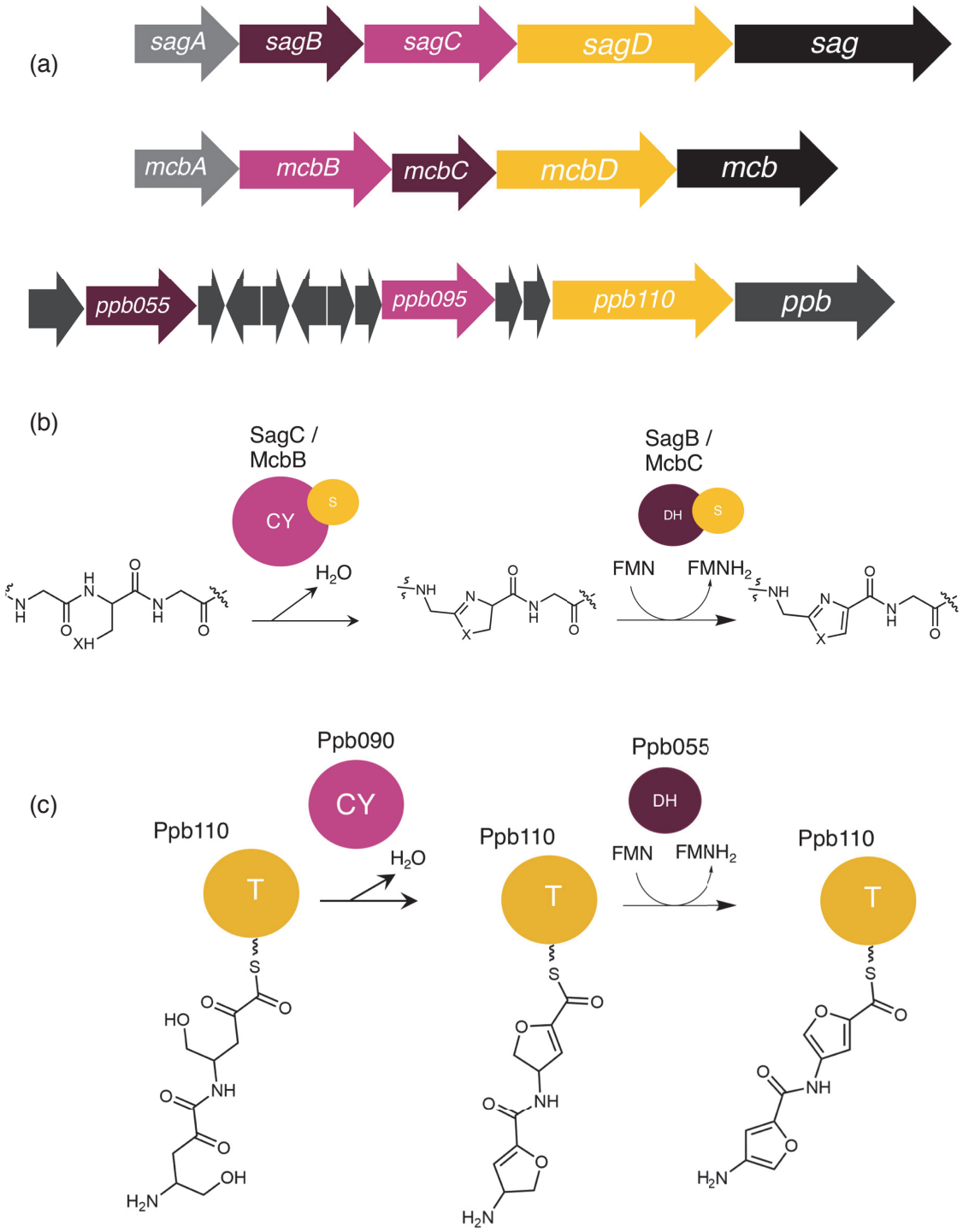


Figure 17. Cyclodehydratase and dehydrongenase reactions. **(a)** genetic organisation of the Sag and Mcb clusters responsible for streptolysin microcin biosynthesis, respectively. Similar genes present in the *ppb* cluster. **(b)** the cyclodehydratase and dehydrogenase catalysed reactions in the streptolysin and microcin pathways. **(c)** potential route for furan incorporation in proximicin biosynthesis via cyclodehydratases and dehydrogenase enzymes.

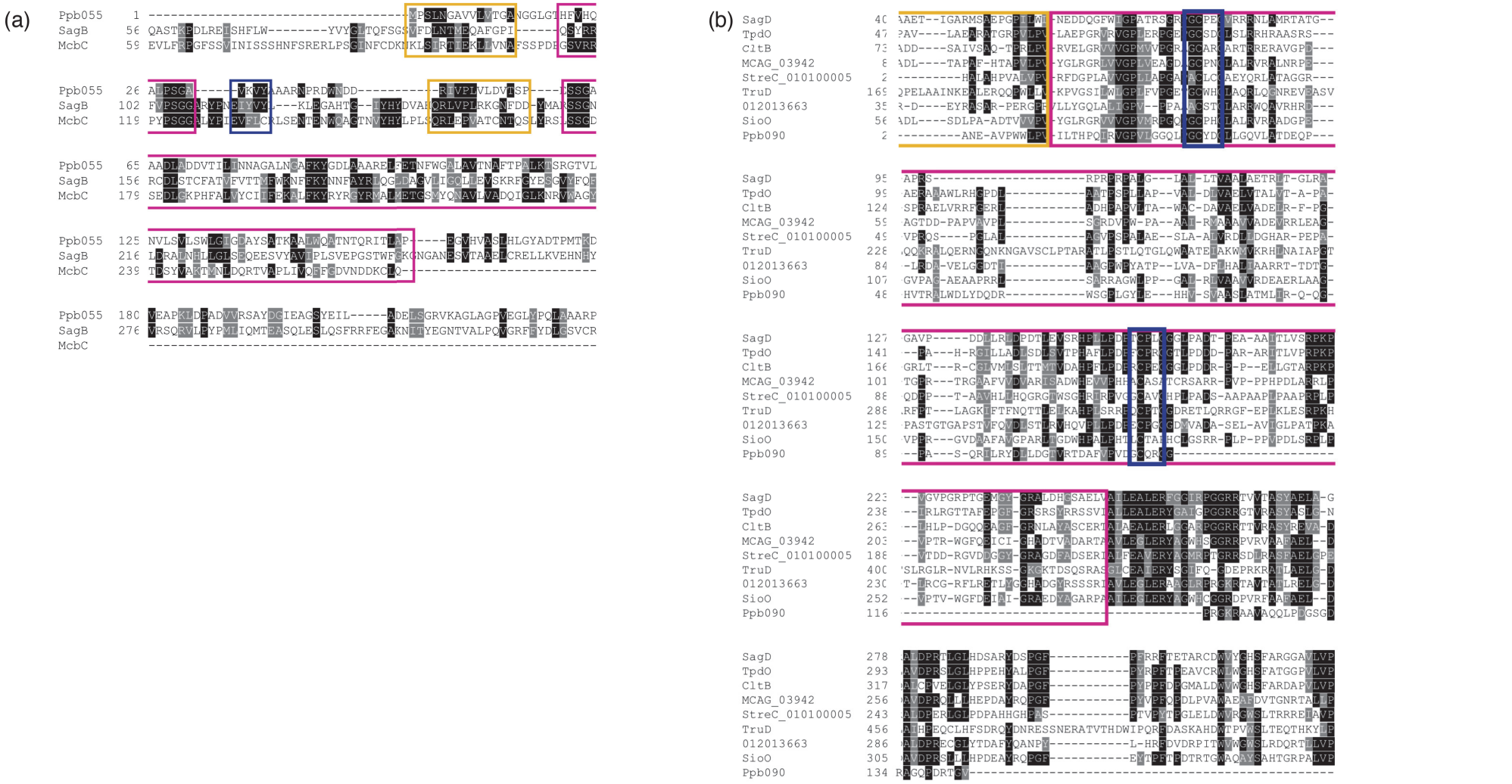


Figure 18. Multiple sequence alignment of cyclodehydratases and dehydrongenases from the Sag and Mcb biosynthetic clusters with genes in the *ppb* cluster. Sag and Mcb cluster is responsible for streptolysin and microcin biosynthesis, respectively. **(a)** Alignment of dehydrogenases *sagB* and *mcbC* with *ppb055*. **(b)** Alignment of many cyclodehydratases and *ppb090*. Core domains shown to be involved in catalysis are highlighted with pink and yellow boxes. Blue boxes denote core residues identified in the proteins.

precursor peptide or the docking protein are present in the *ppb* cluster (Fig. 18); however, these would not be required as the peptide precursor is produced via the NRPS system, and the A-domain would act as an enzymatic scaffold, anchoring the substrate in the correct orientation for cyclisation, respectively (Fig. 17c & Fig. 18). Importantly, the conserved residues shown to be vital for activity, are present in both Ppb090 and Ppb055. Previous studies on cyclodehydratase reactions has shown that they are conducted by trimeric synthetase complex, usually annotated as a C protein (cyclodehydratase), a D protein (docking) and a B protein (dehydrogenase) – in some species the C and D protein are fused (e.g. TruD). As above, these annotations were questioned by Dunbar et al., (2012) as they demonstrated that it is the D protein that displays ATP-dependent cyclodehydratase activity in the absence of the cognate C protein; the presence of the C protein increasing the reaction 1,000-fold. The ATP-binding site of the D protein was later identified as a conserved and novel motif in D proteins (Dunbar et al., 2012). Structural alignment of Ppb090 with TruD (Fig. 19) reveals that it is a C protein Ppb090 also lacks the ATP-binding site present in some C proteins such as MccB. The lack of ATP binding site is interesting as (i) ATP is known to be essential for these cyclization reactions and (ii) to date, no D protein homologue has been identified in the *ppb* gene cluster. The requirement of ATP has been well documented; it is needed by TruD to produce a highly reactive intermediate which decomposes to thiazoline by rapid elimination of AMP following the attack on ATP by the hemiorthamide intermediate displacing pyrophosphate. It is plausible that a similar reaction occurs in proximicin biosynthesis, but due to the lack of ATP binding domain, the phosphate must be added by another process. This means that the *ppb* cluster theoretically contains the genes responsible for post-transcriptional modifications, potentially with the ability to cyclise precursor residues that result in the introduction of the furan groups. This theory is based largely on the sequence, and by extensive structural homology exhibited by genes in the *ppb* cluster and trimeric synthetase complexes utilised in the production of secondary metabolites in bacteria. As discussed, despite sequence similarities in core residues, there are several discrepancies, making it extremely important to support this bioinformatic work with experimental research. If an equivalent system is present, it would mean that the molecule activated by the initial A-domain in the peptide producing chain accepts – potentially iteratively – a linear, and not a pre-cyclised compound.

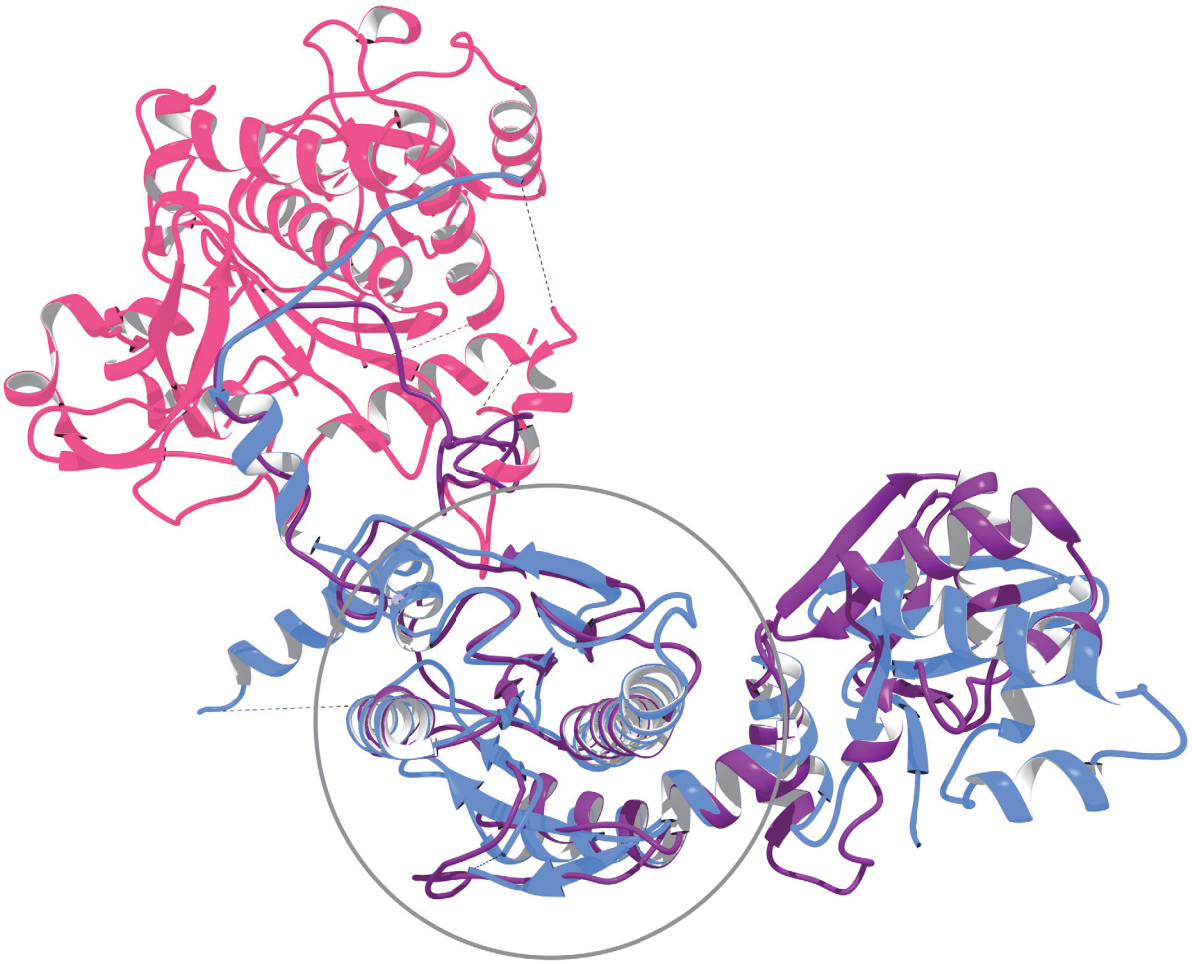


Figure. 19. Structural prediction (i-TASSER) of Ppb090 (purple) overlaid and aligned with TruD (C domain in blue, D domain in pink). The conserved region of the C proteins is circled.

The proposed activity of the McbB and SagC enzymes makes it possible to predict the likely chemistry of this linear precursor (Fig.17) It may also be possible that the cyclodehydratases and dehydrogenase are involved in precursor synthesis, which is then accepted into the NRPS system in the heterocycle form. This is unlikely, however, it has been shown that these enzymes require the substrate to be in a specific orientation for cyclisation to proceed, and – as there is no docking protein homolog, the only reasonable explanation is that this job must be done by the NRPS proteins.

To determine which of the two proposed routes to proximicin biosynthesis is the one utilised by the *Verrucosispora* species, experimental analysis of the A-domain proteins must be done. By confirming the chemical nature– whether a linear molecule or a heterocycle containing molecule – of the initially accepted precursor by the A-domains present in *ppb* will allow the outlining of the route to proximicin biosynthesis.

2.4.6 Future applications of NGS

The research described here reports the sequencing, assembly, annotation and the outlining of a biosynthetic route responsible for novel chemistry incorporation. From this relatively straight forward pipeline, an entire microbial genome was assembled; as the ability and large scale implementation of sequencers continues to increase exponentially, as with the bioinformatical ability to handle the data produced, the potential applications of this technology will be unmatched. As previously discussed, antimicrobial resistance is an ever-looming threat to the sustainability of the human race, and so it is of paramount importance that antimicrobial compounds with novel cellular targets are discovered. NGS platforms offer an unparalleled glimpse into all potential clusters present in an organism, revealing novel chemistry millions of years in the making. Here, a biosynthetic cluster with potentially high pharmaceutical value, was identified and characterized using free to use programs, in a facile manner; resulting in a high-quality output. This technology is no longer reserved for the excessively funded and those groups with large computing capabilities. The latest advancements leading to wide scale applicability of NGS has allowed advancements in novel antimicrobial drug discovery. A large amount of the research is currently focused on mining genomes for compound biosynthetic

clusters which are silent; to identifying the molecular and enzymatic mechanisms which underlie the biosynthesis of a compound lead, first the genes responsible must be identified.

Chapter 2. Genome sequencing, assembly, annotation of *Verrucosispora sp.* str. MG37, and identification of putative proximicin cluster

2.5. References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, *25*(17), pp.3389-3402.

Al-Mestarihi, A.H., Garzan, A., Kim, J.M. and Garneau-Tsodikova, S., 2015. Enzymatic evidence for a revised congocidine biosynthetic pathway. *ChemBioChem*, *16*(9), pp.1307-1313.

Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data.

Arcamone, F., Penco, S., Orezzi, P., Nicolella, V. and Pirelli, A., 1964. Structure and synthesis of distamycin A. *Nature*, *203*(4949), pp.1064-1065.

Arnison, P.G., Bibb, M.J., Bierbaum, G., Bowers, A.A., Bugni, T.S., Bulaj, G., Camarero, J.A., Campopiano, D.J., Challis, G.L., Clardy, J. and Cotter, P.D., 2013. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Natural product reports*, *30*(1), pp.108-160.

Auch, A.F., Von Jan, M., Klenk, H.-P., Göker, M. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Standards in Genomic Sciences* 2:117-134, 2010

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M. and Meyer, F., 2008. The RAST Server: rapid annotations using subsystems technology. *BMC genomics*, *9*(1), p.75.

Bachmann, B.O. and Ravel, J., 2009. Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. *Methods in enzymology*, *458*, pp.181-217.

Bailly, C. and Chaires, J.B., 1998. Sequence-specific DNA minor groove binders. Design and synthesis of netropsin and distamycin analogues. *Bioconjugate chemistry*, *9*(5), pp.513-538.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D. and Pyshkin, A.V., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, *19*(5), pp.455-477.

Bashir, A., Klammer, A.A., Robins, W.P., Chin, C.S., Webster, D., Paxinos, E., Hsu, D., Ashby, M., Wang, S., Peluso, P. and Sebra, R., 2012. A hybrid approach for the automated finishing of bacterial genomes. *Nature biotechnology*, *30*(7), pp.701-707.

Baspinar, A., Cukuroglu, E., Nussinov, R., Keskin, O. and Gursoy, A., 2014. PRISM: a web server and repository for prediction of protein–protein interactions and modeling their 3D complexes. *Nucleic acids research*, *42*(W1), pp.W285-W289.

Bentley, R., 1999. Secondary metabolite biosynthesis: the first century. *Critical reviews in biotechnology*, *19*(1), pp.1-40.

Bentley, S.D., Chater, K.F., Cerdeno-Tarraga, A.M., Challis, G.L., Thomson, N.R., James, K.D., Harris, D.E., Quail, M.A., Kieser, H., Harper, D. and Bateman, A., 2002. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3 (2). *Nature*, *417*(6885), pp.141-147.

Blin, K., Medema, M.H., Kazempour, D., Fischbach, M.A., Breitling, R., Takano, E. and Weber, T., 2013. antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. *Nucleic acids research*, *41*(W1), pp.W204-W212.

Bolger, A.M., Lohse, M. and Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), pp.2114-2120.

Britton, S. and Palacios, R., 1982. Cyclosporin A–usefulness, risks and mechanism of action. *Immunological reviews*, *65*(1), pp.5-22.

Brown, S.D., Klingeman, D.M., Lu, T.Y.S., Johnson, C.M., Utturkar, S.M., Land, M.L., Schadt, C.W., Doktycz, M.J. and Pelletier, D.A., 2012. Draft genome sequence of Rhizobium sp. strain PDO1-076, a bacterium isolated from Populus deltoides. *Journal of bacteriology*, *194*(9), p.2383.

Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.A., Barrell, B.G. and Parkhill, J., 2005. ACT: the Artemis comparison tool. *Bioinformatics*, *21*(16), pp.3422-3423.

Chiang, Y.M., Szewczyk, E., Davidson, A.D., Keller, N., Oakley, B.R. and Wang, C.C., 2009. A gene cluster containing two fungal polyketide synthases encodes the biosynthetic pathway for a polyketide, asperfuranone, in Aspergillus nidulans. *Journal of the American Chemical Society*, *131*(8), pp.2965-2970.

Cimermancic, P., Medema, M.H., Claesen, J., Kurita, K., Brown, L.C.W., Mavrommatis, K., Pati, A., Godfrey, P.A., Koehrsen, M., Clardy, J. and Birren, B.W., 2014. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*, *158*(2), pp.412-421.

Cortes, J., Haydock, S.F., Roberts, G.A., Bevitt, D.J. and Leadlay, P.F., 1990. An unusually large multifunctional polypeptide in the erythromycin-producing polyketide synthase of Saccharopolyspora erythraea. *Nature*, *348*(6297), pp.176-178.

Darling, A.E., Treangen, T.J., Messeguer, X. and Perna, N.T., 2007. Analyzing patterns of microbial evolution using the mauve genome alignment system. *Comparative Genomics*, pp.135-152.

DeLano, W.L., 2002. PyMOL.

Donadio, S., Staver, M.J., McAlpine, J.B., Swanson, S.J. and Katz, L., 1991. Modular organization of genes required for complex polyketide biosynthesis. *Science*, *252*(5006), pp.675-679.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B. and Bibillo, A., 2009. Real-time DNA sequencing from single polymerase molecules. *Science*, *323*(5910), pp.133-138.

Endo, A., 2010. A historical perspective on the discovery of statins. *Proceedings of the Japan Academy, Series B*, *86*(5), pp.484-493.
Finotello, F., Lavezzo, E., Fontana, P., Peruzzo, D., Albiero, A., Barzon, L., Falda, M.,

Di Camillo, B. and Toppo, S., 2011. Comparative analysis of algorithms for whole-genome assembly of pyrosequencing data. *Briefings in bioinformatics*, *13*(3), pp.269-280.

Ferrarini, M., Moretto, M., Ward, J.A., Šurbanovski, N., Stevanović, V., Giongo, L., Viola, R., Cavalieri, D., Velasco, R., Cestaro, A. and Sargent, D.J., 2013. An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. *BMC genomics*, *14*(1), p.670.

Fiedler, H.P., Bruntner, C., Riedlinger, J., Bull, A.T., Knutsen, G., Goodfellow, M., Jones, A., Maldonado, L., Pathom-Aree, W., Beil, W. and Schneider, K., 2008. Proximicin A, B and C, novel aminofuran antibiotic and anticancer compounds isolated from marine strains of the actinomycete Verrucosispora. *Journal of Antibiotics*, *61*(3), p.158.

Finlay, A.C., Hochstein, F.A., Sobin, B.A. and Murphy, F.X., 1951. Netropsin, a new antibiotic produced by a Streptomyces. *Journal of the American Chemical Society*, *73*(1), pp.341-343.

Finotello, F., Lavezzo, E., Fontana, P., Peruzzo, D., Albiero, A., Barzon, L., Falda, M., Di Camillo, B. and Toppo, S., 2011. Comparative analysis of algorithms for whole-genome assembly of pyrosequencing data. *Briefings in bioinformatics*, *13*(3), pp.269-280.

Goto, Y., Li, B., Claesen, J., Shi, Y., Bibb, M.J. and van der Donk, W.A., 2010. Discovery of unique lanthionine synthetases reveals new mechanistic and evolutionary insights. *PLoS biology*, *8*(3), p.e1000339.
Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P. and Tiedje, J.M., 2007. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *International journal of systematic and evolutionary microbiology*, *57*(1), pp.81-91.

Greninger, A.L., Streithorst, J., Chiu, C.Y. and Miller, S., 2017. First Complete Genome Sequence of Corynebacterium riegelii. *Genome announcements*, *5*(13), pp.e00084-17.

Greninger, A.L., Cunningham, G., Joanna, M.Y., Hsu, E.D., Chiu, C.Y. and Miller, S., 2015. Draft genome sequence of Mycobacterium elephantis strain Lipa. *Genome announcements*, *3*(3), pp.e00691-15.

Greninger, A.L., Streithorst, J., Chiu, C.Y. and Miller, S., 2016. First draft genome sequences of Neisseria sp. strain 83E34 and Neisseria sp. strain 74A18, previously identified as CDC eugonic fermenter 4b species. *Genome announcements*, *4*(6), pp.e01277-16.

Gross, H., Stockwell, V.O., Henkels, M.D., Nowak-Thompson, B., Loper, J.E. and Gerwick, W.H., 2007. The genomisotopic approach: a systematic method to isolate products of orphan biosynthetic gene clusters. *Chemistry & biology*, *14*(1), pp.53-63.

Guillen, Y., Casadellà, M., García-de-la-Guarda, R., Espinoza-Culupú, A., Paredes, R., Ruiz, J. and Noguera-Julian, M., 2016. Whole-genome sequencing of two Bartonella bacilliformis strains. *Genome announcements*, *4*(4), pp.e00659-16.

Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G., 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), pp.1072-1075

Hall, T.A., 1999, January. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. In *Nucleic acids symposium series* (Vol. 41, No. 41, pp. 95-98). [London]: Information Retrieval Ltd., c1979-c2000..

Hendrickson, L., Davis, C.R., Roach, C., Aldrich, T., McAda, P.C. and Reeves, C.D., 1999. Lovastatin biosynthesis in Aspergillus terreus: characterization of blocked mutants, enzyme activities and a multifunctional polyketide synthase gene. *Chemistry & biology*, *6*(7), pp.429-439.

Ibrahim, A., Yang, L., Johnston, C., Liu, X., Ma, B. and Magarvey, N.A., 2012. Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (iSNAP) discovery. *Proceedings of the National Academy of Sciences*, *109*(47), pp.19196-19201.

Iftime, D., Jasyk, M., Kulik, A., Imhoff, J.F., Stegmann, E., Wohlleben, W., Süssmuth, R.D. and Weber, T., 2015. Streptocollin, a type IV lanthipeptide produced by Streptomyces collinus Tü 365. *ChemBioChem*, *16*(18), pp.2615-2623.

Ikeda, H., Ishikawa, J., Hanamoto, A., Shinose, M., Kikuchi, H., Shiba, T., Sakaki, Y., Hattori, M. and Omura, S., 2003. Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. *Nature biotechnology*, *21*(5), p.526.

Jin, Z., 2011. Muscarine, imidazole, oxazole, and thiazole alkaloids. *Natural product reports*, *28*(6), pp.1143-1191.
Kardos, N. and Demain, A.L., 2011. Penicillin: the medicine with the greatest impact on therapeutic outcomes. *Applied microbiology and biotechnology*, *92*(4), p.677.

Keating, T.A., Suo, Z., Ehmann, D.E. and Walsh, C.T., 2000. Selectivity of the yersiniabactin synthetase adenylation domain in the two-step process of amino acid activation and transfer to a holo-carrier protein domain. *Biochemistry*, *39*(9), pp.2297-2306.

Koren, S., Harhay, G.P., Smith, T.P., Bono, J.L., Harhay, D.M., Mcvey, S.D., Radune, D., Bergman, N.H. and Phillippy, A.M., 2013. Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome biology*, *14*(9), p.R101.

Krzywinski, M.I. et al., 2009. Circos: An information aesthetic for comparative genomics. Genome research.

Ku, C., Lo, W.S., Chen, L.L. and Kuo, C.H., 2014. Complete genome sequence of Spiroplasma apis B31T (ATCC 33834), a bacterium associated with May disease of honeybees (Apis mellifera). *Genome announcements*, *2*(1), pp.e01151-13.

Laureti, L., Song, L., Huang, S., Corre, C., Leblond, P., Challis, G.L. and 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*, *108*(15), pp.6258-6263.

Lautru, S., Deeth, R.J., Bailey, L.M. and Challis, G.L., 2005. Discovery of a new peptide natural product by Streptomyces coelicolor genome mining. *Nature chemical biology*, *1*(5), p.265.

Lautru, S., Song, L., Demange, L., Lombès, T., Galons, H., Challis, G.L. and Pernodet, J.L., 2012. A Sweet Origin for the Key Congocidine Precursor 4-Acetamidopyrrole-2-carboxylate. *Angewandte Chemie International Edition*, *51*(30), pp.7454-7458.

Lee, S.W., Mitchell, D.A., Markley, A.L., Hensler, M.E., Gonzalez, D., Wohlrab, A., Dorrestein, P.C., Nizet, V. and Dixon, J.E., 2008. Discovery of a widely distributed toxin biosynthetic gene cluster. *Proceedings of the National Academy of Sciences*, *105*(15), pp.5879-5884.

Levy, S.E. and Myers, R.M., 2016. Advancements in next-generation sequencing. *Annual review of genomics and human genetics*, *17*, pp.95-115.

Li, Y.M., Milne, J.C., Madison, L.L., Kolter, R. and Walsh, C.T., 1996. From peptide precursors to oxazole and thiazole-containing peptide antibiotics. *Science*, *274*(5290), pp.1188-1193.

Li, M.H., Ung, P.M., Zajkowski, J., Garneau-Tsodikova, S. and Sherman, D.H., 2009. Automated genome mining for natural products. *BMC bioinformatics*, *10*(1), p.185.

Malpartida, F. and Hopwood, D.A., 1984. Molecular cloning of the whole biosynthetic pathway of a Streptomyces antibiotic and its expression in a heterologous host. *Nature*, *309*(5967), pp.462-464.

Matsuda, Y., Mitsuhashi, T., Quan, Z. and Abe, I., 2015. Molecular basis for stellatic acid biosynthesis: a genome mining approach for discovery of sesterterpene synthases. *Organic letters*, *17*(18), pp.4644-4647.

Mardis, E.R., 2017. DNA sequencing technologies: 2006-2016. *Nature protocols*, *12*(2), pp.213-218.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, *17*(1), pp.pp-10.

Mayorga, M.E. and Timberlake, W.E., 1990. Isolation and molecular characterization of the Aspergillus nidulans wA gene. *Genetics*, *126*(1), pp.73-79.

McIntosh, J.A., Donia, M.S. and Schmidt, E.W., 2010. Insights into heterocyclization from two highly similar enzymes. *Journal of the American Chemical Society*, *132*(12), pp.4089-4091.

Medema, M.H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M.A., Weber, T., Takano, E. and Breitling, R., 2011. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic acids research*, *39*(suppl_2), pp.W339-W346.

Meguro, A., Tomita, T., Nishiyama, M. and Kuzuyama, T., 2013. Identification and characterization of bacterial diterpene cyclases that synthesize the cembrane skeleton. *ChemBioChem*, *14*(3), pp.316-321.

Milne, J.C., Eliot, A.C., Kelleher, N.L. and Walsh, C.T., 1998. ATP/GTP hydrolysis is required for oxazole and thiazole biosynthesis in the peptide antibiotic microcin B17. *Biochemistry*, *37*(38), pp.13250-13261.

Nakano, C., Oshima, M., Kurashima, N. and Hoshino, T., 2015. Identification of a New Diterpene Biosynthetic Gene Cluster that Produces O-Methylkolavelool in Herpetosiphon aurantiacus. *ChemBioChem*, *16*(5), pp.772-781.

Nikolenko, S.I., Korobeynikov, A.I. and Alekseyev, M.A., 2013. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC genomics*, *14*(1), p.S7.

O'Connell, J., Schulz-Trieglaff, O., Carlson, E., Hims, M.M., Gormley, N.A. and Cox, A.J., 2015. NxTrim: optimized trimming of Illumina mate pair reads. *Bioinformatics*, *31*(12), pp.2035-2037.

Ōmura, S., Ikeda, H., Ishikawa, J., Hanamoto, A., Takahashi, C., Shinose, M., Takahashi, Y., Horikawa, H., Nakazawa, H., Osonoe, T. and Kikuchi, H., 2001. Genome sequence of an industrial microorganism Streptomyces avermitilis: deducing the ability of producing secondary metabolites. *Proceedings of the National Academy of Sciences*, *98*(21), pp.12215-12220.

Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R. and Fonstein, M., 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic acids research*, *33*(17), pp.5691-5702.

Overbeek, R., Olson, R., Pusch, G.D., Olsen, G.J., Davis, J.J., Disz, T., Edwards, R.A., Gerdes, S., Parrello, B., Shukla, M. and Vonstein, V., 2013. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic acids research*, *42*(D1), pp.D206-D214.

Perry, B.J., Fitzgerald, S.F., Kröger, C. and Cameron, A.D., 2017. Whole-Genome Sequence and Annotation of Salmonella enterica subsp. enterica Serovar Enteritidis Phage Type 8 Strain EN1660. *Genome announcements*, *5*(4), pp.e01517-16.

Prjibelski, A.D., Vasilinetc, I., Bankevich, A., Gurevich, A., Krivosheeva, T., Nurk, S., Pham, S., Korobeynikov, A., Lapidus, A. and Pevzner, P.A., 2014. ExSPAnder: a universal repeat resolver for DNA fragment assembly. *Bioinformatics*, *30*(12), pp.i293-i301.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P. and Gu, Y., 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, *13*(1), p.341.

Rackham, E.J., Grüschow, S., Ragab, A.E., Dickens, S. and Goss, R.J., 2010. Pacidamycin biosynthesis: identification and heterologous expression of the first uridyl peptide antibiotic gene cluster. *ChemBioChem*, *11*(12), pp.1700-1709.

Ribeiro, F.J., Przybylski, D., Yin, S., Sharpe, T., Gnerre, S., Abouelleil, A., Berlin, A.M., Montmayeur, A., Shea, T.P., Walker, B.J. and Young, S.K., 2012. Finished bacterial genomes from shotgun sequence data. *Genome research*, *22*(11), pp.2270-2277.

Roh, H., Uguru, G.C., Ko, H.J., Kim, S., Kim, B.Y., Goodfellow, M., Bull, A.T., Kim, K.H., Bibb, M.J., Choi, I.G. and Stach, J.E., 2011. Genome Sequence of the Abyssomicin and Proximicin-Producing Marine Actinomycete Verrucosispora maris AB-18-032. *Journal of bacteriology*, pp.JB-05041.

Röttig, M., Medema, M.H., Blin, K., Weber, T., Rausch, C. and Kohlbacher, O., 2011. NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. *Nucleic acids research*, *39*(suppl_2), pp.W362-W367.

Schmieder, R. and Edwards, R., 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, *27*(6), pp.863-864.

Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), pp.2068-2069.

Shibata, T.F., Maeda, T., Nikoh, N., Yamaguchi, K., Oshima, K., Hattori, M., Nishiyama, T., Hasebe, M., Fukatsu, T., Kikuchi, Y. and Shigenobu, S., 2013. Complete genome sequence of Burkholderia sp. strain RPE64, bacterial symbiont of the bean bug Riptortus pedestris. *Genome announcements*, *1*(4), pp.e00441-13.

Skinnider, M.A., Dejong, C.A., Rees, P.N., Johnston, C.W., Li, H., Webster, A.L., Wyatt, M.A. and Magarvey, N.A., 2015. Genomes to natural products prediction informatics for secondary metabolomes (PRISM). *Nucleic acids research*, *43*(20), pp.9645-9662.

Söding, J., Biegert, A. and Lupas, A.N., 2005. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research*, *33*(suppl_2), pp.W244-W248.

Tuncbag, N., Gursoy, A., Nussinov, R. and Keskin, O., 2011. Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nature protocols*, *6*(9), p.1341.

Utturkar, S.M., Klingeman, D.M., Land, M.L., Schadt, C.W., Doktycz, M.J., Pelletier, D.A. and Brown, S.D., 2014. Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. *Bioinformatics*, *30*(19), pp.2709-2716

Van Der Voort, M., Meijer, H.J., Schmidt, Y., Watrous, J., Dekkers, E., Mendes, R., Dorrestein, P.C., Gross, H. and Raaijmakers, J.M., 2015. Genome mining and metabolic profiling of the rhizosphere bacterium Pseudomonas sp. SH-C52 for antimicrobial compounds. *Frontiers in microbiology*, *6*.

Watts, R., Clunie, G., Hall, F. and Marshall, T. eds., 2009. *Rheumatology*. Oxford University Press, USA.

Weber, T., Charusanti, P., Musiol-Kroll, E.M., Jiang, X., Tong, Y., Kim, H.U. and Lee, S.Y., 2015. Metabolic engineering of antibiotic factories: new tools for antibiotic production in actinomycetes. *Trends in biotechnology*, *33*(1), pp.15-26.

Weber, T. and Kim, H.U., 2016. The secondary metabolite bioinformatics portal: Computational tools to facilitate synthetic biology of secondary metabolite production. *Synthetic and Systems Biotechnology*, *1*(2), pp.69-79.

Wolter, F.E., Schneider, K., Davies, B.P., Socher, E.R., Nicholson, G., Seitz, O. and Süssmuth, R.D., 2009. Total Synthesis of Proximicin A– C and Synthesis of New Furan-Based DNA Binding Agents. *Organic letters*, *11*(13), pp.2804-2807.

Wyatt, M.A., Wang, W., Roux, C.M., Beasley, F.C., Heinrichs, D.E., Dunman, P.M. and Magarvey, N.A., 2010. Staphylococcus aureus nonribosomal peptide secondary metabolites regulate virulence. *Science*, *329*(5989), pp.294-296.

Yang, G.E., Rose, M.S., Turgeon, B.G. and Yoder, O.C., 1996. A polyketide synthase is required for fungal virulence and production of the polyketide T-toxin. *The Plant Cell*, *8*(11), pp.2139-2150.

Zachow, C., Jahanshah, G., de Bruijn, I., Song, C., Ianni, F., Pataj, Z., Gerhardt, H., Pianet, I., Lämmerhofer, M., Berg, G. and Gross, H., 2015. The novel lipopeptide poaeamide of the

endophyte Pseudomonas poae RE* 1-1-14 is involved in pathogen suppression and root colonization. *Molecular Plant-Microbe Interactions*, *28*(7), pp.800-810.

Zamble, D.B., McClure, C.P., Penner-Hahn, J.E. and Walsh, C.T., 2000. The McbB component of microcin B17 synthetase is a zinc metalloprotein. *Biochemistry*, *39*(51), pp.16190-16199.

Zazopoulos, E., Huang, K., Staffa, A., Liu, W., Bachmann, B.O., Nonaka, K., Ahlert, J., Thorson, J.S., Shen, B. and Farnet, C.M., 2003. A genomics-guided approach for discovering and expressing cryptic metabolic pathways. *Nature biotechnology*, *21*(2), pp.187-190.

Zhang, Y., 2008. I-TASSER server for protein 3D structure prediction. *BMC bioinformatics*, *9*(1), p.40.

Zhang, W., Ostash, B. and Walsh, C.T., 2010. Identification of the biosynthetic gene cluster for the pacidamycin group of peptidyl nucleoside antibiotics. *Proceedings of the National Academy of Sciences*, *107*(39), pp.16828-16833.

Chapter 3. NRPS Adenylation Characterisation

3.1 Introduction

3.1.1 Importance and structure of NRPS systems

The study of the of adenylating enzymes dates back to the 1940's, with Lipman et al's., (1944) discovery of the activation of biological carboxylates as thioesters with Coenzyme A (CoA). McElroy et al., (1967) was first to note this group contained proteins far beyond the Acyl/Aryl-CoA Synthetases to include proteins which catalyse the production of an Acyl-adenylate intermediate, which is then followed by further modification in a second partial reaction. The term 'ANL superfamily' was later coined, and is comprised of the Acyl/Aryl-CoA Synthetases, firefly luciferases and the A-domains of NRPS's. An overview of the reactions catalysed by this superfamily is given in Fig. 20. This group comprises a mechanistically diverse superfamily with low sequence homology (~ 20% identity), but share a homologous structure and a conserved mechanistic step – the adenylation partial reaction. This family is distinct from other adenylate forming enzymes – such as the acyl-tRNA synthases – and is typically split into three sub-families, however, areas of structural and activity overlap exists. This enzymatic similarity is important to note when determining A-domain specificity as it can give insight into the structure and mechanism potentially responsible.

3.1.2 A-domains as 'gate-keepers'

NRPS systems and their possibly limitless potential for novel antimicrobial and antifungal compound isolation has been the focus of much research, specifically into the mechanics and protein structures which govern substrate specificity. NRPS adenylating domains catalyse the conserved primary partial reaction – activation of the amino acid substrate with ATP; this is then transferred to the pantetheine cofactor of the neighbouring thiolation domain followed by peptide bond formation. To understand and harness this potential avenue for novel compound discovery, the enzymatic mechanics must first be elucidated. The adenylation domains are considered to be the 'gate-keepers' of the reaction, and dictate the chemical nature of building blocks incorporated into the elongating peptide chain. As interest peaked

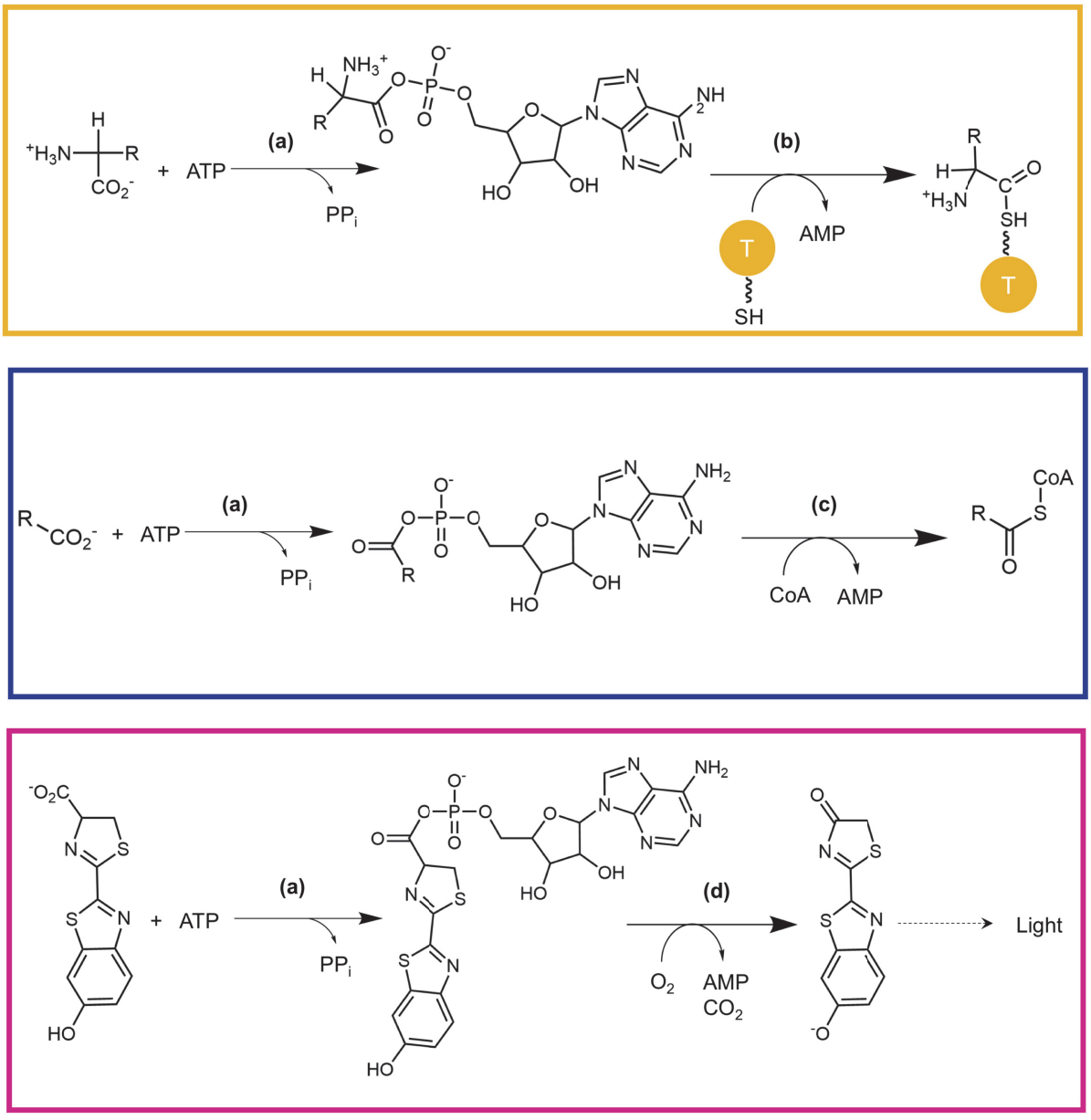


Figure 20. The three reactions catalysed by members of the ANL superfamily of enzymes. All reactions include an initial partial reaction – the adenylate forming reaction **(a)** with ATP to form an acyl-, amino acyl- or aryl-adenylate, with the release of inorganic pyrophosphate (PP_i). The adenylate intermediate then reacts in a second partial reaction, resulting in the release of AMP. For NRPS-adenylation enzymes (yellow), a pantetheine thiol group from the Thiolation (T) domain attacks the carboxylate carbon displacing the AMP leaving group **(b)** This is similar to that of the Acyl-CoA synthetases (blue), in which Coenzyme A (CoA) results in the displacement of AMP **(c)** Finally, the luciferases (pink) result in the production of the intermediate luciferyl-adenylate undergoes oxidative decarboxylation resulting in the formation of an intermediate compound that decomposes to yield a photon of light **(d)**.

in NRPS-system and further members were identified, several research groups proposed conserved sequences motifs which exert control over substrate selection. This is based on the logical assumption that A-domains with identical or similar activity would likely have a similar mechanism, and hence, protein structure and active site. This idea was first investigated by de Crecy-Lagard etal., (1995) who began isolating potential substrates based on sequence homology of NRPS-adenylation domains across a 200 aa region comprising the A3 to A6 motifs (de Crecy-Lagard et al., 1995), which was later refined to a substrate binding pocket - ~100 aa stretch between core motifs, A4 and A5 with 10 well defined principal residues (Stachelhaus et al., 1999). These residues represent 10 amino acids responsible for binding the substrate to be activated. Work by Stachelhaus et al., (1999) extensively studies these residues, summarized in Table 9, to determine their requirement, variability and function. Of these, some are described as 'wobble' positions –of high flexibility with respect to amino acid usage – for example positions 239 and 278 which are known to have a major role in distinguishing between sidechains of possible substrates, and some are described as invariant. One of the latter motifs - Asp_{235} - is involved in the stabilization of the α-amino group on the substrate, and is very highly conserved within the ANL superfamily. This position has been shown to have a variability of ~3% (Stachelhaus et al., 1999) and where variability is observed, it is typically in luciferases and Acyl-CoA synthases enzymes and not NRPS-adenylation domains. This is because most adenylation domains have an α-amino acid substrate, and so this residue is required to direct the amino acid side chain into the binding pocket; hence, mutations lead to complete enzyme inactivity. These 'guidelines' for correlation between substrate conferring residues and activity are not upheld by all NRPS systems, especially those involved in non α-amino acid incorporation; hence, the choreography that delivers the substrate bound to the T-domain is evidently more complex and not highly conserved across all the NRPS systems.

Table 9. The ten residues present in NRPS-adenylation domains which comprise the active site. They are responsible for substrate specificity, and their variability based on studies conducted on 160 different A-domains by Stachelhaus et al., (1999). The number of different amino acids found at each position is noted, and the proportional occurrence of hydrophobia, polar, acidic and basic sidechains. Position 235 highlighted blue, is thought to be completely invariant within the NRPS A-domains, and positions in pink represent highly variable, 'wobble' positions which are though to variate to accommodate substrate sidechains.

	Residue position**									
	235	236	239	**278**	**299**	301	322	330	331	517
Variability (%)	**4**	16	16	**39**	**52**	13	26	23	26	0
Number of aa used	**3**	11	13	**16**	**14**	6	12	7	16	1
Hydrophobic (%)	**3**	93	59	**41**	**77**	93	65	93	51	0
Polar (%)	**1**	3	24	**36**	**19**	7	16	7	28	0
Acidic (%)	**96**	1	11	**7**	**3**	0	11	0	15	0
Basic (%)	**0**	3	6	**16**	**1**	0	8	0	6	100

**GrsA numbering

3.1.3 Novel adenylating activity

Congocidine is a pyrrole-amide compound and one such example of a novel adenylation domain structure and mechanism. Pyrrole-amides are a family of natural products which contain one or several pyrrole-2- carboxamides which confers novel DNA binding and antimicrobial activity, and are typically synthesised by *Streptomyces* and related Actinobacteria. Congocidine, first discovered by Finlay et al., (1951), is produced by *S. netropsis*, and the biosynthetic cluster responsible for production has been well defined and the enzymatic route established. The gene cluster – denoted as *cgc* - contains 3 NRPS-like enzymes, in addition to other proteins responsible for precursor synthesis and regulation. Of the present adenylation domains one is atypical –Cgc18; this NRPS-adenylating domain contains a mutation at the 235 position, from an aspartate to serine (Juguet et al., 2009); this residue is known to be involved in activity and substrate selection, and it was initially proposed that due to a specific unique substitution in a substrate-specifying residue, Cgc18 was responsible for the novel pyrrole incorporation. This was shown to be incorrect however, and the enzymatic integration of the novel structure was in-fact due to non-NRPS adenylating enzyme Cgc3* (Al-Mestarihi et al., 2015). The elucidation of the congocidine biosynthetic route is pertinent to proximicin research as they have similar structures and both are formed by the iterative introduction of 2,4-disubstituted heterocycles, pyrrole and furans, respectively. Hence, it is assumed that their biosynthetic routes would be similar. This was supported by the identification of an adenylation domain in the Ppb cluster with the same substitution pattern in Cgc18; however, the exchange of adenylation domains coming into question. Despite the extensive research into active site residues of adenylation domains and the assembly of ‘guidelines’ which control substrate selection, the route to proximicin biosynthesis appears to be completely novel. Determining how this selection of uncharacterised NRPS-proteins work together to result in proximicin production will likely greatly increase the repertoire of knowledge into the boundaries known to govern adenylation domain activity. To do this, the novel enzymatic machinery to which activity cannot be bioinformatically assigned, substrates must be tested *in vivo* using assays which monitor adenylation activity.

3.1.4 Assessing A-domain activity

Previous work focused on the characterisation of A-domain activity, have largely used the radioactive ATP-[^{32}P] pyrophosphate (PP_i) exchange assay (Fig. 21b), which monitors aminoacyl-AMP formation (Santi et al., 1974, Otten et al., 2007). This approach is widely employed as it has been shown to maintain a high level of accuracy and be able to detect even low activity; however, many facilities are not equipped to handle radioactive materials in a high-throughput manner required for these studies. And so increasingly these methods have been replaced by simpler, non-radioactive assays; however, investigation into comparing approaches has not been widely done. Characterisation of the A-domains responsible for proximicin synthesis presented a good opportunity to test the suitability of these newer methods, in comparison to the traditional PP_i exchange assay. The malachite green assay is a newly outlined approach by McQuade et al., (2009) which monitors A-domain activity, it is reported to give results comparable to $^{32}PP_i$, circumventing the requirement for radioactive materials. The malachite green assay utilises the PP_i produced during aminoacyl-AMP formation to produce orthophosphate (P_i) via an inexpensive pyro-phosphotase which is then quantified using a malachite green reagent (Fig. 21c). Although done in differing ways, both the traditional exchange and malachite green assay measure adenylation domain activity, by the reversible phosphodiester bond cleavage and the rate of PP_i release, respectively. Despite differences, comparison of the enzyme kinetics for both methods (K_m and K_{cat}) appear to be very similar to previous reports (McQuade et al., 2009); however, these are few and relatively limited studies, and so further investigation is required. This research here gives a good opportunity to conduct an unbiased report on the applicability of the malachite green assay to this research field. The discovery of an accurate and time-efficient activity assay would facilitate the discovery of new A-domains and further the determination of the enzyme mechanics responsible for A-domain activity and specificity, both of which would propel combinatorial formation of novel NRPs.

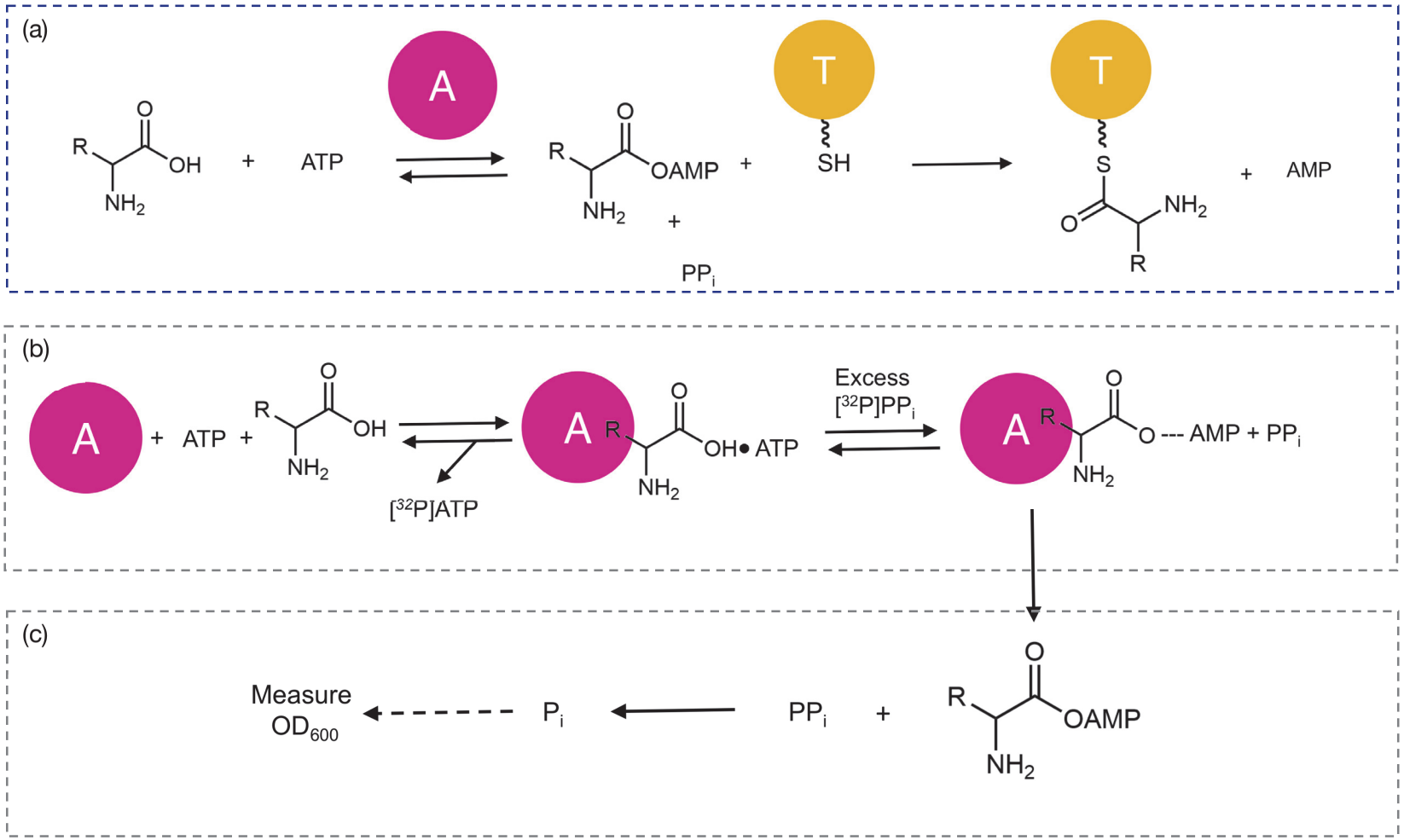


Figure 21. Illustrating the NRPS adenylation domain reaction, and how the ^{32}P exchange and malachite green assay monitor it. (a) the NRPS-adenylation domain ATP-dependent reaction yielding a T-domain tethered substrate and AMP and PP_i. **(b**) The ^{32}P exchange assay monitors the reverse adenylation reaction and the generation of [^{32}P]ATP. **(c)** In contrast, the malachite green assay utilised a pyrophosphatase enzyme to convert PP_i to P_i, which is measured using a colour change of malachite green reagent with the P_i.

3.1.5 Dependence on MbtH proteins

Work conducted on the intricate enzyme mechanics responsible for A-domain activity began decades ago, with the recognition of conserved sequences shown to be have a role in specificity; and with the discovery that many important compounds are NRPS in origin, the focus on these enzyme systems has increased dramatically. One critical recent finding is the dependence of many A-domains on MbtH-like co-proteins (Drake et al.,2007, Lautru et al., 2007); these are small (~70 amino acid) proteins which contain a conserved three tryptophan residue and are named after the primarily discovered MbtH protein from the mycobactin biosynthesis gene cluster (Drake et al., 2007). Recent research has demonstrated the ability of these proteins to influence protein production levels, A-domain activity (Drake et al., 2007) and solubility (Heemstra et al., 2009). With the advent of cheap and rapid sequencing methods, the extent of the presence of these small proteins in association with many secondary metabolite clusters, specifically NRPS genes, has been revealed. For example, the *Streptomyces* producers of the antibiotics daptomycin and A54145 have single MbtH-like proteins, DptG and LptG respectively, located downstream of the NRPS genes (Miao et al., 2005; Miao et al., 2006). This pattern is mirrored in many economically important compounds, such as antibiotics vancomycin and teicoplanin (Stegmann et al., 2006). However, despite their apparent importance, their function has remained largely obscure. Early work questioned a direct role in catalysis (Gehring et al., 1998), but more recent studies have shown they lack motifs indicative of catalysis (Drake et al., 2007), hence suggesting a role as facilitators or chaperones. The solution of the MbtH structure gave further insight into their function (Buchko et al., 2010): they have an intrinsically disorganised C-terminus which is known to be generally associated with binding of multiple partners and functional diversity. Furthermore, the overall conserved shape of MbtH-like proteins is described as ‘flat with a two-sided arrowhead’ (Drake et al., 2007). The MbtH-like protein from the pyoverdine pathway shows the localisation of the invariant tryptophan residue on one face of the arrowhead, inferring that the other face must be variable (Drake et al., 2007). This implies the possibility of the MbtH interacting with two proteins – one interaction on each face of the arrowhead – one of which is more variable in structure than the other. Where the MbtH-specific conserved tryptophan residue sits suggest that it is this side which interacts with the NRPS-proteins, exerting its function. Several studies have suggested possible

chaperone roles, including the facilitation of modification of A/T-domain bound amino acids, as well as direct roles in amino acid activation which is abolished when the MbtH-like protein is absent (Felnagle et al., 2010). Interestingly, it has also been shown that some heterologous MbtH-like proteins can partially complement missing pathway-specific MbtH-like proteins, but the extent and mechanism responsible for this is largely still unknown. This extensive research importantly demonstrates that not all MbtH-like proteins have equivalent activities, and hence, a number of mechanistic questions still remains. Further NRPS exploitation requires that these are resolved and so additional work into characterising the function and enzymatic role in NRPS-systems is required. During elucidation of the putative proximicin biosynthetic cluster in Chapter 2 (Table 4), a singular MbtH-like protein was identified to likely be involved in proximicin biosynthesis- *ppb125*. Due to the established requirement in many NRPS systems, the necessity of MbtH-like protein in proximicin biosynthesis was investigated; specifically, if production was reliant on the proximicin-specific MbtH and which, if any, of the A-domains had a dependence on MbtH-like proteins for activity and solubility.

3.1.6 Aim of Research

The aim of this research is to determine the route by which proximicin is biosynthesised in its producers *V. maris* AB18-032 and *V. sp.* str. MG37– and hence, conclude the origin of the furan moiety. Furans are economically important compounds, and a simple and inexpensive approach to large scale production would be advantageous for many reasons. As previously discussed, many genes within the *ppb* cluster resemble that determined to be involved in the biosynthesis of congocidine. This allows the putative function for *ppb* genes to be partially implied, however, experimental evidence is required to explicitly conclude the route to proximicin. As discussed in Chapter 2, ambiguity exists concerning the route to furan incorporation, leading to the subsequent production of four potential routes:

I) Ppb120 works in a similar way to Cgc18, activating a linear molecule, and once tethered is then cyclised via cyclohydratase-mediated formation of heterocycles **(outlined in chapter 2: 2.3.5);**

II) Ppb120 is responsible for furan activation, and similarities in core residues exhibited in Ppb120 and Cgc18 does not have such an effect on specificity as previously reported **(outlined in chapter 2: 2.3.5);**

III) Ppb120 works in a similar way to Cgc18, activating a linear molecule and a separate A-domain is responsible for furan activation; or

IV) Ppb120 is not responsible, and another ppb A-domain activates the furan-containing precursor in proximicin biosynthesis.

The research here outlines the production of recombinant *E. coli* strains for Ppb A-domain protein expression and purification. The successfully purified Ppb120 protein was used to conclude that the furan moiety in proximicin is incorporated pre-amide bond formation, and that Ppb120 is a highly promiscuous enzyme which can activate a range of 2,4-disubstituted pyrroles. Pyrrole-containing molecules were substituted for furan compounds due to issues involved with the synthesis of 2,4-disubstituted furans; the similar chemistry between the two families of compounds was concluded to be enough to infer activity of furans. It was also shown that the proximicin MbtH-like protein –Ppb125 – was required for Ppb120 activity, as well as solubility of other NRPS proteins. Furthermore, it was determined that it could not be partially complemented by a non-native MbtH protein – TioT. Ppb220 was shown to be completely inactive, and may mark the end of the proximicin cluster, as well as the remnants of a larger, ancient cluster. A-domains – Ppb195 and Ppb210 – were not successfully purified in an active form, and hence, the system which governs the differentiation between different proximicins – A, B and C - was not established. However, vital information on A-domain studies, in regard to the requirement of MbtH-like proteins, linker regions and chemical additives were reported, which will inform future similar work.

Chapter 3. NRPS Adenylation Characterisation

3.2 Materials and Methods

3.2.1 Media and Reagents

LB media and agar was routinely used for growing up *E. coli* strains: 10 g Tryptone, 5 g Yeast extract, 10g NaCl and 7.5 g agar (optional) and 950 mL of water added and the pH adjusted to 7.0 using NaOH and the volume brought up to 1 L, with ddH_2O followed by autoclaving. For growing *V. maris* AB18-032 and *V. sp.* str. MG37 strains, SSG media, first described by Fiedler et al., (2008): 10 g soluble starch, 10 g glucose, 10 g glycerol, 2.5 g corn-steep powder, 5 g Bacto peptone, 2 g yeast extract, 1 g NaCl and 3 g $CaCo_3$, dissolved in 1 L of tap water and adjusted to pH. 7.3 with NaOH before sterilization.

Through the course of this research, issues encountered required many constructs involving many vector types and DNA fragment lengths of the same gene, created via different primer sets. Where possible I have included tables outlining exactly what genotype shorthand is referring to aid in understanding. For sake of clarity, when referring to DNA constructs they will contain the vector utilised in the name i.e. p28 (pET28) however, the resultant strains produced encompassing this vector will not contain this.

3.2.2 Verrucosispora maris AB18-032 and V. sp. str. MG37 *gDNA extraction*

Genomic DNA was extracted from *V. maris* AB18-032 and *V. sp.* str. MG37 using the cetyltrimethyl ammonium bromide (CTAB) method. 30 mL cultures of *V. sp.* str. *MG37* were grown up in SSG media at 28°C for 16 hrs and harvested at 16,000X g and re-suspended in 5 mL TE25S buffer (25 mM Tris-HCl pH. 8; 25 mM EDTA pH. 8 and 0.3 M sucrose). Lysozyme was added to final concentration 2 mg/mL and incubated for 1 hr at 37°C. Proteinase K and SDS were added to final concentration 0.18 mg/mL and 0.5% respectively, and incubated for 1 hr with occasional inversion at 55°C. NaCl added to final concentration 0.8 M and mixed; CTAB/NaCL added and mixed followed by incubation at 55°C for 10 mins followed by cooling to 37°C. Chloroform/isoamyl alcohol added and mixed my inversion for 30 min. Centrifuge at

13,500 X g at 20°C for 15 mins. Supernatant decanted and 0.6 volumes of isopropanol added and mixed; after 3 min DNA spooled and rinsed in ethanol and air dried before dissolving in 1 mL TE buffer (1 mM Tris-HCl; 1 mM EDTA pH. 8) at 55°C. Extracted gDNA was confirmed by running products on 1% agarose gel stained with ethidium bromide. Genomic DNA was quantified using Qubit 2.0 Fluorometer according to ThermoFisher guidelines.

3.2.3 Initial studies I – Vector construction

pOPINF_ppb vector construction

Primers were designed to amplify the entire proximicin NRPS A- domain containing proteins in *V. maris* AB18-032: *ppb120* (Primer # 11 & 12), *ppb195* (Primer # 13 & 14), *ppb210* (Primer # 15 & 16) and *ppb220* (Primer # 17 & 18); primers were designed with In-Fusion™ extensions. Amplification of proximicin genes *ppb120*, *ppb195*, *ppb210* and *ppb220* was done using MyFi taq polymerase. The PCR mixture (25 µL) contained 2X MyFi™ High-Fidelity Polymerase Master-Mix (Bioline), primers (0.5 µM each), template DNA (25 ng) and DMSO (3%). The PCR was performed using MyFi™ Thermal Cycling Procedure. Annealing temperature was varied depending on T_m of primers used – varying from 55-67°C. PCR was checked by gel electrophoresis, using standard DNA visualisation protocol. PCR products were gel purified and eluted into 50 µL water, using the GenElute™ PCR Clean-Up Kit (SIGMA NA1020) according to Sigma guidelines. PCR fragments were then quantified using Qubit 2.0 Fluorometer according to ThermoFisher guidelines.

The pOPINF plasmid (Addgene) was transformed into NEB® 5-alpha competent *E. coli* cells, using the five-minute NEB transformation protocol (cat. no C2987H), with an additional outgrowth step (referred to hence as the Adapted Five-Minute Protocol): NEB® 5-alpha competent *E. coli* cells were allowed to thaw on ice ~5 min, and ~25 ng of DNA was added and mixed with the cells and placed on ice for 2 mins. The mixture was heat shocked at 42°C for exactly 30 seconds and placed on ice for 2 mins. 950 mL of LB media was added and incubated at 37°C for 45 mins. 100% of transformants were plated on LB plates with relevant antibiotic selection – here 30 µg/mL ampicillin, overnight at 37°C. 50 mL cultures were used to prepare high quantities of pure plasmid using the GenElute™ Plasmid Miniprep Kit (SIGMA PLN10), according to the supplier guidelines. The plasmid was then linearized by

restriction digest enzymes *Nco*I and *Kpn*I according to NEB Double Digest Finder protocol and gel purified using GenElute™ Gel Extraction Kit (SIGMA NA1111), according to manufactures protocol, and eluted into 100 µL water. The pOPINF backbone and PCR fragments were then annealed using an In-Fusion reaction: ~80 ng PCR product and ~100 ng pOPINF backbone were mixed with 2 µL 5X In-Fusion HD enzyme Premix (TakaraBio), and water added to 10 µL. The reaction mix was incubated at 50°C for 15 min. This mix was used directly for transformation into NEB® 5-alpha competent *E. coli* cells using the adapted five-minute adapted protocol. The following day, when visible colonies were present, colony PCR was performed: ~ 8-10 colonies/ treatment were picked using a sterile wooden pick and simultaneously suspended in 10 µL ddH_2O and spread on a LB plate with appropriate antibiotics. The mixture was boiled in a thermo-cycler for 10 minutes at 98°C. The resultant solution was used as the source of DNA for a PCR. For colony PCR, MyFi Taq Master Mix was used with standard MyFi conditions, along with correlating primers used in vector construction. Strains showing the correct plasmid determined by correct fragment length were purified using GenElute™ Plasmid Miniprep Kit (SIGMA PLN10), according to the supplier guidelines and transformed into E. cloni Express BL21(DE3) *E. coli* expression strain (Lucigen) using the Adapted Five-Minute Protocol.

Control vector

A control plasmid containing a well characterised A-domain– Ncpb_A_4 – on pET28a was generously provided by Dr. Garneau-Tsodikova. This adenylation is well activated by isoleucine, and its activity is very highly characterized, and so this was used as a control for both induction strategies and the malachite green adenylation domain activity assay. This plasmid was transformed using the Adapted Five-Minute Protocol into E. cloni Express BL21(DE3) *E. coli* expression strain (Lucigen) and its presence confirmed by colony PCR using Ncpb_A_4 specific primers – Primer # 19 & 20.

Protein expression and purification

A singular proximicin protein– Ppb120 – was chosen to be used for preliminary protein expression, purification and assay work, alongside the control – Ncpb_A_4. A routine protocol was developed for small protein expression and purification studies:

Ppb120 and Ncpb_A_4 were grown up in LB media, supplemented with ampicillin (50 µg/mL) and kanamycin (50 µg/mL), respectively. 25 mL cell cultures were grown to OD_{600}. 0.4 at 37°C /200 rpm. 1 mL un-induced sample removed, spun at 16,000 X g for 15 mins, supernatant removed and the cell pellet frozen at -20°C. Cells were induced with IPTG final concentration 1mM and left at 37°C with 200 rpm shaking overnight. 1 mL of culture removed and spun at 16,000 X g for 15 mins and the supernatant removed. The un-induced pellet was mixed 1:4 and the induced 1:20 with 2X SDS loading dye and boiled for 5 min. The samples were briefly spun at 16,000 X g and 5 µL loaded onto a pre-stained SDS-PAGE gel, ran at 200V for 40 mins. The gel was imaged using the BioRad imager. Proteins were purified using His-Pur™ Ni-NTA Spin Columns, 0.2 mL (Thermo Scientific, cat. no. 88224) according to the manufactures directions.

3.2.4 Initial studies II - Malachite Green assay

Assay development

To ensure the malachite green reagent and assay worked accurately, the assay was run with varying amounts of phosphate and varying amounts of PP_i. These were carried out in 398-well plates with each reaction (100 µL) containing: Tris–HCl (pH 7.5, 50 mM), NaCl (100 mM), $MgCl_2$ (10 mM), inorganic pyrophosphatase (0.2 U/mL) and ATP (5 mM). To check the applicability malachite green reagent concentrations (0 – 1000 pmol) of phosphate (potassium phosphate monobasic) was added and the OD_{630} measured. To ensure that the enzymatic part of the assay worked, a similar test was done but with varying levels of PP_i (0 – 100 µM).

Control Adenylation domain: Activity of Ncpb_A_4 control

The well characterized adenylation domain protein - Ncpb_A_4 – was used to test the efficacy of the P_i detection assay; this was performed in 396-well plates. All 20 proteogenic amino acids were tested to determine substrate specificity. Reactions (40 µL) containing Tris-HCl (50 mM, pH 7.5) $MgCl_2$ (10 mM), NaCl (100 mM), inorganic pyrophosphatase (0.2 U/mL)], ATP (5 mM), amino acid in water (6mM) and adenylation domain protein (1 µM) were performed at 25°C. The assay was performed as follows: in a 384 plate, 10 µL of amino acid (in H_2O, 6 mM) was added in duplicate, followed by the addition of the adenylation domain protein (Ncpb_A_4) and inorganic pyrophpsophatase [in 1X reaction buffer: Tris-HCl (50 mM, pH 7.5)

$MgCl_2$ (10 mM), NaCl (100 mM)]. The reactions were initiated by addition of ATP (5 mM), incubated for 2, 4, 6, 8, 12, 16, 24, 32 and 60 min. Reactions were quenched with the addition of malachite green reagent (10 µL). After 15 min of colour development, the liberated P_i concentration was measured by reading absorbance at 600 nm. The experiments were carried out in duplicate for each substrate concentration with a negative control (no adenylation domain protein).

Activity of pOPINF_120 using malachite green assay

The activity of Ppb120 was tested as above using the purified Ppb120 instead of Ncpb_A_4, all proteogenic amino acids, as well as an array of pyrrole containing compounds were tested, as well as demonstrated substrates of similar A-domain Cgc18 – guanidinoacetate (GA) (Fig. 22).

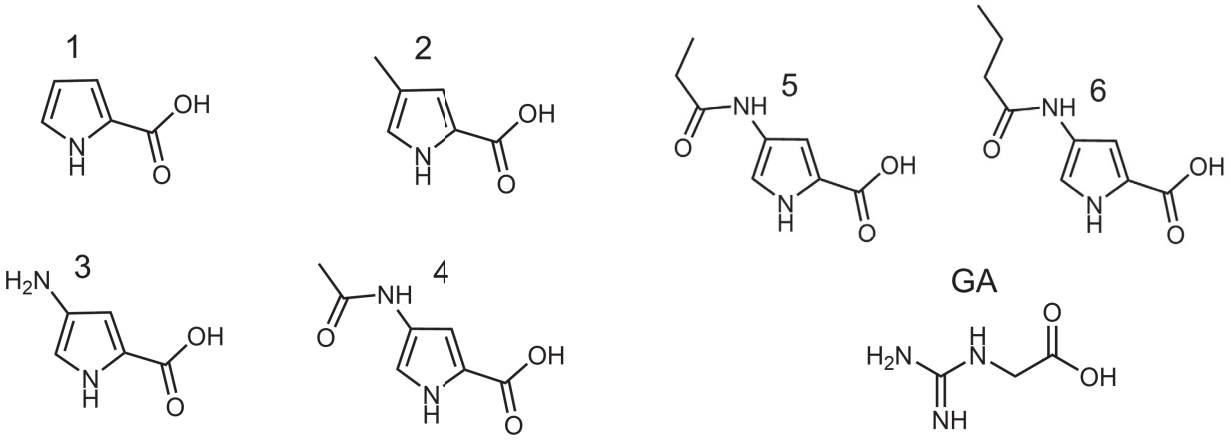


Figure 22. Pyrrole containing analogues of furan-containing precursors.
Compounds with the same substituted pattern as the predicted substrate involved in proximicin biosynthesis. GA is the substrate of Cgc18A, the adenylation domain containing the same amino acid substitution as Ppb120, and hence, potential alternative substrate.

3.2.5 *Chemical synthesis of pyrrole containing precursors*

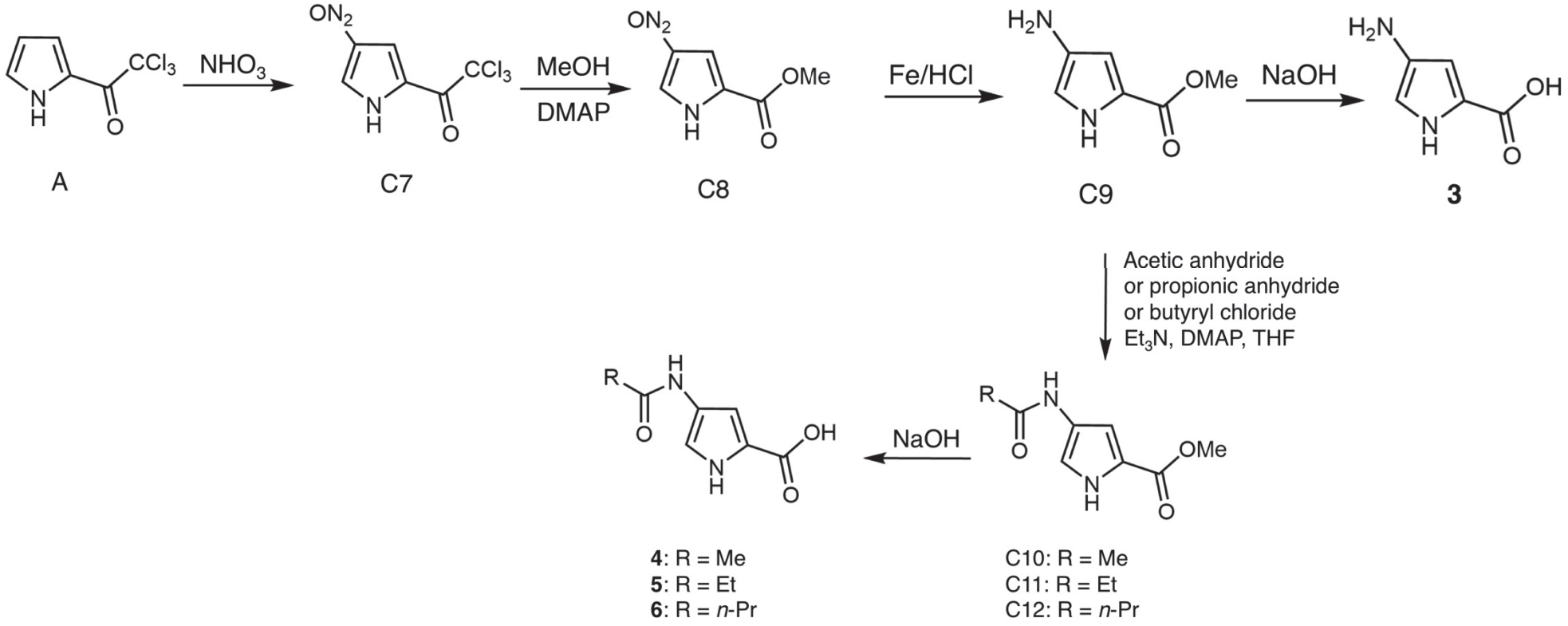


Figure 23. Chemical synthesis route for pyrrole-containing precursors. The chemical synthesis scheme for the production of pyrrole-containing precursors 3-6, from A – 2-(Trichloroacetyl)pyrrole (Sigma Aldrich 395137).

Reactions outlined in Figure 23.

Synthesis of compound C7

Compound C7 was prepared according to a previously published protocol (Al-Mestarihi et al., 2016). A solution of A (5 g, 23.53 mmol) in Ac_2O (27mL) was cooled to -20°C and fuming NHO_3 (2.83 mL) was added dropwise over a period of 30 min. The mixture was then slowly warmed to room temp for 12 hrs. The mixture was poured into ice-H_2O and filtrate collected, washed several times with toluene, and dried over $MgSO_4$. The residue was re-crystallised using $CHCl_3$:EtOH/95:5 and compound C7 was isolated as an off-white solid (31% yield), this matches the literature which reported a 41% yield (Al-Mestarihi et al., 2016).

Synthesis of compound C8

Compound C8 was prepared according to a previously published protocol (Al-Mestarihi et al., 2016). Compound C7 (0.42g, 1.62 mmol) was dissolved in dry MeOH (4 mL) and a catalytic amount of DMAP (10.5 mg, 5%) added. The reaction was stirred for 12 hrs at 70°C, until all C7 had disappeared. The reaction mixture was evaporated to dryness under reduced pressure to yield C8 (87%) as an off-white solid, which could be used without any further purification, this matches the literature which reported a 93% yield (Al-Mestarihi et al., 2016).

Synthesis of compound C9

Compound C9 was prepared according to a previously published protocol (Al-Mestarihi et al., 2016). A mixture of iron (2.66g, 47.65 mmol) and HCl (0.49 mL, 12 M) in EtOH (16 mL) was stirred at 65°C for 2 hrs. A solution of NH_4Cl (8 mL, 24 % in H_2O) was added to which compound C8 was added over a period of 25 mins. The solution was stirred for 12 hrs at 70°C. The mixture was then cooled to room temp and filtered through celite. The resultant solution was concentrated to dryness under reduced pressure and the residue dissolved in EtOAc. The resulting solution was washed twice in NaOH (50 mL, 1 M) and brine (50 mL) and dried over MgSO4 and evaporated to dryness to yield C9 (40%) as a yellow solid, this matches the literature which reported a 45% yield (Al-Mestarihi et al., 2016).

*Synthesis of compound **3***

Compound C3 was prepared according to a previously published protocol (Al-Mestarihi et al., 2016). Compound C9 (0.1 g, 0.71 mmol) was dissolved at 0°C in MeOH (4 mL), and NaOH (4 mL, 1 M) was added dropwise. The reaction was then stirred at room temp for 12 hrs, then cooled on ice and neutralized with HCl (1M). A precipitate formed immediately and the solid collected via filtration and washed with H_2O and Et_2OH to yield compound **3** as a dark brown solid at a 67% yield. this matches the literature which reported a 67% yield (Al-Mestarihi et al., 2016).

All compounds were synthesised following the previously published protocol, and confirmed using ^1H NMR taken (400 MHz, $(CDC_3)_2SO$). All other substrates tested were kindly supplied by Dr. Garneau-Tsodikova, and the route used also outlined.

3.2.6 Identification of adenylation domains and primer design

Identification of adenylation domain boundaries

BLAST search was used to reveal characterised adenylation domain proteins, these were aligned with proximicin A-domain protein sequences using BioEdit (Hall, 1999), and then the boundaries of the proximicin adenylation domains identified. Sets of primers were designed for each to i) include just the A-domain and ii) include linker regions with neighbouring domains. Many cloning strategies must be employed to create constructs with optimum protein purification and, depending on which cloning strategies was used dictates the modifications these primers underwent. For ease of understanding, Table 10 summarises all primer sets used for vector construction, including the domains present in the resultant protein and strains these vectors were used to produce.

Primer design – pET28a

For pET28a clones, primers were designed to include restriction sites *Nde*I and HindIII in the forward and reverse primers, respectively. Primers were designed by hand, checked to prevent any frame shifts and optimised to be appropriate length, Tm and GC%. An additional stop codon was added to each reverse primer.

Primer design – pET30a_Xa/LIC

For pET30_Xa/LIC cloning, 5’ end adaptor sequences were added to allow Xa/LIC cloning:

Sense primer: 5' GGT ATT GAG GGT CGC – insert-specific sequence 3'

Antisense primer: 5' AGA GGA GAG TTA GAG CCX–insert-specific sequence 3'

Where X is any base.

Primer design – pACYC_Duet1

Primers designed for cloning into pACYC_Duet1 using the restriction sites *Eco*RI and *Hind*III in the forward and reverse primers, respectively.

Table 10. Primer and vector sets used to produce adenylation domain expression strains. To aid clarity, all the different vectors produced, what NRPS domains they contain, primers used for their production and the names of the strains each was used to produce.

Gene name in *ppb* cluster	***ppb120***				***ppb195***				***ppb210***		***ppb220***
Domains		L-A-T-L			L-A-T-Te -L				L-A-T-L		L-A-T-L
Vector	pET28a	pET28a	pET28a	pET28a	pET28a	pET28a	pET30_ Xa/LIC	pET28a	pET28a	pET30_ Xa/LIC	pET28a
Primer #'s	21 & 24	21 & 23	22 & 24	22 & 23	29 & 31	29 & 30	35 & 37	26 & 28	26 & 27	39 & 40	32 & 33
Domains in frag.	LATL	LA	ATL	AT	LATTeL	LAT	LAT SYN	ATL	AT	LATL	LAT
Name of vector	p28_120_ LATL	p28_120_ LA	p28_120_ ATL	p28_120_ AT	p28_195_ LATTeL	p28_195_ LAT	p30_ SYN195_ LAT	p28_210_ ATL	p28_210_ AT	p30_210_ LATL	p28_220_ LAT
Name of strain – with no Mbth	120_LATL	120_LA	120_ATL	120_AT	195_ LATTeL	195_LAT	SYN195_ LAT	210_ATL	210_AT	210_LATL	220_LAT
Name of strain – with TioT	TioT_ 120_LATL	TioT_ 120_LA	TioT_ 120_ATL	TioT_ 120_AT	TioT_195_ LATTeL	TioT_ 195_LAT	TioT_ SYN195 _LAT	TioT_ 210_ATL	TioT_ 210_AT	TioT_ 210_LATL	TioT_ 220_LAT
Name of strain – with Ppb125	125_ 120_LATL	125_ 120_LA	125_ 120_ATL	125_ 120_AT	125_ 195_ LATTeL	125_ 195_LAT	125_ SYN195 _LAT	125_ 210_ATL	125_ 210_AT	125_ 210_LATL	125_ 220_LAT

3.2.7 Amplification of ppb genes

PCR amplification, gel purification and quantification

The PCR mixture (50 µL) contained Phusion® High-Fidelity polymerase (0.10 units, NEB), dNTPs (200 µM), Phusion® High Fidelity buffer (1X), primers (0.5 µM each), *Verrucosispora maris* DNA (50 ng) and DMSO (3%). The PCR was performed using a typical thermal cycling procedure referred to hence as: Phusion Thermo-Cycling Conditions - 1 cycle, 98°C for 30 s; 30 cycles, 98°C for 10 s, 71°C for 30 s and 72°C for 3:00 min; and 1 cycle 72°C for 10 min. Annealing temperature was varied depending on T_m of primers used. PCR was checked using gel electrophoresis using the standard technique. PCR fragments of the correct size were gel purified using GenElute™ Gel Extraction Kit, according to Sigma Aldrich guidelines. Fragments were eluted into 50 µL of ddH_2O and quantified using the Qubit 2.0 Fluorometer according to ThermoFisher guidelines.

Thermo-cycling conditions

Basic PCR condition were kept as Phusion® Thermo-Cycling Conditions but adapted to increase yields. The annealing temperature was varied from 54-72°C in 1°C increments, depending on the T_m of the primers being used. Initial denaturation was increased from 98°C to 103°C in 1°C increments; the time was increased from 3 min to 8 min in 30 sec increments.

PCR additives

Both enzymes: 2X MyFi™ High-Fidelity Polymerase Master-Mix (Bioline, error rate: 3.5 $X10^{-6}$) and HF Phusion® (error rate: $9.5X10^{-7}$) were used. Mg^{2+} concentration was increased form 1.4 mM – 4 mM in 0.2 mM increments. DMSO was increased from 3% to 10% in 1% increments. Betaine was added, in concentration 0.25 M to 1.25 M in 0.25 M increments.

Slow-Down PCR conditions

A master mix was created with final concentrations: KCl (50 mM); Tris-HCl (10 mM, pH 8.8); Nonidet-P40 (0.08%); $MgCl_2$ (1.5 mM); dATP, dCTP, dTTP (all 200 mM); 7-deaza-2'deoxyguanosin (150 mM) (NEB); GTP (50 mM); and Taq polymerase (0.01 U) (Fementas, cat. no. EP0281). To 43 µL of master mix, each Primer (final concentration 100 mM) and DNA (final amount: ~80 ng) was added, with ddH_2O to

make a total volume of 50 µL. The slowdown PCR conditions are as follows: amplifications are run for 48 cycles with 30 s denaturation at 95°C, 30 s annealing with a progressively lowered temperature from 70°C to 53°C (at a rate of 1°C every third cycle) and a primer extension of 40 s, followed by 15 additional cycles with an annealing temperature of 58°C. The cycling conditions were altered to have a lower ramp rate of 2.5°C/s and a cooling rate of 1.5°C/s for reaching an annealing temperature, with 48 cycles.

Hot-start PCR conditions

Phusion® Hot Start Flex 2X Master Mix (BioLabs, cat. no. M0536L) was used. The PCR mixture (50 µL) contained Phusion® HF Polymerase containing an aptamer based inhibitor (0.0025 units, NEB), dNTPs (0.5 µM each), 5X HF Phusion® buffer (1X), primers (0.5 µM each), *V. maris* AB18-032 DNA (50ng) and DMSO (3%). The PCR was performed using Phusion Thermo-Cycling Conditions.

3.2.8 Design and construction of synthetic ppb195

Design of SYNppb195

The A-domain protein ppb195 was codon optimised for *E. coli* K-12, successful optimization was checked by analysing the sequence for: a high codon adaptation index (>0.8), codon frequency index of >30%. A synthetic construct was produced in three G-block fragments. Primers were designed for both: the amplification of each synthetic G-Block and amplification of the joined synthetic fragments with pET30a_Xa/LIC adapter arms.

Gibson assembly of SYNppb195

Gibson assembly was used to combine the fragments of the synthetic *ppb195*. Overlap PCR was preformed using a 25 µL volume reaction: Q5 2X master mix (1X), primers (#'s 35-38) (0.5 µM) and G-block fragments 1 and 2 (each 0.4 ng/µL). With Q5 specific conditions: 1 cycle, 98°C for 30 s; 15 cycles, 98°C for 10 s, 71°C for 30 s and 72°C for 1:00 min; and 1 cycle 72°C for 2 min. The fragment and linearized pET30a_Xa/LIC vector were assembled: pET30 (100 ng), *SYNppb195* amplicon (80 ng), NEBuilder assembly master mix (10 µL). 5 µL of ligation reaction was used to transform *E. coli* NEBalpha using the Adapted Five-Minute Protocol. Strains were checked for the presence of a correct construct using the described colony PCR

method using primer #'s 35 & 37; correct strains were grown up in 10 mL LB media with 50 µg/mL kanamycin, plasmids were extracted according to Sigma Aldrich GenElute Mini-prep kit and checked using sequencing.

3.2.9 Construct preparation

pET28a vector production: restriction digest; purification; quantification, ligation and transformation:

Restriction digest mixture (50 µL) contained purified PCR DNA (1 µg), RE buffer (1X NEB CutSmart buffer, cat. no. B7204S), *Hind*III-HF and *Nde*I-HF enzymes (10 units, NEB); and incubated at 37°C overnight. Fragments were gel purified using GenElute™ Gel Extraction Kit (SIGMA NA1111) according to Sigma Aldrich instructions, and eluted in 50 µL ddH_2O. DNA quantity determined by 2.0 Fluorometer according to ThermoFisher guidelines. Ligation mixture (20 µL) contained 10 X T4 ligase buffer (5X, NEB), linearized pET28a (50 ng), insert DNA (37.5 ng) and T4 DNA ligase (400 U, NEB). Incubated at 37°C overnight. 3 µL of ligation reaction was used to transform *E. coli* TOP10 (ThermoFisher Scientific) using the Adapted Five-Minute Protocol. Strains were checked for the presence of a correct construct using colony PCR with corresponding primer sets as previously described in insert amplification; strains were grown up in 10 mL LB media at 37°C with 50 µg/mL kanamycin, plasmids were extracted using GenElute™ Plasmid Miniprep Kit (SIGMA PLN10), according to the supplier guidelines, and checked using sequencing.

pET30_Xa/LIC vector production: Ligation

For pET30_Xa/LIC cloning, the Novagen vector kit was used (Novagen, cat. no. 70073-3). T4 DNA polymerase treatment of target insert, annealing into supplied pET30_Xa/LIC vector and transformation into NovaBlue GigaSingles™ Competent Cells (cat. no. 71127) was done according to Novagen (User protocol TB184). Strains were checked for the presence of a correct construct using the described colony PCR method with corresponding primer sets used for fragment generation; correct strains were grown up in 10 mL LB at 37°C media with 50 µg/mL chloramphenicol, plasmids were extracted according to Sigma Aldrich GenElute Mini-prep kit and checked using sequencing.

pACYC_Duet vector production: restriction digest; purification; quantification and, ligation:

Restriction digest mixture (50 µL) contained *ppb125* purified PCR DNA (1 µg), RE buffer (NEB CutSmart buffer 1X), *Hind*III-HF and *Eco*R1-HF enzymes (10 units); and incubated at 37°C overnight. Fragments were gel purified using GenElute™ Gel Extraction Kit, according to Sigma Aldrich guidelines. DNA quantity determined using Qubit 2.0 Fluorometer according to ThermoFisher guidelines. Ligation mixture (20 µL) contained ligase buffer (5X NEB T4 ligase buffer), linearized pACYC (50 ng), digested *ppb125* insert DNA (37.5 ng) and T4 DNA ligase (400 U, NEB). Incubated at 37°C overnight. 3 µL of ligation reaction was used to transform *E. coli* TOP10 (ThermoFisher Scientific) using the Adapted Five-Minute Protocol. Strains were checked for the presence of a correct construct using the described colony PCR method; correct strains were grown up in 10 mL LB media with 50 µg/mL chloramphenicol, plasmids were extracted according to Sigma Aldrich GenElute Mini-prep kit and checked using sequencing.

3.2.10 Production of competent E. coli expression strain with mbtH inactivation

Production of expression strains

Three expression strains were produced, all starting from *E. coli* BL21(DE3)ΔybdZ. This strain was generously provided by Dr. Garneau-Tsodikova; this strain has the chromosomal copy of *ybdZ* (*E. coli* MbtH-like protein) inactivated and replaced by a chloramphenicol resistance marker and hence, the native *E. coli* MbtH-like protein will not interfere with future applications.

A vector containing TioT - an active MbtH protein involved in the biosynthesis of Thiocoraline - was also kindly supplied by Dr. Garneau-Tsodikova to allow testing of non-native MbtH-like proteins effect on proximicin A-domain activity. pET28a_TioT, in addition to *ppb*-specific MbtH-like protein – pACYC_Duet_ppb125 were transformed into BL21(DE3)ΔybdZ to create strains for the A-domain vectors to be transformed into.

Production of competent expression cells and testing transformation efficiency of BL21(DE3)ΔybdZ_pACYC_ppb125

A typical procedure to produce cell competency was used for all expressions trains, but efficiency only tested for *BL21(DE3)ΔybdZ_pACYC_ppb125* : a starter culture of the strain was grown in LB supplemented with relevant antibiotics, was grown overnight at 37°C; this was diluted 1:100 into 1 L fresh LB with antibiotics and grown to OD_{600}. 0.4, cells were harvested at 3000 X g for 15 min at 4°C in ice cold centrifuge bottles. The pellet was re-suspended in 100 mL ice cold 100 mM $MgCl_2$. Cells were harvested again at 2000 X g for 15 min at 4°C. The pellet was re-suspended in ice cold 100 mM $CaCl_2$. The suspension was incubated on ice for ~1 hour. Cells were then harvested at 2000 X g for 15 min at 4°C. The pellet was re-suspended in ice cold 85 mM $CaCl_2$, 15% glycerol. The cells were harvested at 1000 X g for 15 mins at 4°C. The pellet was re-suspended in 2 mL ice cold 85 mM $CaCl_2$, 15% glycerol. 50 µL aliquots were snap frozen in liquid nitrogen and stored at -80°C.

To test the efficacy of the transformation strain: 100 pg of control pUC19 DNA was transformed into 50 µL of each of the expression strain cells using the Adapted Five-Minute Protocol. The transformation efficiency was calculated by the equation:

TE = Colonies/µg/Dilution

Colonies = the number of colonies counted on the plate

µg = the amount of DNA transformed expressed in µg

Dilution = the total dilution of the DNA before plating

Transformation

All vector constructs were transformed into expression strains:

i. *E. coli* BL21(DE3)ΔybdZ (with no MbtH-like protein);
ii. *E. coli* BL21(DE3)ΔybdZ_pACYC_ppb125 (with proximicin-specific MbtH-like protein), and
iii. *E. coli* BL21(DE3)ΔybdZ_pET28a_TioT (non-specific MbtH-like protein).

This was done using the Adapted Five-Minute Protocol. Strains were checked for the presence of the correct constructs using the described colony PCR method.

3.2.11 Protein purification

Proximicin biosynthetic gene constructs: p28_120_LATL, p28_220_LAT and p30_210_LATL were tested for their expression and protein solubility in three conditions: no MbtH protein, TioT and with proximicin specific MbtH – gene *ppb125*, co-expressed. The same purification method was used: cells were grown up in 10 mL LB media starter cultures, supplemented with $MgCl_2$ (10 mM), kanamycin for pET30a_Xa/LIC (50 µg/mL) and chloramphenicol for pET28a (50 µg/mL) constructs. 25 mL cell cultures were grown to OD_{600}. 0.4 at 37°C /200 rpm, and then cooled at 16°C for 1 hr. 1 mL un-induced sample removed, spun at 16,000 X g for 15 mins, supernatant removed and the cell pellet frozen at -20°C. Cells were induced with IPTG final concentration 1 mM and left at 16°C with 200 rpm shaking overnight. 1 mL of culture removed and spun at 16,000 X g for 15 mins and the supernatant removed. The un-induced pellet was mixed 1:4 and the induced 1:20 with 2X SDS loading dye and boiled for 5 min. The samples were briefly spun at 16,000 X g and 5 µL loaded onto an SDS-PAGE gel, ran at 200 V for 40 mins. The gel was stained with cromassie blue and de-stained overnight.

Large scale protein purification: 125_120_LATL and 125_220_LAT

To get large quantities of purified protein, 60 mL starter cultures of 125_120_LATL and 125_220_LAT with kanamycin (50 µg/mL) and chloramphenicol (50 µg/mL) were grown overnight at 37°C; these were diluted 1:100 into fresh LB with antibiotics and supplemented with $MgCl_2$ (10 mM) and grown at 37°C /200 rpm to OD_{600}. 0.4 then cooled at 16°C for 1 hr. IPTG was added to final concentration 1 mM and grown overnight at 16°C /200 rpm. The cell culture was spun 5000 rpm/ 4°C / 10 mins and the pellet washed in water and then lysis buffer (40 mM imidazole, 20mM tris-HCl pH. 8.0, 400 mM NaCl, 10% glycerol) and then re-suspended in ~ 40 mL lysis buffer. DDT and DMSF (both final concentration 1 mM) were added. Cell suspension was sonicated for 6 cycles of pulses for 5 seconds and 20 % amplitude with 5 second gap, and then centrifuged 16,000 rpm/ 4°C / 30 min. The supernatant removed and moved to a new tube and centrifuged again. 250 µL/L of culture of Ni-NTA beads (HisPur Ni-NTA resin, ThermoFisher) were prepared by washing twice with diH_2O. The supernatant was mixed with the prepared beads and batch bound on a shaking platform 2 hrs / 4°C. The bead solution was added to a column and the liquid drained; the beads were then washed with 10 X 10mL fractions of wash buffer

(40 mM imidazole, 20 mM tris-HCl pH. 8.0, 400 mM NaCl, 10% glycerol) and the flow through collected. The protein was eluted using 3 X 5 mL elution buffer; wash and elution fractions run on an SDS-PAGE gel and fractions containing purified protein of the correct MW were dialysed using 3.5 K MWCO tubing (SnakeSkin Dialysis tubing, ThermoFisher). The protein was dialysed in three steps, in 2 L dialysis solution firstly for 3 hrs / 4°C, then the dialysis solution refreshed and left overnight at 4°C and then for a further 3 hrs in fresh dialysis solution, supplemented with glycerol (10%). Proteins were concentrated using a regenerated cellulose membrane column (Amicon Ultra-15) to a concentration 50-100 µM, flash frozen in liquid nitrogen and stored at -80°C.

Troubleshooting protein purification of insoluble 125_210

Different length fragments of *ppb210* were used to try and resolve solubility issues: *ppb210* strains: 125_210_LATL, 125_210_AT and 125_210_ATL were grown up in various conditions changing temperature, OD_{600} of induction, IPTG concentration and presence of non-ionic detergents (Triton-X100, NP-40 and IGEPACa-40) to determine if solubility could be achieved. Cells were grown to OD_{600} of various points between 0.2 and 0.6 at 16°C and 25°C, cooled to 16°C if required and induced with 1 mM IPTG and grown overnight at 16°C. Cells were harvested at 5000 rpm/ 4°C / 10 mins and the pellet re-suspended in either lysis buffer or non-ionic detergents and left for 60 min at 4°C on a shaking platform. The non-ionic detergents used were Triton-X100, NP-40 and IGEPACa-40 along with 6 M urea being used as a control. All tested conditioned summarised in Table 11. Cell suspension was sonicated for 6 cycles of pulses for 5 seconds and 20% amplitude with 5 second gap, and then centrifuged 16,000 rpm/ 4°C / 30 min. The supernatant removed and moved to a new tube and centrifuged again. The supernatant was removed, and labelled as Sol. 1. The pellet was re-suspended in detergent and left again for 60 min at 4°C with shaking. The suspended insoluble pellet was then centrifuged 16,000 rpm/ 4°C / 30min, the supernatant removed again and labelled Sol. 2. The pellet was kept for further analysis, along with both Sol. 1 and Sol. 2 for viewing on SDS-page.

Table 11. Conditions for optimizing protein solubility. Induction conditions and detergent treatment combinations for increasing the solubility of the *ppb210* adenylation domain.

Strain	Grown Temp (°C)	Induction Temp (°C)	Induced OD_{600}	Cell pellet re-suspended	Insoluble pellet re-suspended
125_210_LATL	37	16	0.6	Lysis buffer	NP-40
125_210_LATL	37	16	0.6	Lysis buffer	Triton-X100
125_210_LATL	37	16	0.6	Lysis buffer	IGEPACa-40
125_210_LATL	37	16	0.6	Lysis buffer	6M Urea
125_210_LATL	25	16	0.2	Lysis buffer	NP-40
125_210_LATL	25	16	0.6	Lysis buffer	Triton-X100
125_210_LATL	25	16	0.2	Lysis buffer	IGEPACa-40
125_210_LATL	25	16	0.2	NP-40	NP-40
125_210_LATL	25	16	0.6	NP-40	NP-40
125_210_LATL	25	16	0.2	Triton-X100	Triton-X100
125_210_LATL	25	16	0.6	Triton-X100	Triton-X100
125_210_LATL	25	16	0.2	IGEPACa-40	IGEPACa-40
125_210_LATL	25	16	0.6	IGEPACa-40	IGEPACa-40
125_210_LATL	37	16	0.6	Lysis buffer	NP-40
125_210_LATL	37	16	0.6	Lysis buffer	Triton-X100
125_210_AT	37	16	0.6	Lysis buffer	IGEPACa-40
125_210_AT	37	16	0.6	Lysis buffer	6M Urea
125_210_AT	25	16	0.2	Lysis buffer	NP-40
125_210_AT	25	16	0.6	Lysis buffer	Triton-X100
125_210_AT	25	16	0.2	Lysis buffer	IGEPACa-40
125_210_AT	25	16	0.2	NP-40	NP-40
125_210_AT	25	16	0.6	NP-40	NP-40
125_210_AT	25	16	0.2	Triton-X100	Triton-X100
125_210_AT	25	16	0.6	Triton-X100	Triton-X100
125_210_AT	25	16	0.2	IGEPACa-40	IGEPACa-40
125_210_AT	25	16	0.6	IGEPACa-40	IGEPACa-40
125_210_ATL	37	16	0.6	Lysis buffer	NP-40
125_210_ATL	37	16	0.6	Lysis buffer	Triton-X100
125_210_ATL	37	16	0.6	Lysis buffer	IGEPACa-40
125_210_ATL	37	16	0.6	Lysis buffer	6M Urea
125_210_ATL	25	16	0.2	Lysis buffer	NP-40
125_210_ATL	25	16	0.6	Lysis buffer	Triton-X100
125_210_ATL	25	16	0.2	Lysis buffer	IGEPACa-40
125_210_ATL	25	16	0.2	NP-40	NP-40
125_210_ATL	25	16	0.6	NP-40	NP-40
125_210_ATL	25	16	0.2	Triton-X100	Triton-X100
125_210_ATL	25	16	0.6	Triton-X100	Triton-X100
125_210_ATL	25	16	0.2	IGEPACa-40	IGEPACa-40
125_210_ATL	25	16	0.6	IGEPACa-40	IGEPACa-40

3.2.12 Radioactive adenylation domain activity assay

Determination of substrate specificity and kinetic parameters by ATP-[^{32}P]PP$_i$ exchange assay

To determine substrate specificity, ATP-[^{32}P]PP$_i$ reactions (100 µL) containing Tris–HCl (pH 7.5, 75 mM), $MgCl_2$ (10 mM), tris-(2-carboxyethylphosphine) (TCEP, pH 7.0, 5 mM), ATP (5 mM), amino acid/pyrrole (5 mM), and [^{32}P]PP$_i$ (1 mM, 84.12 Ci/mM) were performed at 25 °C. The reactions were started by the addition of the ppb_22120 (1 µM) and incubated for up to 3 h before quenching with charcoal suspensions (500 µL, 1.6% [w/v] activated charcoal, 4.5% [w/v] $Na_4P_2O_7$, and 3.5% [v/v] perchloric acid in H_2O). The charcoal was pelleted by centrifugation prior to being washed twice with the wash solution (500 µL, 4.5% [w/v] $Na_4P_2O_7$ and 3.5% [v/v] perchloric acid in H_2O), resuspended in H_2O (500 µL), and counted by liquid scintillation (Beckman LS6500, Beckman Coulter, Fullerton, CA, USA).

Time course determination

Time course of the active substrates was determined using the same reaction parameters, but a single substrate in duplicate, quenching each reaction at a time point: 10, 20, 40, 60, 120 and 240 min.

For the determination of $K_{m,ex}$ and $k_{cat,ex}$ for formation of pyrrole-AMP, reactions (100 µl) containing Tris–HCl (pH 7.5, 75 mM), $MgCl_2$ (10 mM), TCEP (pH 7.0, 5 mM), ATP (5 mM), [^{32}P]PP$_i$ (1 mM, 84.12 Ci/mM), and varying concentrations of pyrrole (0.00, 0.025, 0.05, 0.10, 0.175, 0.25, 0.50, 1.00, 1.75, 2.50 and 5.0 mM) were done at 25°C. The reactions were initiated by the addition of ppb_22120 (5 µM) and were stopped after 5 min. The experiments were carried out in duplicate for each substrate concentration with a negative control (no enzyme). This was repeated with higher substrate concentrations: 0, 0.25, 0.5, 1.0, 1.75, 2.5, 5, 10, 17.5, 25 and 50mM.

3.2.13 Comparing radioactive phosphate vs. malachite green assay

Activity of Ppb120 when co-purified with Ppb125

Using the same protocol as previously outlined for the malachite green assay, purified 120_LATL protein - which had been co-expressed and purified with Ppb125, was tested for activity. The assay was performed as previously outlined.

3.2.14 Re-testing pOPINF construct activity with MbtH

Production of pOPINF_120_125 strain, and purifying protein

pOPINF_*120* was transformed into BL21(DE3)ΔybdZ_pACYC_ppb125 using the Adapted Five-Minute Protocol, and small scale protein expression and purification was done, according to the previously outlined method.

Activity of pOPINF_ppb120

The activity of purified Ppb120 when co-expressed with Ppb125 using the pOPINF vector was tested as with previous A-domains, all proteogenic amino acids as well as an array of pyrrole containing compounds.

3.3 Results

3.3.1 V. maris AB18-032 and V. sp. str. MG37 *gDNA extraction*

High quality DNA from *V. maris* AB18-032 and *V. sp.* str. MG37 was successfully extracted (Fig. 24) – and PCR showed correct bands for proximicin gene cluster genes confirming the correct organism.

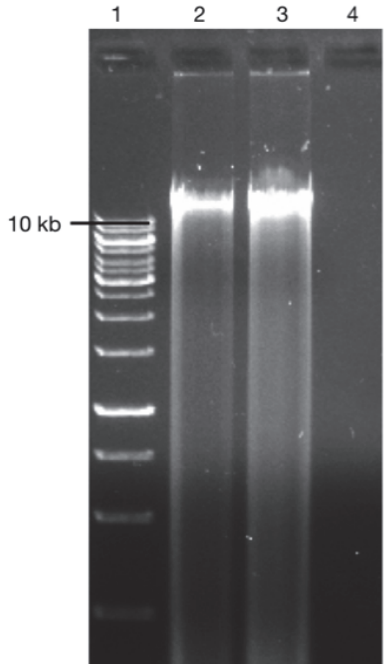


Figure 24. *Verrucosispora* spp. genomic DNA extraction. 1% agarose gel showing CTAB extracted genomic DNA. **1.** GeneRuler 1Kb ladder **2.** *V. maris* AB18-032, **3.** *Verrucosispora sp. str. MG37*

3.3.2 Initial studies I - Vector construction

pOPINF_ppb vector construction

DNA fragments of complete *ppb120*, *210* and *220* genes were successfully amplified using PCR; optimization was required to get a high enough quantity of product for restriction digest. *ppb195* was not able to be amplified using MyFi polymerase. PCR fragments were successfully gel purified (Fig. 25); concentration of resultant fragments was >100 ng/µL. The empty pOPINF vector was effectively transformed into BL21 *E. coli* cells, purified, restriction digested and gel purified with a final concentration of >108 ng/µL (Fig. 25). Ligation of *ppb* fragments into pOPINF and subsequent transformation into NEBalpha *E. coli* cells was confirmed using colony PCR. This showed successful incorporation of *ppb* genes into pOPINF. Correct frame and complete fragments were confirmed using sequencing.

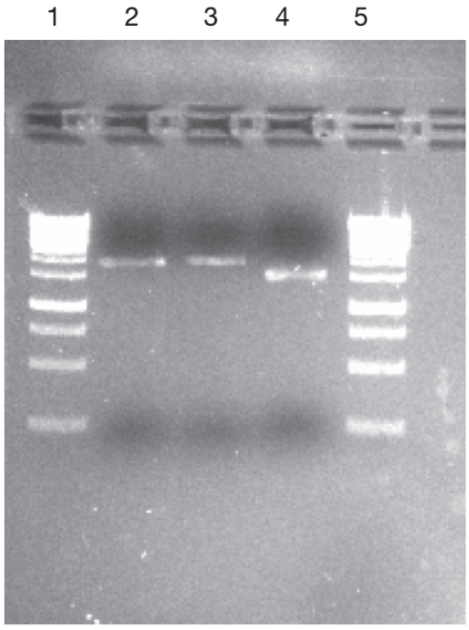


Figure 25. *ppb* gene amplicons for ligation into pOPINF. 1 % agarose gel showing the successful gel purification of DNA fragments containing entire *ppb* genes generated with primers containing In-Fusion modifications to allow efficient ligation into pOPINF. **1.** 1Kb ladder **2.** *ppb120* **3.** *ppb210* **4.** *ppb220* **5.** 1Kb ladder.

Control Vector

Ncpb_A_4 was successfully transformed into BL21 *E. coli* cells. This was confirmed using colony PCR, with Ncpb_A_4 specific primers giving the correct 1500 bp product.

Protein expression and purification of Ppb_120 and Ncpb_A_4

Initial protein induction studies demonstrated a low induction rate - specifically for Ppb_120 - but a high purification factor (Fig. 26). A large amount of both proteins was required – Ncpb_A_4 for assay development and Ppb120 for activity studies - with ~1 µM for each assay run/substrate, and so optimisation of protein induction was done.

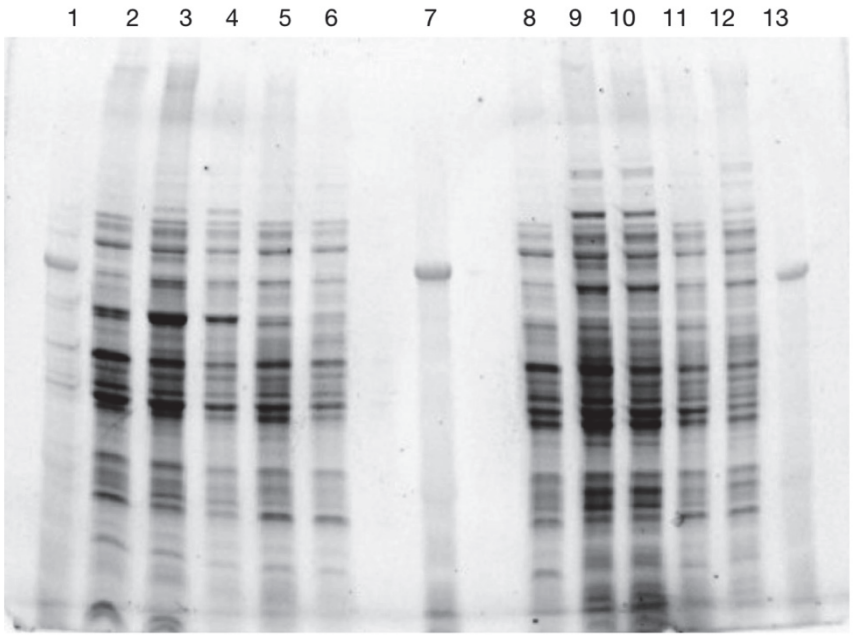


Figure 26. Successful purification of Ncpb_A4 and Ppb120. SDS-PAGE protein gel showing purification of Ncpb_A_4 (Lane 2-7) and pOPINF_ppb_120 (Lane 8-13) before optimization. Lane **1.** BR ladder **2.** FT **3.** Wash_1 4. Wash_2 **5.** Wash_3 **6.** Wash_4 **7.** Elution_1 Ncpb_A_4. **8.** FT **9.** Wash_1 **10.** Wash_2 **11.** Wash_3 **12.** Wash_4 **13.** Elution_1 Ppb120

Induction of Ncpb_A_4 was successfully increased, leading to higher purified protein yields (Fig. 27); this was done by using fresh transformants each purification and using a slow IPTG induction technique – inducing and growing overnight at 18°C. Induction rates for pOPINF_ppb120 failed to increase despite changes in IPTG concentration, growth and induction temp and using fresh transformants; it was always present in the soluble fraction and so solubility was not investigated. To by-

pass this issue, large quantities of culture were used (~16 L) and purified to get a large enough amount of protein for succeeding assays (Fig. 28).

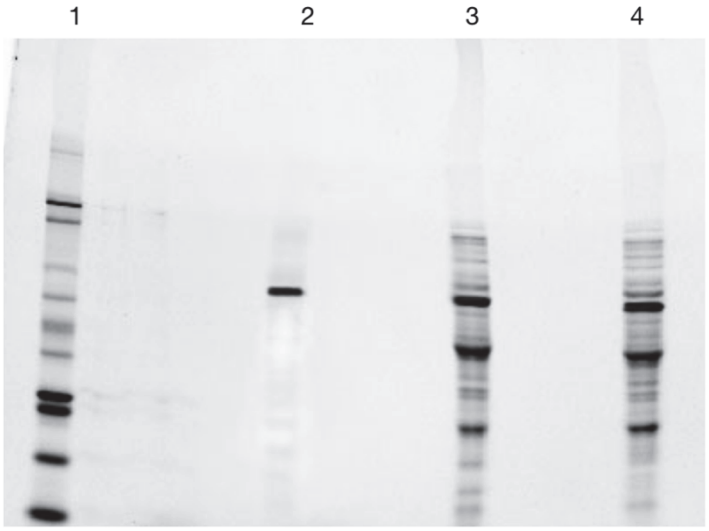


Figure 27. Purification of Ncpb_A_4The SDS-PAGE of Ncpb_A4 optimised purification, showing very high proportion of all desired protein in the purified fraction **1.** BR ladder **2.** Elution **3.** Wash_2 **4.** Wash_1

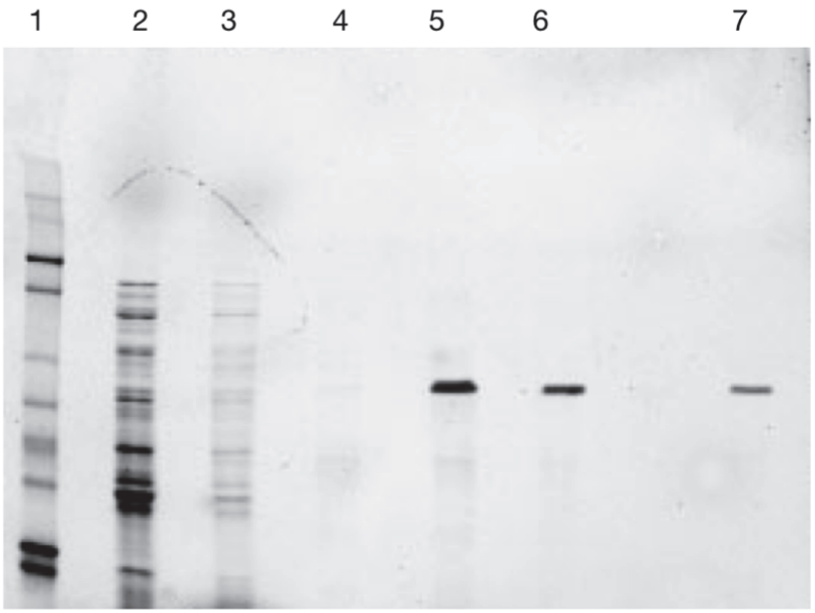


Figure 28. Successful purification of Ppb120. The SDS-PAGE of pOPINF_ppb120 optimised purification, showing very high proportion of all desired protein in the purified fraction, and little remaining in the insoluble fraction. **1.** BR ladder **2.** Wash_1 **3.** Wash_2 **4.** Wash_3 **5.** Concentrated fraction 1 **6.** Concentration fraction 2 **7.** Concentrated fraction 3.

3.3.3 *Chemical synthesis of pyrrole containing precursors*

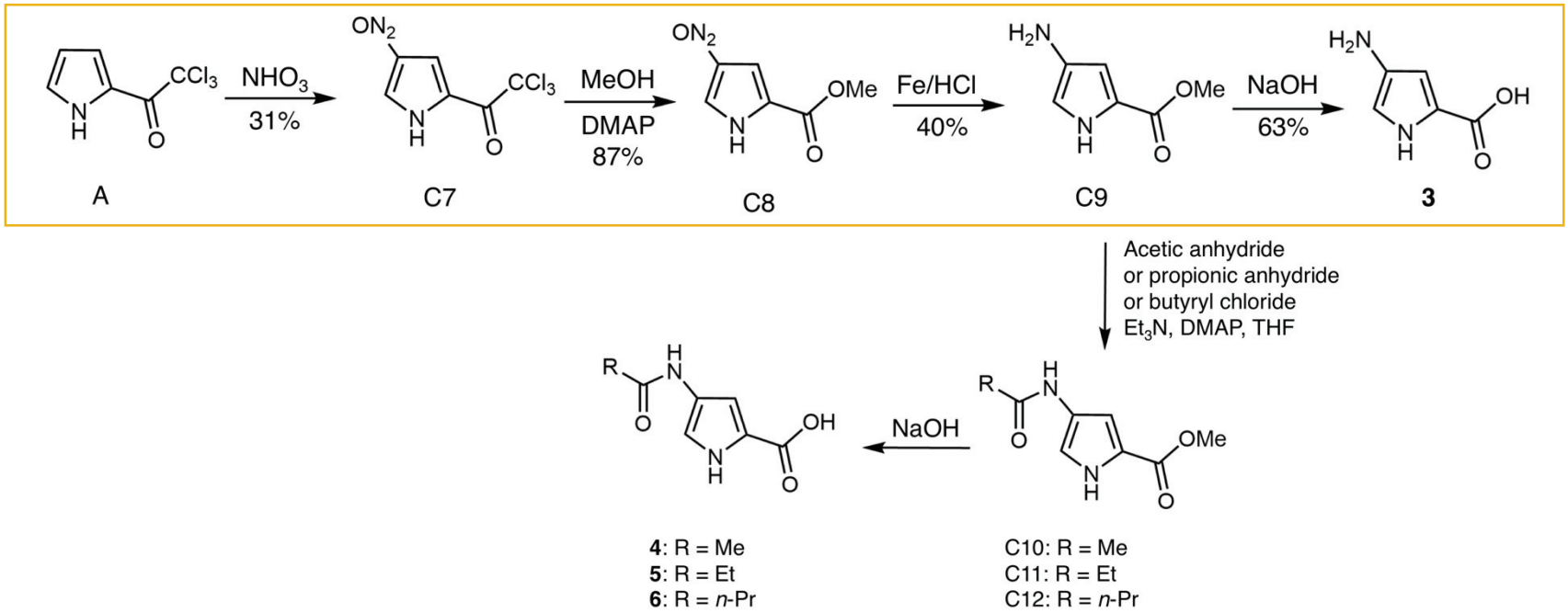


Figure 29. Chemical synthesis of pyrrole-containing precursors. Compound 3 was yielded from the chemical synthesis route from A to C7, C8 then C9 to yield compound 3, outlined in yellow. The other compounds for substrate activity tested were kindly supplied by Dr. Garneau-Tsodikova.

To produce a large array of chemically relevant potential precursor molecules to be tested, the 2,4 di-substitution pattern on the heterocycle was essential. An initial literature review and a preliminary study into producing 2,4 substituted furans was done (data not shown), and were unsuccessful. Major issues encountered were primarily associated with high cost associated with potential starting compounds and subsequent over substitution of the heterocycle ring. This was supported by research conducted by Garneau-Tsodikova et al., (2017) [personal correspondence]. This high cost/low yields associated made it not feasible for the large assay application for which they are required. To circumvent this issue, it was decided that the pyrrole analogue would be used in its place. The high-yielding synthesis routes to 2,4-disubstituted pyrroles are well characterized (Wolter et al., 2009), and the scheme utilized by Al-Mestarihi et al. (2015) was used as the basis of the research here.

3.3.4 Initial studies II - Malachite Green assay development

To ensure the assay was reproduced correctly, and was able to monitor the level of liberated phosphate at a useful level, different concentrations of phosphate was used and the change in absorbance recorded (Fig. 30). The assay gave a linear output to increasing phosphate concentrations, successfully confirming its application as a phosphate detection assay. To test the enzyme component of the assay (conversion of PP_i to P_i via inorganic pyrophosphotase) (Fig. 31), PP_i was added at known concentration. It was shown that the increasing level PP_i gave a linear correlation with absorbance, showing the assay worked and was suitable to monitor A-domain activity.

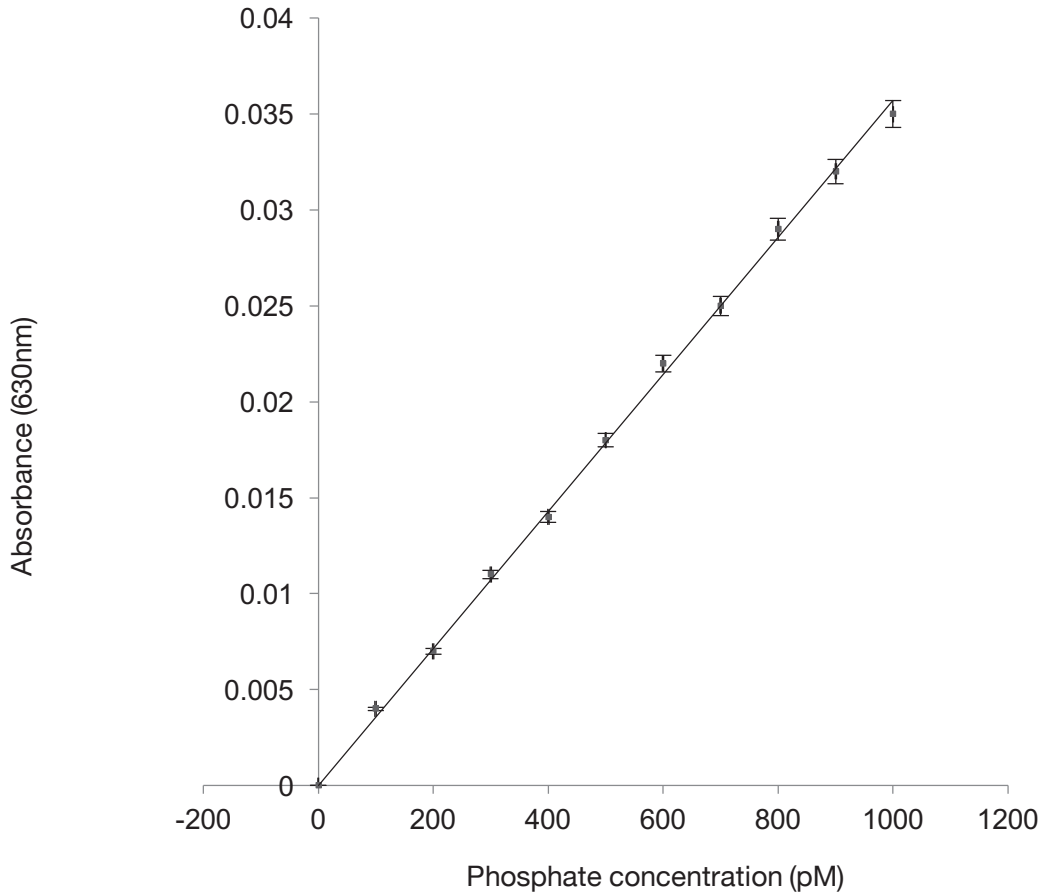


Figure 30. Assessing malachite green assay applicability to monitor A-domain activity. The linear relationship between phosphate concentration and absorbance (630 nm), demonstrating the phosphate assay correctly works. $R^2 = 0.999$. Error bars show SE.

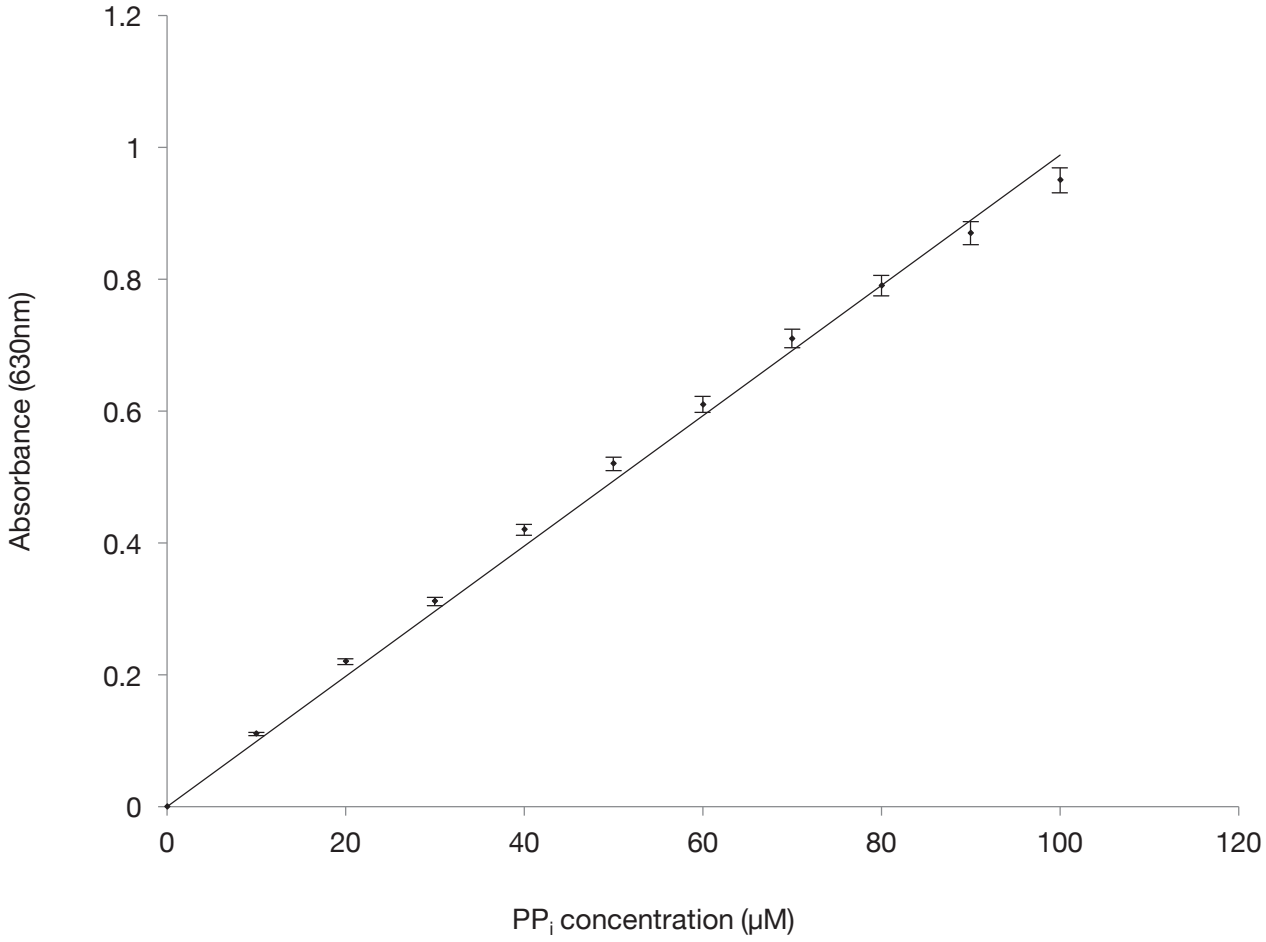


Figure 31. Assessing malachite green assay applicability to monitor A-domain activity. The positive linear relationship between PP_i concentration and absorbance (630 nm), showing the enzyme component of the assay works correctly. $R^2 = 0.9954$. Error bars show SE.

Activity of Ncpb_A4 control

The purified Ncpb_A_4 protein was used in the malachite green assay and tested again with all 20 proteogenic amino acids. As shown in Fig. 32 19 amino acids: isoleucine, leucine and valine were shown to have a significant A-domain activity shown by an increase in absorbance, in comparison to other amino acids. The amino acids included in the graph are those with R-groups with similar chemical structure or properties of the known substrates to show the level of specificity of the A-domain. The difference in absorbance between 1 hr and 16 hrs is due to residual P_i reacting with the malachite reagent, not more PP_i being produced as the addition of the reagent stops the enzymatic conversion of PP_i to P_i.

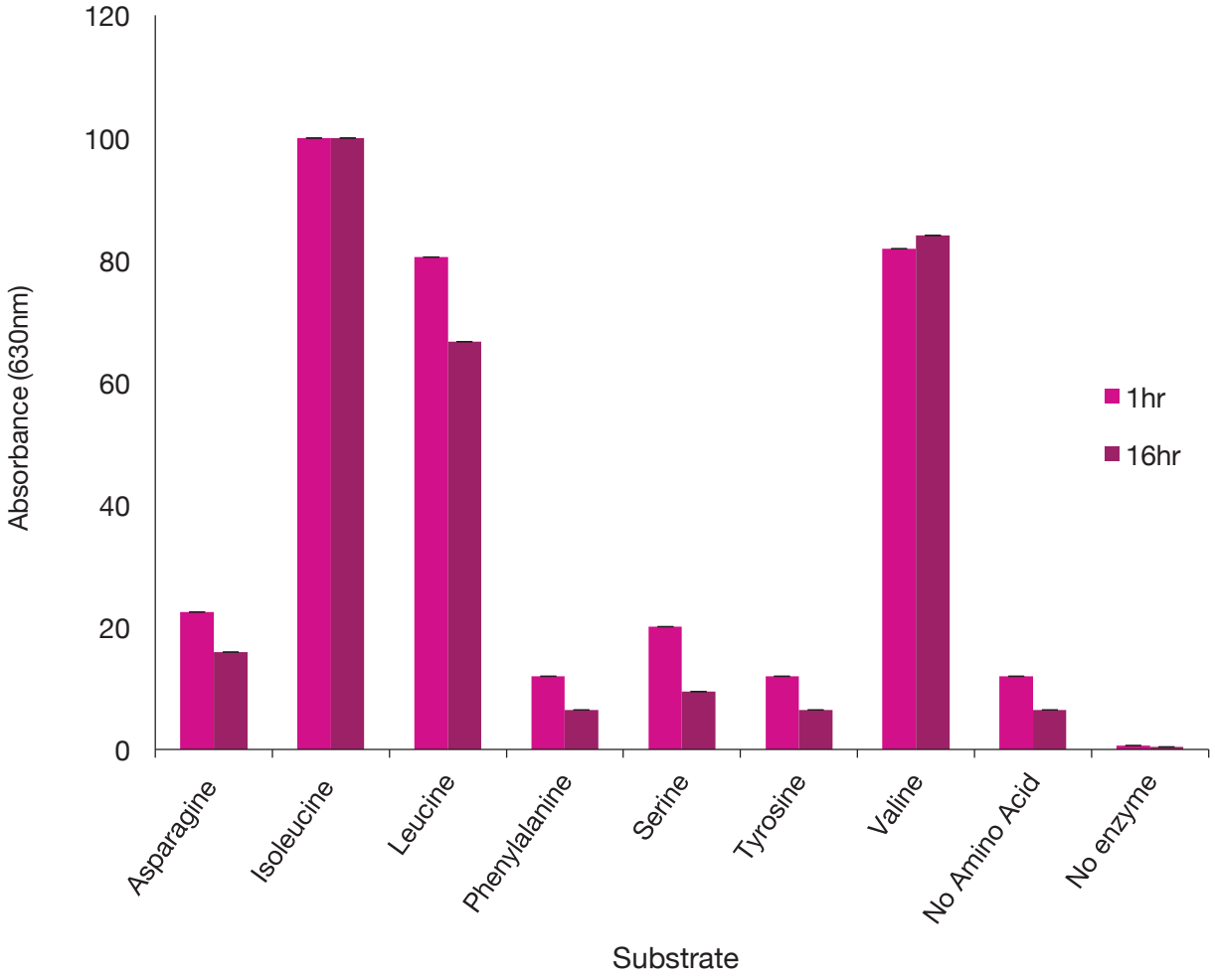


Figure 32. Ncpb_A_4 activity with a selection of amino acids assessed by the malachite green assay. A significant increase in absorbance can be seen with amino acids isoleucine, leucine and valine. Error bars represent standard error.

Activity of pOPINF_ppb120

Ppb120 showed no significantly increased activity with any amino acid or acetylated pyrrole (Fig. 33). This inactivity with all substrates was determined to be most likely due to the protein being inactive, this was supported by previously low yields of protein induction and purification. To circumvent this issue, all the adenylation domains were re-analysed and new constructs designed.

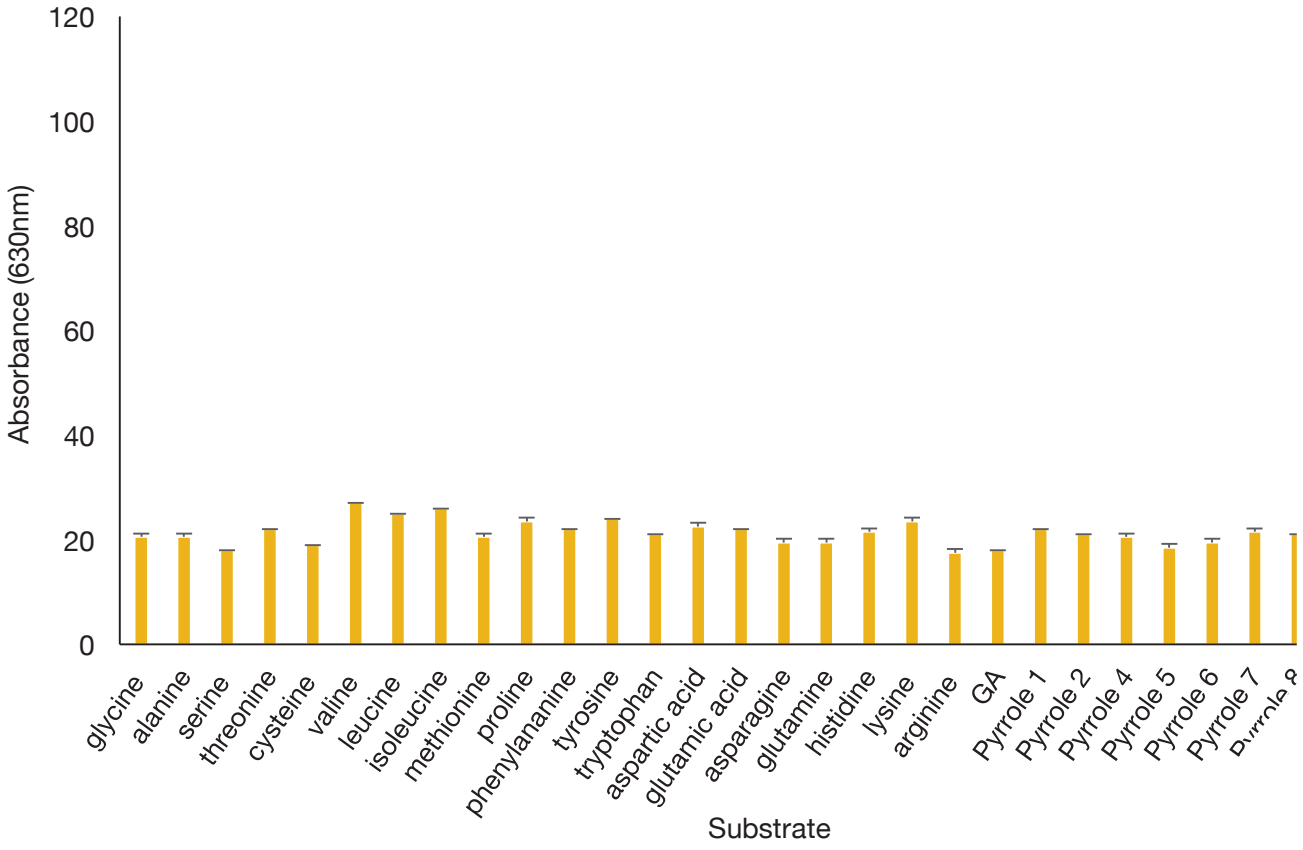


Figure 33. Ppb120 adenylation domain activity assessed by malachite green assay. pOPINF_ppb120 activity with all proteogenic amino acids and a selection of pyrrole containing compounds. Error bars represent standard error

3.3.5 Identification of adenylation domains and primer design

Changing approach to NRPS protein cloning

After issues were encountered with the activity of Ppb proteins when using the pOPINF vector system, it was decided that – taking heed from other adenylation domain work – the A-domains specifically would be targeted and not the entire protein as done in initial studies. Starting over gave us the option to change vector system to one more robust, commonly used in successful adenylation domain studies and so the pET vector system was chosen. This change was also due to the increasing realisation that a MbtH-co protein would likely be important in the *ppb* A-domain activity as *ppb120* was shown to be inactive in the initial studies. And so, it was important to chose a vector system which could be used in tandem with another vector system. pACYC_Duet gave an interest opportunity to incorporate the *ppb* genes and MbtH-like protein onto the same vector. This was initially attempted, but induction of *ppb* genes was greatly reduced (results not shown). When discussing this issue with other research groups, a similar problem was encountered and so a combination of pET and pACYC_duet was used. To identify just the adenylation domains, the exact boundaries were determined.

Identification of adenylation domain boundaries

Adenylation domain boundaries were successfully confirmed using known, well characterised domain features. Adenylation, thiolation, condensation and linker region-specific motifs were highlighted and their similarity to well characterised NRPS systems to identify any sequence level reasons for inactivity or unique specificity.

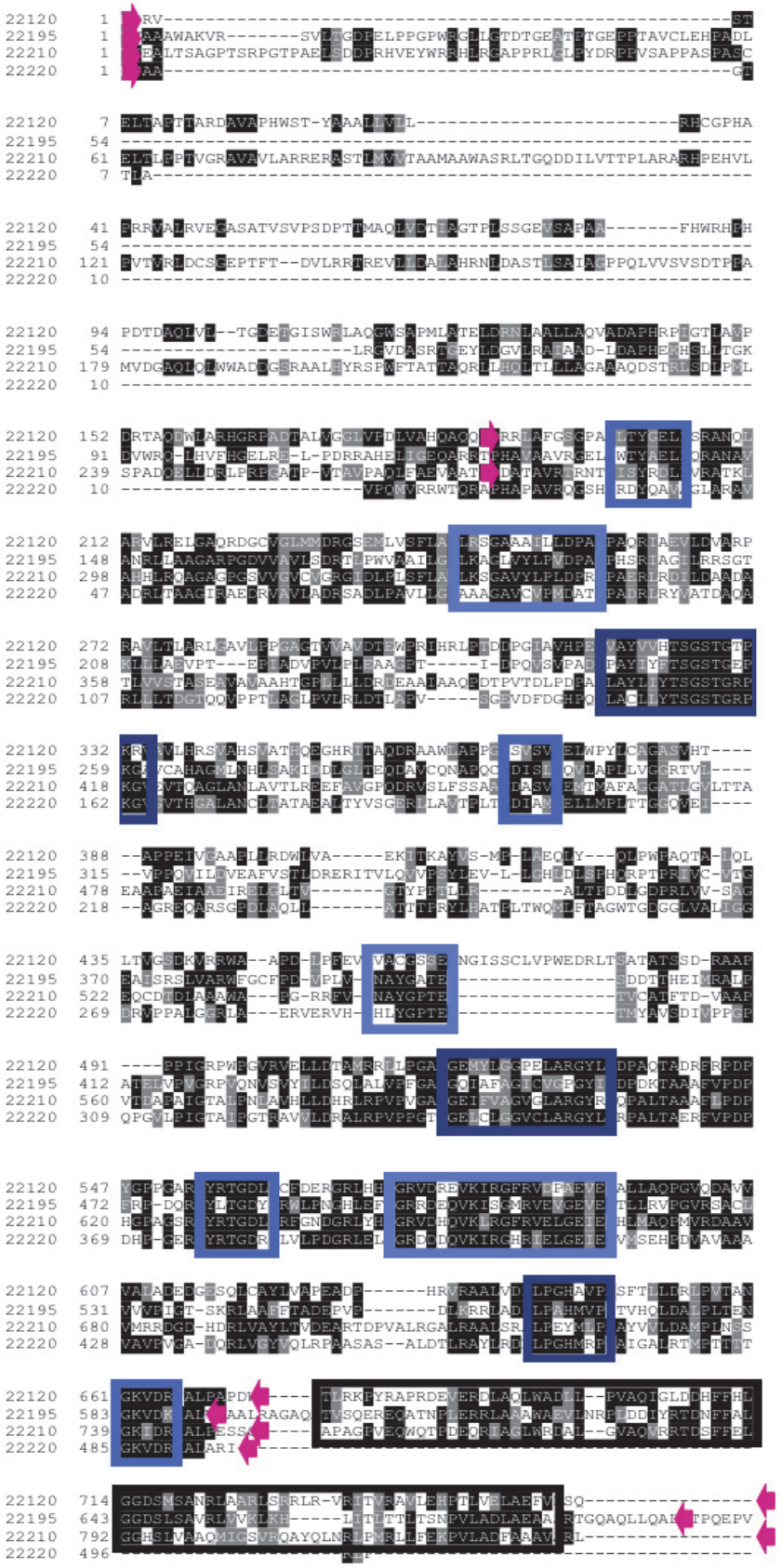


Figure 34. Multiple sequence alignment of *ppb* NRPS-genes, showing core domains and primer sets for amplification of A-domains. Amino acid alignment of all NRPS-related genes in the proximicin gene cluster. Blue boxes show core adenylation domain motifs; black boxes show other NRPS motifs, and pink arrows show placement of PCR primers.

For ease of understanding, gene fragments and the resultant constructs and expressed proteins are named according to the domains present, for example, adenylation, thiolation and linker region. Table 12 below shows the primer sets designed, named by the amino acid residue number they begin at, which were used to generate the fragments. When referring to a DNA fragment the *ppb* prefix will be used, and when referring to cognate protein, it will be absent. For example, ppb120_AT is the DNA fragment produced and inserted into a vector, to yield 120_AT protein. Further, the vector system utilised will proceed the gene name. Figure 35 gives an overview of results from PCR to A-domain activity assays.

Table 12. Overview of *ppb* adenylation domain constructs produced. List of identified proximicin adenylation domain PCR primers, and base pair product size. L – Linker A – Adenylation T – Thiolation

	Fwd Primer	F Primer #	Rev Primer	R Primer #	Product size (bp)	Domains present
ppb120	120_F	21	120_R	24	2322	L-A-T-L
	120_F	21	120_672R	23	2064	L-A
	120_185F	22	120_R	24	1758	A-T-L
	120_185F	22	120_672R	23	1500	A-T
ppb195	195_F	29	195_R	31	2119	L-A-T-Te-L
	195_F	29	195_591R	30	1819	L-A-T
ppb210	210_271F	26	210_R	28	1751	A-T-L
	210_271F	26	210_750R	27	1419	A-T
	210_F	25	210_R	28	2111	L-A-T-L
ppb220	220_F	32	220_R	33	1444	L-A-L

For pET28a, primers were successfully designed for all *ppb* adenylation domains, with the NedI restriction site in the forward primer, and the HindIII and a stop codon in the reverse primer. For *ppb210*, pET30_Xa/LIC primers were successfully designed, in the same location as for the pET28a primers, but instead of the restriction sites, Xa/LIC arms were added. Primers were designed for the entire *ppb125* gene, restriction sites: EcoRI and HindIII added to the forward and reverse primers respectively.

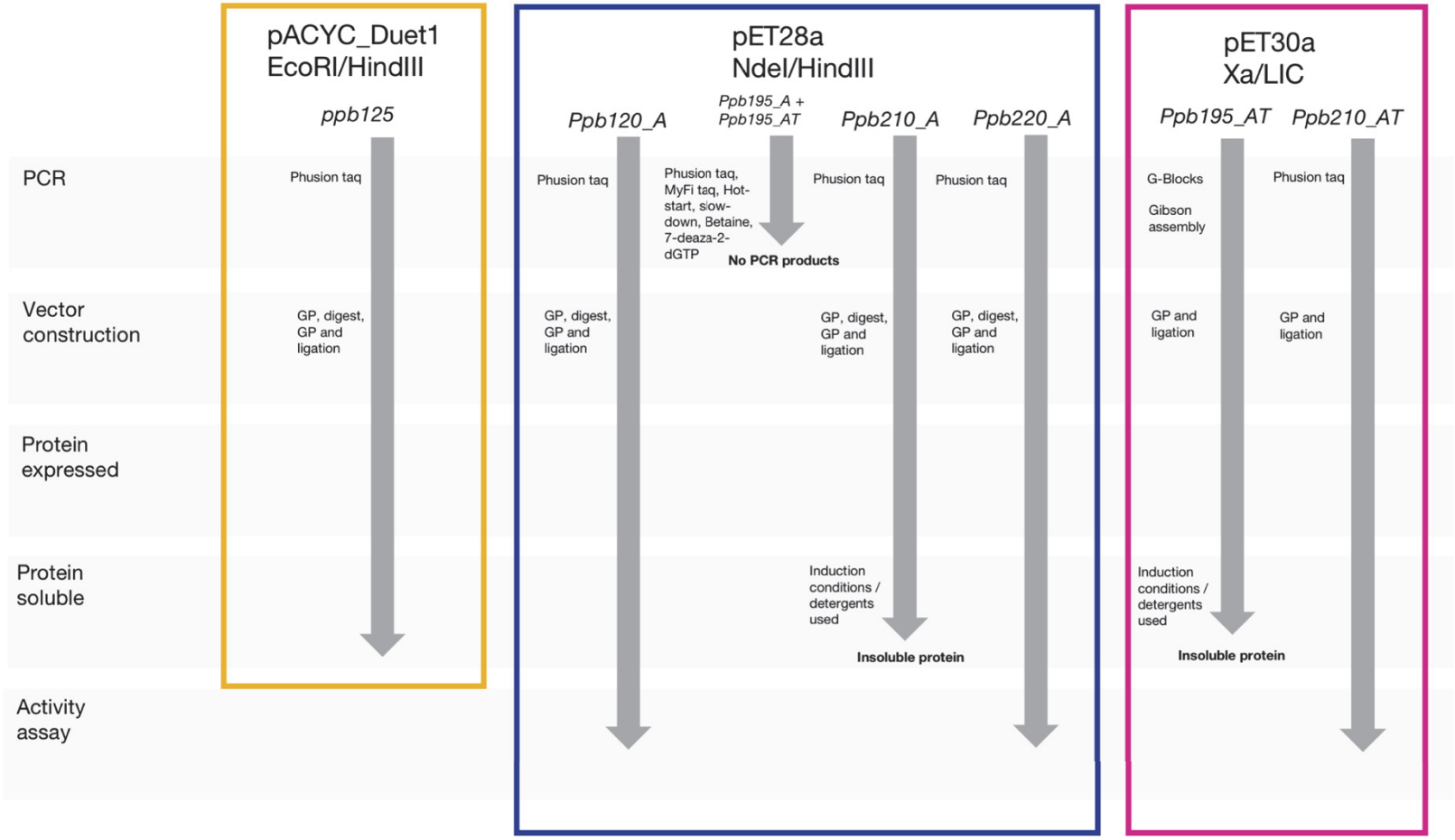


Figure 35. Illustration the overall success of each vector construct. An overall summary of the different vectors produced, and issues encountered with certain constructs.

3.3.6 *Amplification of ppb genes*

PCR amplification, gel purification and quantification

Initial attempts were made to clone all *ppb* genes into pET28a; products were successfully generated for fragments as shown in Table 13 and Figure 36. Proximicin gene *ppb120* generated products for all primer sets, however combination 120_A-T-L gave a lower yield and so was not taken forward to ligation. *ppb195* was not successfully amplified using PCR. All successful amplifications were successfully purified (Fig. 37).

Table 13. Overview of *ppb* gene fragments produced for construction of adenylation domain vectors. Summary of PCR method alterations to generate high enough product yields for restriction digests.

	Fwd Primer	Rev Primer	Product size (bp)	Domains in product	Lane in Fig. 36	Alterations to PCR method	µg/mL of purified product Fig. 36
ppb120	F	R	2322	L-A-T-L	2	Ex. Time ext. to 2:30 min	101
	F	672R	2064	L-A	3	Ex. Time ext. to 2:30 min	125
	185F	R	1758	A-T-L	4	None	67
	185F	672R	1500	A-T	5	None	139
ppb195	F	R	2119	L-A-T-Te-L	8	Unsuccessful PCR	
	F	591R	1819	L-A-T	9		
ppb210	271F	R	1751	A-T-L	6	None	59
	271F	750R	1419	A-T	7	None	81
	F	R	2111	L-A-T-L	/	None	132
ppb220	F	R	1444	L-A-L	10	None	99

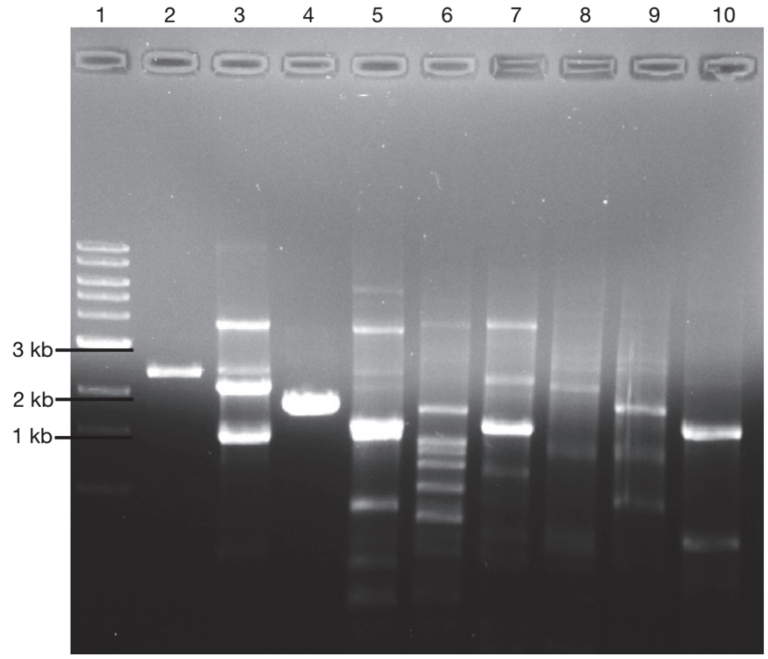


Figure 36. PCR of *ppb* genes. 1. 1Kb DNA ladder **2.** 120_LATL **3.** 120_LA **4.** 120_ATL **5.** 120_AT **6.** 210_ATL **7.** 210_AT **8.** 195_LATTeL **9.** 195_LAT and **10.** 220_LAL

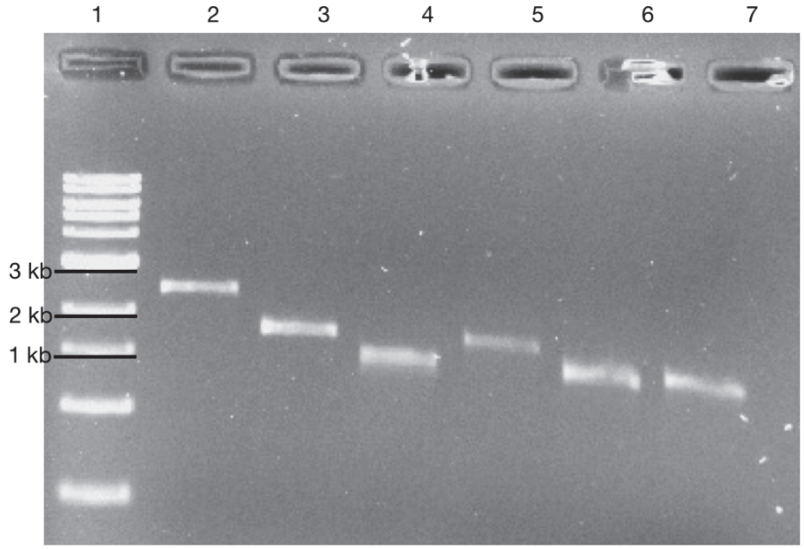


Figure 37. Gel purified *ppb* gene fragments. 1. 1Kb DNA ladder; **2.** 120_LATL **3.** 120_LA **4.** 120_ATL **5.** 210_ATL; **6.** 210_AT and **7.** 220_LAL.

Troubleshooting *ppb195* amplification

The proximicin gene *pbb195* was extremely difficult to amplify using PCR – primers were designed for both pET28a and pET30_Xa/LIC primer sets. Products of the correct size were generated, but at very low yields with many non-target products (Fig. 38). To increase this, many PCR cycling conditions and additives, as well as enzyme and pET vector systems were tested, the results are summarised in Table 14. Overall, the best results were achieved by a combination of: PCR cycle optimization, higher DMSO concentration (>7%) and using MyFi™ polymerase. This resulted in a lower amount of off-target amplicons, as shown in the difference between Fig. 38 and 39. However, any fragment cloned into a vector had mutations present in the DNA sequence. It was thus decided to avoid issues that a synthetic version of *ppb195* would be designed and synthetically produced.

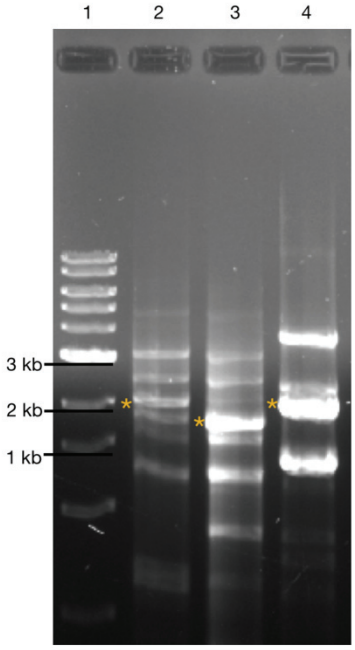


Figure 38. Trouble-shooting of PCR of *ppb195*. The highest amount of product achieved Phusion polymerase - using 10% DMSO. **1.** 1 Kb DNA ladder **2.** P30_195_LAT **3.** P30_195_AT **4.** P30_195_LATTe. The yellow asterisk denotes the correct product size.

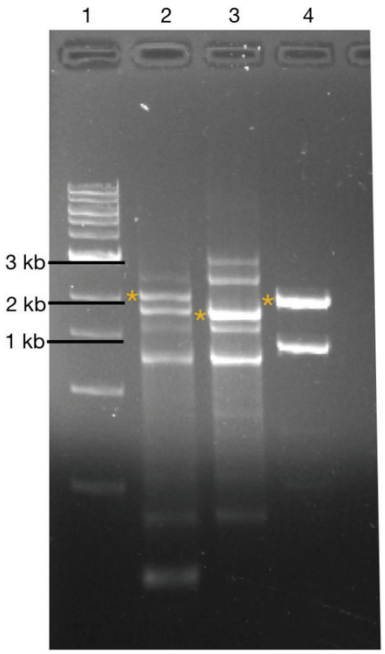


Figure 39. Trouble-shooting of PCR of *ppb195* I. The highest amount of product achieved MyFi Taq polymerase and 7% DMSO. **1.** 1 Kb DNA ladder **2.** P30_195_LAT **3.** P30_195_AT **4.** P30_195_LATTe. The yellow asterisk denotes the correct product size.

Table 14. Summary of PCR troubleshooting for *ppb195* adenylation domain fragments. Combinations of thermos-cycling conditions, various additives, enzymes and PCR type utilized to increase yield of high-GC product *ppb195* in comparison to standard Phusion manufacturers guidelines. Condition combinations and fragment length which saw an increase in yield are in bold/blue.

Cond. Type	Condition	p28_195 L-A-T-Te-L	p28_195 L-A-T	p28_195 L-A-T-Te-L	p30_195 L-A-T
Thermo-cycling	Increasing denaturation temp from 98 – 103 °C in 1°C increments	Less amount of even mismatch products	Less amount of even mismatch products	Less amount of even mismatch products	Less amount of even mismatch products
	Increasing denaturation time 3 – 8 min in 30 sec increments	Less amount of even mismatch products	Less amount of even mismatch products	Less amount of even mismatch products	Less amount of even mismatch products
	[Mg^{2+}]1.4 mM-4 mM in 0.2 mM increments	**Desired product present at [Mg^{2+}] 2.2 – 3 mM Low yield**	**Desired product present at [Mg^{2+}] 2.0 – 3 mM Low yield**	**Desired product present at [Mg^{2+}] 2.0 – 3 mM Low yield**	**Desired product present at [Mg^{2+}] 2.0 – 3 mM Low yield**
Additives	**DMSO 3% - 10% in 1% increments**	**Desired product from 6%-10% Low yield**	**Desired product from 7%-10% Low yield**	**Desired product from 7%-10% Low yield**	**Desired product from 7%-10% Low yield**
	Betaine 0.25 M – 1.25 M in 0.25 M increments	No effect	No effect	No effect	No effect
	7-deaza-2-dGTP (25 mM)	No effect	No effect	No effect	No effect
Enzyme	High-Fidelity Phusion® buffer	Miss-match products	Miss-match products	Miss-match products	Miss-match products
	MyFi™ polymerase	**Clean desired products**	**Clean desired products**	**Clean desired products**	**Clean desired products**
PCR type	Hot-Start	No effect	No effect	No effect	No effect
	Slow-Down	Miss-match products	Miss-match products	Miss-match products	Miss-match products

3.3.7 Design and construction of synthetic ppb195

Design of SYNppb195

When re-designing *ppb195* primers, it became evident that it has an extremely high GC content surrounding the adenylation domain, particularly at the beginning of *ppb195* where the A-domain is located (Fig. 40), has a GC content of ~80%. Due to the inability to produce a correct *ppb195* DNA fragment, a synthetic *E. coli* optimised *ppb195* was successfully designed into three fragments. The codon adaptation index (>0.8), codon frequency index of >30%, GC% and number of negative CIS and repeat elements were all improved with optimization – to allow for efficient expression of protein (Table 15).

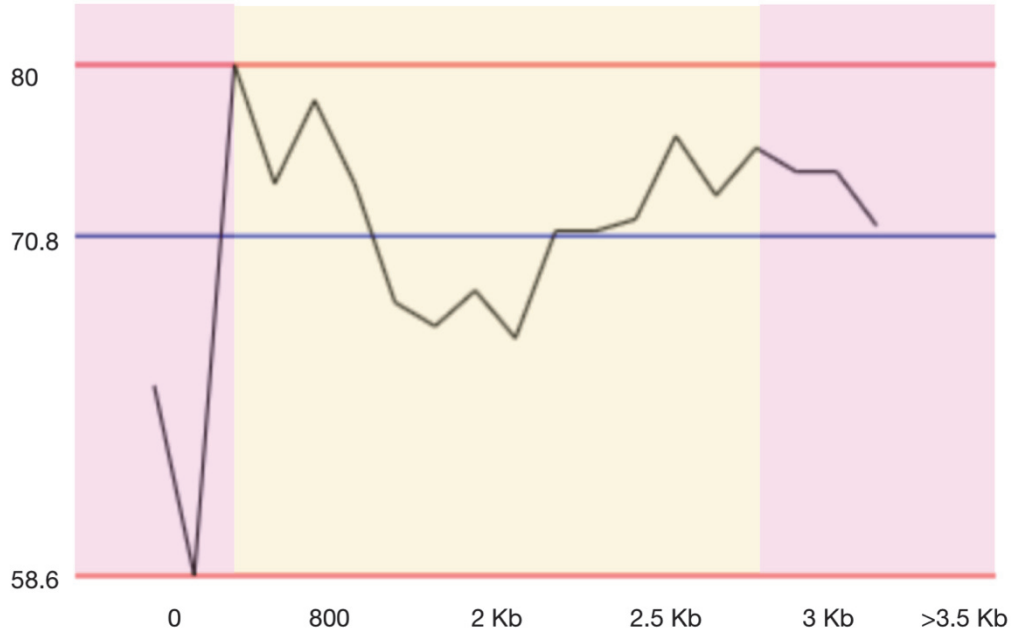


Figure 40. GC% across the entire native *V. maris* AB18-032 *ppb195* gene. Exceptionally high GC content across large spans of the gene explains previously encountered issues with PCR.

Table 15. Assessing the success of *E. coli* optimization of *ppb195*. A lower average GC% content was observed – the main aim of synthesizing *ppb195*, but also the occurrence of rare codons was lowered, which should aid in *E. coli* expression.

	Codon adaptation index	Codon frequency index	GC content (%)	Negative CIS	Negative repeat elements
WT*ppb195*	0.68	12	73.9	1	2
SYN*ppb195*	0.98	29	56.6	0	0

Fragments were designed to incorporate overlapping regions to allow Gibson assembly. Primers were designed for amplification with Xa/LIC over hang arms for ligation into pET30. The synthetic design of *ppb195* was done in a such a way to include the A domain in fragment A, the T domain in fragment B and linker regions in fragment C.

Gibson assembly of SYNppb195

Issues with Gibson assembly of all three fragments were encountered; however, fragments A and B were successfully joined, confirmed by sequencing. As it is fragments A and B which hold the functional regions of the protein (Fig. 41), this was initially deemed enough for activity of Ppb195.

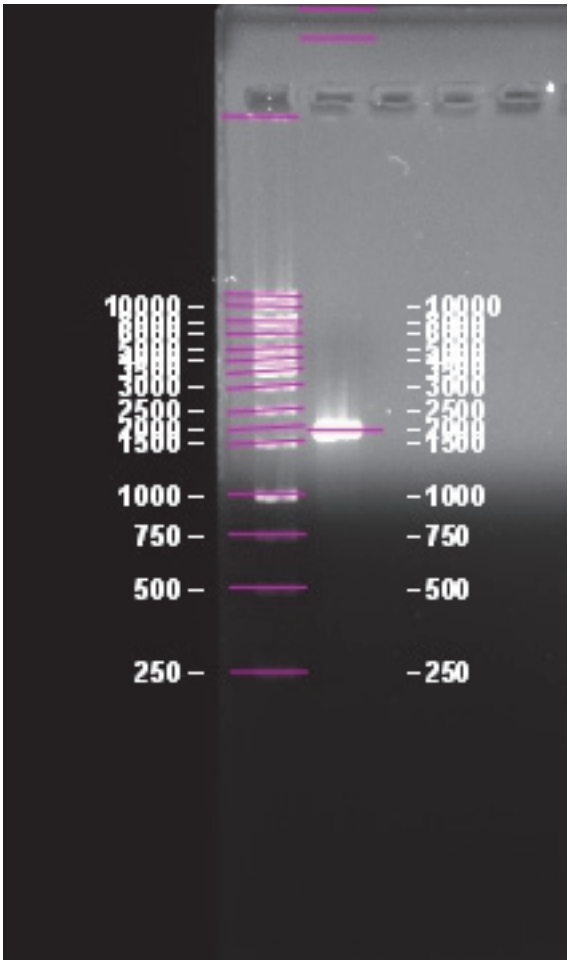


Figure 41. Assembly of synthetic *ppb195* G-Blocks 1 & 2 containing the A-domain. Successful assembly of *ppb195* synthetic (*SYNppb195*) fragments A and B giving a resultant product of ~1900bp.

3.3.8 Construct preparation

pET_*ppb* construct production

pET28a was supplied by Novagen, and digested successfully using NdeI and HindIII. The vector backbone was gel purified to give a high purity and quality DNA fragment. The gene fragment ligation into pET28a had the highest efficiency when left at 37°C overnight, and so, this was used as standard procedure. *ppb* gene purified fragments were successfully cloned into pET28a to produce constructs: p28_120_AT, p28_120_A, p28_210_A, p28_210_AT and p28_220_LAL.

In addition, p30_SYNppb195, p30_210_A and p30_210_AT were all successfully produced. *ppb125* was also successfully ligated into pACYC_Duet1. All were initially confirmed via colony PCR and then sequencing.

3.3.9 Expression strain production

Production of competent expression strain

pACYC_Duet_ppb125 was successfully transformed into BL21(DE3))ΔybdZ confirmed by colony PCR. Competent cells were successfully produced and 50 µL aliquots frozen in liquid nitrogen and stored at -80C. The transformation efficiency using the control plasmid pUC19 was calculated to average at 3.9 $x10^{10}$ cfu/ µg DNA, which is reasonable for a plasmid of that small size.

Transformation into expression strains

All generated *ppb*-containing expression vectors were successfully transformed into:

i. *E. coli* BL21(DE3)ΔybdZ (with no MbtH-like protein);
ii. *E. coli* BL21(DE3)ΔybdZ_pACYC_ppb125 (with proximicin-specific MbtH-like protein), and
iii. *E. coli* BL21(DE3)ΔybdZ_pET28a_TioT (non-specific MbtH-like protein).

The transformation efficiency was generally lower, ranging from an average of 0.9 – 3.1 $x10^{10}$ cfu/ µg plasmid DNA, likely due to the increased plasmid sizes.

3.3.10 *Protein purification*

Co-expression with no MbtH protein or with non-native MbtH-like protein: TioT

Initial induction studies were done to test protein expression and solubility with either a) no MbtH protein and b) with a non-native TioT MbtH protein (Fig. 42 & 43). Protein construct combinations: TioT_210_ATL; TioT_220_A and TioT_SYN195_LA displayed no increased expression when TioT was present in comparison to when no MbtH was present, or a change in solubility. Conversely, TioT_120_AT, TioT_210_LATL and TioT_210_AT had a lower level of expression when TioT was present, showing the presence inhibited expression. Of those which did express, none were purified successfully. All *ppb* constructs had large quantities of insoluble desired protein – were seen in the insoluble pellet. Due to these results, it was deemed that there was some level of MbtH specific to the proximicin biosynthetic route, and so a non-native MbtH protein was not able to do this.

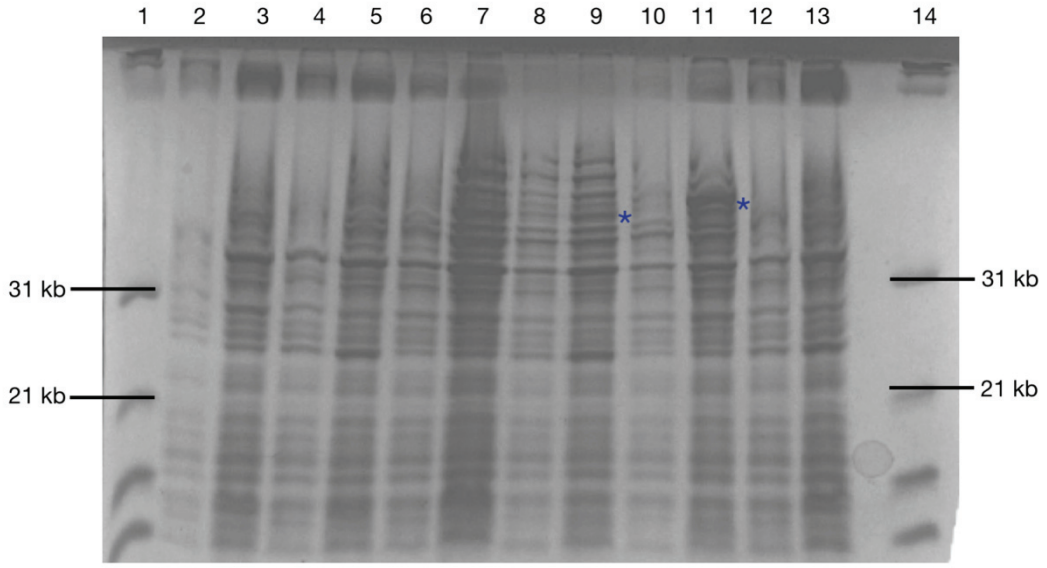


Figure 42. SDS-PAGE of Ppb A-domain proteins initial induction studies with either no MbtH-like protein co-expressed, or co-expressed with TioT. 1. Sigma BR ladder **2.**120_LATL – **3.** 120_LATL + **4.** TioT _120_LATL – **5.** TioT_120_LATL + **6.** 120 AT – **7.** 120_AT + **8.** TioT_120 AT - **9.** TioT_120_AT + **10.** 210_ATL – **11.** 210_ATL + **12.** TioT_210_ATL – **13.** TioT_210_ATL +. A blue asterisk denotes the increased presence of protein in the induced culture in the left hand lane, of the correct size. +/- = induced/non-induced, respectively.

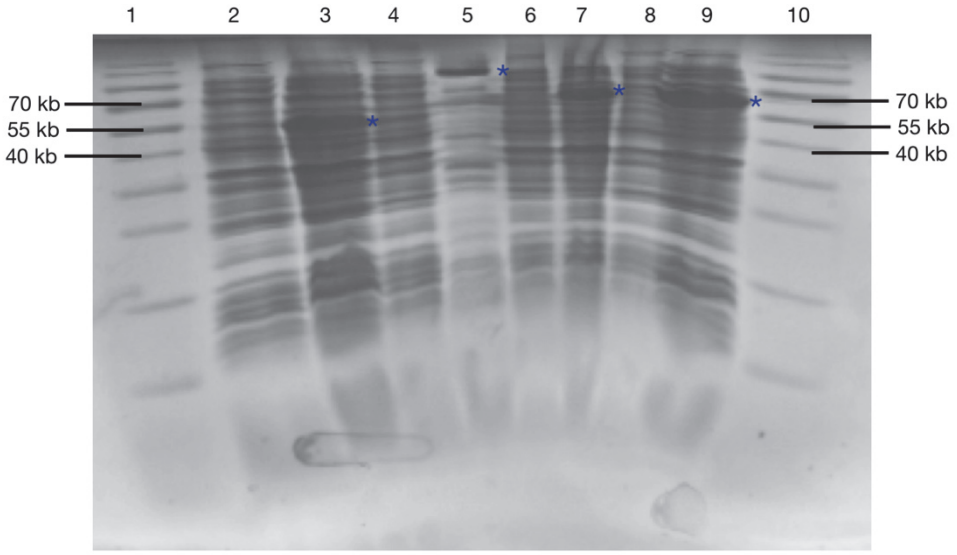


Figure 43. SDS-PAGE of Ppb A-domain proteins initial induction trial. 1. Cont. Sigma BR ladder **2.** 210_AT– **3.** 210_AT + **4.** TioT_210_AT – **5.** TioT_210_AT + **6.** 220_LATL – **7.** 220_LATL+ **8.** TioT_220_LATL – **9.** TioT_220_LATL + **10.** BR ladder. A blue asterisk denotes the increased presence of protein in the induced culture in the left hand lane, of the correct size. +/- = induced/non-induced, respectively.

Co-expression with native Ppb125 MbtH-like protein

All strains tested – 125_120_AT, 125_220_LAL, 125_210_AT, 125_210_ATL, 125_210_LATL and 125_SYN195 – showed a markedly improved, high levels of induction when co-expressed with proximicin-specific MbtH protein Ppb125. Large quantities of purified soluble protein were achieved with proteins 120AT and 220A. However, protein constructs: 210_AT, 210_LATL and SYN195_LA remained insoluble, and 210_ATL was only partly soluble - most of the expressed protein remained in the insoluble pellet. For *ppb210* A-domain constructs, this was partly resolved in two ways: a) designing a vector with the more of the linker regions on - 210_ATL, and b) the addition of detergents in the purification process (see later results).

Large scale protein purification

Large scale purification and concentration of – 125_120_AT and 125_220_A was successful, giving high yields of purified proteins shown in Figure 44 and 45, respectively.

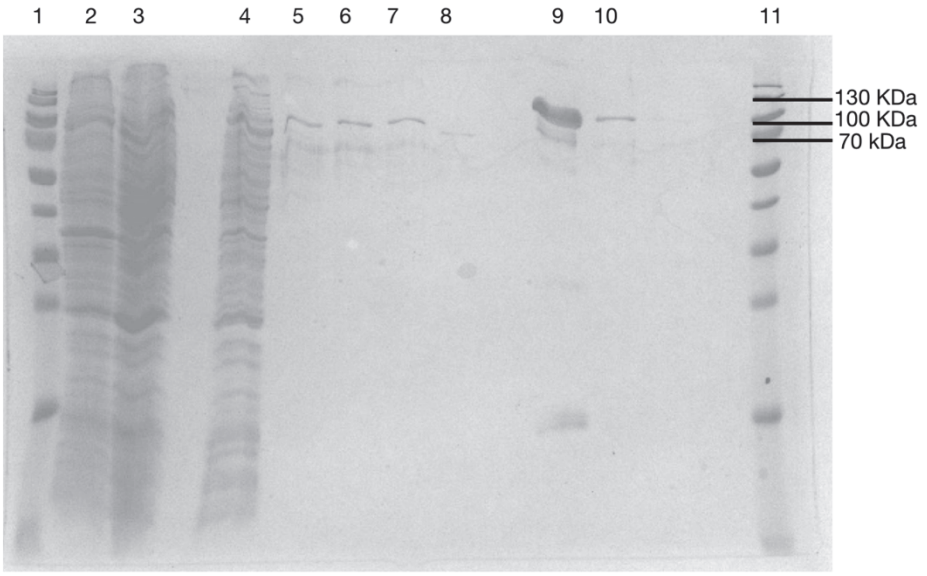


Figure 44. Induction and purification of 125_120_AT. The SDS-PAGE shows: **1.** BR ladder **2**. Insoluble pellet **3.** Flow-through **4.** Wash 1 **5.** Wash 3 **6.** Wash 5 **7.** Wash 7 **8.** Wash 9 **9**. Elution 1 **10.** Elution 2 **11.** BR ladder

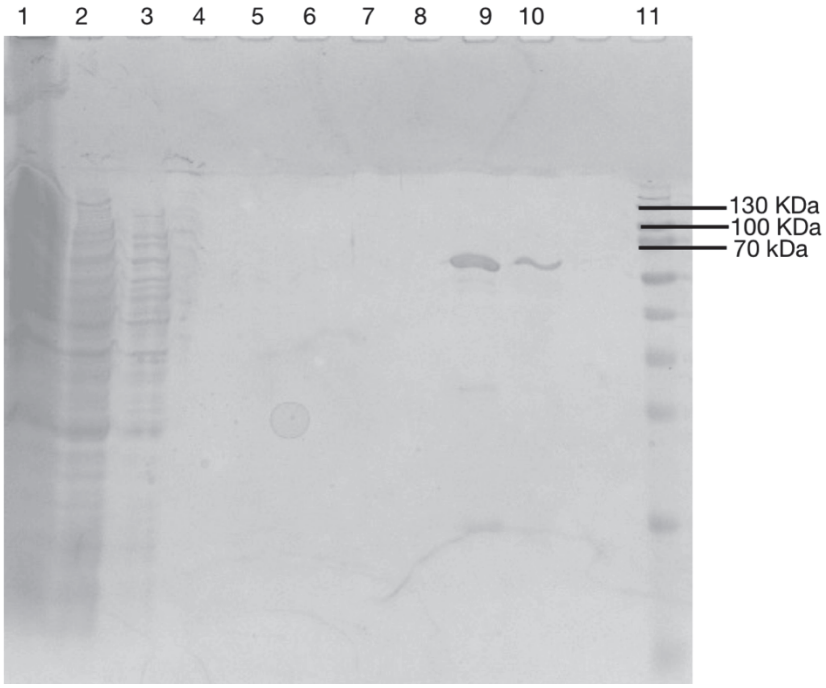


Figure 45. Induction and purification of 125_220_A. The SDS-PAGE shows: **1.** BR ladder **2**. Insoluble pellet **3.** Flow-through **4.** Wash 1 **5.** Wash 3 **6.** Wash 5 **7.** Wash 7 **8.** Wash 9 **9**. Elution 1 **10.** Elution 2 **11.** BR ladder

Troubleshooting protein purification of insoluble 125_210_ATL, 125_LATL and 125_SYN195_LA

After extensive protein solubility testing, 125_210_LATL was shown to produce a small amount of soluble protein which could be purified. The exact conditions to achieve this were a combination of the presence of non-ionic detergents, the best being 25 mM triton X-100 incubated with the insoluble pellet, and a lower growth and induction temperature of 16°C (Figure 46 – 48). This meant the strain was extremely slow growing, and with the low yield, large amounts of culture had to be harvested to get the desired amount of protein required for substrate testing.

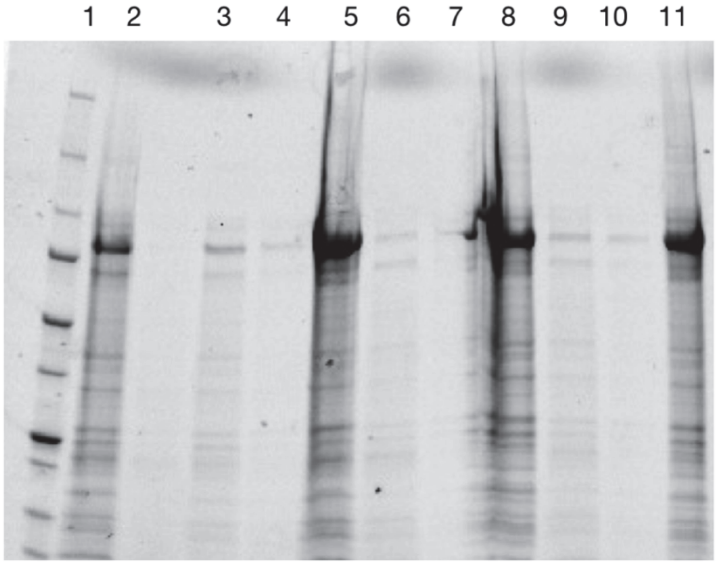


Figure 46. Solubility studies of 125_210_LATL. Temperature prior to induction varied and OD_{600} of induction. **1.** BR ladder **2**. Insoluble pellet control 37/16 OD_{600} 0.4 **3.** 25/16 OD_{600} 0.4 Elution 1 **4.** 25/16 OD_{600} 0.4 Elution 2 **5.** 25/16 OD_{600} 0.4 Elution 3 Urea **6.** 16/16 OD_{600} 0.4 Elution 1 **7.** 16/16 OD_{600} 0.4 Elution 2 **8.** 25/16 OD_{600} 0.4 Elution 3 Urea **9**. 25/16 OD_{600} 0.1 Elution 1 **10.** 25/16 OD_{600} 0.1 Elution 2 **11.** 25/16 OD_{600} 0.1 Elution 3 Urea

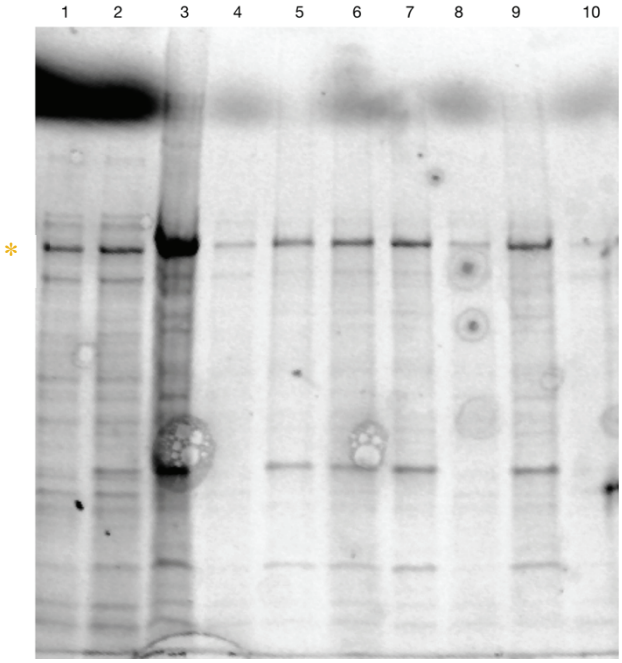


Figure 47. Solubility studies of 125_210_LATL I. Temperature prior to induction varied and OD_{600} of induction. **1.** 37/16 OD_{600} 0.4, 25mM Triton X-100 Elution 1 **2**. 25/16 OD_{600} 0.4, 25mM Triton X-100 Elution 1 **3.** 16/16 OD_{600} 0.4, 25mM Triton X-100 Elution 1 **4.** 37/16 OD_{600} 0.4, 5mM Triton X-100 Elution 1 **5.** 25/16 OD_{600} 0.4, 5mM Triton X-100 Elution 1 **6.** 16/16 OD_{600} 0.4, 5mM Triton X-100 Elution 1 **7.** 37/16 OD_{600} 0.4, 15mM Triton X-100 Elution 1 **8.** 25/16 OD_{600} 0.4, 15mM Triton X-100 Elution 1 **9**. 16/16 OD_{600} 0.4, 15mM Triton X-100 Elution 1. A yellow asterisk marks the correct MW band.

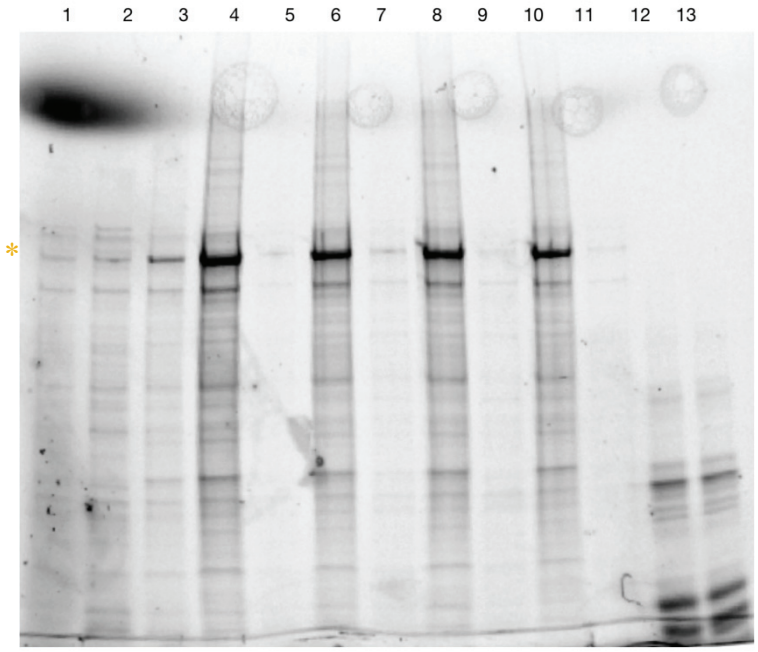


Figure 48. Solubility studies of 125_SYN195_LA. Temperature prior to induction varied and detergent varied. **1** 37/16 OD_{600} 0.4 Elution 1 **2**. 25/16 OD_{600} 0.4 Elution 1 **3.** 16/16 OD_{600} 0.4 Elution 1 **4.** 16/16 OD_{600} 0.4, Elution 1 Urea **5.** 16/16 OD_{600} 0.4, Elution 1 NP-40 **6.** 16/16 OD_{600} 0.4 Elution 1 Urea on pellet **7.** 16/16 OD_{600} 0.4 Elution 1 NP-40 on pellet **8.** 16/16 OD_{600} 0.4 Elution 1 Urea **9**. 16/16 OD_{600} 0.4 Elution 1 Triton X-100 **10.** 16/16 OD_{600} 0.4, Elution 1 Urea on pellet **11.** 16/16 OD_{600} 0.4, Elution 1 Triton X-100 on pellet **12**. 16/16 OD_{600} 0.4, Elution 1 IGEPACa-40 **13**. 16/16 OD_{600} 0.4, Elution 1 IGEPACa-40 on pellet. A yellow asterisk marks the correct MW band.

125_SYN195_LA solubility was not increased with the addition of any detergent or growth/induction conditions. Of the protein which was soluble, when tested using the malachite green adenylation domain assay showed no activity – likely denatured protein (results not shown).

3.3.11 Radioactive adenylation domain activity assay

Determination of substrate specificity

For all purified proteins, all proteogenic amino acids, along with a variety of 2,4-substituted acetylated pyrrole and furans were tested for activity.

For 125_120_AT, of the amino acids, asparagine, glutamine, threonine and valine gave results over the standard threshold for activity (>3000cpm) (Fig. 49). However, the acetylated pyrrole gave a much larger activity response. To further see the activity – a range of substituted pyrroles were tested – with compounds 4 and 6 giving a large amount of activity. This demonstrates both the level of promiscuity 125_120_AT possess, but also how the structure of the compound is detrimental to whether it is activated or not.

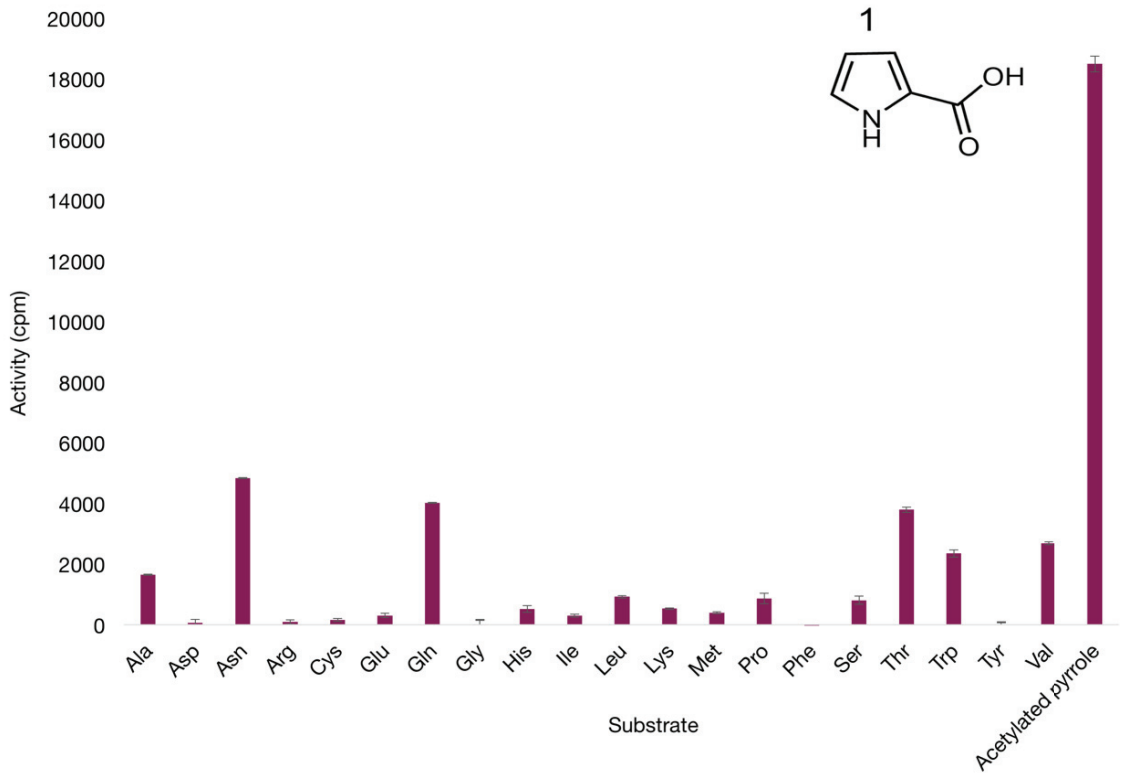


Figure 49. Activity of 125_120_AT purified A-domain determined via radioactive phosphate exchange assay. All proteogenic amino acids were tested in triplicate, with error bars denoting standard error. Initial assays also tested the acetylated pyrrole molecule 1.

Other available acetylated pyrrole containing compounds were tested for activity (Fig. 50), including substrate ‘A’ which is an additional pyrrole derivative tested in later assays (Fig. 51). Compounds 4 and 6 showed high activity, over the 3000cpm threshold for successful activity.

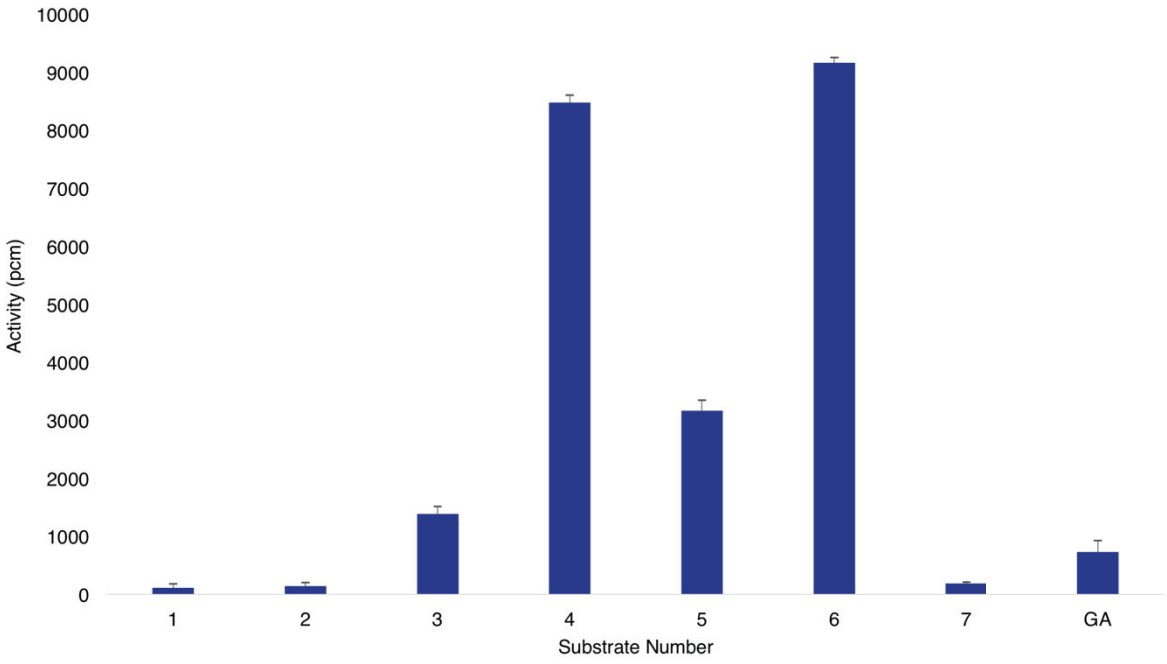


Figure 50. Activity of 125_120_AT purified A-domain determined via radioactive phosphate exchange assay I. 2,4-disubstituted pyrroles were tested, along with GA – the substrate of a similar A-domain. Error bars denoting standard error.

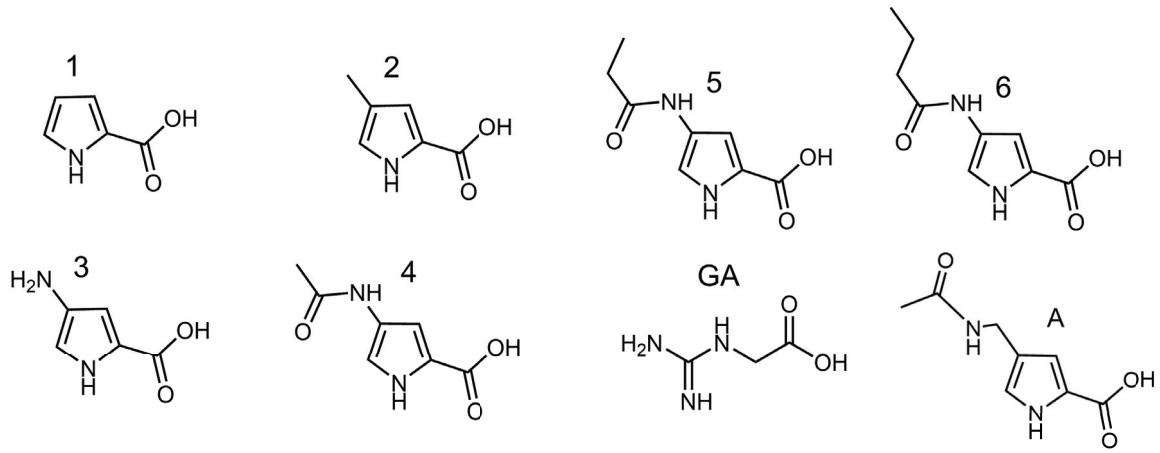


Figure 51. Pyrrole containing analogues of precursors predicted to potentially be involved in proximicin biosynthesis. GA is the substrate of Cgc18A, the adenylation domain containing the same amino acid substitution as Ppb120, and hence, potential alternative substrate.

Kinetic parameters by ATP-[^{32}P]PP_i exchange assay were determined for 125_120_AT, using the compounds which displayed the most activity – compounds 4 and 6. The time course data demonstrated that, although the overall activity of 125_120_AT adenylation domain is high, it is a very slow working enzyme – in comparison to other characterized domains.

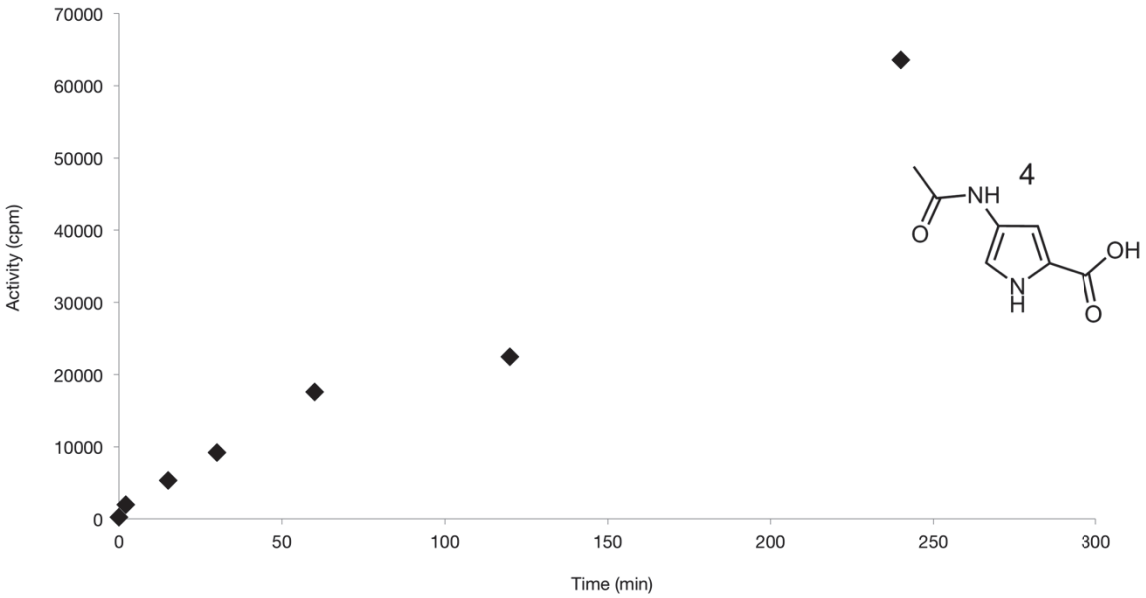


Figure 52. Time course determination of 125_120_AT adenylation domain. Determined using the radioactive phosphate exchange assay, using substrate 4. K_{cat} min^{-1}=0.0204 +/- 0.016, K_m[µM] = 6.2923 +/- 0.9491

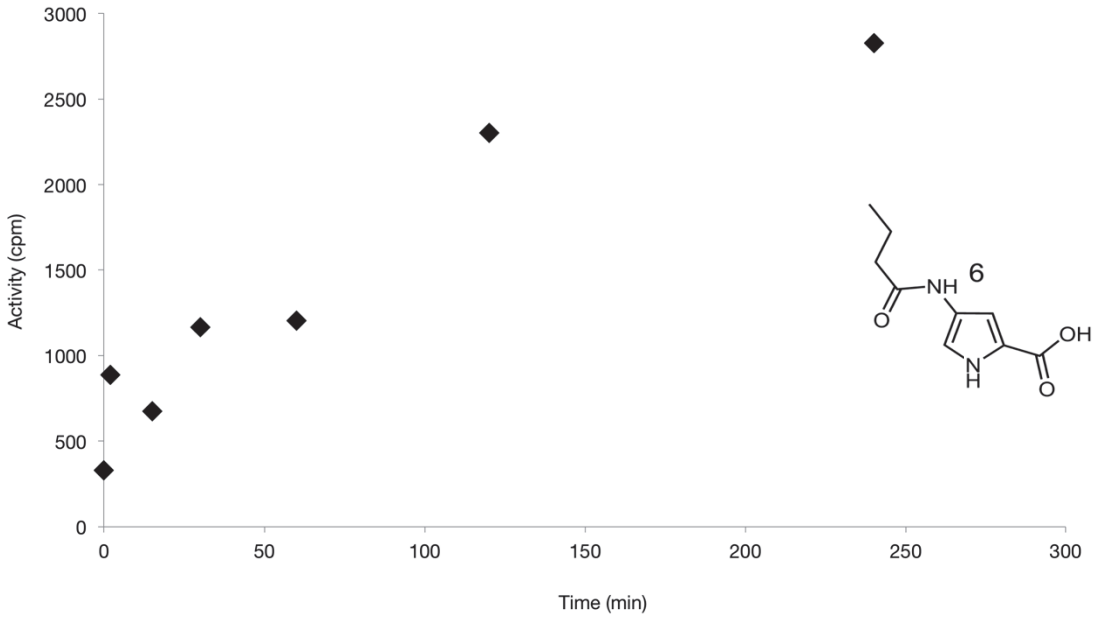


Figure 53. Time course determination of 125_120_AT adenylation domain I. Determined using the radioactive phosphate exchange assay, using substrate 6. K_{cat} min^{-1}=0.0294 +/- 0.014, K_m[µM] = 3.843 +/- 0.821

The enzyme kinetics of 125_120_AT with pyrrole containing compounds 4 and 6, gave the steady state kinetic parameters in Table 16, with Michaelis-Menten kinetic parameters in graphical form in Appendix F.

Table 16. Summary of enzyme kinetics of 125_120_AT

Enzyme	Substrate	K_m [μM]	k_{cat} [min^{-1}]
125_120_AT	4	6.2923 +/- 0.9491	0.0204 +/- 0.016
125_120_AT	6	3.843 +/- 0.821	0.0294 +/- 0.014

Purified adenylation domain protein 125_220_LATL displayed no activity (Fig. 54) with any proteogenic amino acid, pyrrole-containing compounds or furan-containing compound.

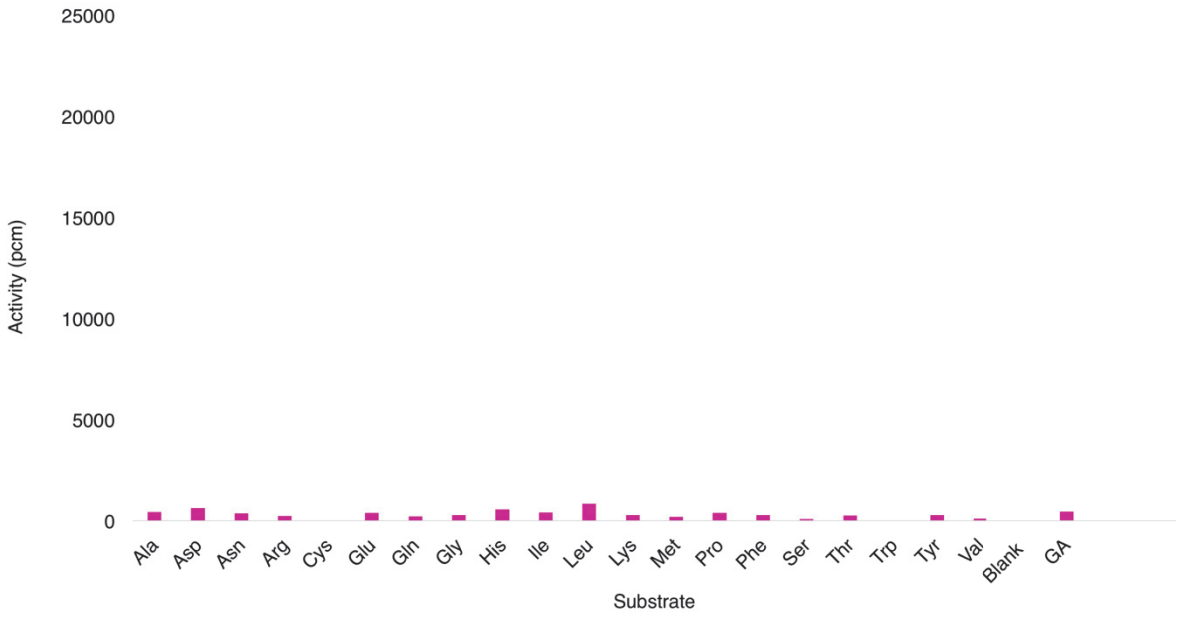


Figure 54. Activity of 125_220_A adenylation domain when tested against all proteogenic amino acids. Done in triplicate using the radioactive phosphate exchange assay.

Purified adenylation domain protein 125_210_LATL displayed no activity towards any proteogenic amino acid (Fig. 55), pyrrole-containing compounds or furan-containing compound – it specifically did not activate the substrate proposed in my proximicin synthetic route.

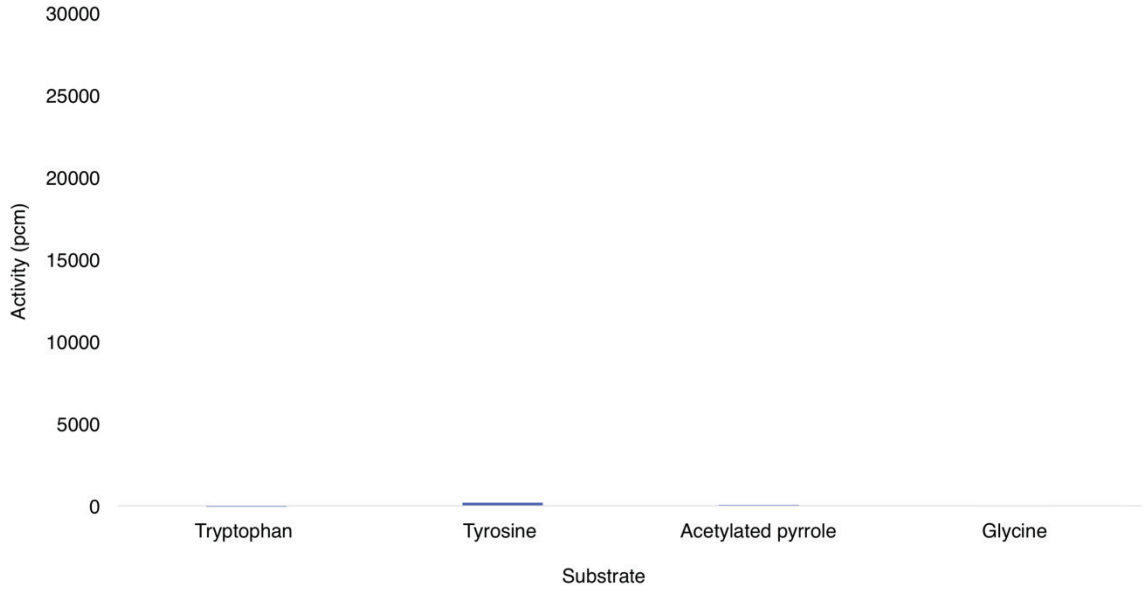


Figure 55. Activity of 125_210_LATL adenylation domain when tested against all proteogenic amino acids. None exhibited any activity with Ppb210 (only predicted compounds shown). Done in triplicate using the radioactive phosphate exchange assay.

3.3.12 Radioactive phosphate vs. malachite green assays

To assess the validity of the malachite green assay, 125_120_AT was tested against all the substrates tested using the malachite green assay. The results show that although some significant activation can be seen, the background noise is so high that it is hard to be sure of the results.

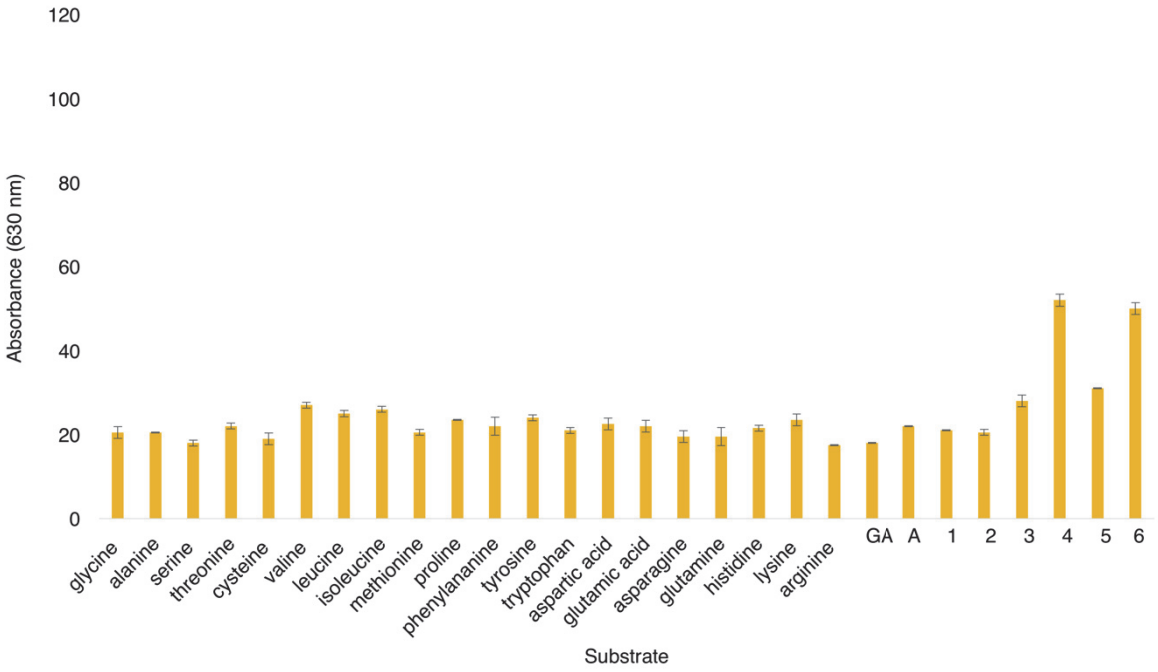


Figure 56. Activity of 125_120_AT determined using the malachite green adenylation domain assay. Error bars represent standard error.

3.3.13 Re-testing pOPINF construct activity with MbtH

Re-testing pOPINF_120 vector with ppb125 MbtH-like protein

The addition of Ppb125 had no effect on the induction level of pOPINF_ppb120, induction can still be seen but it is still at a relatively low rate.

Activity of pOPINF_ppb120 with Ppb125

pOPINF_ppb120_ppb125 showed no significantly increased activity with any amino acid or acetylated pyrrole. This inactivity with all substrates was determined to be most likely due to the protein being inactive, this was supported by previously low yields of protein induction and purification.

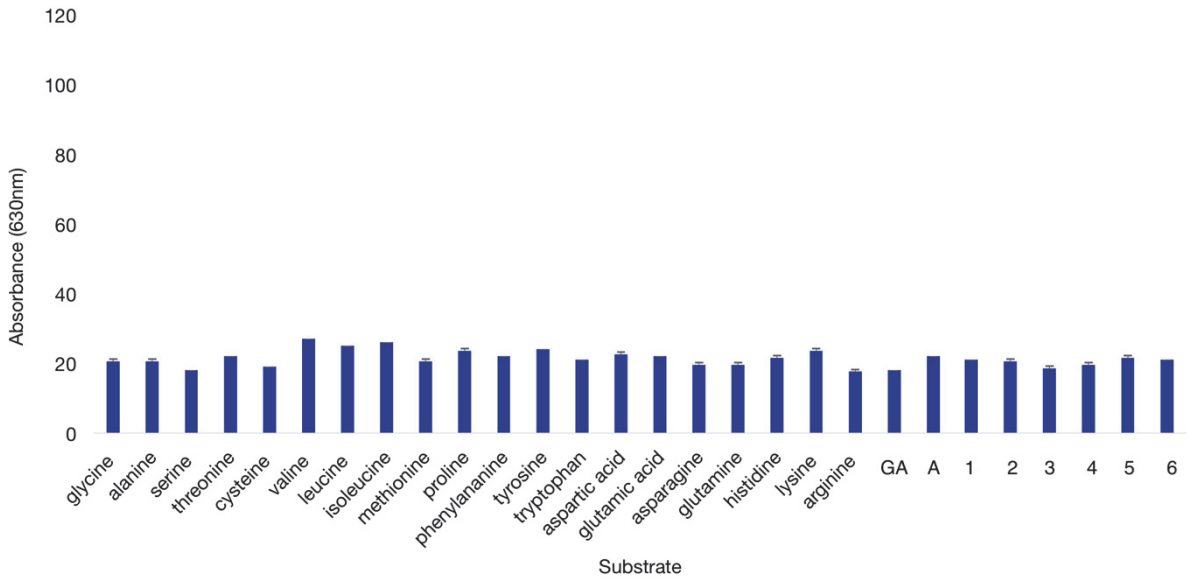


Figure 57. Activity of pOPINF_ppb120 + ppb125. Tested towards all proteogenic amino acids and a selection of pyrrole containing compounds. Determined using the malachite green adenylation domain activity assay. Error bars represent standard error.

3.4 Discussion

3.4.1 Overview of findings

In order to elucidate the proximicin biosynthesis pathway, all identified adenylating enzymes in the putative cluster were investigated - these enzymes dictate the amino acids incorporated into the natural product, providing evidence of the biosynthetic route utilized by *Verrucosispora* spp. The Ppb120 adenylating enzyme was revealed to be inefficient and promiscuous, able to activate a large range of heterocycle-containing substrates, giving support to a previously outlined route for proximicin biosynthesis. This was proven using two competing assay approaches to determine their applicability and accuracy to this type of research. Ppb220 was shown to be completely inactive, which was expected due to previously described bioinformatic analysis of the protein. This gave insights into possible origins of the proximicin cluster and avenues for potential combinatorial chemistry approaches to novel natural product production. Issues were encountered in the recombinant expression of two of the adenylation domains – Ppb195 and Ppb210 – due to their high GC% and solubility, respectively. This research will inform future work into novel furan-containing compound discovery and synthesis, as well as future experimental approaches used in adenylation domain activity elucidation, for example the importance of linker regions between neighboring domains and compound-specific MbtH-like proteins.

3.4.2 Novel activity and unique structure of Pbb120

The unique predicted structure of Ppb120 and the presence of 2,4-disubstituted furan groups is what initially attracted interest in proximicin biosynthesis; it poses an attractive target for novel antimicrobial compound production and combinatorial chemistry. This research has confirmed the novel activity of Ppb120, and discovered a high level of promiscuity in the compounds it can activate. Adenylating activity towards 2,4-disubstituted 5-membered heterocyclic compounds has only been seen previously once before – in congocidine biosynthesis with the AMP-binding super-family protein - Cgc3* which activates 4-aminopyrrole-2-carboxylate (Al-Mestarihi et

al., 2015). Cgc3* is not an NRPS adenylation domain, but a member of the ANL family, and shares high-to-moderate sequence identity with the acyl-CoA synthetases (Al-Mestarihi et al., 2015). To be able to harness the biosynthetic potential of Ppb120 it is important to determine the source of the novel activity; do both adenylating proteins work in the same way to activate this family of molecules, demonstrating common functions of two diverged proteins? Or is it a case of convergent evolution?

3.4.3 Enzyme kinetics of Ppb120

This work describes the adenylating activity of Ppb120 when tested against a range of potential substrates from the 2,4-pyrrole-containing family of compounds. In the outlined biosynthetic route of proximicin, a 2,4-furan precursor is iteratively incorporated into an elongating peptide chain. The complex chemistry which makes furans such an alluring target for investigation, makes them extremely difficult to synthesize; and, despite extensive attempts, the 2,4-furan containing precursor compounds were never successfully synthesized with a high enough purity and hence, could not be tested for activity. However, this gave an exciting opportunity to assess the promiscuity of the Ppb120 enzyme towards other heterocycle substrates containing a similar substitution pattern. Ppb120 was shown to be able to activate a large selection of 2,4-pyrrole molecules; the specific substitution pattern was extremely important for activity, and when absent activity was abolished. The two compounds with the greatest activity – compound 4 and 6 - appear contradictory: the length of carbon chain attached to the 2-carbonyl group increased activity when odd numbered, but even numbered carbon chains tended to decrease activity by a factor of >3. Potential explanations for this unusual result is discussed later, but it highlights the complexity of the enzyme/substrate interactions occurring. The enzyme kinetics, when compared to previous similar work (Sun et al., 2015), demonstrates that overall Ppb120 is active, but slow working protein which supports the theory that pyrroles are not the native substrates, but can be accepted and activated with reduced efficiency. Activity for a compound which exceeds >5000 pcm in the radioactive $^{32}PP_i$ exchange assay is typically categorized as a substrate of the enzyme; interestingly Ppb120 showed activity just over this threshold towards polar hydrophilic α–proteogenic amino acids, these included: asparginine, glutamine, threonine and valine. This result is extremely unusual when

considering Ppb120 has such stringent requirements for the compounds it will activate – seen particularly by inactivity towards compounds with even carbon lengths. As the research effort was focused on identifying the A-domain responsible for furan-precursor activation, this result was not regarded as being significant in the scope of this investigation, and hence will not be further discussed here. It should be noted, however, that it was concluded that this apparent activity was most likely an artifact of the assay methodology, and any genuine activation was at a very low and inefficient level, but should be considered in further research into the promiscuous activity of Ppb120 as an exciting prospect for unnatural natural product combinatorial biosynthesis.

3.4.4 *Asp_{235} substitution*

To exploit the novel activity exhibited by Ppb120, it is important to identify the enzymatic architecture responsible. Ppb120 forms part of the ANL family of adenylating enzymes, which is composed of the acyl- and aryl-CoA synthases, firefly luciferases, and the adenylation domains of NRPS's; it is a mechanistically diverse enzyme superfamily. Members catalyze two partial reactions: i) the initial adenylation of a carboxylate substrate with ATP to form an acyl-AMP intermediate and, ii) most commonly, the formation of a thioester. All ANL enzymes share a conserved mechanistic step – the adenylation partial reaction- and are structurally homologous, but share only ~20% sequence identity. Due to the breadth and importance of the reactions they catalyze, the ANL family have been subject to extensive study, with much work centered on the understanding of their enzymatic architecture. Early studies described Asp_{235} as strictly invariant (Stachelhaus et al., 1999); however, examples of non-conforming NRPS-adenylation domains have been discovered. One of interest and previously discussed is Cgc18 that is involved in the biosynthesis of the antibiotic congocidine, and has the same substitution pattern as Ppb120: $Asp_{235} \rightarrow Ser_{235}$. This radical substitution from a very acidic to polar-hydrophilic amino acid in an apparent 'invariant' residue seemed an extremely good reason to explain the proposed adenylation domain activity of Cgc18 and Ppb120 i.e. the selection and activation of pyrroles and furans, respectively (Table 17). This hypothesis was shared by Juguet et al., (2009) when the initial congocidine pathway was outlined. It was thought that the Cgc18 adenylation domain was responsible for the activation of a pyrrole-containing molecule, 4-aminopyrrole-2-

carboxylate – a molecule not previously seen to be activated by an adenylation domain; a novel enzyme activating a novel compound. This would support a homologous route to furan activation in Ppb120 – a molecule very similar to that activated by Cgc18. However, the congocidine pathway was later revised (Al-Mestarihi et al., 2015) and it was shown to activate guanidinoacetic acid (GA), and it is a different adenylating protein – Cgc3* – which activates the pyrrole-containing molecule. Cgc3* was shown to begin the Congocidine biosynthetic pathway by the ATP-dependent activation of 4-acetamidopyrrole-2-carboxylic acid. Cgc3* also has a residue substitution: $Asp_{235} \rightarrow Tyr_{235}$. Using the chemical complementarity between the binding pocket and the substrate as the basis for clustering (Rausch et al., 2003; Challis et al., 2007; Reger et al., 2007) and predication (Lautru et al., 2007) of NRPS adenylation domain activity is not a new concept. It is well documented that adenylation domains which activate similar compounds, will likely have a similar binding pocket arrangement; this is indeed the premise used by many NRPS-predicting programs. To this end, it would be sensible to suggest that Cgc3* and Ppb120, as activating similar molecules, would have a similar binding pocket architecture. However, by doing simple phylogenetic studies considering only the specific conferring code, it can be seen that despite similarities in activities, they belong to distinct clades in the ANL family of proteins (Fig. 58). The whole binding pocket consensus residues of Cgc3*, as determined by Stachelhaus et al., (1999) are atypical, with most being completely absent. This is expected as Cgc3* is not a NRPS-adenylating protein, and hence, will not conform to the parameters dictated for that specific group of proteins. However, it is logical to assume that being able to activate the same compounds would be the result of similar enzymatic machinery, but no homology can be found. Because of this, it is hard to propose how the Cgc3* binding pocket is able to accommodate and activate its substrate, and hence, how this relates to Ppb120 activity. Employing this substrate prediction method for novel, non-conforming adenylation domain activity is bottle-necked by the lack of characterization of such enzymes. As the capability of NGS continues to flourish, currently unidentified or silent biosynthetic gene clusters will inevitably add to this sub-family of novel enzymes and increase bioinformatical substrate identification capabilities, as well as how different members of the ANL with distinct structures are able to support the same activity.

Table 17. The core residue which line the A-domain active site in NRPS adenylation domains. Residues of particular interest shown in bold: the aspartate residue previously reported to be invariant and pink: 'wobble-like' positions shown to support high variance between A-domains, and be responsible for arranging side chains of substrates.
* Cgc3* is not an NRPS-adenylation domain, but included for comparisons sake.

Adenylation domain	Substrate specificity	Residue position**									
		235	236	239	**278**	**299**	301	322	330	331	517
Ppb120	2,4-Pyrrole/furans	**S**	S	E	**Y**	**T**	G	A	G	I	K
Cgc18	GA	**S**	V	E	**Q**	**V**	G	E	V	S	K
Cgc3*	Pyrrole	**Y**	P	L	*	*	*	*	*	*	*
DhbE	DHB/salicylate	**P**	L	P	**A**	**Q**	G	V	V	N	K
EntE	DHB/salicylate	**A**	M	P	**L**	**Q**	G	V	V	N	K
Luc	Luciferin	**G**	F	F	**L**	**A**	G	G	S	A	*
GrsA	Phe	**D**	A	W	**T**	**I**	A	A	V	C	K

**GrsA numbering

3.4.5 Similarity with other ANL superfamily enzymes

When looking at the clustering pattern of the specificity conferring code Ppb120 (Fig. 58), it is not closely related to other NRPS-adenylating enzymes: neither Cgc18 which has an identical substitution pattern in a fundamental residue, nor Cgc3* which is shown to activate the same family of compounds. It branches from the other NRPS-adenylation domains at a very early node; it does, however, cluster with the firefly - *Photinus pyralis* – and another luciferase adenylating enzyme. Both enzymes catalyze the activation of a carboxylate substrate by reacting with ATP to form an acyladenylate and PP_i (Fig. 59). Distinct from NRPS systems, luciferase adenylating enzymes activate luciferase to produce a luciferacyl-adenylate compound which undergoes oxidative decarboxylation resulting in the production of a photon of light (Fig. 59a). It is expected to see a certain level of homology as they share a similar initial partial reaction and it can be assumed that they do not share a similar secondary reaction – Ppb120 is not a luciferase. However, they do share similarity when looking directly at the active site residues responsible for substrate binding. The two substrates known to be activated by Ppb120 (Fig. 59b) and the *P. pyralis* luciferase appear to occupy a similar chemical space - they have a notable Maximum Common Substructure (MCS) value of 9 (Fig. 59c). MSC gives a numerical value to common elements shared by two structures; a value of 9 means a path of 9 identical bonds can traced through each molecule, specifically 9/12 (75%) bonds present in the furan-containing compound can be mapped directly onto luciferin – this represents a large amount of structural similarity. The common substructure is interesting as it has a similar structure to GA (Fig 59d) – the substrate of Cgc18, which is the other adenylation domain with the Asp_{235} substitution. Without knowing how the pocket binding interacts with the substrate, it is difficult to make conclusions regarding the similarity between Ppb120, Cgc18 and the luciferases. However, it is possible to theorize that Ppb120 and indeed Cgc18, may represent an early diverging ancestor of the NRPS adenylating enzymes and luciferases, which have evolved different secondary enzymatic reactions but maintained similar substrate activity. Or simply, the Asp_{235} mutation has allowed the NRPS-adenylating enzymes to activate molecules which coincidentally share a high MCS with the luciferase substrate - luciferin. This is not the first time overlap between these two apparently distinct ANL sub-families has been described: GrsA – the phenylalanine

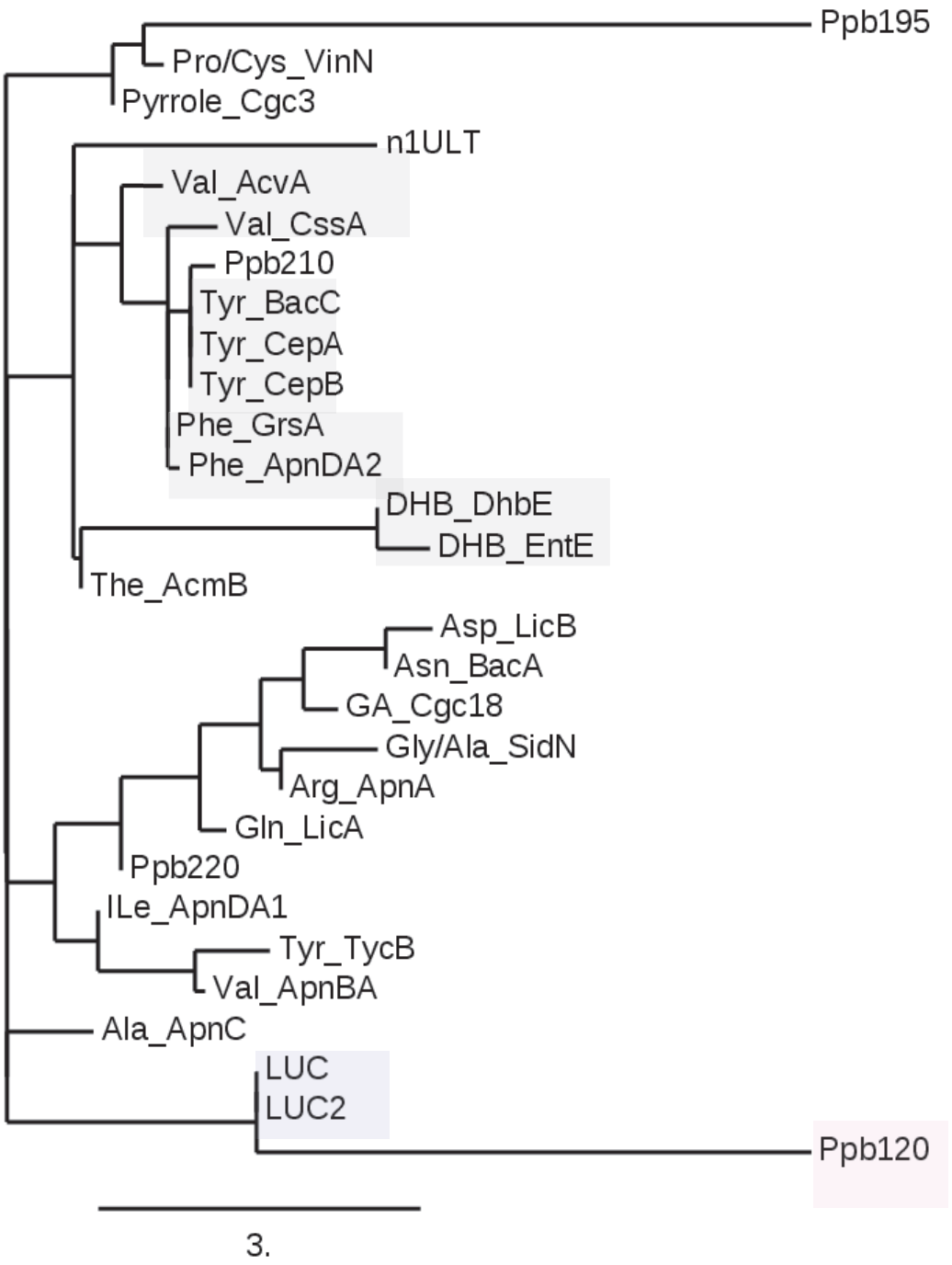


Figure 58. Phylogenetic analysis of the conserved specificity conferring code of NRPS-adenylation domains. Demonstrating how substrate structure can be predicted from the comparison of characterised A-domains and their substrates. Grey shading shows clades of very conserved activity for specific amino acids. Blue shading shows that the luciferases diverge very early from the other family of NRPS adenylation domains, as expected. Pink shading shows the lack of clustering the specificity conferring code of Ppb120, indicating its unique adenylating activity.

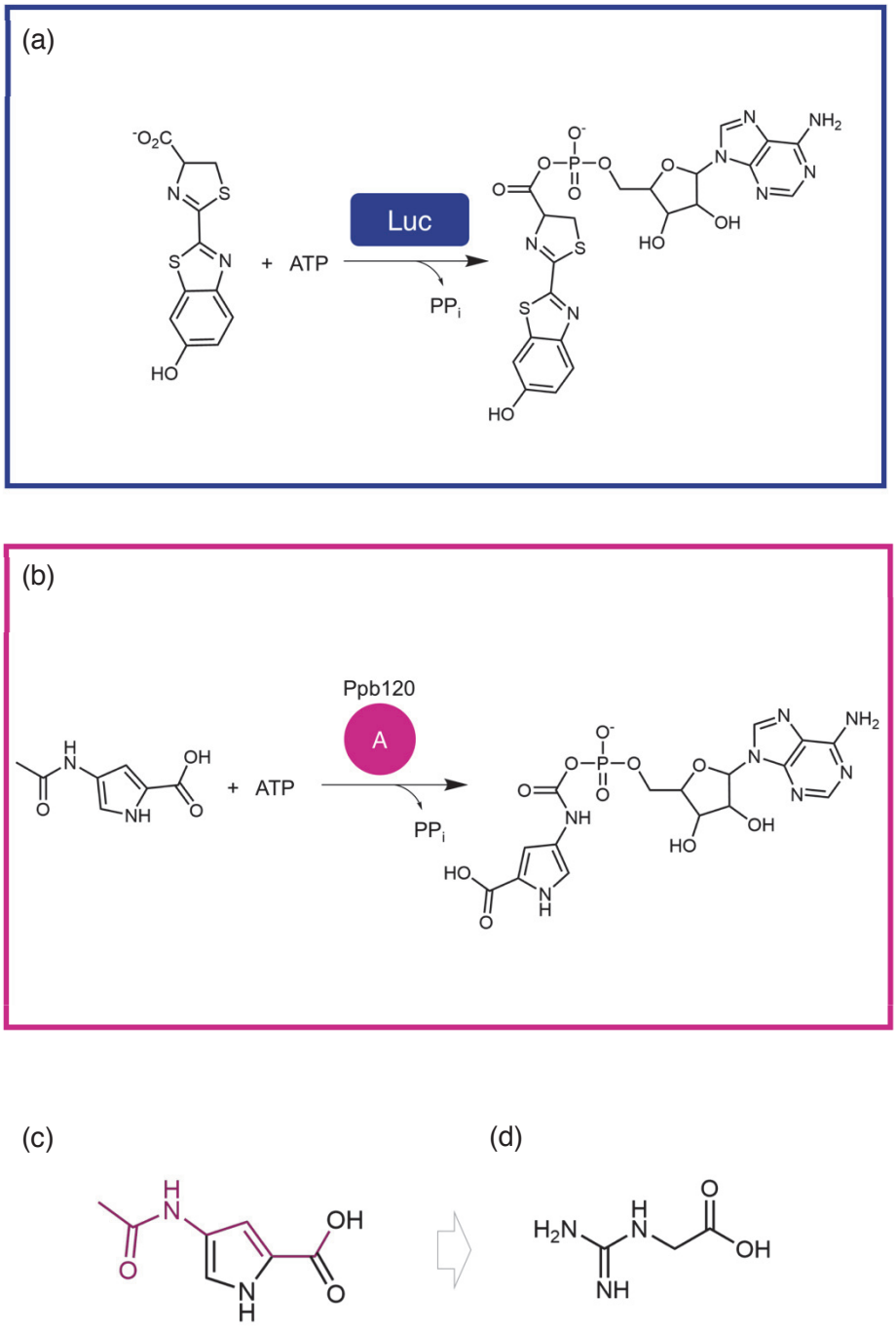


Figure 59. Formation of an acyl-AMP intermediate from a carboxylate and ATP via enzymes representative of two ANL sub-families. (a) The adenylation reaction of luciferin, catalysed by Firefly luciferase. **(b)** The adenylation reaction shown to be catalysed by NRPS-protein Ppb120 of a heterocycle-containing compound. Despite being shown to activate similar initial substrates, both differ in the second partial reaction they go on to catalyse. **(c)** The conserved chemical structure of both luciferin and 2,4-disubstituted furan highlighted in purple, which shares a similar structure to the demonstrated substrate of Cgc18 – GA **(d)**. This shows potential crossover between families of adenylating enzymes, not previously highlighted.

activating adenylation domain involved in gramicidin production – has a similar three-dimensional structure to that of firefly luciferase (Conti et el., 1997). They do not cluster according to these substrate conferring residues – as they activate different substrates – but they have a conserved tertiary arrangement. This is unusual as structural studies of luciferases has shown that despite a similar overall 2-domain architecture, the C-domain is rotated by ~ 90° in luciferases compared to the orientation seen in NRPS enzymes. To further determine whether Ppb120 has a luciferase-like C-domain orientation, the crystal structure of Ppb120 would have to be solved, and hence, issues surrounding getting a large enough quantity of soluble protein would have to be resolved. Other overlaps between ANL enzymes has been exhibited previously, Al-Mestarihi et al., (2015) stated that Cgc3* has 'high to moderate' sequence similarity with Acyl-CoA synthetases, and has been shown to be able to act as a acyl-CoA synthetase – once the substrate has been activated by ATP in the first partial reaction, Cgc3* could then produce substrate-CoA. As the ANL super family have been shown to accept both CoA an T-domains as acyl accepters and the amino acid sequence similarities between Ppb120 and Cgc3*, it would be interesting to determine whether Ppb120 has this promiscuity in the second partial reaction. The ANL superfamily is a relatively new concept, and the work described here reiterates the need for further investigation into this complex enzyme group and enzymes which potentially occupy activities and structures between sub-families.

3.4.6 A5-core motif altered

The novel adenylation activity of Ppb120 and Cgc3* may be the consequence of other, less investigated residues. As well as the specificity-conferring code previously discussed, Marahiel et al. (1997) outlined ten NRPS adenylating enzyme-specific core motif residues– A1-A10. One of particular interest here is core motif A5, which is responsible for both structure of the enzyme and substrate binding; specifically, the aromatic residue of the conserved tyrosine is stacked against the adenine base of the ATP molecule during binding. Crystal structures have also shown it to form the walls of the active site, and is required for the binding of Mg^{2+} ion for activity (Gullick, 2009). In both Cgc3* and Ppb120, this region is atypical (Fig. 60a); Ppb120 has substitutions in core residues, and in Cgc3* the domain is missing entirely. When analyzing the predicted protein structure produced by iTASSER,

(a)

```
D1GLU5    NGYGPAE
Q70LM7    NIYGPTE
Q08787    NCYGPTE
P0C061    NAYGPTE
Q9I4B7    NGYGPTE
3L8C      NAYGPTE
4ZXJ      NLYGPTE
3ITE      NVYGPTE
3WV4      RMYGQTE
4D4G      NVYGPTE
Cgc18     QEYGATE
Cgc3      -------
22210     NAYGPTE
22195     NAYGATE
22220     HLYGPTE
22120     VACGSSE
```

(b)

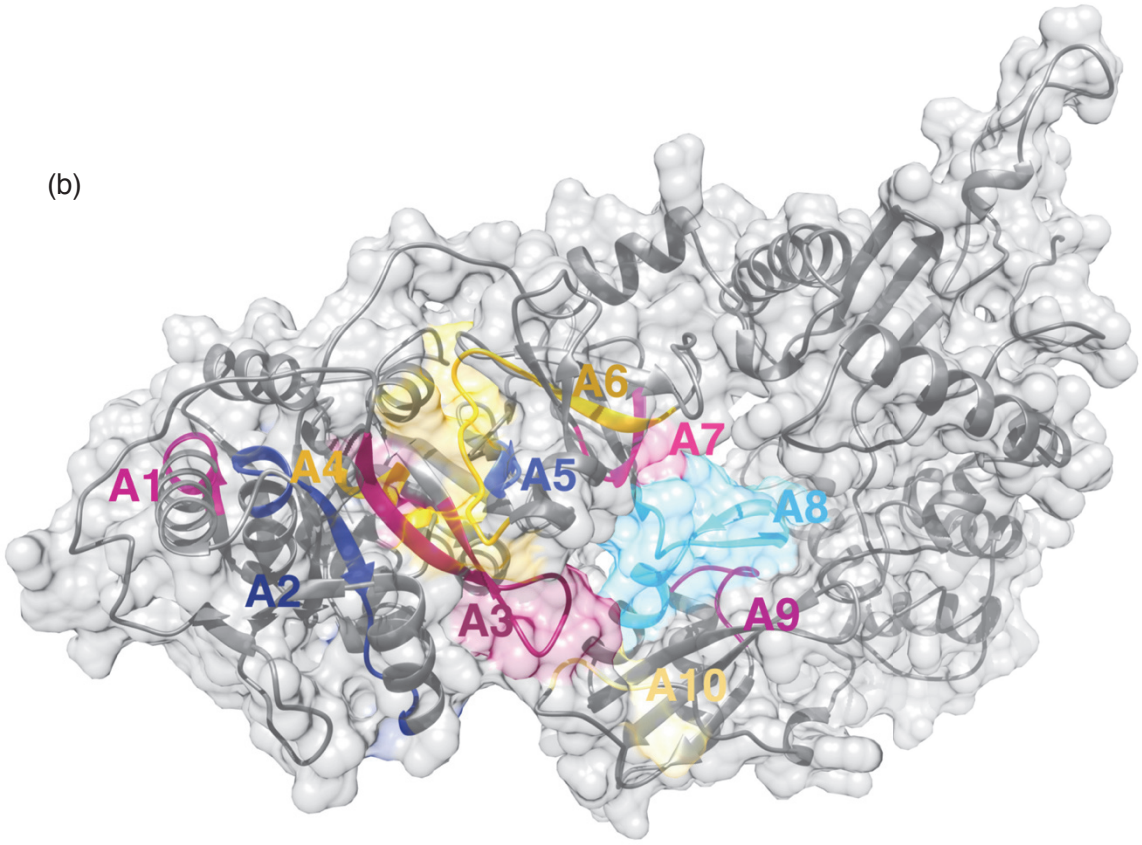


Figure 60. Importance of A5 domain in NRPS-adenylation domains. **(a)** The aligned sequences of domain A5 known to be important in ATP, Mg^{2+} and substrate binding, of many characterised A-domains, with Ppb120 and Cgc3 highlighted to show the lack of conserved residues in either adenylating enzyme in this region – in stark comparison to the highly conserved motif seen in all other A-domains. **(b)** The predicted secondary structure of Ppb120 using iTASSER and rendering done in Chimera with all core domains highlighted, showing the central role A5 (Blue) has and how mutations in this region will likely cause ambiguity in substrate preference and promiscuity.

(Fig. 60b) the area which spans the A5 domain is completely novel in structure – it has very low identity with any solved NRPS-domains. It is known that this stretch of amino acids – between A4 and A5 motifs - form a pocket that is responsible for accommodating the sidechain of the substrate. It is also important for maneuvering the substrate and AMP into the correct conformation for the adenylation reaction to take place. The adenylate is bound in the cleft on the surface of the A domain by means of van der Waals and hydrophobic interactions, in addition to hydrogen bonding (Weber et al., 2002). Any change to the amino acids present or charges which line the binding pocket would detrimentally affect both adenylation activity and substrate specificity. This region is of such importance, that a single amino acid substitution resulting in the introduction of a negative charge can result in a complete enzyme inactivation (Saito et al., 1995). The overall similarity of Ppb120 and Cgc3* is very low, and so it is unlikely – despite their activity towards the same substrates – that they act in a similar mechanism. Rather, it is possible that the lack of this region simply allows the activation of unusual substrates as the confines these residues typically present are absent. Further work into both Cgc3 and Ppb120 to discover the rules which govern their substrate selection and activity is required to be able to appreciate and utilize their naturally exquisite molecular design.

3.4.7 *Novel binding pocket structure*

The adenylation protein GrsA which activated PheA can be considered a precedent for the entire family of adenylating enzymes and allows the binding pockets of other adenylation domains to be determined and assumptions of structure made. Conclusions can be drawn about the mechanisms responsible for the unusual substrate activation patterns demonstrated by Ppb120 by using comparisons to the solved structure of GrsA. An example being the previously discussed conserved Asp_{235} in GrsA is modified to Ser_{235} in Ppb120 (Table 17) a variation from an acidic to polar and highly hydrophilic group. This residue is responsible for the stabilization of the α-amino group of the substrate, but as, unlike GrsA, Ppb120 does not display significant activity towards α-amino acids, this is not required. Other adaptations of the binding pocket can be observed and used to explain the novel activity, such as the high incidence of small side-chain amino acids (Fig. 61). It has been shown that

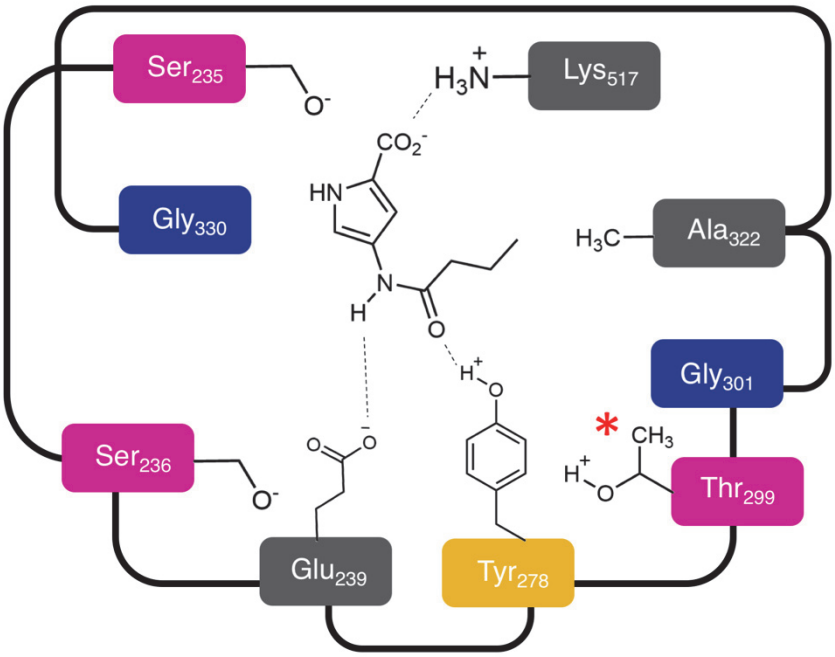


Figure 61. **The binding pocket of the adenylation domain in Ppb120.** Potential interactions with the substrate, dictated by the binding pocket residues are shown. The red asterisk marks an area of potential conflict, leading to the paradoxical substrate activity exhibited by Ppb120 A-domain towards substrates with even numbered carbon chains. Adenylation domain in the style of Challis et al., (2000).

substrate specificity can be relaxed by a factor >5 if large residues are replaced by smaller ones (Stachelhaus et al., 1999); such amino acids - glycine and alanine - represent a large proportion of the top of the active site. The reduction of space occupied by binding pocket residues allows more space to be utilized by the substrate. This supports the experimental findings that Ppb120 is a very promiscuous enzyme, with a large potential substrate pool. In GrsA, the large sidechain of Trp_{239} residue at the bottom of the pocket allows the active side to be successfully separated; this is not the case in Ppb120 which has the relatively small Glu_{239} (Fig. 61). However, neighboring this at the bottom of the of the active site is Tyr_{278} which has a large sidechain, and so we can assume these residues together have a similar structural role. Also, we know that these two residues either separately or in combination, must be responsible for securing the furan-containing compound into the active site. They both represent well described 'wobble-like' positions in the binding pocket, and hence, are most likely responsible for binding unique side chains. One potential way this could be achieved is by hydrogen bonding between the amide group on substrate and the hydroxyl group on Tyr_{278} or more likely due to increased polarity – Glu_{239}; this would ensure that the substrate was in a very specific orientation for activation. The unusual finding that only odd chain lengths are tolerated for high enzyme activity suggests that this dynamic is very fragile, and any slight change to structural shape leads to ineffective activity; the acidic residue Thr_{229} may interfere with the hydroxyl-amide hydrogen bonding if the substrate is not in a very specific orientation i.e. has odd branched R-group chains; the location of this potential conflict is marked with a red asterisk. Until the structure of Ppb120 is solved, this is simply interesting conjecture; the affect that these unusual binding pocket residues exert can, at this point, only be theorized.

3.4.8 Inactivity of Ppb220

Possible end of cluster

As seen from the results, Ppb220 is completely inactive against all proteogenic amino acids and any furan/pyrrole containing molecule tested. Despite containing all the essential residues, *ppb220* lacks many of the highly conserved domains – specifically at the N-terminus of the protein: motifs A1-A7. It was suspected that due to the lack of large spans of required residues, Ppb220 is likely to be completely inactive; this is supported by the experimental results. The missing or obscured core

regions are known to be vital in the adenylating activity and structure of the protein, and so inactivity was expected. Although differences between outputs given by secondary metabolite finding programs, all included *ppb220* as part of the *ppb* cluster. As *ppb220* lies at the far boundary of the ppb cluster, and no genes beyond it have a function in the proposed proximicin biosynthetic route. I suggest *ppb220* is not part of the proximicin biosynthetic gene cluster, and the boundaries of the cluster have been over-estimated. To confirm this, gene deletion studies must be undertaken, inactivating or deleting genes sequentially until proximicin production is stopped.

Potential ancient products

The *ppb220* gene was regarded as part of the *ppb* cluster as it has characteristics of an active adenylating enzyme: it contains at least some core motifs and structural anchors that secondary metabolite programs search for to identify biosynthetic enzymes. The non-functioning *ppb220* gene may represent the remnants of a once active A-domain, this suggests that the *ppb* genes were once responsible for the production of a larger product which would of likely contained a furan moiety. This is further supported by the presence of a partial neighboring biosynthetic cluster – denoted as PNB (Fig. 62)- upstream of *ppb220*, which some programs such AntiSMASH predict as being part of a singular NRPS cluster responsible for a much

Table 18. Analysis of the NRPS-domains present in the PNB cluster. The specificity conferring code of A-domains present and proposed substrate determined by NRPS2Predictor. Genes names based on the published *V. maris* AB18-032 genome.

Gene	Domains present	Consensus sequence	Proposed substrate
22280	T	/	/
22285	C-T	/	/
22290	T	/	/
22305	C-A_1-T-C-A_2-T-Te	A_1 DVPHFSLV	Serine
		A_2 DLGGLGGI	Ornithine

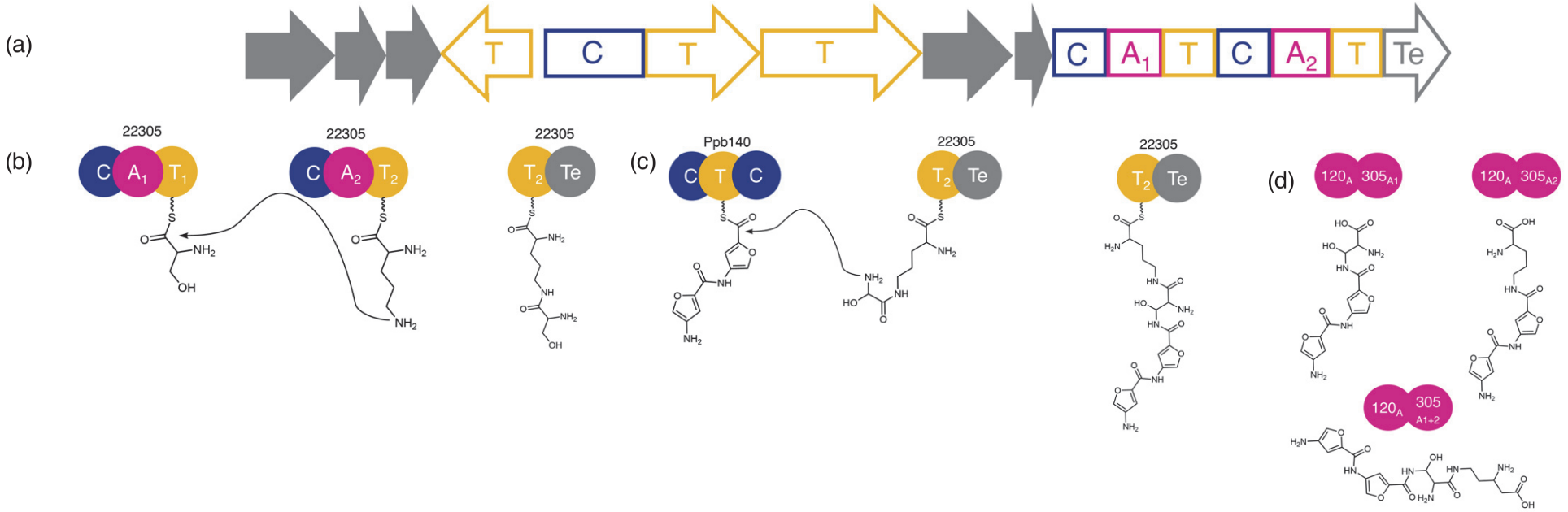


Figure 62. The ancient NRPS cluster upstream of *ppb* - PNB cluster (a) the genomic organisation of the PNB cluster, showing the abundance of T-domains, potentially responsible for iteratively added substrates. **(b)** the structure of PNB-alone compounds **(c)** the structure of compounds of potential ancient compounds, if Ppb and PNB clusters were once combined into a single cluster. **(d)** some structures if Ppb/PNB clusters were combined, with potential novel DNA binding and antimicrobial activity.

larger compound. The structure of likely compounds produced from this larger cluster are predicted to contain heterocycles, with residues similar to tryptophan and tyrosine present – reminiscent of proximicin, as well as a chain three amino acid chain. Mutations leading to the inactivity of Ppb220 could represent a disjointing event of this cluster, stopping production of the larger peptide, resulting in proximicin production. We dismissed this in the search for the proximicin biosynthesis genes, as this only requires the incorporation of a maximum of three precursors and hence many of these NRPS-enzymes would be redundant. Evaluation of this secondary cluster reveals 4 NRPS-related proteins, shown in table, and a small selection of peptide tailoring enzymes. Only a singular NRPS-like gene of PNB– *22305* - contains any adenylating domains, of which it contains two. As shown in Table 18, using a NRPS-predictor program they were identified as potentially activating serine and ornithine. We cannot confirm whether these enzymes have intact machinery capable of producing a compound which has yet to be identified, if the cluster is simply silent or if these genes represent the remnants of a large pathway once in association with the *ppb* cluster. To determine which it is, potential PNB-alone products can be hypothesized and production searched for (Fig.62). Trying to identify any compounds produced by the native *Verrucosispora* spp. *in vivo* will be potentially very challenging: the abundance of standalone T/C-domains within the cluster, suggests that the compounds produced may contain subunits iteratively added to the extending peptide chain. Also, the vast array of tailoring enzymes such as aminotransferases and dehydrogenases present suggest a potentially large chemical space to explore. It would be interesting to determine whether it was once part of a larger single biosynthetic cluster, as this would indicate a whole array of enzymes capable of modifying precursors/peptide chains containing the furan moiety. Regardless of the origins and historical functions of these enzymes, they offer a potentially large array of biosynthetic machinery for exploitation. The pharmaceutical potential of furan – containing molecules has been extensively discussed, and any other 2,4-furan containing molecules could be of significance. Expanding the chemical space of this compound family, and in turn, the potential antimicrobial, antifungal or/and DNA binding activity would be extremely valuable. It is by their very modular architecture that NRPS systems lend themselves to modification; this has been demonstrated extensively with domains being inactivated and interrupted to result in the production of novel non-natural

natural products. This approach could be used to: I) replace *ppb220* with an active adenylation domain, or repair *ppb220* by CRISPR e.g. change the residues that cause inactivity II) delete the Te domain to see whether the elongating peptide chain would be moved to *22305* and III) co-purify with a range of tailoring enzymes to determine their affect.

3.4.9 Insolubility of Ppb210 and Ppb195

Despite efforts, both adenylation domain proteins Ppb210 and Ppb195 could not be effectively purified in an active form; this is an issue commonly encountered during adenylation domain activity studies. Adenylation domains have complex structures and the requirement of many co-compounds for solubility and activity; with advances in scientific capabilities and the full realization of the potential these bacterial systems offer, NRPS clusters will become more widely studied and hence, understood. From the work outlined here, much can be gathered which could inform future research attempts. Of important note is: the presence of magnesium ions (Mg^{2+}); compound specific MbtH-like proteins and, linker regions; these factors played an apparent key role in proximicin adenylate enzyme solubility and activity.

Mg^{2+} Concentration

Magnesium has been shown to be essential for activity in all adenylating enzymes, although some have been shown to be able to utilize Mn^{2+}. Proximicin adenylating enzymes were a not an exception to this rule, in the absence of Mg^{2+} in both culturing media and activity assays, the enzymes were inactive (results no shown). The presence of this divalent metal ion has been shown to neutralize the charge of the ATP, stabilize the transition state and neutralize the leaving pyrophosphate group. How it accomplishes this, the concentration required and its coordination geometry, varies across the ANL super family and even within sub-families. Crystal structures of NRPS enzymes have revealed the binding of Mg^{2+} via a β - γ coordination; in this way, Mg^{2+} also has an important role in ensuring that the correct protein structure is present for adenylation to take place. This has been seen in NRPS A-domain DltA - Mg^{2+} cannot bind in the active site and hence allow adenylation, unless a conformational change has occurred and the protein is in the correct orientation for activity (Yonus et al., 2008). This blocking of activity is the result of a salt bridge displacing the Mg^{2+} preventing activity due to bonding

interactions between the anionic carboxylate of an aspartic or glutamic acid with the cationic ammonium from a lysine or arginine residue. From previously solved structures it can be seen that this bonding occurs at the Glu_{297} and Arg_{407} residues in the adenylating DltA protein (Yonis et al., 2008); all the proximicin adenylation enzymes, however, contain Pro_{407}. The absence of the cognate base prevents any salt bridges forming, and so it is more likely that the relationship between metal ions and Ppb enzymes is more akin to that seen in Acyl-CoA synthases. This subfamily of enzymes has been shown to have an equivalent 407 residue replacement from a basic to a non-polar hydrophobic amino acid which results in an open active site structure where Mg^{2+} can freely enter (Hisanaga et al., 2004). The stringent requirement of Mg^{2+} shown here supports the idea a homologous system being utilized - the presence of a divalent metal ion is required for activity and not a conformational role. However, crystal structures with bound Mg^{2+} would be required to conclude this is the case for Ppb120 and other Ppb enzymes.

Compound-specific MbtH

Many bacterial gene clusters encoding NRPS systems involved in natural production also code for a small protein which contains three specific conserved tryptophan residues (Felnagle et al., 2010). It has been shown that they can be essential chaperones for NRPS-dependent natural product synthesis (Drake et al., 2007; Lautru et al., 2007), and also have roles in protein production and solubility (Heemstra et al., 2009), and play a vital part as activators of acyl-adenylate formation (Felnagle et al., 2010 and Zhang et al., 2010). Further work has also shown that MbtH-like proteins can operate in different biosynthetic clusters of the same species and across species (Boll et al., 2011). A MbtH-like protein was identified as being part of the proximicin cluster – Ppb125, and its requirement in proximicin biosynthesis was explored. The results show that Ppb125 was required for activity and solubility of all Ppb adenylating enzymes, including the novel-substrate activating enzyme Ppb120. When replaced with a non-native MbtH protein – the MbtH-like protein TioT – solubility was greatly reduced, and any protein purified was inactive. This shows a level of specificity of the proximicin biosynthetic pathways towards Ppb125. It is possible that Ppb125 could promote adenylate-forming activity by either altering the structure of the adenylation domain resulting in the adoption of a catalytically competent formation or by the prevention of the alternate

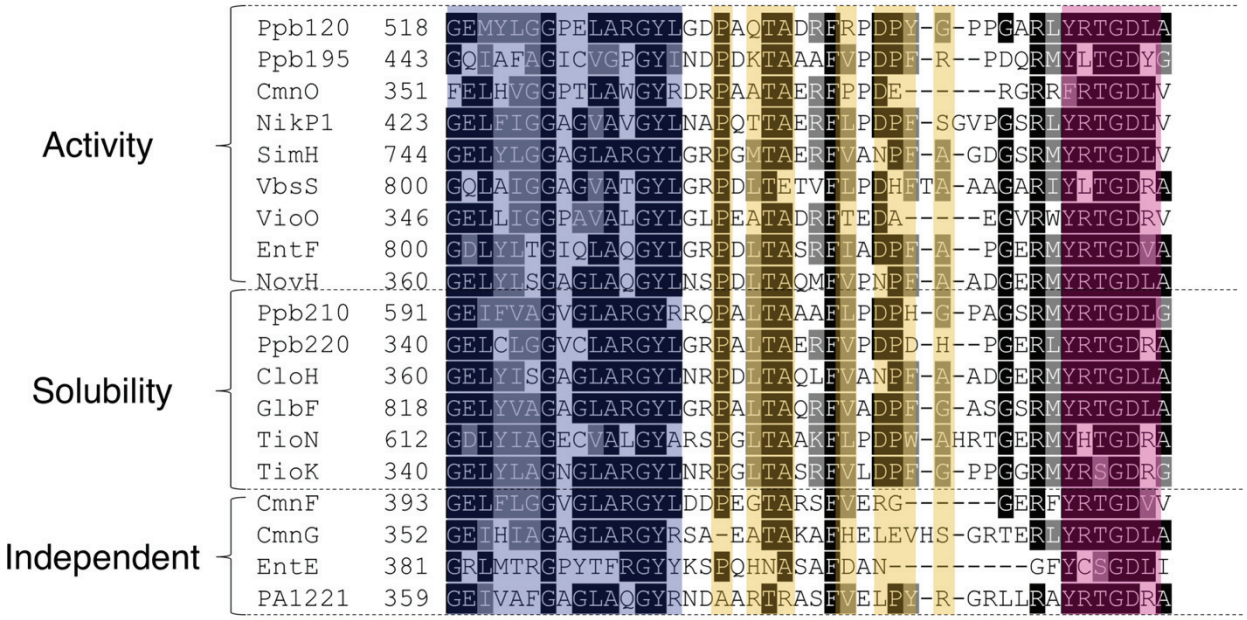


Figure 63. Multiple sequence alignment of characterised NRPS adenylating enzymes, known to have different reliance on their cognate MbtH-like protein. Either for adenylation activity, protein solubility or no dependency. A-domain core motif A6 (blue) and A7 (pink). Any conserved residues predicted to play a role in MbtH dependence is highlighted in yellow. Previously, only four residues have been highlighted, but here it can be seen that at least 10 amino acid positions can be used to predict the reliance on MbtH-like proteins, specifically if a MbtH-like protein will be required or not; the role MbtH-like protein will have is harder to discern.

catalytic orientation. An important caveat is that we cannot conclude the level dependency *ppb*-NRPS enzymes have on Ppb125; Ppb195 and Ppb210 were never purified in an active state, and so it cannot be determined whether Ppb125 alters their activity, although it can be concluded that it enhances their protein production and solubility. It would be very useful to determine an amino acid sequence component of this relationship. It has been established that only specific residues can accommodate the distinctive stacked tryptophan residues of a MbtH-like protein: both alanine and proline at the 817 position are compatible. This region forms part of the A6 motif, and all *ppb* A-domains have this present, supporting a MbtH-like protein dependence. However, some A-domains have this ability to potentially accommodate an MbtH-protein, yet show no MbtH binding or enhancement (Felnagle et al., 2010). A preliminary bioinformatic search was conducted to identify other regions of similarity between A-domains identified to have a MbtH-like protein dependence, that was distinct from those which do not require MbtH (Fig. 63). It was shown that the region surrounding the A6 domain could be used to predict whether an A-domains activity was enhanced via the requirement of MbtH for acyl-adenylate formation, solubility or for an unknown purpose. Overall, all A-domains which require MbtH interaction have a highly conserved region consisting of Pro_{817}, Thr_{820}, Ala_{821} and Arg_{823} which is not present in MbtH-independent A-domains (Fig. 63). A-domains which require MbtH-proteins for solubility had a non-polar hydrophobic residue at the 819 position, - typically Leu_{819}, and when required for acyl-adenylate formation activity this was more likely to be polar. If this is correct, as Ppb210 has the non-polar Leu_{819} residue, issues encountered regarding its solubility may have been MbtH-related but issues concerning activity, theoretically, were not. Similarly, Ppb120 and Ppb195 have polar residues at this position and so it can be concluded that they require Ppb125 for activity as well as solubility. Miller et al., (2016) demonstrated that Pro_{817} was essential in determining the dependency on an MbtH-protein. Here, we extend that to include a selection of residues which span the region between A6 and A7 core motifs. It is apparent that this area is vital in the binding of MbtH proteins, and further work is required to determine what mechanism is responsible for this.

Importance of linker regions

Many issues were encountered regarding protein solubility and activity which were resolved by the incorporation of the entire NRPS protein, the A-domain complete with sub-domains and linker regions. Previous similar work (Juguet et al., 2009; Al-Mestarihi et al., 2014; Al-Mestarihi et al., 2015) looking into A-domain substrate activity and biosynthetic route elucidation, did not highlight any issues surrounding solubility with respect to linker regions - all were described as being cloned individually as stand-alone A-domains. It is difficult to determine whether issues were not encountered, or just simply not reported; either, the importance of linker regions is becoming increasingly apparent. It has been shown that while some mutations between neighboring units are well tolerated (Doekel et al., 2008), there is a progressively large amount of work suggesting they contain areas vital for activity (Yu et al., 2013; Beer et al., 2014). It was shown that Ppb210 was completely insoluble and inactive when not cloned as the whole native protein; potential explanation for this result comes from previous work on Acyl-CoA synthases which demonstrates the restriction of subdomain rotation when mutations were introduced into an A-T linker region (Reger et al., 2007; Wu et al., 2009). A-domain activity was abolished due to this point mutation in the hinge region leading to the enzyme being frozen in a singular configuration. Other work found that there is a proline-rich region following the A10 motif found only in A-T domains which, if altered, leads to diminished compound production (Miller et al., 2014). It can be hence concluded that these linker regions are required to form a specific conformation for activity which is partially controlled by these proline residues. This explains issues encountered in Ppb210 purification: if the protein is not in its natural configuration, it is less likely to be soluble. And also why issues weren't encountered for other A-domain - Ppb220 for example, which is a stand-alone A-domain. This work into the importance of linker regions is significant for future similar A-domain activity studies, but equally for potential combinatorial biosynthesis endeavors.

3.4.10 Malachite vs. Phosphate assays

The modular nature of NRPS systems offers an attractive opportunity for a combinatorial chemistry approach to novel compound production; to utilize this, the substrate pool of adenylation domains must be established in a high throughput and accurate way. As interest in the pharmaceutical potential of NRPS systems, ways to

identify A-domain substrate profiles has increased. The traditional cumbersome radioactive assays have been replaced by cheaper, quicker and simpler alternatives. It is unclear whether these replacement assays have a similar sensitivity and accuracy. This study describes the comparison between two of these such assays: the radioactive phosphate assay and the new malachite green assay; these assess A-domain activity by measuring phosphodiester bond cleavage and rate of release of PP_i, respectively. The creators of the latter (McQuade et al., 2008) show that the enzyme kinetics of both are comparable, and similar to that found in the literature; however, this was not supported by work done here. Not only were the two substrate profiles of Ppb120 very different, the kinetics – e.g. K_m values - for both assays were very dissimilar. Both assay types had strengths and weaknesses. The radioactive phosphate assay represents the classical method for detecting A-domain activity, although higher in complexity of set up, is very sensitive and has a low level of background noise. It is restricted to measuring the formation of isotopically labelled ATP from aminoacyl-AMP reacted with an excess of [^{32}P] PP_i which means it is extremely substrate specific. It is because of this that it has been effectively used for many A-domain activity investigations (Garneau et al., 2005; Pfleger et al.,2007; Singh et al., 2007). One issue encountered, apart from the obvious restrictions dictated by the use of a radioactive substance, was the occurrence of false-positive results. This issue has been previously highlighted by McQuade et al., (2008), and arises due to inefficient washing steps involved in experimental procedure. However, undertaking experiments in triplicate – as done here - which is extremely time consuming and expensive, can rectify this issue. The malachite green test alleviates many of the hurdles presented by the [^{32}P] PP_i assay; it is cheaper and less complex, but a sacrifice is made regarding sensitivity. As seen in the results, the background level of activation is very high – despite rigorous attempts to lower outside phosphate contamination; this results in any, even high activity being diminished and low activity being completely eliminated. Despite both pyrrole compounds being correctly identified as substrates, their level of activity in comparison to other tested compounds is much reduced, in comparison to that seen in the [^{32}P] PP_i assay. From the results shown here, a combination to exploit each assays strengths would be suggested – using the simple malachite green assay to determine broad ranges of potential substrates followed by the [^{32}P] PP_i to determine specific enzyme kinetics. This would be of particular use when testing the

promiscuity of synthetically altered or relaxed A-domain enzymes when employing a combinatorial chemistry approach to novel compound discovery.

A combination approach, using both assay types would be the ideal solution to the issues discussed regarding each individual assay type; however, some research groups simply do not have the facilities or funding to incorporate radioactive work and hence, a radiation laboratory into their work-flow, irrespective of any methods to lower its use. Ideally, a method would be devised which has the efficacy of the radioactive phosphate exchange assay, with the cost, ease and safety of the malachite green assay. This issue has obviously been encountered by previous groups working on adenylation domain activity elucidation as other techniques, for example the large molecule mass spectrometry method (LM-MS), have been implemented (Dorrestein et al., 2006). This approach uses the mass change exhibited when the covalent bond is formed between the A-domain and the substrate. Major advantages of this include the ability to screen multiple potential substrates in parallel, as the activated substrate can be determined by the specific mass change displayed; and, its utilization in determining the order and timings of tailoring reactions that may also take place. It has been shown to have a similar level of accuracy as the radioactive ^{32}P assay (Dorrestein et al., 2006); however, similarly, it requires some extremely costly equipment, making it a not acceptable approach for researchers without easy access to a high resolution Mass Spectrometer.

As it stands currently, A-domain activity studies and hence, NRPS cluster elucidation and novel compound production is bottle-necked by the lack of easily accessible activity assays resulting in their utilization solely in large, well-funded and well-equipped labs. To try to ease this bias I have designed an enzyme assay, similar to that of the malachite green assay, but precluding any issues surrounding phosphate contamination. This approach – termed the Luciferase Assay, and outlined in Figure 64 – exploits the luciferase enzymes ability to utilize ATP to yield light; this will not be the first time this enzyme has been used as a molecular biology tool due its bioluminescent properties, but the first in this A-domain activity application. The assay will work in theory, in a similar way to the phosphate exchange as it monitors the use of ATP during activation of the substrate. However, for the luciferase assay, it will be the chemical conversion and hence the depletion,

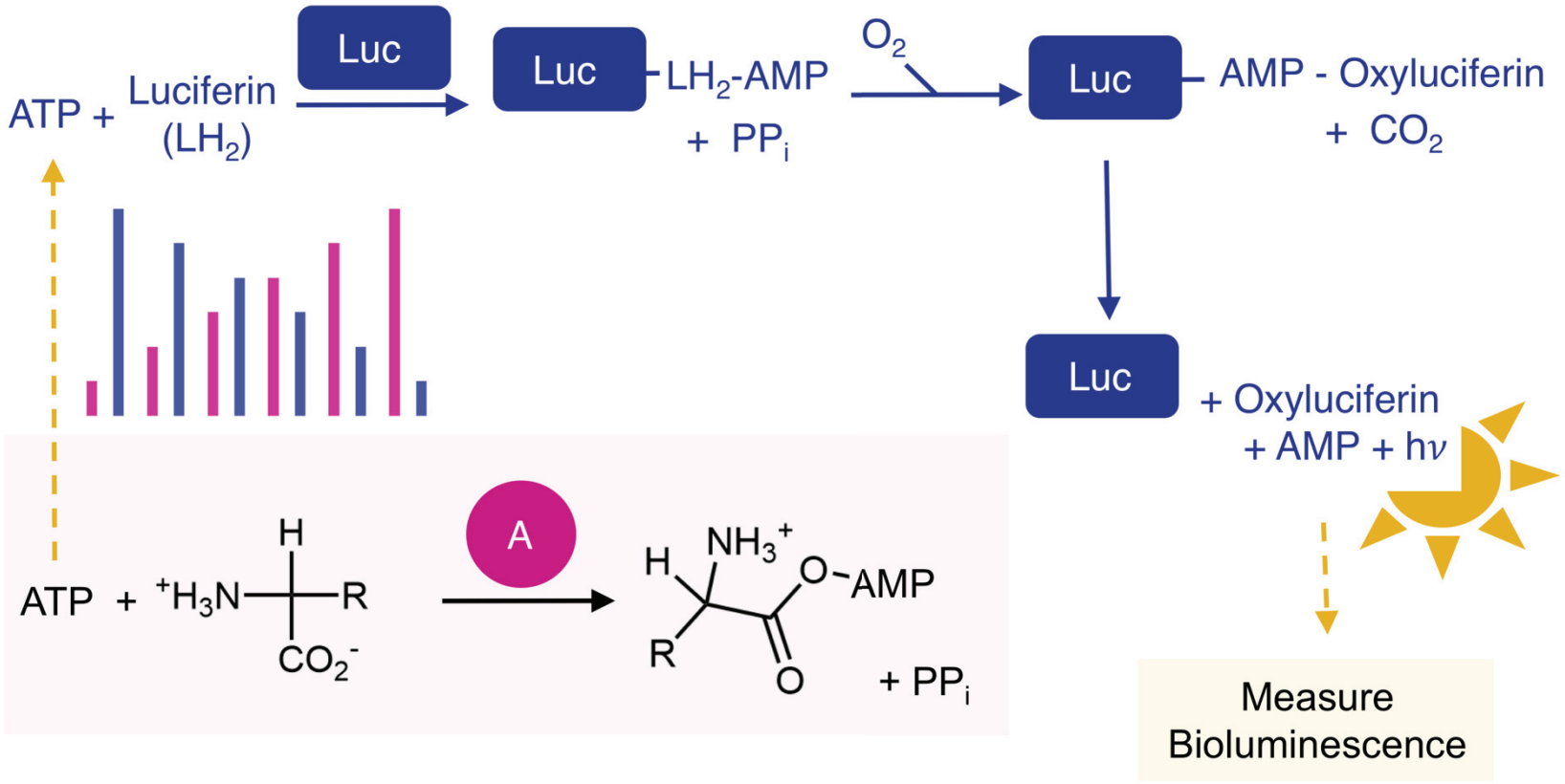


Figure 64. Scheme of proposed luciferase assay for adenylation domain activity. The normal luciferase reaction is highlighted in blue, showing luciferase (Luc) catalysing the reaction of luciferin and ATP to yield light via a two-step reaction. Highlighted in pink is the NRPS adenylation domain activity. The graph illustrates when the NRPS adenylation reaction is in progress (i.e. the compound is a substrate for the A-domain), there will be less ATP to power the luciferase reaction, and hence less bioluminescence measured.

of ATP during adenylation activity that will be measured. A correlation will be first established between bioluminescence and ATP concentration, and then when the light produced of the final reaction is measured, we could determine whether ATP concentration has been decreased, in comparison to the known concentration in the initial reaction. Any lowering of bioluminescence in comparison to the start – and hence lowering of ATP concentration – will confirm adenylation domain activity. The luciferase enzyme assay, in stark contrast to malachite green, will not have an issue with contamination and background noise: there is virtually no non-specific light production induced in the cellular environment and so we can be certain any exhibited will be a result of luciferase activity on ATP. One potential encounterable issues would be that luciferase enzymes themselves are a member of the ANL superfamily – and hence, have the ability to adenylate luciferin in the presence of ATP. This may present an issue if any compounds being tested have a similar structure as luciferin, and hence, have the ability to be substrates to both the A-domain being tested and luciferase. This can be overcome either by (a) avoiding the use of this method when any molecules of similar chemistry to luciferin is being tested, (b) ensuring the activity of the substrate against luciferase is nil before starting, or preferably (c) design a luciferase enzyme with the adenylation activity promiscuity restricted resulting in a very high specificity. The former two schemes are less desirable as they require excluding molecules and additional assay runs, respectively, both of which would detrimentally reduce the applicability of this method to A-domain-substrate assays; hence, the concept of producing a luciferase with *only* activity for luciferin is superior. Branchini et al., (2015) has demonstrated the ease and tolerance of *Photinus pyralis* luciferase to genetic manipulation, and so this approach is likely to be the best potential route to assay application. I propose that *P. pyralis* luciferase is the initial enzyme tested for this application, in addition to its bioluminescent-conferring regions being well described and so giving already set boundaries for future manipulations to reduce promiscuity, it has been shown to maintain a very highly specific dependence on the luciferase catalyzed light emission process on ATP. Furthermore, engineered protein lines of this luciferase with two-fold enhance specific activity and 1.4-fold greater bioluminescence have already been reported (Branchini et al., 2014). The

luciferase assay described here, presents an exciting opportunity for determining, and further exploring the governing mechanics of A-domain substrate specificity. We currently sit on the edge of an era in which the identification of compounds with novel antimicrobial and antifungal targets is going to be of overwhelming importance; this cheap and easily implementable assay format is hence especially timely, when considering the accelerated rate at which assembly lines that defy previous classification – such as Ppb described here – are being discovered.

3.5. References

Al-Mestarihi, A.H., Villamizar, G., Fernández, J., Zolova, O.E., Lombó, F. and Garneau-Tsodikova, S., 2014. Adenylation and S-methylation of cysteine by the bifunctional enzyme TioN in thiocoraline biosynthesis. *Journal of the American Chemical Society*, *136*(49), pp.17350-17354.

Al-Mestarihi, A.H., Garzan, A., Kim, J.M. and Garneau-Tsodikova, S., 2015. Enzymatic evidence for a revised congocidine biosynthetic pathway. *ChemBioChem*, *16*(9), pp.1307-1313.

Aron, Z.D., Dorrestein, P.C., Blackhall, J.R., Kelleher, N.L. and Walsh, C.T., 2005. Characterization of a new tailoring domain in polyketide biogenesis: the amine transferase domain of MycA in the mycosubtilin gene cluster. *Journal of the American Chemical Society*, *127*(43), pp.14986-14987.

Babbitt, P.C., Kenyon, G.L., Martin, B.M., Charest, H., Slyvestre, M., Scholten, J.D., Chang, K.H., Liang, P.H. and Dunaway-Mariano, D., 1992. Ancestry of the 4-chlorobenzoate dehalogenase: analysis of amino acid sequence identities among families of acyl: adenyl ligases, enoyl-CoA hydratases/isomerases, and acyl-CoA thioesterases. *Biochemistry*, *31*(24), pp.5594-5604.

Beer, R., Herbst, K., Ignatiadis, N., Kats, I., Adlung, L., Meyer, H., Niopek, D., Christiansen, T., Georgi, F., Kurzawa, N. and Meichsner, J., 2014. Creating functional engineered variants of the single-module non-ribosomal peptide synthetase IndC by T domain exchange. *Molecular BioSystems*, *10*(7), pp.1709-1718.

Boll, B., Taubitz, T., and Heide, L. (2011) Role of MbtH-like proteins in the adenylation of tyrosine during aminocoumarin and vancomycin biosynthesis. *J. Biol. Chem.* **286**, 36281-36290

Branchini, B.R., Southworth, T.L., Fontaine, D.M., Davis, A.L., Behney, C.E. and Murtiashaw, M.H., 2014. A Photinus pyralis and Luciola italica chimeric firefly luciferase produces enhanced bioluminescence. *Biochemistry*, *53*(40), pp.6287-6289.

Branchini, B.R., Southworth, T.L., Fontaine, D.M., Kohrt, D., Talukder, M., Michelini, E., Cevenini, L., Roda, A. and Grossel, M.J., 2015. An enhanced chimeric firefly luciferase-inspired enzyme for ATP detection and bioluminescence reporter and imaging applications. *Analytical biochemistry*, *484*, pp.148-153.

Buchko, G.W., Kim, C.Y., Terwilliger, T.C. and Myler, P.J., 2010. Solution structure of Rv2377c-founding member of the MbtH-like protein family. *Tuberculosis*, *90*(4), pp.245-251 Challis, G.L., Ravel, J. and Townsend, C.A., 2000. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chemistry & biology*, *7*(3), pp.211-224.

Conti, E., Stachelhaus, T., Marahiel, M.A. and Brick, P., 1997. Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. *The EMBO journal*, *16*(14), pp.4174-4183.

De Crécy-Lagard, V., Marliere, P. and Saurin, W., 1995. Multienzymatic non ribosomal peptide biosynthesis: identification of the functional domains catalysing peptide elongation and epimerisation. *Comptes rendus de l'Academie des sciences. Serie III, Sciences de la vie*, *318*(9), pp.927-936.

Doekel, S., Coeffet-Le Gal, M.F., Gu, J.Q., Chu, M., Baltz, R.H. and Brian, P., 2008. Non-ribosomal peptide synthetase module fusions to produce derivatives of daptomycin in Streptomyces roseosporus. *Microbiology*, *154*(9), pp.2872-2880.

Dorrestein, P.C., Blackhall, J., Straight, P.D., Fischbach, M.A., Garneau-Tsodikova, S., Edwards, D.J., McLaughlin, S., Lin, M., Gerwick, W.H., Kolter, R. and Walsh, C.T., 2006. Activity screening of carrier domains within nonribosomal peptide synthetases using complex substrate mixtures and large molecule mass spectrometry. *Biochemistry*, *45*(6), pp.1537-1546.

Drake, E.J., Cao, J., Qu, J., Shah, M.B., Straubinger, R.M. and Gulick, A.M., 2007. The 1.8 Å crystal structure of PA2412, an MbtH-like protein from the pyoverdine cluster of Pseudomonas aeruginosa. *Journal of Biological Chemistry*, *282*(28), pp.20425-20434.

Felnagle, E.A., Barkei, J.J., Park, H., Podevels, A.M., McMahon, M.D., Drott, D.W. and Thomas, M.G., 2010. MbtH-like proteins as integral components of bacterial nonribosomal peptide synthetases. *Biochemistry*, *49*(41), pp.8815-8817.

Finlay, A.C., Hochstein, F.A., Sobin, B.A. and Murphy, F.X., 1951. Netropsin, a new antibiotic produced by a Streptomyces. *Journal of the American Chemical Society*, *73*(1), pp.341-343.

Fischbach, M.A. and Walsh, C.T., 2006. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. *Chemical reviews*, *106*(8), pp.3468-3496.

Garneau, S., Dorrestein, P.C., Kelleher, N.L. and Walsh, C.T., 2005. Characterization of the formation of the pyrrole moiety during clorobiocin and coumermycin A1 biosynthesis. *Biochemistry*, *44*(8), pp.2770-2780.

Gehring, A.M., Mori, I. and Walsh, C.T., 1998. Reconstitution and characterization of the Escherichia coli enterobactin synthetase from EntB, EntE, and EntF. *Biochemistry*, *37*(8), pp.2648-2659.
Gulick, A.M., 2009. Conformational dynamics in the Acyl-CoA synthetases, adenylation domains of non-ribosomal peptide synthetases, and firefly luciferase. *ACS chemical biology*, *4*(10), pp.811-827.

Heemstra Jr, J.R., Walsh, C.T. and Sattely, E.S., 2009. Enzymatic tailoring of ornithine in the biosynthesis of the Rhizobium cyclic trihydroxamate siderophore vicibactin. *Journal of the American Chemical Society*, *131*(42), pp.15317-15329.

Hisanaga, Y., Ago, H., Nakagawa, N., Hamada, K., Ida, K., Yamamoto, M., Hori, T., Arii, Y., Sugahara, M., Kuramitsu, S. and Yokoyama, S., 2004. Structural basis of the substrate-specific two-step catalysis of long chain fatty acyl-CoA synthetase dimer. *Journal of Biological Chemistry*, *279*(30), pp.31717-31726.

Juguet, M., Lautru, S., Francou, F.X., Nezbedová, Š., Leblond, P., Gondry, M. and Pernodet, J.L., 2009. An iterative nonribosomal peptide synthetase assembles the pyrrole-amide antibiotic congocidine in Streptomyces ambofaciens. *Chemistry & biology*, *16*(4), pp.421-431.

Kopp, F. and Marahiel, M.A., 2007. Where chemistry meets biology: the chemoenzymatic synthesis of nonribosomal peptides and polyketides. *Current opinion in biotechnology*, *18*(6), pp.513-520.

Lautru, S., Deeth, R.J., Bailey, L.M. and Challis, G.L., 2005. Discovery of a new peptide natural product by Streptomyces coelicolor genome mining. *Nature chemical biology*, *1*(5), p.265.

Lautru, S., Oves-Costales, D., Pernodet, J.L. and Challis, G.L., 2007. MbtH-like protein-mediated cross-talk between non-ribosomal peptide antibiotic and siderophore biosynthetic pathways in Streptomyces coelicolor M145. *Microbiology*, *153*(5), pp.1405-1412.

Lipmann, F., 1944. Enzymatic synthesis of acetyl phosphate. *Journal of Biological Chemistry*, *155*(1), pp.55-70.

Marahiel, M.A., Stachelhaus, T. and Mootz, H.D., 1997. Modular peptide synthetases involved in nonribosomal peptide synthesis. *Chemical reviews*, *97*(7), pp.2651-2674.

Mareš, J., Hájek, J., Urajová, P., Kopecký, J. and Hrouzek, P., 2014. A hybrid non-ribosomal peptide/polyketide synthetase containing fatty-acyl ligase (FAAL) synthesizes the β-amino fatty acid lipopeptides puwainaphycins in the Cyanobacterium Cylindrospermum alatosporum. *PloS one*, *9*(11), p.e111904.

McElroy W., DeLuca M. and Travis J., 1967. Molecular uniformity in biological catalyses. The enzymes concerned with firefly luciferin, amino acid, and fatty acid utilization are compared. Science. 157:150–160.

McQuade, T.J., Shallop, A.D., Sheoran, A., DelProposto, J.E., Tsodikov, O.V. and Garneau-Tsodikova, S., 2009. A nonradioactive high-throughput assay for screening and characterization of adenylation domains for nonribosomal peptide combinatorial biosynthesis. *Analytical biochemistry*, *386*(2), pp.244-250.

Miao, V., Coeffet-LeGal, M.F., Brian, P., Brost, R., Penn, J., Whiting, A., Martin, S., Ford, R., Parr, I., Bouchard, M. and Silva, C.J., 2005. Daptomycin biosynthesis in Streptomyces roseosporus: cloning and analysis of the gene cluster and revision of peptide stereochemistry. *Microbiology*, *151*(5), pp.1507-1523.

Miao, V., Brost, R., Chapple, J., She, K., Coëffet-Le Gal, M.F. and Baltz, R.H., 2006. The lipopeptide antibiotic A54145 biosynthetic gene cluster from Streptomyces fradiae. *Journal of Industrial Microbiology and Biotechnology*, *33*(2), pp.129-140.

Miller, B.R., Sundlov, J.A., Drake, E.J., Makin, T.A. and Gulick, A.M., 2014. Analysis of the linker region joining the adenylation and carrier protein domains of the modular nonribosomal peptide synthetases. *Proteins: Structure, Function, and Bioinformatics*, *82*(10), pp.2691-2702.

Miller, B.R., Drake, E.J., Shi, C., Aldrich, C.C. and Gulick, A.M., 2016. Structures of a nonribosomal peptide synthetase module bound to MbtH-like proteins support a highly dynamic domain architecture. *Journal of Biological Chemistry*, *291*(43), pp.22559-22571.

Otten, L.G., Schaffer, M.L., Villiers, B.R., Stachelhaus, T. and Hollfelder, F., 2007. An optimized ATP/PPi-exchange assay in 96-well format for screening of adenylation domains for applications in combinatorial biosynthesis. *Biotechnology journal*, *2*(2), pp.232-240.

Pfleger, B.F., Lee, J.Y., Somu, R.V., Aldrich, C.C., Hanna, P.C. and Sherman, D.H., 2007. Characterization and analysis of early enzymes for petrobactin biosynthesis in Bacillus anthracis. *Biochemistry*, *46*(13), pp.4147-4157.

Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W. and Huson, D.H., 2005. Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic acids research*, *33*(18), pp.5799-5808.

Reger, A.S., Carney, J.M. and Gulick, A.M., 2007. Biochemical and crystallographic analysis of substrate binding and conformational changes in acetyl-CoA synthetase. *Biochemistry*, *46*(22), pp.6536-6546.

Saito, M., Hori, K., Kurotsu, T., Kanda, M. and Saito, Y., 1995. Three conserved glycine residues in valine activation of gramicidin S synthetase 2 from Bacillus brevis. *The Journal of Biochemistry*, *117*(2), pp.276-282.

Santi, D.V., Webster, R.W. and Cleland, W.W., 1974. [49] Kinetics of aminoacyl-tRNA synthetases catalyzed ATP-PP i exchange. *Methods in enzymology*, *29*, pp.620-627.

Singh, G.M., Vaillancourt, F.H., Yin, J. and Walsh, C.T., 2007. Characterization of SyrC, an aminoacyltransferase shuttling threonyl and chlorothreonyl residues in the syringomycin biosynthetic assembly line. *Chemistry & biology*, *14*(1), pp.31-40.

Stachelhaus, T., Mootz, H. and Marahiel, M. (1999). The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chemistry & biology*: *6*(8), pp.493-505.
Stegmann, E., Rausch, C., Stockert, S., Burkert, D. and Wohlleben, W., 2006. The small MbtH-like protein encoded by an internal gene of the balhimycin biosynthetic gene cluster is not required for glycopeptide production. *FEMS microbiology letters*, *262*(1), pp.85-92.

Sun, X., Li, H., Alfermann, J., Mootz, H.D. and Yang, H., 2014. Kinetics profiling of gramicidin S synthetase A, a member of nonribosomal peptide synthetases. *Biochemistry*, *53*(50), pp.7983-7989.

Weber, T. and Marahiel, M.A., 2001. Exploring the domain structure of modular nonribosomal peptide synthetases. *Structure*, *9*(1), pp.R3-R9.

Wu, R., Reger, A.S., Lu, X., Gulick, A.M. and Dunaway-Mariano, D., 2009. The mechanism of domain alternation in the acyl-adenylate forming ligase superfamily member 4-chlorobenzoate: coenzyme A ligase. *Biochemistry*, *48*(19), pp.4115-4125.

Yonus, H., Neumann, P., Zimmermann, S., May, J.J., Marahiel, M.A. and Stubbs, M.T., 2008. Crystal Structure of DltA IMPLICATIONS FOR THE REACTION MECHANISM OF

NON-RIBOSOMAL PEPTIDE SYNTHETASE ADENYLATION DOMAINS. *Journal of Biological Chemistry*, *283*(47), pp.32484-32491.

Yu, D., Xu, F., Gage, D. and Zhan, J., 2013. Functional dissection and module swapping of fungal cyclooligomer depsipeptide synthetases. *Chemical Communications*, *49*(55), pp.6176-6178.

Zhang, W., Heemstra Jr, J.R., Walsh, C.T. and Imker, H.J., 2010. Activation of the pacidamycin PacL adenylation domain by MbtH-like proteins. *Biochemistry*, *49*(46), pp.9946-9947.

Chapter 4. CRISPR/Cas gene editing in *Verrucosispora* spp.

4.1 Introduction

Proximicins represent a family of both medically and industrially relevant secondary metabolites (Fiedler et al., 2008). The intriguing structure, action and biological activity of these compounds make this small class of natural products an exciting starting point for a combinatorial biosynthesis strategy to novel molecule development. So far, the putative biosynthetic gene cluster responsible for proximicin production has been identified and putatively characterised in the two producers: *V. maris* AB18-032 *and V. sp.* str. MG37; the cluster harbours a NRPS system with a number of additional genes related to post-tailoring, regulation and resistance. Extensive analysis of the adenylation domains present led to the formation of two competing routes for the production of the novel chemistry characteristic of the proximicin family. With the aims of fully understanding the proximicin pathway, providing insights into the formation of the 2,4-disubstituted furan group, and allowing the generation of novel proximicin derivatives, the *ppb* cluster needs to be investigated further.

4.1.1 Secondary Metabolite Cluster analysis in Actinomycetes

Actinomycetes represent a well-studied order of bacteria, known to harbour the capacity to produce many biologically significant compounds, including many well-known herbicides, chemotherapeutics and immunosuppressants such as bialaphos, doxorubicin and rapamycin, respectively (Berdy, 2005; Berdy, 2012; Hwang et al., 2014). It is also estimated that they are responsible for nearly two-thirds of the natural antimicrobial drug compounds currently in application (Bentley et al., 2002). The recent advent of NGS technologies leading to modern genome mining studies (Blin et al., 2014) indicate that they harbour a huge as yet unexploited potential to produce secondary metabolites with novel structures (Weber et al., 2015); this yielded the recent renaissance in the investigation of Actinomycetes. The ultimate aim of these research efforts is typically two-fold: (i) to metabolically engineer strains to express novel biosynthetic pathways in high enough concentrations for

identification, and (ii) to delete genes to confirm the clusters responsible for biosynthesis, to aid in the large-scale production of medically important compounds (Lee et al., 2009). Both of these approaches share a common challenge: the necessity to genetically manipulate strains; this is much more difficult in Actinomycetes in comparison to model organisms such as *E. coli* or *Saccharomyces cerevisiae.* It is typically associated with the characteristically high GC content, in addition to diverse genomic contents present in Actinomycete genomes, but also inpart to their complex growth requirements. Traditional genetic manipulation attempts in these organisms typically used RecA mediated double crossover events (Kieser et al., 2000); this approach exploits the complex enzymatic machinery responsible for DNA repair in bacteria, at the core of which is the DNA recombinase protein RecA. The complementary DNA to the gene to be disrupted is provided by a non-replicative or temperature sensitive plasmid, and the homologous recombination event results in specific gene interference (Fig. 65a). The efficacy of this approach has increased rapidly in the past decade, by using templates with long homology regions, such as the ReDirect protocol developed by Gust et al., (2004) which modified the highly efficient Red-mediated genetic manipulation approach popular in *E. coli* studies to be utilised in *Streptomyces* spp., using 39 bp homology arms to result in high precision gene disruptions (Fig. 65b). Further, the meganuclease – I-SceI - from *S. cerevisiae* mitochondria has been codon optimised for use in Actinomycetes by two groups, leading to unmarked targeted deletions in *Streptomyces* spp. Tu6071 (Siegl et al., 2010) and *S. coelicolor* (Fernandez-Martinez et al., 2014) by RecA-mediated homologous recombination to repair the double-strand breaks catalysed by I-SceI (Fig. 65c). Other approaches, such as the exploitation of large serine recombinases involved in site-specific integration of actinophages into Actinomycete chromosomes, have been intensively investigated (Zhang et al., 2010; Zhang et al., 2013; Myronovsky et al., 2014; Du et al., 2015); however, they all share a collective obstacle: the protocols are time consuming and labour intensive (Cobb et al., 2014). For instance, it took six years to construct an *S. coelicolor* strain deficient in subteleomeric DNA and PKS/NRPS gene clusters (Zhou et al., 2012). To fully exploit the potential harboured within the

genomes of Actinomycetes for the discovery of biologically active compounds, efficient genome manipulation techniques must be developed.

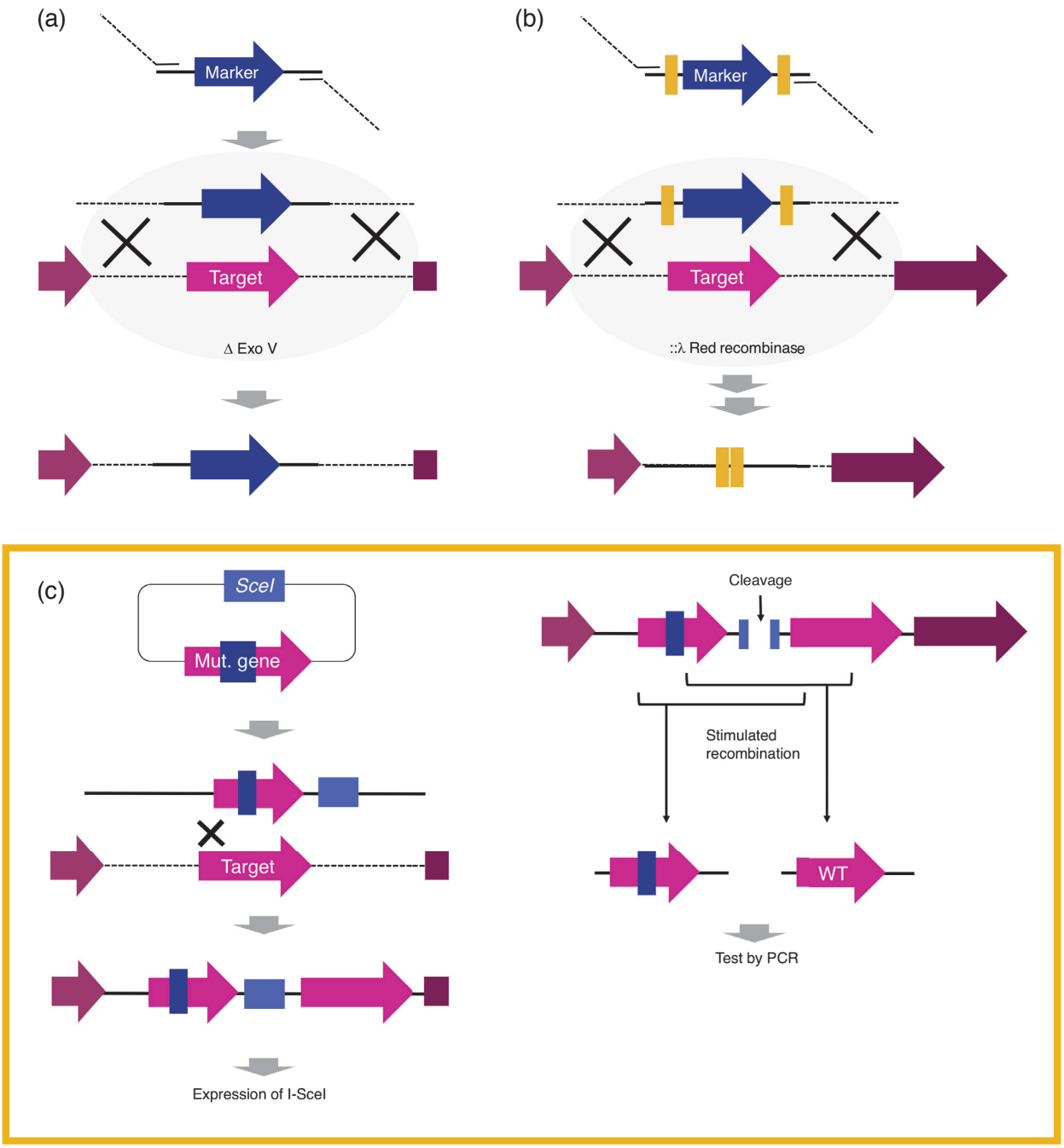


Figure 65. Previous approaches to bacterial gene editing, and specifically Actinomycete gene editing. (a) RecA mediated double cross over event. **(b)** Red-mediated genetic manipulation approach. **(c, outlined in yellow)** Utilisation of meganuclease SceI to produce double-strand breaks leading to RecA-mediated repair.

4.1.2 Introduction to CRISPR/Cas systems

Recent advances in the understanding of clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated proteins (Cas) systems have revolutionised biological and biotechnological research. These prokaryote-derived systems – specifically CRISPR/Cas9 – have been rapidly adopted for genome editing in eukaryotes (Cho et al., 2013; Hruscha et al., 2013; Ren et al., 2013, and Shan et al., 2013), facilitating some of the most important advances in science of the past decade (Fig. 66). The latest applications of CRISPR/Cas9 are ranging from treating genetic diseases (Bassuk et al., 2016; Nelson et al., 2016), to the introduction of beneficial traits into crops, while circumventing strict GMO regulations (Waltz, 2016). In addition to these, and arguably the most remarkable, is the approval of the first in-human CRISPR trial to modulate patient immune cells for cancer research treatment (Cyranoski, 2016; Kaiser et al., 2016). Just as Sanger et al., (1977) bore the genomic revolution with the advent of DNA sequencing, the realisation of the potential held within the CRISPR/Cas9 system (Ran et al.,2013), has unlocked the genome editing era.

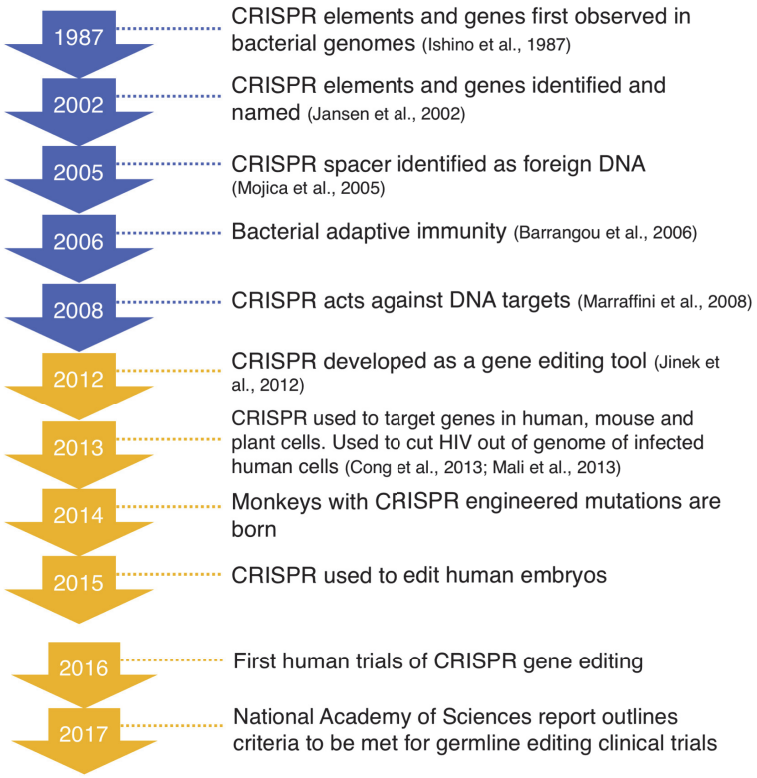


Figure 66. History of CRISPR/Cas achievements and landmark studies.

While the potential of CRISPR/Cas as a genome editing tool was not realised until recently, the knowledge of CRISPR systems has been building for decades. The initial discovery of a genetic element constituting a CRISPR locus by Ishino et al., (1987) began the 30-year progression of understanding, culminating in its applications today. Initial bioinformatical studies of CRISPR/Cas9 systems unveiled their function in the prokaryotic adaptive immune system (Mojica et al., 2000; Jansen et al., 2002; Pourcel et al., 2005). This led to a huge paradigm shift as it was previously agreed that prokaryotes had the ability for only innate immune systems, comprised of restriction modification systems and exonucleases. In contrast to the adaptive immunity exhibited in jawed vertebrates, as unicellular organisms, prokaryotes must store the whole immunological information in a single cell. This complex ability was resolved with the discovery of CRISPR loci, in which the immunological memories are engraved by the CRISPR/cas9 system; these loci consist of short (typically 21-40 bp) palindromic repeats, separated by non-repetitive sequences (spacers) (Barrangou et al., 2007; Yosef et al., 2012). Each spacer represents a 'scar' created by, and completely matching to, a heterologous DNA sequence that once invaded the host (Barrangou et al., 2007); this array of 'scars' is bordered by *cas* genes which use the information stored in the spacer regions to conduct adaptive immunity (Jansen et al., 2002). This is done in three phases outlined in Figure 67: adaption, expression and interference.

I. Adaption is characterised by the sampling a stretch of exogenous DNA bases (protospacer), encountered through various events such as viral infection; this is done by a set of Cas proteins, potentially in assistance with other systems. In many systems, a specific protospacer adjacent motif (PAM) is required for protospacer selection; this is an adjacent, stereotyped sequence motif, a few bases in length. This protospacer is then inserted into the host DNA at the CRISPR array loci, which is done is a sequential, methodical order.
II. Expression follows by the expression of the CRISPR array as a single transcript resulting in precursor CRISPR RNA (pre-crRNA) which is then cleaved by Cas and ribonuclease proteins to produce a functional

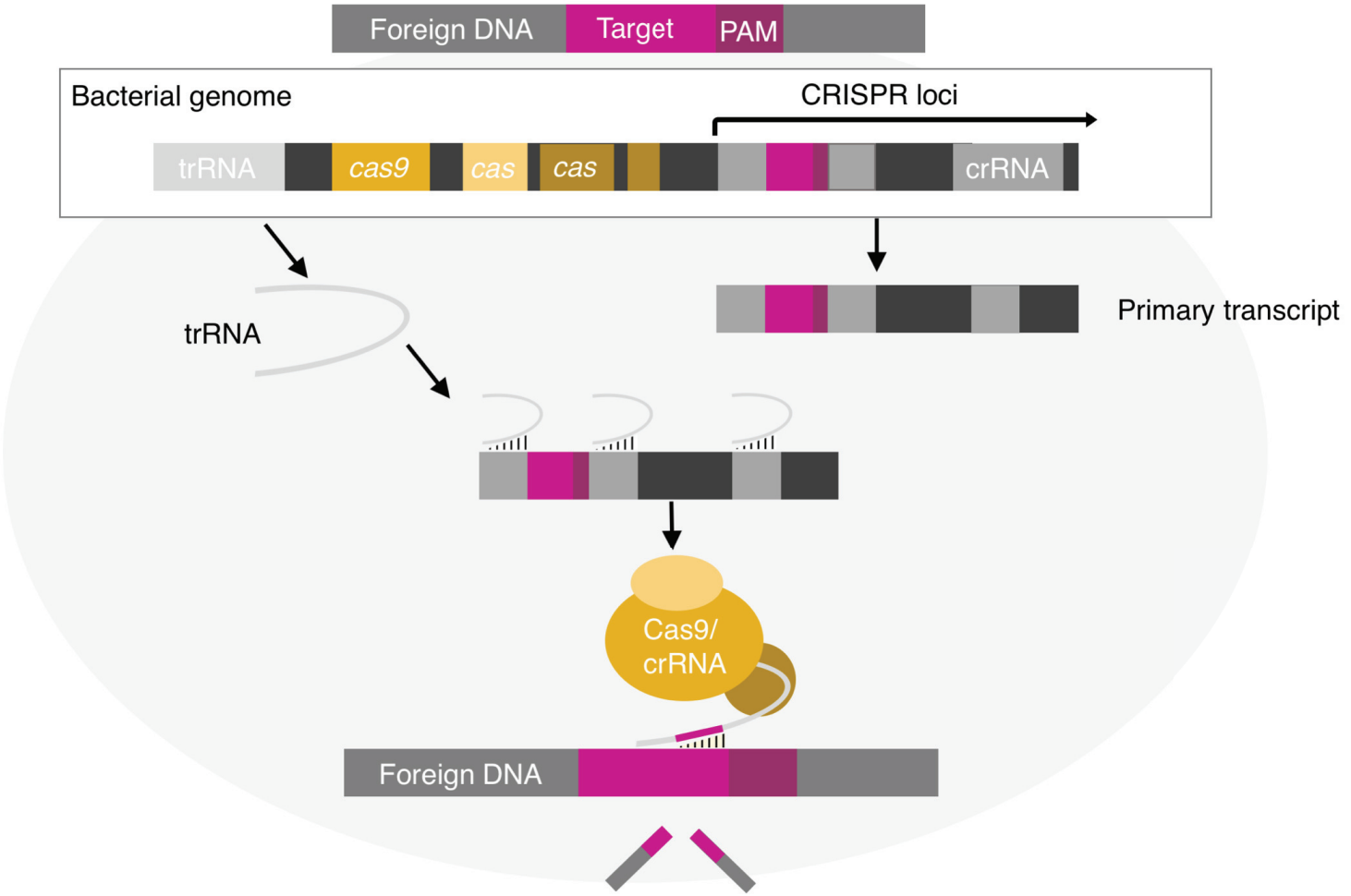


Figure 67. Native CRISPR/Cas acquired resistance. When foreign DNA is encountered a sample with an adjacent PAM sequence is inserted into the CRISPR loci. When it is re-encountered, the CRIPSR array is expressed as a single primary transcript. This is processed by other Cas proteins, which then produces a crRNP complex by binding to Cas9 and other components (yellow) which is then guided to the foreign DNA leading to foreign DNA attack.

III. CRISPR RNA (crRNA) which consists of a sequence corresponding to the single spacer and an adjacent single repeat. The crRNA then binds to other components present in the CRISPR/cas9 system leading to the formation of the effector CRISPR ribonucleoprotein (crRNP) complex.

IV. Interference denotes the guiding of the crRNP complex by the bound crRNA to the target protospacers, and hence the degradation of the exogenous DNA targets. The CRISPR arrays in the host genome remain a scar of a previous attack from exogenous DNA, and hence, allow a faster response to subsequent assaults.

Large sequencing projects have resulted in many bacterial genomes being sequenced (Choi & Lee, 2016); using this information, it has become apparent that CRISPR/cas systems are diverse across the domain. Attempts at classification of these systems has largely been focused on bioinformatical categories (Makarova et al., 2015), which are accumulating and adapting, as the full extent of variance present in CRISPR/cas systems is realised. However, they typically fall into three classes: type I, type II and type II depending on the presence of Cas3, Cas9 and Cas10 respectively (Makarova et al., 2011). Despite CRISPR/cas systems being of prokaryote origin, the bacterial focus of these systems has primarily been in their characterisation and classification, with the aim of their utilisation in eukaryotes. The ‘CRISPR big bang’ in eukaryotes currently overshadows its bacterial counterparts, however, the potential to advance bacterial-based research and industry, is steadily advancing.

Genome editing studies is the primary focus of CRISPR/cas9 application; the first bacterial genome editing in this way was reported in 2013 (Jiang et al., 2013), leading onto many similar reports in following years. Much of this work focuses on the exploitation of type II systems, which are distinct from the other classes in the requirement of a dual RNA complex consisting of crRNA and transactivating RNA (trRNA) (Deltcheva et al., 2011; Jinek et al., 2012). During the final stage of CRISPR/cas9 a cas9:dual RNA complex searches for and binds to a PAM sequence located on the target DNA, while simultaneously inspecting the 5’ region for

complementarity to the protospacer of its crRNA, leading to the initiation of a double stranded break in the DNA (Sternburg et al., 2014). This guided the creation of many CRISPR/cas9 engineering plasmids containing all the enzymatic machinery required for CRISPR/cas9 genome editing, and an insert site for incorporation of sequences matching the genomic locus of interest. Further adaptation of the type II system for genome engineering, led to the fusion of the crRNA and trRNA into a single synthetic guide RNA (sgRNA) transcript, removing the necessity of pre-crRNA processing (Cobb et al., 2015). The application of these technologies toward seamless bacterial genome engineering is advancing rapidly, however, some economically important orders and families remain largely un-investigated.

4.1.3 CRISPR/Cas9 engineering of Actinomycete genomes

As previously discussed, Actinomycetes and *Streptomyces* spp. in particular, represent a group with high industrial and medicinal importance; however, the application of CRISPR/Cas9 technologies towards this bacterial class has not yet been fully determined. Importantly, *Streptomyces* spp. harbour the ability to repair double stranded breaks in DNA via non-homologous end joining (NHEJ), a trait rarely exhibited by bacteria. This means, that unlike other prokaryotic CRISPR/Cas facilitated gene editing approaches, *Streptomyces* spp. can use a system more akin to that of eukaryotic systems. Bacteria without this ability utilise a functionally different system, in which CRISPR/Cas is used as a selective pressure opposed to the primary gene editing apparatus; this maintains much of the difficulties associated with previous approaches, and solutions are discussed in depth in future prospects. To exploit this exclusive attribute of *Streptomyces* spp., Cobb et al., (2015) designed a *Streptomyces* specific CRISPR/Cas9 gene deletion vector system, consisting of two plasmids pCRISPomyces-1 and 2; among their notable features include a codon-optimised *cas9* from *Streptomyces pyogenes*; a *Bbs*I-flanked lacZ cassette to allow Golden Gate assembly of spacer sequences, a temperature sensitive pSG5 origin, and importantly, a *Xba*I site for incorporation of an editing template. The addition of an editing template via recombination-driven repair circumvents issues previously encountered concerning the substantial and unpredictable excision of DNA during CRISPR/Cas9 activity, to instead result in a

pre-determined, precise genome modification. It has been demonstrated by implementation of the pCRISPomyces system, along with similar systems (Tong et al., 2015), has the ability to create site-specific alterations to Actinomycete genomes; however, not all families have been examined. Elucidation of the proximicin biosynthetic cluster, specifically (i) confirming the stage of furan incorporation and (ii) the boundaries of the *ppb* cluster, present an exciting opportunity to test the applicability of CRISPR/cas9 technology in the Micromonosporaceae family, to which *Verrucosispora* spp., belong.

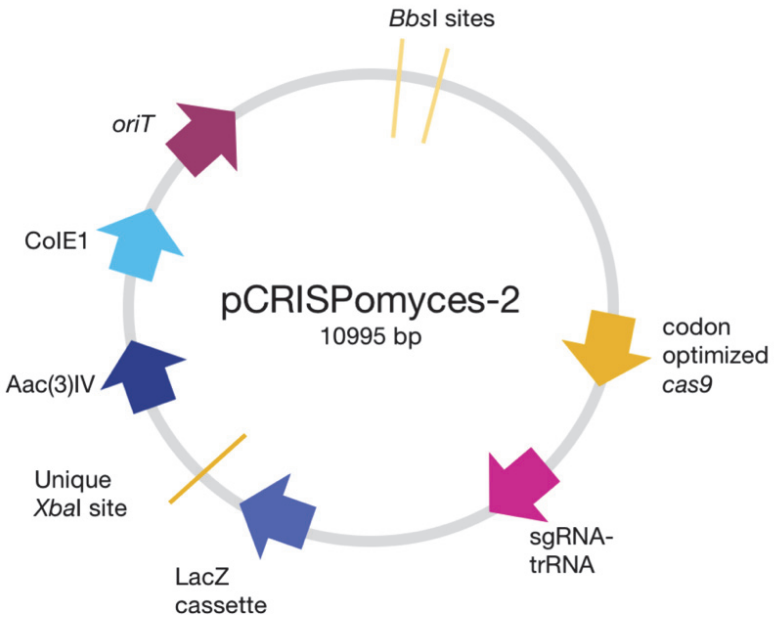


Figure 68. pCRRISPomyces-2 vector. *Streptomyces* specific pCRISPR-2 vector designed by Cobb et al., (2015) for effective and efficient gene editing via CRISPR/Cas9

4.1.4 Proof of Concept study in Verrucosispora spp.

Here we report the first instance of CRISPR/Cas9 gene editing technology being applied to *Verrucosispora sp.* str. MG37 the first member of the Micromonosporaceae family to have its genome modified utilising this approach. The gene responsible for the characteristic orange pigmentation exhibited by *Verrucosispora* spp., was initially targeted resulting in an off-white strain, capable of sporulation. The action of the CRISPR/Cas9 system resulted in a sickly *V. sp.* str. MG37strain, incapable of liquid fermentation; this made the extraction of high quality gDNA for sequencing difficult, and so, successful gene deletion was

confirmed using PCR. Once the applicability of CRISPR/Cas9 technology was established in *V. sp.* str. MG37, preliminary attempts into *ppb* cluster gene deletion were completed; this resulted in an *V. sp.* str. MG37 strain with a large deleted region encompassing *ppb120.*

4.2 Material and Methods

4.2.1 Media and Reagents

2 X YT broth: 16 g Tryptone, 10 g Bacto Yeast Extract, 5 g NaCl, add ~900 mL of diH_2O and adjust to pH. 7.0 with NaOH. Top up to 1 L and autoclave. MS (mannitol Soya Flour agar): 20 g Mannitol; 20g Soya Flour; 100 mM $CaCl_2$ and 13 g Agar, add 1 L of diH_2O and autoclave.

4.2.2 sgRNA design

Targets for CRISPR gene editing were chosen: for the initial studies a phenotypic target was chosen. To determine the genes responsible for the orange pigment exhibited by both *Verrucosispora* spp., BLAST homology search was done using genes well characterised to be responsible for orange pigment in related species. After the proof of concept study, genes identified as being involved in proximicin biosynthesis were to be sequentially deleted. For the scope of this research, only a single *ppb* gene was targeted – *ppb120*.

For both target genes, a 20 nt protospacer was chosen, with priority given to those which fulfil specific guidelines:

- The 3' protospacer adjacent sequence (PAM) must be NGG, where N is any nucleotide.
- Sequences with purines occupying the last four (3') bases of the protospacer.
- Sequences on the non-coding strand.
- Sequences in which the last 12 nt of protospacer + 3 nt PAM are unique in the genome

For each protospacer, two 24 nt oligonucleotides were designed consisting of a 5' 4nt sticky end and 20 nt spacer sequence. The sticky ends being ACGC on the forward and AAAC on the reverse primer.

4.2.3 Oligonucleotides Annealing and Spacer insert into pCRISPomyces-2

Both oligonucleotides (oligos) were re-suspended to 100 μM in water. Oligos were

annealed using the following ligation reaction: forward and reverse oligos (final concentration 5 μM) with HEPES (27 mM, pH7.8) to a final volume of 100 μL. The annealed spacer was the ligated into the pCRISPomyces-2 backbone using Golden Gate assembly as follows: pCRISPomyces-2 (100 ng); annealed spacer (0.5 μM or 10-fold dilution of annealed oligo stock); T4 ligase buffer (NEB, 5X); T4 ligase (NEB, 400U), BbsI (NEB, 10 U) and water added to 20 μL. Mixed and assembled in a thermocycler with the cycle conditions: 9 cycles of 37°C for 10 mins then 16°C for 10 mins; 50°C for 5 mins; 65°C for 20 mins and 4°C forever.

4.2.4 Transformation of pCRISPR-sgRNA into E. coli

3 μL of each annealing reaction was used to transform NEB® 5-alpha competent *E. coli* cells, using the adapted five-minute NEB transformation protocol (cat. no C2987H), with an additional outgrowth step (referred to hence as the Adapted Five-Minute Protocol): NEB® 5-alpha competent *E. coli* cells were allowed to thaw on ice ~5 min, and ~25 ng of DNA was added and mixed with the cells and placed on ice for 2 mins. The mixture was heat shocked at 42°C for exactly 30 seconds and placed on ice for 2 mins. 950 mL of LB media was added and incubated at 37°C for 45 mins. 100% of transformants were plated on LB plates with relevant antibiotic selection – here 30 μg/mL apramycin, overnight at 37°C. Blue/White screening was used to aid in successful transformant identification – 0.05 M IPTG and 0.005 mg of X-Gal in DMSO was spread on the LB/ 50 µg/mL apramycin plates and allowed to dry prior to plating. White colonies were selected and PCR checked for correct insert. 20 mL of sgRNA containing plasmid strains were grown up O/N at 37°C with 200 rpm, cultures were used to prepare high quantities of pure plasmid using the GenElute^{TM} Plasmid Miniprep Kit (SIGMA PLN10), according to the supplier guidelines. The plasmid was then linearized by restriction digest enzymes *Xba*I according to the NEB protocol: Restriction digest mixture (50 µL) contained DNA (1µg), RE buffer (1X NEB CutSmart buffer, cat. no. B7204S), restriction enzyme (10 units, NEB); and incubated at 37°C overnight. and gel purified using GenElute^{TM} Gel Extraction Kit (SIGMA NA1111), according to manufactures protocol, and eluted into 50 μL water.

4.2.5 PCR of repair template

Two 1kb ‘arms’ of genomic region for the repair template were produced by PCR using 30 nt overlaps between the two arms and *Xba*I cut sites at the opposite ends.

PCR was done using to amplify the arms separately: the PCR mixture (50 µL) contained Phusion® High-Fidelity polymerase (0.10 units, NEB), dNTPs (200 µM), Phusion® High Fidelity buffer (1X), primers (0.5 µM each), *Verrucosispora sp.* str. MG37 DNA (50ng) and DMSO (3%). The PCR was performed using a typical thermal cycling procedure – Phusion Thermo-Cycling Conditions - 1 cycle, 98°C for 30 s; 30 cycles, 98°C for 10 s, 71°C for 30 s and 72°C for 3:00 min; and 1 cycle 72°C for 10 min. Annealing temperature was varied depending on T_m of primers used. PCR was checked using gel electrophoresis using the standard technique. PCR fragments of the correct size were gel purified using GenElute™ Gel Extraction Kit, according to Sigma Aldrich guidelines. Fragments were eluted into 50 µL of ddH_2O and quantified using the Qubit 2.0 Fluorometer according to ThermoFisher guidelines.

The two 1Kb arms were then spliced together using overlap-extension PCR. Each PCR was subjected to a temperature regimen similar to the following: initial denaturation at 98°C for 2 min, denaturation at 94°C for 30 s, annealing at 60°C for 30 s, extension at 68°C for 90 s/kb for 30–32 cycles, with a final extension of 68°C for 10 min. PCR was checked using gel electrophoresis using the standard technique. PCR fragments of the correct size – *ppb120* the repair template was ~2.5 Kb - were gel purified using GenElute™ Gel Extraction Kit, according to Sigma Aldrich guidelines. Fragments were eluted into 50 µL of ddH_2O and quantified using the Qubit 2.0 Fluorometer according to ThermoFisher guidelines.

The resultant 2.5 Kb repair template were then digested using *Xba*I to create compatible overlaps with the sgRNA-plasmid for ligation, using the NEB protocol as previously described, and gel purified using GenElute™ Gel Extraction Kit (SIGMA NA1111), according to manufactures protocol, and eluted into 100 *µ*L water.

4.2.6 Ligation of repair template in pCRISPR-sgRNA to produce pCRISPR-sgRNA-RT

The digested pCRISPR-sgRNA plasmid and the digested repair template were ligated using the overhangs created during restriction digest. Ligation mixture (20 µL) contained 10 X T4 ligase buffer (5X, NEB), linearized pCRISPR-sgRNA (50ng), repair template (37.5 ng) and T4 DNA ligase (400 U, NEB). Incubated at 37°C overnight. 3 µL of ligation reaction was used to transform *E. coli* TOP10

(ThermoFisher Scientific) using the Adapted Five-Minute Protocol. Transformants were plated on LB with 30 μg/mL apramycin, kanamycin and chloramphenicol, overnight at 37°C and then checked for the presence of a correct construct using colony PCR using primers located on the pCRISPomyces-2 backbone, as previously described. Strains showing the correct PCR amplicon, were grown up in 10 mL LB media at 37°C with 50 µg/mL apramycin, plasmids were extracted using GenElute^{TM} Plasmid Miniprep Kit (SIGMA PLN10), according to the supplier guidelines, and checked using sequencing. Purified plasmids, as well as plasmids not containing the RT – pCRISPR-sgRNA, were transformed into *E. coli* conjugation strain ET12567_pUZ8002 which was chosen for all conjugation reactions; pCRISPR-sgRNA and pCRISPR-sgRNA-RT was transformed into ET12567_pUZ8002 using the Adapted Five-Minute Protocol and checked using colony PCR for correct insertion of pCRISPR-sgRNA and pCRISPR-sgRNA-RT to create ET-pCRISPR-sgRNA and ET- pCRISPR-sgRNA-RT, respectively.

4.2.7 Conjugation into Verrucosispora sp. str. MG37

10 mL of LB supplemented with kanamycin and chloramphenicol (both 30 μg/mL) was inoculated with and ET-pCRISPR-sgRNA and ET- pCRISPR-sgRNA-RT and grown overnight at 37°C. Cells were diluted 1:100 into 10 mL fresh LB with the same selection, and grown for ~4 hrs at 37°C until the OD_{600}. 0.4. The cells were harvested and washed twice in 10 mL LB and re-suspended in 1 mL LB. ~ 10^8 of *V. sp.* str. MG37 spores were re-suspended in 500 μL of 2 X YT broth; these were heat shocked at 65°C for 10 mins and allowed to cool at RT. 0.5 mL of the *E. coli* suspension and 0.5 mL of the spores were mixed and spun briefly; the supernatant decanted and re-suspended in any residual liquid. 10^{-1} to 10^{-4} dilutions were made in a total of 100 μL of water, and plated on MS agar with 10 mM $MgCl_2$ without any antibiotics and incubated at 30°C for 20 hrs. Each plates were overlaid with 1mL of water containing 0.5 mg naladixic acid and 1 mg apramycin, using a spreader to distribute the antibiotic solution evenly; incubation at 30°C continued for 7 days or until colonies appeared. Single colonies were plated on MS agar containing naladixic acid and apramycin, 25 μg/mL and 50 μg/mL, respectively for 7 days or until colonies appeared, and gDNA extracted.

To determine the best time/temperature combination for heat shock in *Verrucosispora* spp., varying combinations were first tested and the resulting spores analysed under a light microscope and the extent of sporulation determined. Spores were analysed using a haemocytometer using a volume of 25 nL as the counting area, which typically had ~200 spores present using the spore stock (200,000 spores/ uL). Temperatures between 40 – 100°C were tested with time varying between 5 – 30 minutes.

4.2.8 Genomic DNA extraction

Many techniques for gDNA extraction from exconjugants for PCR and sequencing were tested. CTAB, Sigma-Aldrich kit (cat. No. NA2110) with bead-beating and the potassium hydroxide Ethylene Diamine Tetraacetic Acid (KOH-EDTA) method.

CTAB

Genomic DNA was extracted from *V. sp.* str. MG37 exconjugants using the cetyltrimethyl ammonium bromide (CTAB) method. 30 mL GYM media was inoculated with exconjugants from MS plates with single colonies at 7 days and grown at 28°C for 3 days and harvested at 16,000X g and re-suspended in 5 mL TE25S buffer (25 mM Tris-HCl pH. 8; 25 mM EDTA pH. 8 and 0.3 M sucrose). Lysozyme was added to final concentration 2 mg/mL and incubated for 1hr at 37°C. Proteinase K and SDS were added to final concentration 0.18 mg/mL and 0.5% respectively, and incubated for 1hr with occasional inversion at 55°C. NaCl added to final concentration 0.8 M and mixed; CTAB/ NaCl added and mixed followed by incubation at 55°C for 10 mins followed by cooling to 37°C. Chloroform/isoamyl alcohol added and mixed my inversion for 30 min. Centrifuge at 13,500X g at 20°C for 15 mins. Supernatant decanted and 0.6 volumes of isopropanol added and mixed; after 3 min DNA spooled and rinsed in ethanol and air dried before dissolving in 1mL TE buffer (1 mM Tris-HCl; 1 mM EDTA pH. 8) at 55°C. Extracted gDNA was confirmed by running products on 1% agarose gel stained with ethidium bromide. Genomic DNA was quantified using Qubit 2.0 Fluorometer according to ThermoFisher guidelines.

Sigma-Aldrich kit with bead beating

Prior to the use of the Sigma-Aldrich gDNA extraction kit (cat. No. NA2110), colonies were re-suspended in 200 μL of water, and 0.8 g of 0.1mm glass beads were added. Cells were disrupted for 30 seconds at an agitation speed of 4,500 rpm and then harvested at 10,000X g for 3 minutes and the supernatant transferred to a new tube. This was then used for gDNA extraction by Sigma-Aldrich gDNA extraction kit. Extracted gDNA was confirmed by running products on 1% agarose gel stained with ethidium bromide. Genomic DNA was quantified using Qubit 2.0 Fluorometer according to ThermoFisher guidelines.

Potassium hydroxide-EDTA method (Sun et al., 2014)

A single exconjugant was transferred to a micro centrifuge tube containing 27 μL 10 mM Tris-EDTA (1 mM, pH. 7.6) and then 3 μL 0.4 M KOH-10mM EDTA added and incubated at 70°C for 5 min. 3 μL Tris-HCl (pH. 4.0) was added and this was used directly for PCR amplification. The quality of the gDNA extracted was checked by running products on 1% agarose gel stained with ethidium bromide and by using it in a MyFi PCR reaction using 16S primers to check the quality of the method. PCR mixture (25 μL) contained 2XMyfi^{TM} High-Fidelity polymerase master-mix (Bioline), primers (0.5 μM each), template DNA (25 ng) and DMSO (3%). The PCR was performed using a specified MyFi^{TM} Thermal Cycling Procedure: 1 cycle, 95°C for 60 s; 30 cycles, 95°C for 15 s, 71°C for 15 s and 72°C for 3:00 min; and 1 cycle 72°C for 15 min. 5 μL of each product was run on 2% agarose gel with ethidium bromide.

4.2.9 Checking for gene disruption

The extracted gDNA was used for PCR to check for successful gene disruption: PCR mixture (25 μL) contained the outlined Myfi^{TM} procedure. Primer sets were designed for each gene being disrupted at varying intervals from the site of the sgRNA – 0.75 Kb, 2Kb, 5Kb and 7 Kb. Extension time for each PCR was altered depending on the expected size of product.

4.3 Results

4.3.1 Identification of target genes

Genes for phenotypic qualities exhibited by *V. sp.* str. MG37 were determined using BLAST search with characterised genes known to contribute to colour in another related spp. This led to the identification of one gene - *VAB05470* – which is predicted to responsible in the biosynthesis of the orange pigment of *Verrucosispora* species. The target gene involved in proximicin biosynthesis chosen was *ppb120*, as it is predicted to being vital in proximicin biosynthesis and hence, its absence, should result in a drastic and easily detectable reduction in proximicin production.

4.3.2 Design of sgRNA's

Summarised in Table 19, a single protospacer was designed for *VAB05470*, designed to delete the amino acid sequence: PAVLAA and cause a frameshift resulting in an early stop codon. Three were designed for *ppb120* to increase the likelihood of producing a high efficiency disruption plasmid. All were designed to delete a single amino acid resulting in frameshift (Table 19). Each contained the adjacent PAM sequence, occupied the non-coding strand, and were unique in the *V.*

Table 19. Design of synthetic guide RNA's for gene editing. Pigmentation gene in *V. MG37 VAB05470,* and *ppb* gene *ppb120*. Pink denotes the 3' Purines and in bold is the PAM sequence.

	Protospacer	PAM	Non-coding	Unique	Purines
VAB05470_1	CCGGGTGTACGCCGCG **AGGACGG**	Y	Y	Y	100%
Ppb_120_1	CGGCGATCCGCTGTGC **CGGATGG**	Y	Y	Y	75%
Ppb_120_2	ACCCGCAGGGCGACCT **GGCGTGG**	Y	Y	Y	75%
Ppb_120_3	CGTCGTCCGACCGGAC **GGTACGG**	Y	Y	Y	75%

sp. str. MG37 genome. Attempts were made to have purines occupying the last four (3') bases, but this was not 100 % achieved in all. Primers were successfully designed to incorporate these protospacers with stated over hangs.

4.3.3 Production of pCRISPR-sgRNA-RT

All protospacers were successfully assembled and ligated into pCRISPomyces-2. Sequencing determined that for *VAB05470* the sgRNA inserted correctly; but for *ppb120*, only 2/3 were correctly inserted. A repair template for *VAB05470* could not be successfully amplified, however, it was deemed as non-essential at these initial stages as any disruption resulting in pigment alteration was confirmation of CRISPR/cas9 application in *Verrucosispora* spp. For *ppb120,* a ~ 2.5 Kb repair template was successfully amplified, annealed (Fig. 69), digested and inserted to produce pCRISP-120sgRNA-RT. Production of correct plasmid was tested by (i) restriction digest and the correct size insert was seen; and (ii) sequencing which was correct.

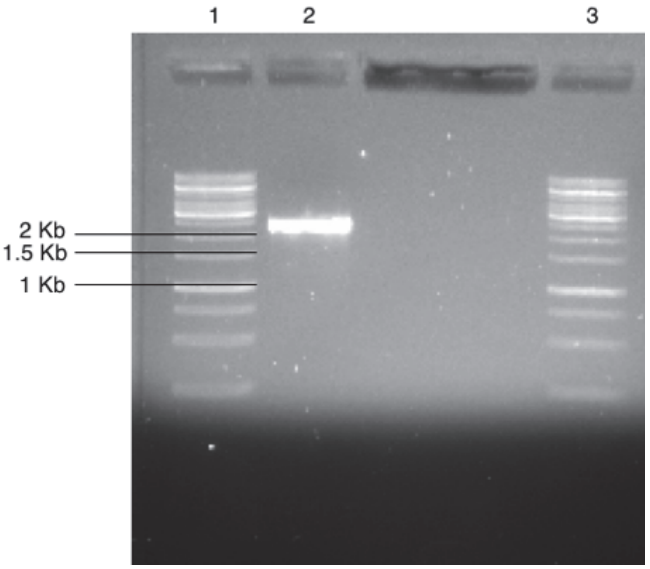


Figure 69. *ppb120* 2.5 Kb repair template. **1.** A fragment 2.5 Kb in size which spans the *ppb120* gene produced by PCR and gel purified.

4.3.4 Genomic DNA extraction

All three genomic DNA extraction methods – CTAB, Sigma-Aldrich kits with bead beating and KOH-EDTA - failed to produce high quality DNA for sequencing. CTAB was difficult to implement as the disrupted *V. sp.* str. MG37 strains – even those with inserted repair templates – were difficult to grow in liquid culture, resulting in very low yields of DNA – too low for the Qubit to register a quantity. DNA was run on agarose gels to assess the quality, and the result was typically a complete lack of gDNA being extracted. A combination of Sigma-Aldrich kits with bead beating did successfully yield genomic DNA, however, it contained many low MW bands – suggesting degradation due to the abrasive bead beating (Fig. 70). This quality was enough to allow PCR confirmation of gene disruption and where necessary PCR products were sequenced for confirmation. Shown in Fig. 70, the concentration of DNA extracted from *V. sp.* str. MG37*ΔVAB0547* in comparison to WT *V. sp.* str. MG37. Despite designed for PCR amplification of DNA, the KOH-EDTA-PCR technique did not yield any amplicons, and so it was assumed that little gDNA was extracted, if any at all. This was tested against already extracted gDNA, with the same treatment done to ensure it wasn't the chemical e.g. KOH preventing the PCR reaction from working (Fig. 71).

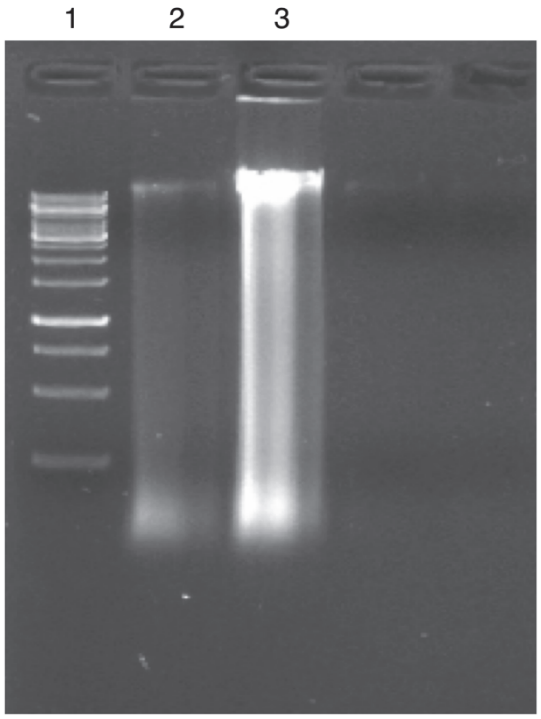


Figure 70. Genomic DNA extraction from *V.* sp. str. MG37 by a combination of bead beating and Sigma Aldrich gDNA kit. **1.** 1 Kb GeneRuler ladder **2.** *V. MG37 ΔVAB0547* **3.** WT *V. MG37*

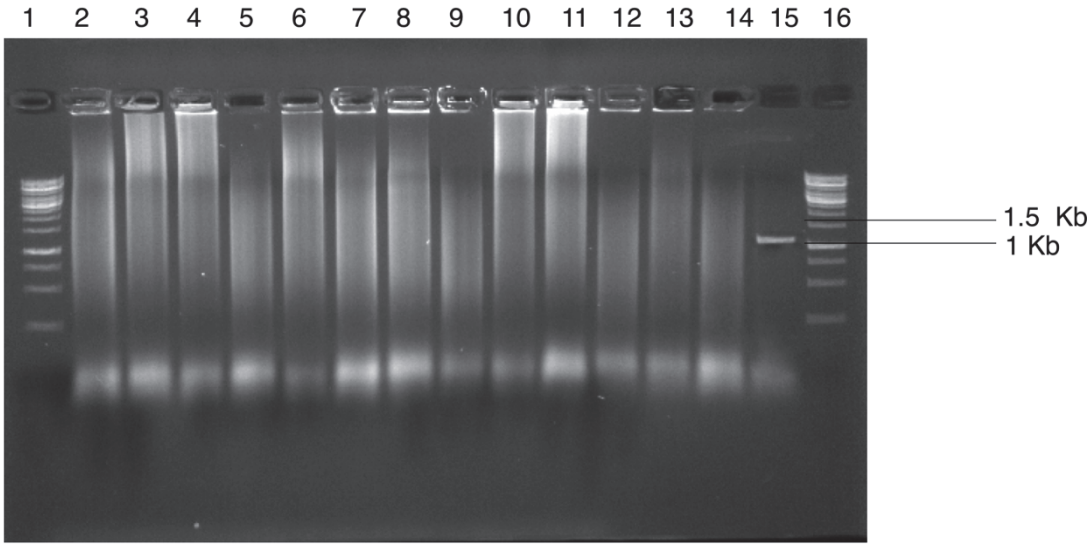


Figure 71. PCR products of KOH-EDTA-PCR technique using 16S primers. 1. 1 Kb GeneRuler Ladder **2-14.** Extracted gDNA using KOH-EDTA, **15.** CTAB-extracted WT gDNA **16.** 1 Kb GeneRuler ladder.

4.3.5 Interruption of phenotypic gene - VAB05470

Exconjugants isolated in *V. sp.* str. MG37Δ*VAB05470* disruption studies were shown to have altered pigmentation, in comparison to WT *V. sp.* str. MG37. This can be seen in Figure 72 after 6 days of incubation; single isolated colonies of *V. sp.* str. MG37Δ*VAB0547* strains were very slow growing in comparison to WT *V. sp.* str. MG37and showed sporulation only after ~21 days. They show similar morphology characteristic to *Verrucosispora* spp., with the loss of the vibrant orange pigment resulting in a creamy/orange colour. In addition to slow growth, the texture of colonies was 'stickier' in comparison to WT *V. sp.* str. MG37, suggesting a potentially substantial deletion in the absence of a repair template.

As mentioned, despite extensive attempts, gDNA was never extracted. PCR confirmation was done (Fig. 73), which showed a region encompassing the *VAB0547* gene which was 5 Kb was reduced to ~ 1.2 Kb, and primer sets designed to amplify 3.5 and 2 Kb around *VAB0547* did not give any on target amplification in *V. sp.* str. MG37 Δ*VAB0547.*

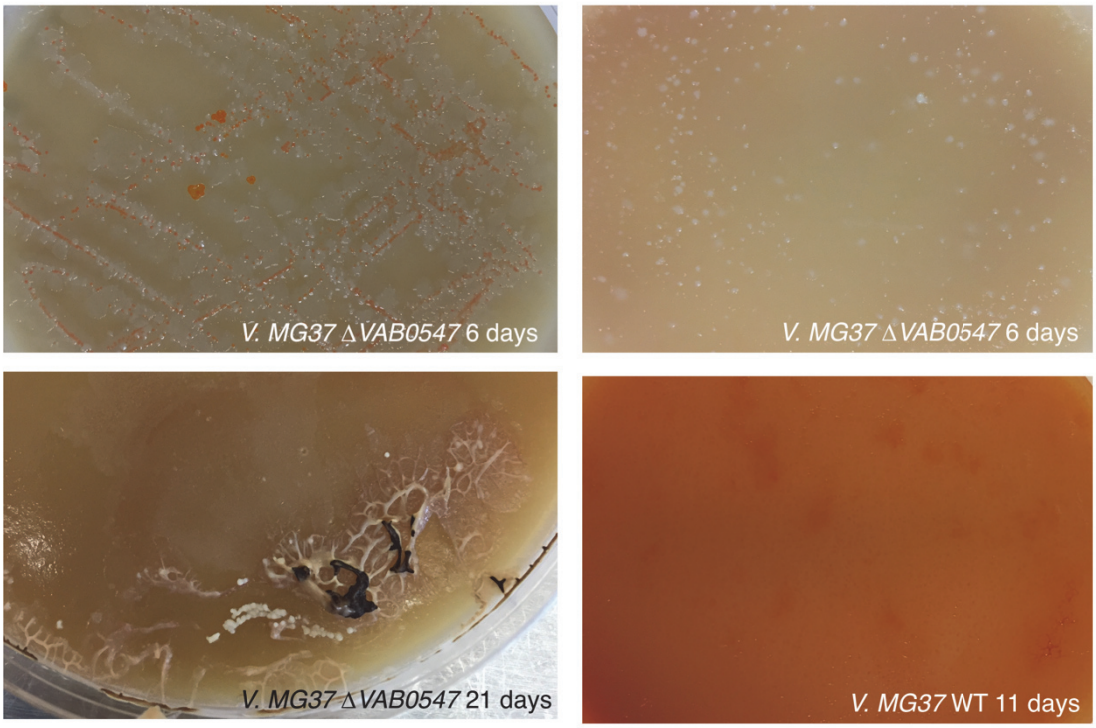


Figure 72. Change in phenotypic traits and growth in *VAB05470* deletion strains. **Top left** and **right**. *V. MG37* Δ*VAB0547* using 55C and 65C heat-shock respectively. **Bottom left**. *V. MG37* Δ*VAB0547* retained the ability to sporulation – dark spores being produced after 21 days. **Bottom right**. Colour of WT *V.* sp. str. MG37

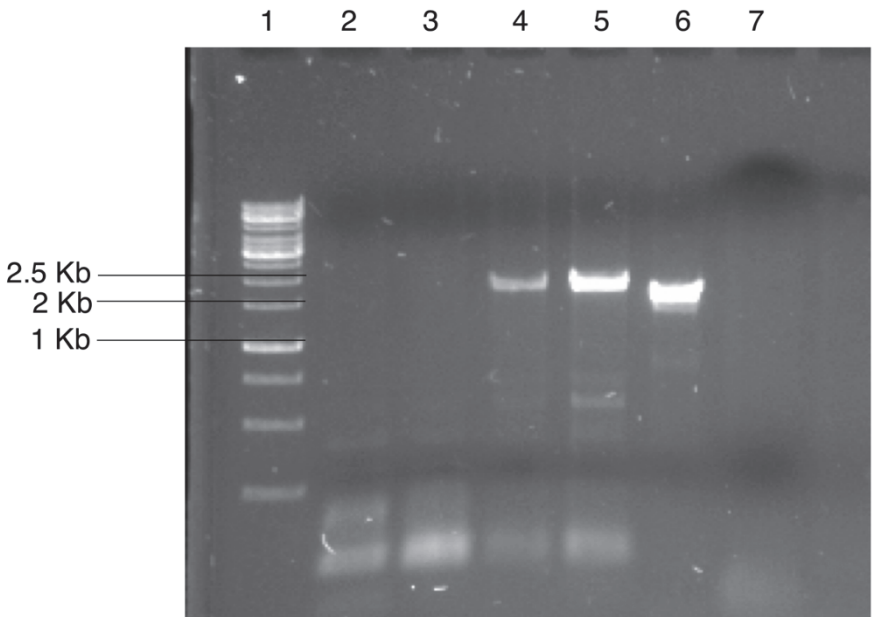


Figure 73. Confirming the deletion of *VAB05470* gene in *V.* sp. str. MG37. 1. *V. MG37* WT 5 Kb **2.** *V. MG37* WT 3.5 Kb **3.** *V. MG37* WT 2 Kb **4.** *V. MG37* Δ*VAB0547* 5Kb reduced to 1.2. **5.** *V. MG37* Δ*VAB0547* 3.5 Kb with no products. **6.** *V. MG37* Δ*VAB0547* 2Kb with no products. **7.** 1 Kb GeneRuler Ladder.

4.3.6 Disruption of proximicin biosynthetic gene - ppb120

Both pCRISP-120sgRNA and pCRISP-120sgRNA-RT were successfully produced and conjugated into *V. sp.* str. MG37. Extraction of genomic DNA was once again an issue, and PCR was used to confirm gene deletion – in pCRISP-120sgRNA a ~ 7 Kb region was deleted and pCRISP-120sgRNA-RT the correct ~ 2.5 Kb repair template was present (Fig. 74). Initial PCR of pCRISP-120sgRNA, using primer sets to amplify 0.75, 2 and 5 Kb regions surround *ppb120* showed no products - demonstrating a large region of DNA has been disrupted (Fig. 75). Only primer sets which spanned a 5 Kb region surrounding *ppb120* gave products of ~500 bp. Six pCRISP-120sgRNA-RT exconjugants were isolated and demonstrated consistent apramycin resistance, however, PCR confirmed only 3/6 had the repair template present, and one of those had a smaller template present. Sequencing of the PCR products confirmed the all of the repair templates were correct with a product of ~2.5 Kb; hence the region missing in the smaller template amplicon (~ 2 Kb), must be located in the centre of the template – the region not resolved by sequencing. Due to time constraints and a collaborating group taking on this work, analysis of proximicin production in these strains was not done.

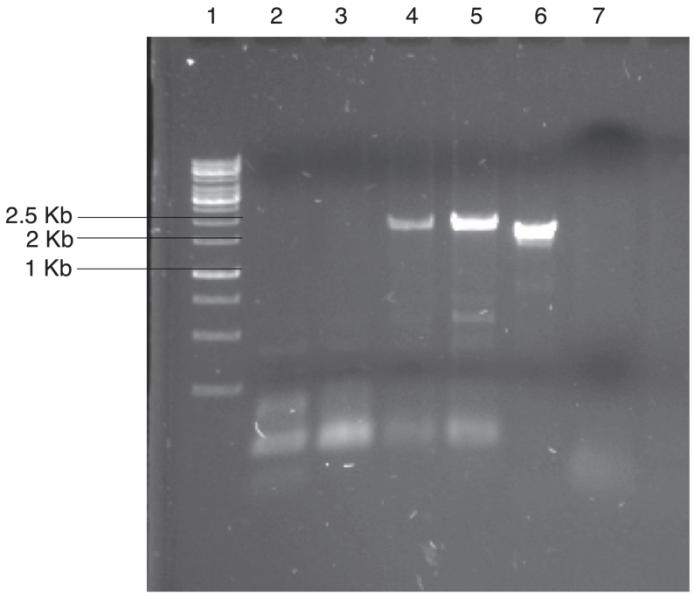


Figure 74. Successful confirmation of the repair template in 3/6 *V.* sp. str. MG37 Δ*ppb120* exconjugants. **1.** 1 Kb GeneRuler DNA Ladder **2-4.** *V. MG37 Δppb120* exconjugants 1-3 with no RT present **5 and 6.** *V. MG37 Δppb120* exconjugants 4 and 5 with correct size RT **7.** *V. MG37 Δppb120* exconjugant 6 with a smaller RT than expected.

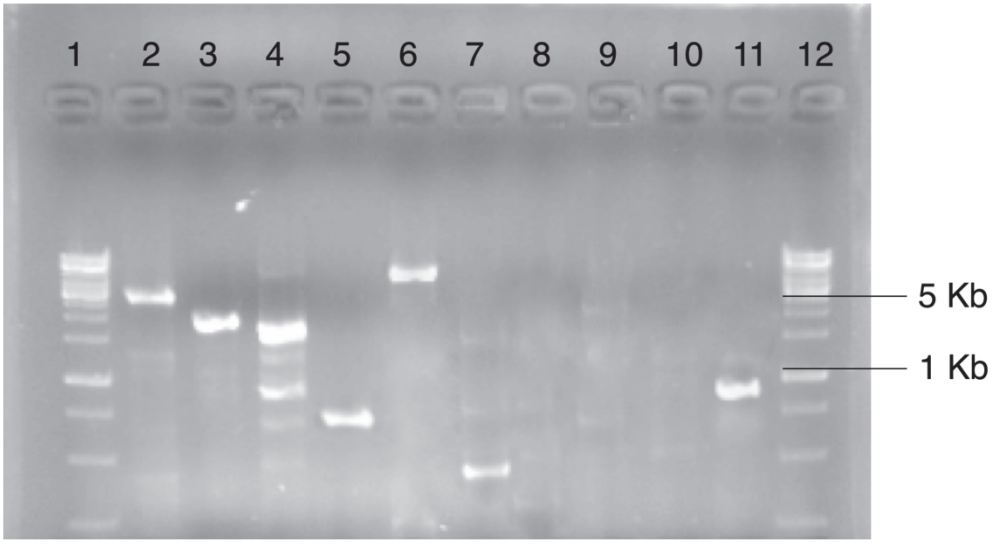


Figure 75. Successful PCR confirmation of gene deletion in *V.* sp. str. MG37 Δ*ppb120* exconjugants with no repair template. Primer sets are designed to amplify fragments of 0.75 Kb, 2Kb, 5Kb and 7 Kb regions surrounding *ppb120*. **1.** 1 Kb GeneRuler DNA ladder**. 2.** 5 Kb Primers WT **3.** 3.5 Kb primers WT **4.** 2 Kb primers WT **5.** 0.75 Kb primers WT **6.** 7 Kb Primers WT **7.** 5 Kb Primers Δ*ppb120* showing ~500 bp product **8.** 3.5 Kb primers Δ*ppb120* **9.** 2 Kb primers Δ*ppb120* **10.** 0.75 Kb primers Δ*ppb120* **11.** 7 Kb primers Δ*ppb120* now a ~1 Kb product.

4.3.7 Heat shock optimisation

Table 20. HS combinations tested in optimisation. Temperature and time analysed for ability to initiate germination, using the percentage of spores appearing to germinate, counted using a haemocytometer. Colours denote extent of germination – white – no effect of HS (below 25 %); blue: low induction of sporulation (25-50%); Yellow: very high sporulation initiation (51-75%); Pink: Very high sporulation (>75%)

Time	Temperature (°C)							
	40	**50**	**55**	**60**	**65**	**70**	**80**	**100**
30 sec	0	0	0	1-25	25-50	25-50	1-25	1-25
60 sec	0	0	0	25-50	25-50	> 51	25-50	1-25
90 sec	0	0	0	25-50	> 51	> 51	25-50	1-25
2 min	0	0	0	> 51	> 51	> 51	25-50	1-25
3 min	0	0	1-25	> 51	> 51	> 51	25-50	1-25
5 min	0	0	1-25	> 51	> 75	> 51	25-50	1-25
7 min	0	0	1-25	> 51	> 75	> 51	25-50	1-25
10 min	0	0	1-25	> 75	> 75	> 51	25-50	1-25

The extent of germination was determined by the number of spores in 0.25 nL which appeared to have begun germination, and the range noted. As shown in Table 20 the lower and upper limits of sporulation were determined - temperatures of 50°C or below did not elicit any sporulation response, regardless of the exposure duration, and temperatures over 80°C were shown to be detrimental after exposure of over 30 seconds. The optimal temp/time combination were shown to be at 60 – 65°C for times longer than 5 minutes; 65°C for 10 mins was chosen as the routine heat shock protocol. When testing conditions around the optimal, it became clear that only small changes the duration and temperature of the HS had a large effect on the number of *V. sp.* str. MG37shown to be not disrupted after conjugation, as shown in Figure 76.

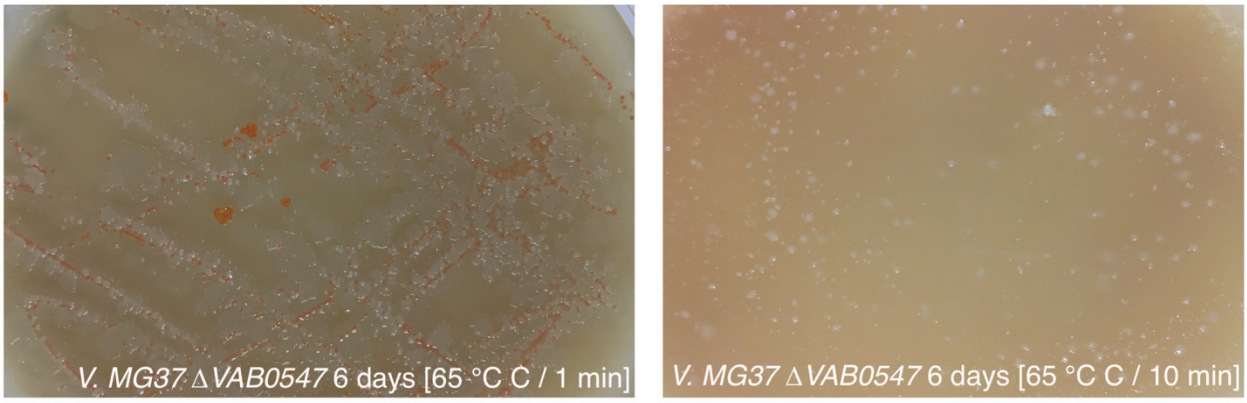


Figure 76. Heat shock optimisation. Photograph of two CRISPR-targeted V. MG37 strains showing the number of exconjugants with successful gene editing – pale colour colonies when heat shock was applied for 1 minute (Left) and 10 minutes (Right).

4.4 Discussion

4.4.1 Overview of findings

We report for the first time the application of CRISPR/Cas gene editing technology in the Micromonosporaceae family of bacteria. The initial proof-of-concept study outlines the deletion of the orange pigment-conferring gene in *V. sp.* str. MG37, to result in a sickly, cream coloured strain with the correct genomic region missing confirmed by PCR. Using the CRISPR/Cas9 system, double stranded breaks were introduced and repaired via the native and highly error prone non-homologous end joining (NHEJ) pathway which resulted in large, random gene inactivation, leading to substantial areas of gene interruption. Subsequent goals to interrupt genes predicted to be essential in proximicin biosynthesis were attempted, however, issues were encountered regarding the rapid large scale screening of *V. sp.* str. MG37 exconjugants and Cas9 toxicity. Here, I discuss these problems and potential resolutions in depth, in conjunction with future directions of CRISPR/Cas9 in Actinomycetes, with the aim of yielding novel biologically active compounds.

4.4.2 Efficient gDNA isolation

CRISPR/cas signifies a breakthrough in gene editing technology, however, issues of its implementation in some bacterial families has been encountered (Lou et al., 2016). Although overcoming many problems concerning traditional gene deletion strategies, one caveat remains: the requirement to screen exconjugants. This is an issue encountered in every outlined gene-disruption method, as it relates to complex genomic DNA extraction required for screening transformants; this is exacerbated by intricate growth requirements of many bacteria, especially marine Actinomycetes (summarised in Fig. 77). To allow the utilisation of CRISPR/Cas to its fullest potential in these classes, these issues must be resolved; two potential approaches discussed are: improving routine culturing techniques for viable growth, and increasing the quality and quantity of DNA from extraction methods.

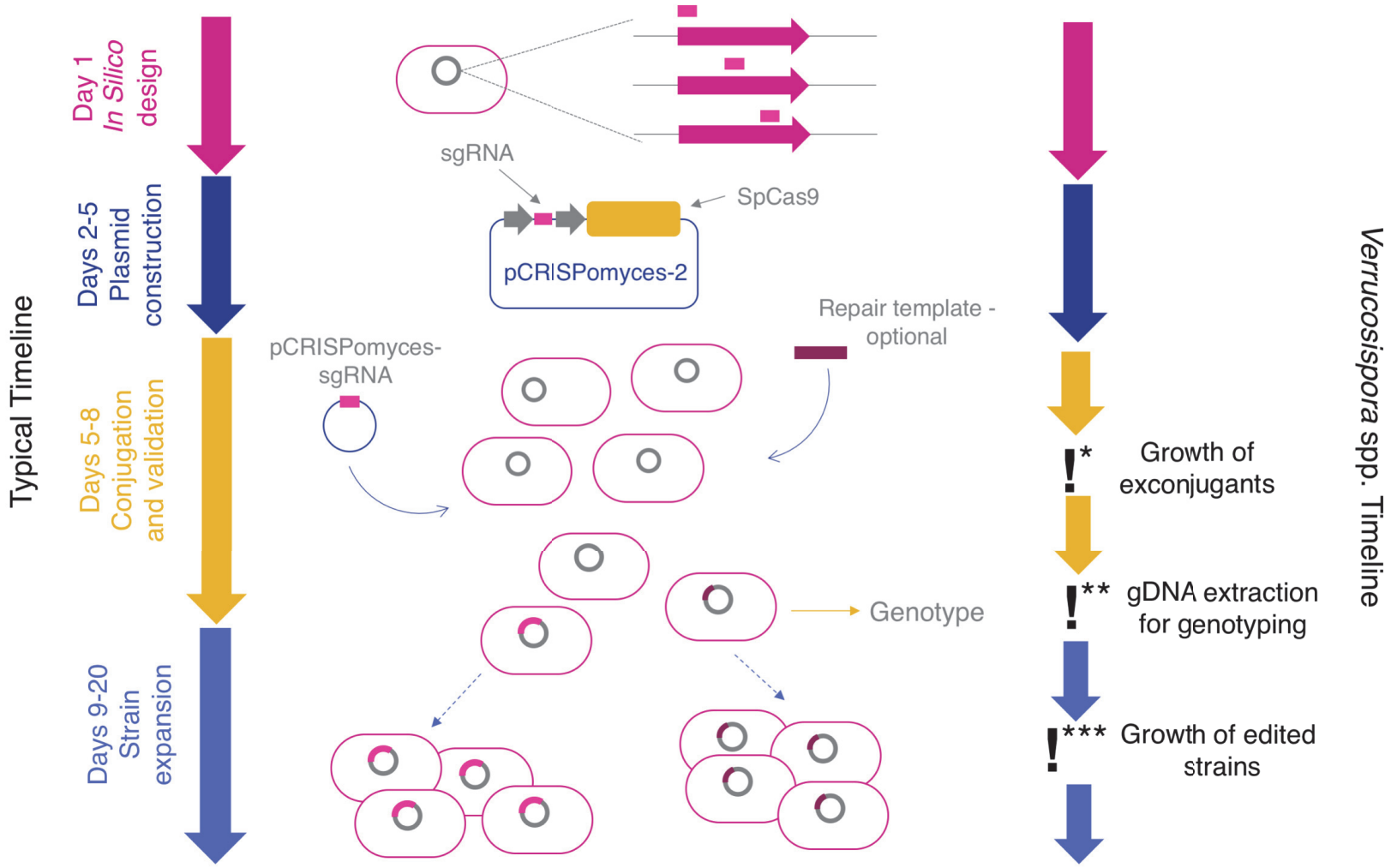


Figure 77. Issues encountered during CRISPR/Cas application in *Verrucosispora* spp. The typical timeline for CRISPR/Cas gene editing applications is ~20 days, from designing sgRNA's to having successfully screened edited strains. Issues were encountered in the application in *Verrucosispora* sp. st. MG37 due to the long growth steps required and cumbersome and largely inefficient genomic DNA extraction methods. Here, I outline potential solutions to the problems encountered here. Adapted, in part, from Ran et al., (2013).

Improving culturing techniques

Actinomycetes represent a large group of scientifically important microbes and they have been shown to colonise almost every environment on Earth (Bukhari et al., 2013). They typically occupy complex niches in their natural habitats with intricate biotic and abiotic requirements; hence, it is difficult to recreate this environment in a lab setting (Joint et al., 2010). This has led to the long-standing observation that standard microbiological techniques only isolate a minuscule proportion of the microbial diversity present. *Verrucosispora spp*. are no exception to this obstacle; *V. maris* AB18-032 and *V. sp.* str. MG37 were isolated from deep sea sediments in the Sea of Japan and in Norway, respectively (Fiedler et al., 2008). The remote nature of these habitats make it difficult to recreate and yield viable, long-running growth in a laboratory setting. This is not a recently appreciated barrier, with Staley & Konopka (1985) dubbing it the 'great plate count anomaly' with reference to the several orders of magnitude difference between the number of colonies which grew on laboratory media, in comparison to the total number which could be exhibited by epifluorescence microscopy, specifically in relation to marine isolates. This discrepancy is typically attributed to lack of knowledge regarding organic substrates or concentrations that would have been present in the natural habitat, owing to the so few marine bacteria in culture. However, other factors may affect growth, such as: disruption of the sensitive biotic factors; high concentrations of substrates required for detectable growth may be toxic or large scale growth may result in compound production leading to self-toxicity, or the slow growing nature of the bacteria may lead to inevitable contamination before they have had time to thrive (Stewart, 2012). Alain & Querellou (2009) reviewed this problem of culturability, and culminated in a principal resolution being increased periods of growth to allow the transition from a dormancy to growing state, likely due to the requirement of *de novo* synthesis of enzymes which enable growth on synthetic medium. Fiedler et al., (2008) was the first to attempt *Verrucosispora* spp. growth and metabolite studies, resulting in the development of SSG media, with the aim of optimum *Verrucosispora* spp., culture. As we do not know the chemical composition of the natural environment occupied by *Verrucosispora* spp., it is impossible to assess the resemblance of SSG. SSG and other *Streptomyces* specific culturing media were analysed here (Kieser, 2000), all

resulted in very slow growing bacteria with a lag phase of ~11 days and sporulation not occurring until >28 days. It is difficult to determine whether this was due to poor substrate combinations, or if *Verrucosispora* spp., simply represent a very sedentary microbe. Irrespective of the cause, this led to issues concerning the ability to test large ranges of exconjugants for successful CRISPR/cas mediated gene editing. By the time exconjugants had grown up to a level capable of gDNA extraction, the cultures had typically been contaminated due to the lack of antibiotics necessitated by the conjugation protocols (Fig. 77). This was only exacerbated by off target gene deletions resulting in sickly, slower growing strains that were easily out competed. For the majority of the research conducted here, these issues were largely negated by using a phenotypically active protein as the deletion target, and so strains harbouring a successful gene interruption could be rapidly identified. However, for future work into the proximicin gene cluster, further work into either (i) improving growth conditions to better mimic the *Verrucosispora* natural environment or (ii) better extraction methods resulting in high quality gDNA from smaller, ideally single colony, cultures.

To stimulate the natural habitat of *Verrucosispora* spp. to enable optimal growth, it would be to advantageous to study the environment, to uncover which of the parameters are important for the growth. This could be achieved by moving a portion of the sediment into the laboratory; some groups are even going as far as culturing the bacteria back in the natural habitat, and then re-isolating it. This can also be done by placing the bacteria into a membrane-lined diffusion chamber which permits the movement of nutrients and growth factors, but not bacteria; it was shown that this yielded micro-colony production of up to 40%, in comparison to the same inoculum on standard laboratory plates, which had a recovery of 0.05% (Kaeberlein et al., 2002; Bollmann et al., 2007). Although this would not applicable for large-scale deletion screening, it does demonstrate the powerful tool the environment itself represents, which could be exploited. Other studies have shown how sediment taken from the habitat can serve as the basis for incubation conditions within the lab (Morris et al., 2002; Rappe et al., 2002; Logares et al., 2009; Song et al., 2009); using the same premise as before, with the convenience of in-lab facilities. With the

current scientific abilities, it is not feasible to determine the complex relationships and needs a bacterium has with its neighbouring microbes and its surroundings, and so maintaining its natural environment to such a high degree could greatly increase culturability. This is not limited to increasing growth rates for screening following CRISRP/Cas exconjugants. It has previously been shown that bacteria undergoing environmental stress are less amenable to genetic manipulation and conjugation (Kieser, 2000; Joint et al., 2010), meaning efficacy of gene editing may too be increased. It could be possible to use an amalgamation of approaches presented and to reveal a potentially feasible method: culture *Verrucosispora* spp., in the membrane lined pouches, in aquarium-like conditions containing sediment from their natural habitat– and harvest shortly prior to conjugation (Kaeberlein et al., 2002). This would use unstressed bacteria for the genomic alteration aspects of the research, potentially increasing its efficacy. While also increasing growth rates following conjugation to allow increased gDNA extraction. These methods are already in practice with the steady deviation from synthetic media in favour of more nature-mimicking approaches as discussed here, establishing the significance of the standard environment of economically important microbes. This area of research is constantly expanding, with most current efforts aimed at culture of non-culturable bacteria for novel compound identification. However, if these bacteria are to be utilised for such a purpose, issues surrounding in-lab growth for large scale screening and gDNA extraction need to be addressed.

Improving gDNA extraction techniques

As shown here, high quality genomic DNA extraction of small *V. sp.* str. MG37 colonies for sequencing was not achieved. The major paradox encountered was: with increasingly harsh chemical and mechanical treatment to yield higher concentrations, came lower quality gDNA, indicated by the increased occurrence of low molecular weight bands showing DNA degradation. As previously discussed, this could be resolved by better culturing methods of *Verrucosispora* strains resulting in larger biomass production, or as investigated here – by different extraction methods. As researchers have encountered similar issues regarding large-scale transformant screening, the advent of new techniques has begun, shifting away from more

traditional, time intensive approaches (Kumar et al., 2010). This problem is further demonstrated in the research here as Actinomycetes possess the characteristic robust cell wall composition which makes it difficult to efficiently lyse the cells, while leaving the DNA unaffected. This has led to the design of Actinomycete-specific genomic DNA extraction protocols, one such being the potassium hydroxide (KOH) approach used here (Sun et al., 2014). This utilises the envelope degrading capacity of KOH to allow the leaching of gDNA, which can then be used as a PCR template. Sun et al., (2014) first described this alkaline lysis method and demonstrated its successful application to produce clean PCR fragments using well characterised house-keeping genes in many Actinomycete species; I was not able to recreate this here using *WT V. sp.* str. MG37, or any *V. sp.* str. MG37 gene deletion strains. This is likely owing to the extended duration required for *V. sp.* str. MG37 growth, as it was shown that cultures older than 7 days were difficult to disrupt for DNA amplification (Sun et al., 2014). The large-scale success of this method amongst other Actinomycete strains, suggests it is a viable option for screening projects, and potentially *Verrucosispora* strains if issues surrounding slow growth patterns are resolved. *Verrucosispora* spp. genomic DNA was extracted using either the CTAB method or a lysis following mechanical cell wall disruption; the former method is not feasible for screening projects as, although it produces high quality DNA, is time intensive and requires substantial biomass due to an inefficient process. Mechanical disruption of the cell wall by bead beading followed by chemical lysis, allowed DNA for PCR template to be extracted; however, the extracted DNA was not of high enough quality for sequencing. Other Actinomycete-specific methodologies, for example the utilisation of liquid nitrogen to help break the cell wall to allow better lysozyme action (Kumar et al., 2010), or bead beating in combination with detergents such as Chelex, to better protect the DNA resulting in the ‘perfect combination lysis’ (Jaffe et al., 2002), should be investigated. The ability to rapidly and efficiently assess exconjugants for gene editing events is likely to become an increasingly prominent issue in upcoming years as the bacterial applications of CRISPR/cas technology broadens.

4.4.3 Bacterial CRISPR/Cas delivery systems

The explosion of CRISPR/cas applications in eukaryotic cells has over-shadowed that of prokaryotes, resulting in a comparatively low amount of well characterised bacterial CRISPR/cas delivery systems. CRISPR/Cas techniques used in eukaryotes cannot simply be repurposed for application in bacteria; this is due to the non-homologous end joining (NHEJ) repair pathway in bacteria not being particularly robust (Fig. 78), in addition to homology-directed repair being functionally ineffective (Fig. 79). This has led to the production of bacteria-specific CRISPR/cas plasmids which incorporate phage genetic systems to allow efficient homologous recombination in bacteria. This means that CRISPR/Cas gene editing works in a functionally different way in comparison to eukaryotic approaches. Instead of the innate CRISPR/cas pathway seen in bacterial acquired immunity, bacterial gene editing utilises phage-driven recombination to produce gene modification and then

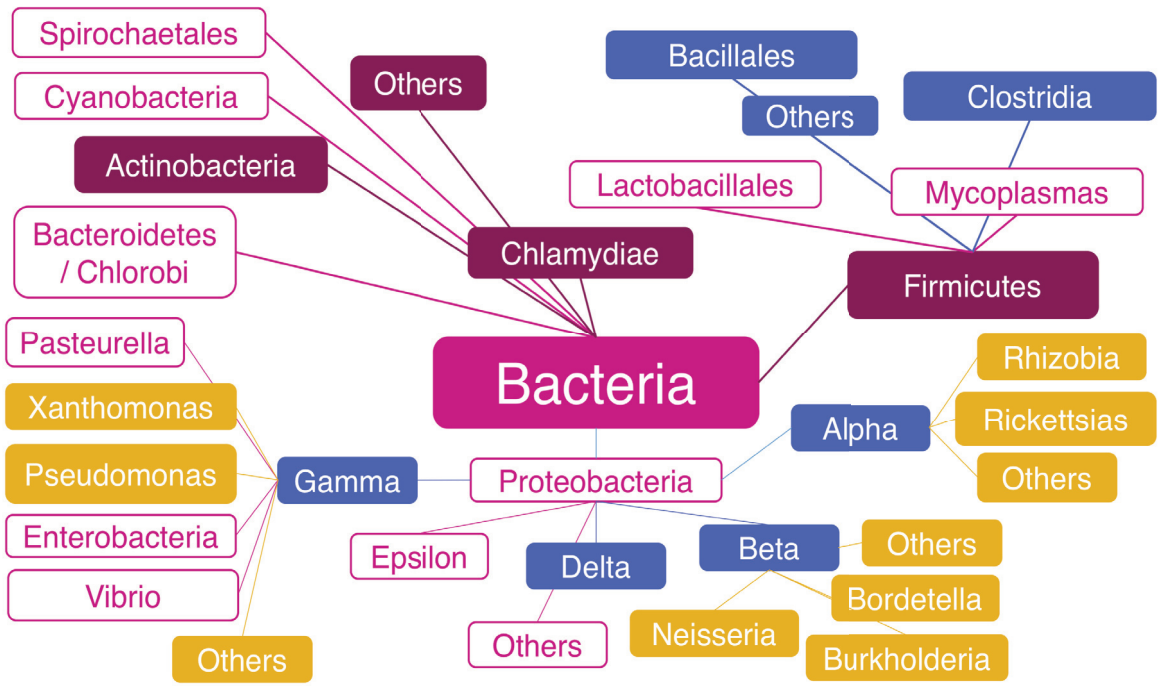


Figure 78. Distribution of the NHEJ Ku protein in Bacteria. A phylogenetic tree showing the major phyla, classes and families of bacteria. Dark pink boxes show phyla that contain homologues of Ku – including Actinobacteria (*Verrucosispora* spp.); Blue boxes denote classes or orders that contain homologues of Ku and yellow boxes indicate families that contain homologues of Ku. Important to note, is that there is no instance in which Ku homologues are present in all species within a phylum or class. Adapted from Bowater & Doherty (2006)

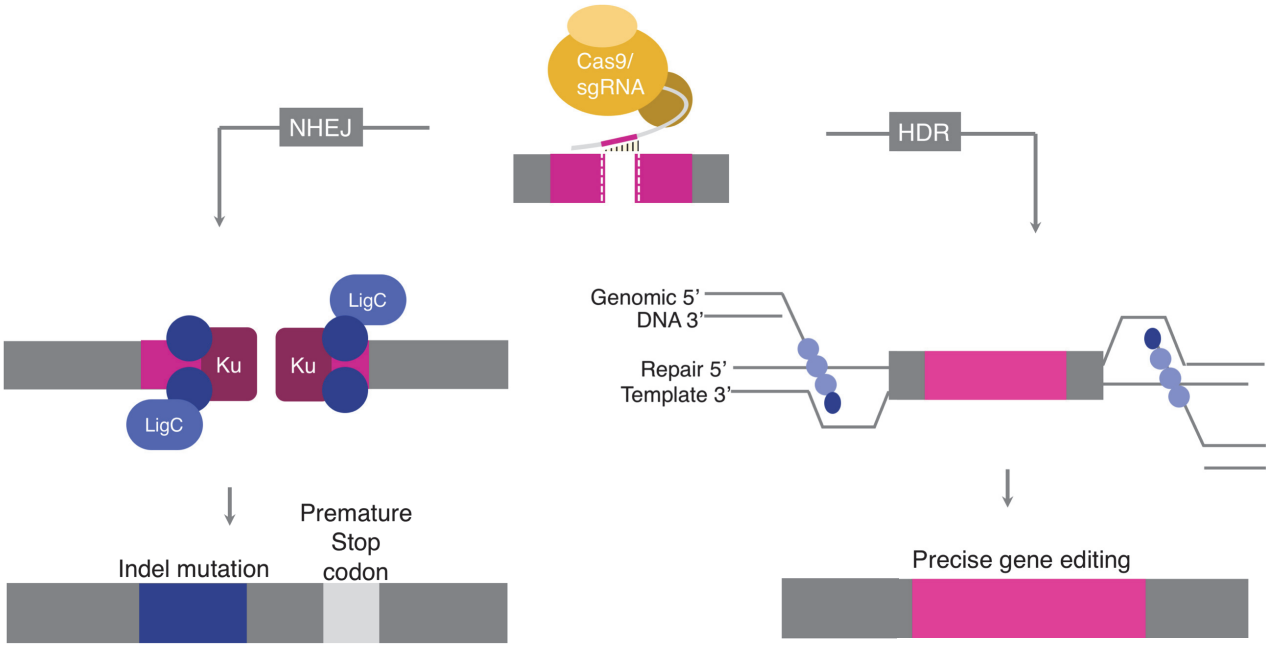


Figure 79. Double stranded break repair in prokaryotes promotes gene editing. DSBs created by Cas9 complex (yellow) can be done in two ways: the error-prone NHEJ pathways or a repair template can be supplied – by plasmid – to leverage the homology dependent (HDR) repair pathway. Adapted from Ran et al., (2014).

CRISPR/Cas is used to exert a selection pressure for modification, by degrading any WT gene-containing DNA. This system is distinct from those used in eukaryotes as CRISPR is not the primary editing force, but instead, a way of selecting against non-edited microbes resulting in a powerful negative selection system. A current review of marketed CRISPR/Cas delivery techniques designed specifically for gene editing in bacteria, shows that there are 4 plasmid-based systems available. These include: pCRISPomyces (Cobb et al., 2015); dual-RNA:Cas9 (Jiang et al., 2013); noSCAR system (Reische & Prather, 2015) and the two plasmid system designed by Li et al., (2015). Each consists of a codon optimised Cas9 gene, an insert site for custom spacers and repair templates, typically spread over two plasmids, each with a selectable marker and temperature sensitivity site. The latter three methods utilise CRISPR as a negative selection force (Fig. 70); interestingly, the system employed here, pCRISPomyces designed by Cobb et al., (2015) works in a method more akin to eukaryotic approaches. This is because it is designed specifically for *Streptomyces* (Fig. 78), which contain homologous NHEJ machinery, and hence, are more recombinogenic that other bacteria. Although the pCRISPomyces system

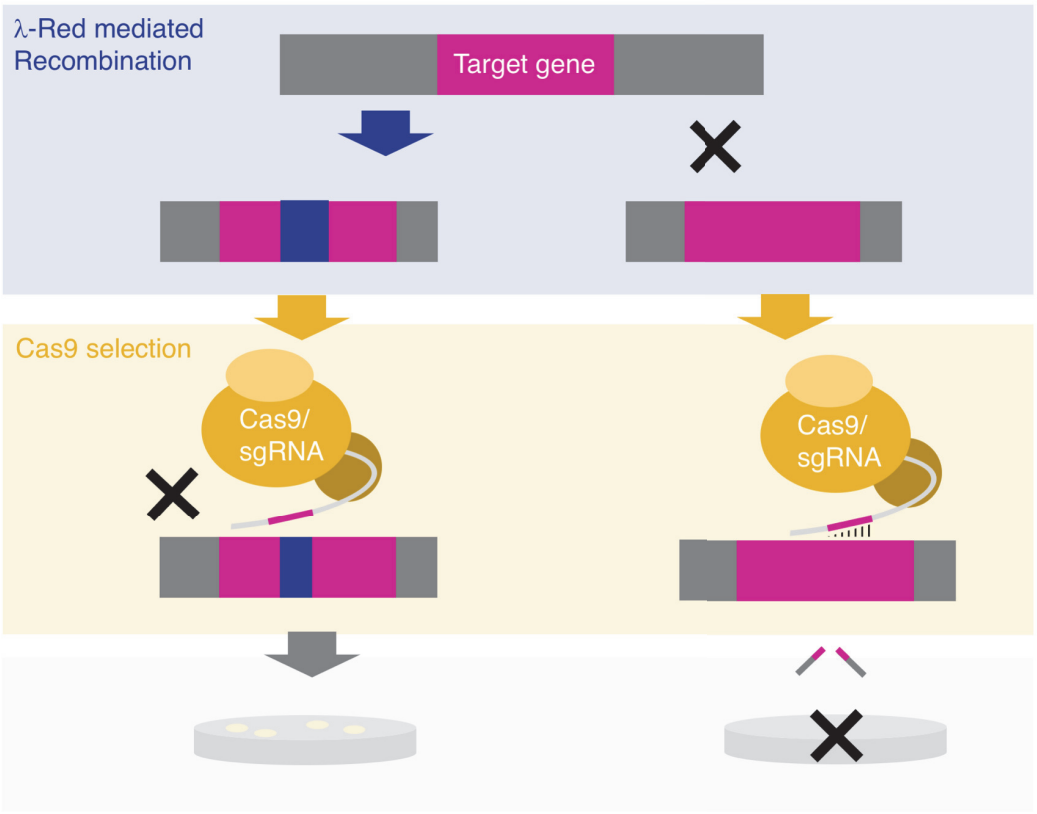


Figure 80. The current typical application of CRISPR/Cas technologies in bacterial systems. Here the CRISPR/Cas is not the driving gene editing force, rather, used to exert selection on those not successfully edited via red-mediated recombination gene deletion. **Left**: Workflow shows the successful editing via Red, and so the homology site present for sgRNA binding is no longer present, and so no Cas9 endonuclease activity can occur, exconjugants grow on the plate. **Right**: Workflow shows unsuccessful gene editing via Red, and so the homology site for the sgRNA remains, and so Cas9 destroys the DNA and so no exconjugants are produced.

successfully employed here to result in genetically modified organisms, issues were encountered regarding the viability of the transformants. It is likely due to two potential reasons: (i) although the NHEJ pathway is present in *Streptomyces*, it is error-prone and results in large regions of deletion around the target site, potentially leading to off-target gene deletion (ii) random Cas9 endonuclease activity across the genome, resulting in 'Cas9 toxicity'. The former can be improved by introducing a repair template to allow homologous directed repair following Cas9 activity (Fig. 79). This has previously shown a dramatic increase in the efficiency of deletions, resulting in 100% frequency, in some cases (Tong et al., 2015). The introduction of a repair template requires the PCR and splicing together of two ~1Kb regions surrounding the target DNA region; issues were encountered in the amplification and

splicing of this region owing to the considerably high GC content of the fragments, resulting in its optimization requiring further investigation. In its absence, large areas of the target regions were deleted in comparison to the designed subtle frameshifts to inactivate a singular gene (Fig 79). As shown here, in addition to the desired colour change, the gene deletion changed many phenotypic characteristics such as making the texture stickier and inhibiting sporulation, suggesting important genes were disturbed. Future work into optimising PCR protocols for high (>75%) GC regions will help rectify these issues, allowing the full potential of CRISPR/cas specificity to be utilised. Future directions of these issues are discussed further in section.

The pCRISPomyces system (Cobb et al., 2015) was used here to exploit the *Streptomyces* specificity; however, despite resulting in successful target gene editing, issues regarding Cas9 toxicity were encountered. This is displayed in the lack of successfully modified viable transformants, in comparison to negative controls done. Each of the marketed systems contain *Streptomyces pyogenes* Cas9 (SpCas9) codon optimised protein under the control of a promoter system, in each case a previously characterised strong promoter is used. The potential of off-target Cas9 action to cause toxicity is not a surprise, considering the extensive evidence of earlier RNA interference-based screening technologies exhibiting non-specific toxicity. It was hence commonly proposed as a potential CRISPR/Cas limitation, prior to direct evidence confirming its existence. Attempts to hinder Cas9 activity towards non-target PAM sites throughout the genome, have centred on overloading the nuclease with non-targeting single guide RNAs (sgRNAs). This approach to control the disruptive Cas9 enzyme, however, fails to prevent the arguably most dramatic off target effect of double stranded breaks, which are fatal towards many cell types. Numerous strategies have been employed to negate this affect, including paired Cas9 nickases containing individual sgRNA's to generate two single stranded breaks (Cho et al., 2014); truncated guides (<20 nts) to prevent similarity with off-target genomic regions without sacrificing on-target efficiency, resulting in a 5,000-fold decrease in undesired mutagenesis (Fu et al., 2014; Tsai et al., 2015) and modified Cas9 enzymes to engineer increased fidelity (Kleinstiver et al., 2016;

Slaymaker et al., 2015). Genome-wide assay experiments to monitor double strand breaks (Crosetto et al., 2013; Kim et al., 2015; Kim et al., 2016) have indicated these strategies successfully limit off target mutagenesis. However, these studies have focused extensively on eukaryotic systems, and the incorporation of these advances in bacterial systems is limited. For use in *Verrucosispora* spp. and other *Streptomyces* strains, pCRISPomyces could be modified to contain an inducible promoter system to lower the overall concentration, and hence, off-target effects of Cas9. This would be straightforward to do by using restriction sites to excise the current system, and replace with a system under strict inducible control such as the XylS/*Pm* or *Pm*ML1-17 systems which have been shown to exhibit extremely low basal expression (Balzer et al., 2013). However, as well as decreasing off target mutations, it will likely lower on-target activity. Although this price is acceptable for small scale research attempts as outlined here, for the implementation of large SMC characterisation resulting in novel compound discovery, other approaches to lowering Cas toxicity, such as intelligent sgRNA design, must be investigated.

4.4.4 Optimal sgRNA design

CRISPR/Cas9 exploitation has the ability to drastically reduce the time and labour associated with gene modification in comparison to traditional techniques; now, focus is shifting to increasing its on-target specificity. Research into hindering Cas9 toxicity has seen in-depth investigation into truncated sgRNA's, along with a building cache of other evidence, it has become increasingly clear that not all apparently analogous sgRNAs work with the same efficiency (Doench et al., 2014; Wang et al., 2014). Just as issues regarding off-target Cas9 activity were encountered in the work outlined here, it is important that research is ongoing to determine criteria that can predict sites favouring high activity and specificity. Currently these can be largely grouped into two considerations: the PAM sequence chosen and the composition of the sgRNA homology region.

There are some required features for a sgRNA to function, one of the earliest recognised is a PAM sequence adjacent to the sgRNA homology region (Shah et al., 2013). The Cas9 protein most commonly utilised in CRISPR/cas9 studies is that of

Streptomyces pyogenes (SpCas9), which has the optimal PAM site: NGG; high GC DNA content, exhibited in many Actinomycetes, mean these PAM sequences are a common occurrence (Heler et al., 2015). While initially appearing advantageous - as there are many potential PAM sequences for sgRNA design, it increases the chances of SpCas9 off target effects. This has led to the engineering of spCas9 proteins with altered PAM specificities, and some design tools offering user-defined PAMs. Kleinstiver et al., (2015) has engineered spCas9 proteins exhibiting better discrimination between NAG and NGA PAMS, leading to reduced Cas9 toxicity. For use in *Verrucosispora spp.,* and other extremely high GC organisms, a more radical approach may be required to achieved high efficiency. One such tactic may involve the focus shift, from SpCas9, to engineering Cas9 specificity from other species which are known to have longer PAM sequences. For example, *Staphylococcus aureus* Cas9 has been shown to recognise NNGRR PAM sites (Ran et al., 2015); the longer sequence confers increased on-target specificity in comparison to spCas9, as well as increasing potential scope for modification for use in other organisms.

While the NGG PAM region is essential for high efficiency gene editing, it does not assure it; the sequence features, independent of PAM, are important for target efficiency (Fu et al., 2014; Doench et al., 2015). One logical variable is the targeting of a region at the 5' end of the gene, resulting in an early frame shifts or stop codons; however, despite the reduced efficacy seen in some cases, the majority show that target sites throughout the gene result in similar disruption rates (Wang et al., 2014; Doench et al., 2015). Research then shifted to decrease non-specific mutagenesis by engineering truncated sgRNAs, the principle being that a shorter stretch of DNA will statistically find partial-homology less often (Doench et al., 2015). Although I am not convinced that the effects seen aren't simply the result of an overall decrease in Cas activity – leading to lower off- but also on-target events – it is nevertheless an interesting observation and, mentioned here for the sake of inclusivity. Another non-PAM specific consideration is the overall GC content of the sgRNA; intermediate GC percentage, outperformed their high and low GC counterparts (Wang et al., 2014). This suggests that either high or low affinities of the sgRNA-target duplexes

negatively impact Cas9 activity. This is further increased if a purine occupies the most PAM-proximal position (Wang et al., 2014). Extensive screening of sgRNA efficacies, as done by Doench et al., (2014) in which the specificity of over 6000 sgRNAs were tested, are forging the way into bioinformatical design of sgRNAs; as this research continues, sequence based activity predication modelling programs are likely to closely follow the currently available sgRNA designing tools.

Considered sgRNA and PAM selection, in parallel with many freely available sgRNA designing tools were used to produce the *ppb120* targeting sgRNAs used in the research here, and yet Cas9 toxicity – shown by 'sickly' or non-viable exconjugants – was still present. This could potentially be resolved by increasing the expressional control of the sgCas9 protein, as discussed. However, this will not resolve issues encountered regarding high GC organisms possessing many NGG PAM sites leading to off target effects. Nor will shortening sgRNA's increase specificity, owing to the same principle. As high GC organisms, such as the *Verrucosispora* strains examined here along with other Actinomycetes, are relied upon for such a large proportion of novel compound discovery, it is vital that CRISPR/Cas technology can be utilised in these organisms. As the number of bacterial genomes sequenced and annotated continues to steadily rise, the real extent of CRISPR/cas9 diversity across prokaryotes will become apparent. Investigation into the systems utilised in Actinomycetes and other families, will likely expose more-suitable CRISPR systems for gene editing applications focused on them. Just as bacteria evolve systems to allow self-resistance to antimicrobial compounds they produce, they too must possess methods which prevent self-DNA degradation from CRISPR/Cas action. A cursory look at the *Verrucosispora maris* AB18-032 genome presents only a single area displaying characteristic features of a CRISPR/Cas loci. This region contains multiple Cas-related genes, as well as 41 foreign spacers, representing previous scars. Further investigation reveals two interesting observations: (i) the repeating regions comprising the *V. maris* CRISPR are completely unique, and (ii) the Cas proteins present show no conserved identity with spCas9 (<8%). These together further suggest that *V. maris* AB18-032 utilises a Cas specificity, and hence sgRNA and PAM sequence requirements, very different to that of sgCas9. This is supported

by work shown here, as when sgCas9 is employed in *V. maris* AB18-032 it results in Cas9 toxicity. Prior to future work in other high GC strains, I suggest the exploitation of CRISPR/cas systems in these organisms to engineer more suitable gene editing machineries, with the aim of lowering research barriers associated with off target DNA mutagenesis affects.

4.4.5 Current uses of bacterial CRISPR/Cas technologies

The renaissance created in by CRISPR/cas technologies is not limited to the gene editing field of biology and biotechnology, with many other approaches being formed from its exploitation. The breadth of uses this technology represents is unparalleled; the ability to advance other areas, distinct from gene editing, is only just starting to become appreciated with efforts being re-directed to areas of bacterial research such as: gene expression modulation; imagine of genomes; phylogenetic studies and vaccination of bacterial strains. These will each be briefly discussed here:

Modulation of bacterial gene expression

Repurposing CRISPR/Cas systems for targetable gene regulation is another common version of the application. The concept of programmable bacterial gene regulation has already been investigated using other forms of controllable DNA binding proteins such as Zinc-finger and transcription activator like effector arrays (Carroll 2014). The application of CRISPR/cas was initially realised with two papers (Bikard et al., 2013; Qi et al., 2013) which similarly used a guided Cas9 protein to enable the blocking of transcription of the target gene, achieved by either blocking movement of the RNA polymerase or preventing binding of RNA polymerase to the promoter sequence, respectively. The latter of these approaches has been shown to be highly effective in yielding gene repression (Bikard et al., 2013; Luo et al., 2015; Tong et al., 2015), whereas the former is better utilised as a gene down regulation tool (Choudhary et al., 2013; Qi et al., 2013; Tong et al., 2015). As investigations have progressed, a whole array of CRISPR/cas expression control capabilities been have been reported; integrating the results produces in a set of heuristics: target promoters for complete gene repression; to loosen control, produce sgRNA's towards the end of the gene and, to lower the control further, target the template

strand at the end of the gene (Cleto et al., 2016; Luo et al., 2015; Qi et al., 1023; Tong et al., 2015). Concurrent to this work, Cas9-mediated activation of target genes has been reported (Bikard et al., 2013) – this was done by fusing known RNA polymerase stabilisers to Cas9 to facilitate complex assembly. This resulted in an up to 23-fold increase of expression depending on the native promoter strength; change in expression was more dramatic for weaker promoter systems (Bikard et al., 2013). Mentioned here is just a summary of some approaches of CRISPR/cas repurposing for gene expression control, many more exist, and are likely to be developed.

Imaging of bacterial genomes

Both bacterial and eukaryotic chromosomes are known to be complex and high organised structures; growing evidence suggests cellular localisation and structure of microbial chromosomes is tightly associated with gene expression (Ryter and Change, 1975; Bryant et al., 2014; Wen and Xiao et al., 2014). This is important information when intending to control the physiology of industrial strains. The characteristic short cell cycle of bacteria dictates the co-occurrence of transcription and replication of chromosome, however, the effect each has on the other is vague; questions regarding the effect chromosomal modification has on replication is still unclear. Current microbial *in vivo* genomic loci tagging methodologies are laborious and time consuming, and this has hindered work in this area. The target specificity and simplicity of Cas9 has lent itself to such an application; the ability to optically tag a specific genomic locus can result in increased insight into the implementation of recombinant gene technologies. DNA imaging via Cas9 uses the same premise as Cas9 gene activation, and is typically done by using a site-specific targeting domain consisting of Cas9::sgRNA with a fluorescent domain attached (Anton et al., 2014; Deng et al., 2015; Ma et al., 2015). As with much of CRISPR/cas applications, this research is largely focused on eukaryotic organisms, but it could be rapidly re-purposed for work in prokaryotes to allow better understanding of chromosomal organisation and dynamics present in bacterial systems.

Phylogenetic studies

Millions of years of evolution and a myriad of niches, has resulted in Bacteria being an almost incomprehensively broad and diverse domain. To be able to exploit the divergent and distinct metabolism and physiology they possess, it is vital to be able to accurately identify strains, a problem intensified by horizontal gene transfer increasing diversity within species. This is a problem encountered in any bioprocess which relies on the characteristics of a specific bacterial strain, such as bio-refineries; similarly, in the race for novel compound discovery, the ability to identify species accurately is vital as to avoid wasting time in previously characterized species. CRISPR/Cas systems have been extensively proven to be a reliable and precise method for genotyping bacterial strains (Cui et al., 2008; Lindenstrauss et al., 2011; Shariat et al., 2015; Sun et al., 2015); this is done by analyzing the unique set of 'scars' a strain has acquired represented in the chronologically ordered spacers incorporated at the CRISPR loci. This allows the diversification of two bacterial populations by a discrete event series in each strain; this allows detailed insight into the history of strains, showing times they existed as a single population – similar spacers at the 3' end – and diversification events – diverging 5' spacers. Vitally, this information can be used to track pathogenic bacteria through human society providing insights into the trajectory of pathogens responsible for outbreaks, giving insights into potential prevention routes.

Vaccination of microbial strains

There are many industrially important bacterial strains, such as diary and pharmaceutical production, in which contamination with a pathogenic bacteria or bacteriophage, can cause catastrophe. Even small infections can be amplified in reactors, resulting in huge economic loss to companies. Methods to prevent this has traditionally been focused on improving the bacterial innate immune system, preventing phage DNA injection or absorption (Sturino & Klaenhammer, 2006). The concept of protecting bacteria from potential enemies was one of the first uses of complete CRISPR/Cas systems (Barrangou et al., 2007). For implementation of this approach, all foreseeable contaminants are predicted and spacers targeting their genome are designed and introduced into the CRISPR array. Issues are

encountered due to high mutation rates exhibited in viruses, and so areas of the genome which are rarely altered should be carefully identified. This approach allows the successful protection of the industrially important strain, and hence, cheaper, more efficient fermentation practices.

4.4.6 The future of bacterial CRISPR/Cas technology

As discussed here, the repurposing of CRISPR/Cas systems to allow specific gene editing, as well as a whole host of other applications, is a powerful tool. The explosion of CRISPR/Cas systems engineered towards eukaryotes continues to rumble on, versatile bacterial tools are evolving at a steady pace. Specialized bacterial CRISPR/Cas approaches represent tools of promising ability; however, issues regarding NHEJ, delivery systems and utilization of host strategies continue. The lack of NHEJ ability characteristic of many bacteria species was not encountered or discussed here in depth, as it is present in the target species;

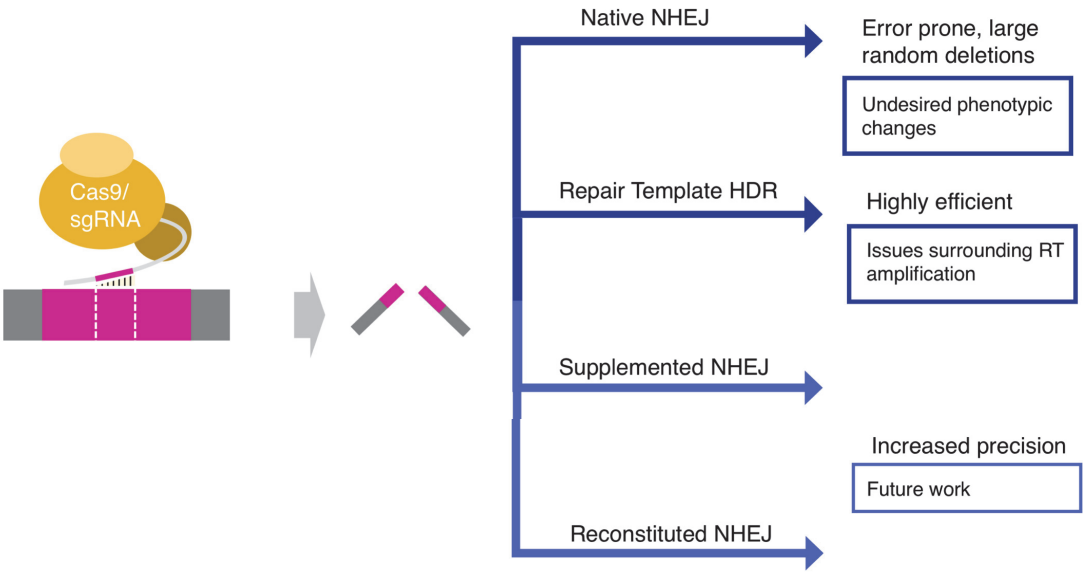


Figure 81. The current issues encountered when utilising CRISPR/Cas technology in bacterial species. CRISPR/Cas systems depend on Non-homologous end joining (NHEJ systems), which are not present in most bacterial species, and when it is present it is not very robust. Dark blue arrows represent currently utilised CRISPR/Cas gene editing applications – either using the inefficient host NHEJ system, or incorporating a repair template to allow homology dependent repair (HDR), both of which encounter issues. In light blue, are the potential ways to overcome these issues explained here: supplementing the proteins required for NHEJ into bacterial strains which do not possess them or have inefficient versions. Another approach would be, as many bacterial species genomes contain NHEJ systems, they could be altered to be under the control of a strong promoter, effectively reconstituting their production.

however, for future large scale applications it must along with other issues, be addressed. The major advantage CRISPR/Cas conveys in comparison to other gene editing techniques, is its ease of use and its adaptability to multiplex approaches. These capabilities are in part owed to NHEJ-dependent double stranded break (DSB) repair systems, which allow template-independent genome editing (Fig. 81). The presence of this system negates the complex requirement of multiple template DNAs for a single loci dictated by homology dependent repair; however, these systems are rare. Currently, this issue has yet to be overcame. One potential approach would be to engineer enzymatic requirements for NHEJ – such as Ku and LigD proteins – to create a companion plasmid used in combination with other bacterial CRISPR gene editing vectors; this would allow gene editing via NHEJ (Fig. 81). It has also been seen that many bacteria once had the genetic ability for NHEJ but it has since been made redundant (Tong et al., 2015), and so simply reconstituting the components on the host genome is another potential research avenue. This was not an issue for the research here as Actinomycetes represent some of the few prokaryotes which contain NHEJ homologs, however, they are extremely error prone, resulting in large unpredictable gene interruptions. The wide scale applicability of CRISPR/Cas gene editing technology in prokaryotes must be resolved. It is interesting to note that ~40% of bacteria contain CRISPR/Cas systems, a percentage which continues to rise as their impact widens; this number largely eclipses the number of bacterial species with NHEJ homologs. Why would these species contain the apparatus for double stranded DNA breaks, without the machinery capable of repairing them? These systems obviously serve some function, which requires further investigation. With the plethora of evolutionary history and diversification, it is hard to imagine that all these systems work in a similar manner. Better understanding of the CRISPR/Cas mechanisms comprising the bacteria domain, facilitated by improved culturing and sequencing techniques, could potentially result in the utilization of the systems present in the target bacteria. As outlined specifically for *Verrucosispora* spp. previously, the idea of switching focus away from spCas9 towards better understanding endogenous systems in target organism, could eventually lead to facile gene editing systems. By simply introducing compatible pre-crRNAs into the target organism, and ensuring their

endogenous systems are actively expressed – potentially by other CRISPR/Cas-driven approach – you could engineer bacteria to target their own DNA. To although routine implementation of these applications much still requires research, primarily, issues surrounding crRNA delivery systems, it stands to demonstrate the potential the already revolutionary CRISPR/cas discovery, encompasses.

Bacteria and other microorganisms provide the basis to much of what we regard as healthcare; their ability to produce such diverse biologically active compounds comes as a result of millennia of evolution. The recent discovery and potential of CRISPR/Cas systems has been adopted by various disciplines; however, bacterial application of these technologies has lagged behind in many respects. Here, we outline the use of CRISPR/cas gene editing technology for use in marine Actinomycete, *Verrucosispora sp.* str. MG37, to aid in novel compound identification. The phenotypic gene responsible for the characteristic orange pigment was targeted, and successfully deleted. Issues surrounding its application are discussed here, in addition to other repurposing possibilities and future prospects of CRISPR/Cas technologies.

4.5. References

Alain, K. and Querellou, J., 2009. Cultivating the uncultured: limits, advances and future challenges. *Extremophiles*, *13*(4), pp.583-594.

Balzer, S., Kucharova, V., Megerle, J., Lale, R., Brautaset, T. and Valla, S., 2013. A comparative analysis of the properties of regulated promoter systems commonly used for recombinant gene expression in Escherichia coli. *Microbial cell factories*, *12*(1), p.26.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A. and Horvath, P., 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, *315*(5819), pp.1709-1712.

Bassuk, A.G., Zheng, A., Li, Y., Tsang, S.H. and Mahajan, V.B., 2016. Precision medicine: genetic repair of retinitis pigmentosa in patient-derived stem cells. *Scientific reports*, *6*.

Bentley, S.D., Chater, K.F., Cerdeno-Tarraga, A.M., Challis, G.L., Thomson, N.R., James, K.D., Harris, D.E., Quail, M.A., Kieser, H., Harper, D. and Bateman, A., 2002. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3 (2). *Nature*, *417*(6885), pp.141-147.

Berdy, J., 2005. Bioactive microbial metabolites. *Journal of Antibiotics*, *58*(1), p.1.

Bérdy, J., 2012. Thoughts and facts about antibiotics: where we are now and where we are heading. *Journal of Antibiotics*, *65*(8), p.385.

Blin, K., Medema, M.H., Kazempour, D., Fischbach, M.A., Breitling, R., Takano, E. and Weber, T., 2013. antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. *Nucleic acids research*, *41*(W1), pp.W204-W212.

Bollmann, A., Lewis, K. and Epstein, S.S., 2007. Incubation of environmental samples in a diffusion chamber increases the diversity of recovered isolates. *Applied and Environmental Microbiology*, *73*(20), pp.6386-6390.

Bowater, R. and Doherty, A.J., 2006. Making ends meet: repairing breaks in bacterial DNA by non-homologous end-joining. *PLoS genetics*, *2*(2), p.e8.

Bukhari, M.A.A., Thomas, A.N. and Wong, N.K., 2013. Culture Conditions for Optimal Growth of Actinomycetes from Marine Sponges. In *Developments in Sustainable Chemical and Bioprocess Technology* (pp. 203-210). Springer US.

Cho, S.W., Kim, S., Kim, J.M. and Kim, J.S., 2013. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nature biotechnology*, *31*(3), pp.230-232.

Cho, S.W., Kim, S., Kim, Y., Kweon, J., Kim, H.S., Bae, S. and Kim, J.S., 2014. Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome research*, *24*(1), pp.132-141.

Choi, K.R. and Lee, S.Y., 2016. CRISPR technologies for bacterial systems: current achievements and future directions. *Biotechnology advances*, *34*(7), pp.1180-1209.

Cobb, R.E., Wang, Y. and Zhao, H., 2014. High-efficiency multiplex genome editing of Streptomyces species using an engineered CRISPR/Cas system. *ACS synthetic biology*, *4*(6), pp.723-728.

Crosetto, N., Mitra, A., Silva, M.J., Bienko, M., Dojer, N., Wang, Q., Karaca, E., Chiarle, R., Skrzypczak, M., Ginalski, K. and Pasero, P., 2013. Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nature methods*, *10*(4), pp.361-365.

Cyranoski, D., 2016. Chinese scientists to pioneer first human CRISPR trial. *Nature News*, *535*(7613), p.476.

Deltcheva, E., Chylinski, K., Sharma, C.M., Gonzales, K., Chao, Y., Pirzada, Z.A., Eckert, M.R., Vogel, J. and Charpentier, E., 2011. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*, *471*(7340), pp.602-607.

Doench, J.G., Hartenian, E., Graham, D.B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B.L., Xavier, R.J. and Root, D.E., 2014. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nature biotechnology*, *32*(12), pp.1262-1267.

Du, D., Wang, L., Tian, Y., Liu, H., Tan, H. and Niu, G., 2015. Genome engineering and direct cloning of antibiotic gene clusters via phage φBT1 integrase-mediated site-specific recombination in Streptomyces. *Scientific Reports*, *5*.

Fernández-Martínez, L.T. and Bibb, M.J., 2014. Use of the meganuclease I-SceI of Saccharomyces cerevisiae to select for gene deletions in actinomycetes. *Scientific reports*, *4*.

Fiedler, H.P., Bruntner, C., Riedlinger, J., Bull, A.T., Knutsen, G., Goodfellow, M., Jones, A., Maldonado, L., Pathom-Aree, W., Beil, W. and Schneider, K., 2008. Proximicin A, B and C, novel aminofuran antibiotic and anticancer compounds isolated from marine strains of the actinomycete Verrucosispora. *Journal of Antibiotics*, *61*(3), p.158.

Fu, Y., Sander, J.D., Reyon, D., Cascio, V.M. and Joung, J.K., 2014. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nature biotechnology*, *32*(3), pp.279-284.

Graham, D.B. and Root, D.E., 2015. Resources for the design of CRISPR gene editing experiments. *Genome biology*, *16*(1), p.260.

Gust, B., Chandra, G., Jakimowicz, D., Yuqing, T.I.A.N., Bruton, C.J. and Chater, K.F., 2004. λ Red-mediated genetic manipulation of antibiotic-producing Streptomyces. *Advances in applied microbiology*, *54*, pp.107-128.

Heler, R., Samai, P., Modell, J.W., Weiner, C., Goldberg, G.W., Bikard, D. and Marraffini, L.A., 2015. Cas9 specifies functional viral targets during CRISPR-Cas adaptation. *Nature*, *519*(7542), p.199.

Hruscha, A., Krawitz, P., Rechenberg, A., Heinrich, V., Hecht, J., Haass, C. and Schmid, B., 2013. Efficient CRISPR/Cas9 genome editing with low off-target effects in zebrafish. *Development*, *140*(24), pp.4982-4987.

Hwang, K.S., Kim, H.U., Charusanti, P., Palsson, B.Ø. and Lee, S.Y., 2014. Systems biology and biotechnology of Streptomyces species for the production of secondary metabolites. *Biotechnology advances*, *32*(2), pp.255-268.

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. and Nakata, A., 1987. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *Journal of bacteriology*, *169*(12), pp.5429-5433.

Jaffe, R.I., Lane, J.D., Albury, S.V. and Niemeyer, D.M., 2000. Rapid extraction from and direct identification in clinical samples of methicillin-resistant staphylococci using the PCR. *Journal of clinical microbiology*, *38*(9), pp.3407-3412.

Jansen, R., Embden, J., Gaastra, W. and Schouls, L., 2002. Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular microbiology*, *43*(6), pp.1565-1575.

Jiang, W., Bikard, D., Cox, D., Zhang, F. and Marraffini, L.A., 2013. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nature biotechnology*, *31*(3), pp.233-239.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A. and Charpentier, E., 2012. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science*, *337*(6096), pp.816-821.

Joint, I., Mühling, M. and Querellou, J., 2010. Culturing marine bacteria–an essential prerequisite for biodiscovery. *Microbial biotechnology*, *3*(5), pp.564-575.

Kaeberlein, T., Lewis, K. and Epstein, S.S., 2002. Isolating" uncultivable" microorganisms in pure culture in a simulated natural environment. *Science*, *296*(5570), pp.1127-1129.

Kaiser, J., 2016. First proposed human test of CRISPR passes initial safety review. *Science*.

Kieser, T.B.M.J., Bibb, M.J., Buttner, M.J., Chater, K.F. and Hopwood, D.A., Practical Streptomyces Genetics. 2000. *Norwich: John Innes Foundation Google Scholar*.

Kim, D., Bae, S., Park, J., Kim, E., Kim, S., Yu, H.R., Hwang, J., Kim, J.I. and Kim, J.S., 2015. Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nature methods*, *12*(3), pp.237-243.

Kim, D., Kim, S., Kim, S., Park, J. and Kim, J.S., 2016. Genome-wide target specificities of CRISPR-Cas9 nucleases revealed by multiplex Digenome-seq. *Genome research*, *26*(3), pp.406-415.

Kleinstiver, B.P., Prew, M.S., Tsai, S.Q., Topkar, V., Nguyen, N.T., Zheng, Z., Gonzales, A.P., Li, Z., Peterson, R.T., Yeh, J.R.J. and Aryee, M.J., 2015. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature*, *523*(7561), p.481.

Kleinstiver, B.P., Pattanayak, V., Prew, M.S., Tsai, S.Q., Nguyen, N.T. and Joung, J.K., 2016. 731. High-Fidelity CRISPR-Cas9 Nucleases with No Detectable Genome-Wide Off-Target Effects. *Molecular Therapy*, *24*, p.S288.

Kumar, V., Bharti, A., Gusain, O. and Bisht, G.S., 2010. An improved method for isolation of genomic DNA from filamentous actinomycetes. *J Sci Eng Tech Mgt*, *2*, pp.10-13.

Lee, S.Y., Kim, H.U., Park, J.H., Park, J.M. and Kim, T.Y., 2009. Metabolic engineering of microorganisms: general strategies and drug production. *Drug Discovery Today*, *14*(1), pp.78-88.

Li, Y., Lin, Z., Huang, C., Zhang, Y., Wang, Z., Tang, Y.J., Chen, T. and Zhao, X., 2015. Metabolic engineering of Escherichia coli using CRISPR–Cas9 meditated genome editing. *Metabolic engineering*, *31*, pp.13-21.

Logares, R., Bråte, J., Heinrich, F., Shalchian-Tabrizi, K. and Bertilsson, S., 2009. Infrequent transitions between saline and fresh waters in one of the most abundant microbial lineages (SAR11). *Molecular biology and evolution*, *27*(2), pp.347-357.

Luo, M.L., Leenay, R.T. and Beisel, C.L., 2016. Current and future prospects for CRISPR-based tools in bacteria. *Biotechnology and bioengineering*, *113*(5), pp.930-943.

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J., Wolf, Y.I., Yakunin, A.F. and Van Der Oost, J., 2011. Evolution and classification of the CRISPR-Cas systems. *Nature reviews. Microbiology*, *9*(6), p.467.

Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J., Charpentier, E., Haft, D.H. and Horvath, P., 2015. An updated evolutionary classification of CRISPR–Cas systems. *Nature Reviews. Microbiology*, *13*(11), p.722.

Mojica, F.J., Díez-Villaseñor, C., Soria, E. and Juez, G., 2000. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Molecular microbiology*, *36*(1), pp.244-246.

Morris, R.M., Rappé, M.S., Connon, S.A. and Vergin, K.L., 2002. SAR11 clade dominates ocean surface bacterioplankton communities. *Nature*, *420*(6917), p.806.

Myronovskyi, M., Rosenkränzer, B. and Luzhetskyy, A., 2014. Iterative marker excision system. *Applied microbiology and biotechnology*, *98*(10), pp.4557-4570.

Nelson, C.E., Hakim, C.H., Ousterout, D.G., Thakore, P.I., Moreb, E.A., Rivera, R.M.C., Madhavan, S., Pan, X., Ran, F.A., Yan, W.X. and Asokan, A., 2016. In vivo genome editing improves muscle function in a mouse model of Duchenne muscular dystrophy. *Science*, *351*(6271), pp.403-407.

Peters, J.M., Silvis, M.R., Zhao, D., Hawkins, J.S., Gross, C.A. and Qi, L.S., 2015. Bacterial CRISPR: accomplishments and prospects. *Current opinion in microbiology*, *27*, pp.121-126.

Pourcel, C., Salvignol, G. and Vergnaud, G., 2005. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology*, *151*(3), pp.653-663.

Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A. and Zhang, F., 2013. Genome engineering using the CRISPR-Cas9 system. *Nature protocols*, *8*(11), pp.2281-2308.

Ran, F.A., Hsu, P.D., Lin, C.Y., Gootenberg, J.S., Konermann, S., Trevino, A.E., Scott, D.A., Inoue, A., Matoba, S., Zhang, Y. and Zhang, F., 2013. Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell*, *154*(6), pp.1380-1389.

Ran, F.A., Cong, L., Yan, W.X., Scott, D.A., Gootenberg, J.S., Kriz, A.J., Zetsche, B., Shalem, O., Wu, X., Makarova, K.S. and Koonin, E.V., 2015. In vivo genome editing using Staphylococcus aureus Cas9. *Nature*, *520*(7546), pp.186-191.

Rappé, M.S., Connon, S.A., Vergin, K.L. and Giovannoni, S.J., 2002. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature*, *418*(6898), p.630.

Reisch, C.R. and Prather, K.L., 2015. The no-SCAR (Scarless Cas9 Assisted Recombineering) system for genome editing in Escherichia coli. *Scientific reports*, *5*, p.15096.

Ren, X., Sun, J., Housden, B.E., Hu, Y., Roesel, C., Lin, S., Liu, L.P., Yang, Z., Mao, D., Sun, L. and Wu, Q., 2013. Optimized gene editing technology for Drosophila melanogaster using germ line-specific Cas9. *Proceedings of the National Academy of Sciences*, *110*(47), pp.19012-19017.

Sanger, F., Nicklen, S. and Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, *74*(12), pp.5463-5467.

Shah, S.A., Erdmann, S., Mojica, F.J. and Garrett, R.A., 2013. Protospacer recognition motifs: mixed identities and functional diversity. *RNA biology*, *10*(5), pp.891-899.

Shan, Q., Wang, Y., Li, J., Zhang, Y., Chen, K., Liang, Z., Zhang, K., Liu, J., Xi, J.J., Qiu, J.L. and Gao, C., 2013. Targeted genome modification of crop plants using a CRISPR-Cas system. *Nature biotechnology*, *31*(8), pp.686-688.

Siegl, T., Petzke, L., Welle, E. and Luzhetskyy, A., 2010. I-SceI endonuclease: a new tool for DNA repair studies and genetic manipulations in streptomycetes. *Applied microbiology and biotechnology*, *87*(4), pp.1525-1532.

Slaymaker, I.M., Gao, L., Zetsche, B., Scott, D.A., Yan, W.X. and Zhang, F., 2016. Rationally engineered Cas9 nucleases with improved specificity. *Science*, *351*(6268), pp.84-88.

Song, J., Oh, H.M. and Cho, J.C., 2009. Improved culturability of SAR11 strains in dilution-to-extinction culturing from the East Sea, West Pacific Ocean. *FEMS microbiology letters*, *295*(2), pp.141-147.

Staley, J.T. and Konopka, A., 1985. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Reviews in Microbiology*, *39*(1), pp.321-346.

Sternberg, S.H., Redding, S., Jinek, M., Greene, E.C. and Doudna, J.A., 2014. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature*, *507*(7490), p.62.

Stewart, E.J., 2012. Growing unculturable bacteria. *Journal of bacteriology*, *194*(16), pp.4151-4160.
Sun, Z., Huang, Y., Wang, Y., Zhao, Y. and Cui, Z., 2014. Potassium hydroxide-ethylene diamine tetraacetic acid method for the rapid preparation of small-scale PCR template DNA from actinobacteria. *Molecular Genetics, Microbiology and Virology*, *29*(1), pp.42-46.

Tong, Y., Charusanti, P., Zhang, L., Weber, T. and Lee, S.Y., 2015. CRISPR-Cas9 based engineering of actinomycetal genomes. *ACS synthetic biology*, *4*(9), pp.1020-1029.

Tsai, S.Q., Zheng, Z., Nguyen, N.T., Liebers, M., Topkar, V.V., Thapar, V., Wyvekens, N., Khayter, C., Iafrate, A.J., Le, L.P. and Aryee, M.J., 2015. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature biotechnology*, *33*(2), p.187.

Waltz, E., 2016. Gene-edited CRISPR mushroom escapes US regulation. *Nature*, *532*(7599), p.293.

Wang, T., Wei, J.J., Sabatini, D.M. and Lander, E.S., 2014. Genetic screens in human cells using the CRISPR-Cas9 system. *Science*, *343*(6166), pp.80-84.

Weber, T., Charusanti, P., Musiol-Kroll, E.M., Jiang, X., Tong, Y., Kim, H.U. and Lee, S.Y., 2015. Metabolic engineering of antibiotic factories: new tools for antibiotic production in actinomycetes. *Trends in biotechnology*, *33*(1), pp.15-26.

Zhang, L., Wang, L., Wang, J., Ou, X., Zhao, G. and Ding, X., 2010. DNA cleavage is independent of synapsis during Streptomyces phage ϕBT1 integrase-mediated site-specific recombination. *Journal of molecular cell biology*, *2*(5), pp.264-275.

Zhang, B., Zhang, L., Dai, R., Yu, M., Zhao, G. and Ding, X., 2013. An efficient procedure for marker-free mutagenesis of S. coelicolor by site-specific recombination for secondary metabolite overproduction. *PloS one*, *8*(2), p.e55906.

Zhou, M., Jing, X., Xie, P., Chen, W., Wang, T., Xia, H. and Qin, Z., 2012. Sequential deletion of all the polyketide synthase and nonribosomal peptide synthetase biosynthetic gene clusters and a 900-kb subtelomeric sequence of the linear chromosome of Streptomyces coelicolor. *FEMS microbiology letters*, *333*(2), pp.169-179.

Chapter Five. Discussion of Research

5.1 Overview of Findings

Proximicins represent an as-yet unexploited chemical repertoire for novel compound discovery, owing to the distinctive 2,4-disubstituted furan group which simultaneously conveys attractive yet synthetically complicated chemistry. Previous work aimed at exploiting this potentially interesting and biologically active chemical scaffold have been hindered by the complex and financially expensive approach associated with heterocycle synthesis. Here, we report the sequencing-based identification of the proximicin biosynthetic cluster permitting the discovery and characterisation of the enzyme putatively responsible for furan group incorporation into the elongating peptide via NRPS biosynthetic enzymes. The adenylation domain responsible for this mechanism contains a previously undescribed structure, refuting previously outlined boundaries thought to confine A-domain specificity and activity. This will inform future bioinformatic studies into NRPS identification and substrate prediction, in addition to providing new guidelines for exploitation of bacterial NRPS enzymology with the aim of novel compound discovery. In the process of this research, other findings became increasingly apparent; primarily, the unambiguous nature of furan incorporation, with it being either pre / post amide bond, and the high level of promiscuity exhibited by Ppb120 adenylation domain capable of furan incorporation. These issues further complicate the ability to validate the route to proximicin biosynthesis; although confirming enzymes likely *capable* of introducing the 2,4-disubstituted furan and novelty in line previous predictions, we cannot be sure this is the actual *in vivo* route. Issues concerning complete biosynthetic route resolution, however, do not negate future exploitation of the uncovered mechanism; the novel enzymology represents a completely unique NRPS A-domain activity and structure, offering a potential molecular tool for innovative compound production.

The techniques and approaches employed in this research were chosen to assess the applicability of cutting edge technical and scientific innovation in combatting the void which currently plagues drug discovery pipelines. The aim was to establish a cost effective and efficient pipeline for microbe genome sequencing and cryptic

NRPS cluster identification, adenylation domain characterisation for biosynthetic route prediction, and gene deletion studies to provide the considered 'gold standard' of gene cluster elucidation. The primary purpose of this research is to inform intelligent novel compound discovery, specifically uncovering biologically active compounds for facile exploitation of their biosynthetic machinery responsible, leading to the ability to construct large compound libraries production. As specific issues encountered in each stage of this research – sequencing, adenylation domain characterisation and CRISPR gene editing – were discussed in preceding chapters, the aim here will be to add context to these findings, explore future research avenues and applications.

5.2 Limitations of Research

At the beginning of this research, a selection of likely potential routes to furan incorporation and hence, proximicin biosynthesis, were developed. The aim was to inform the refinement process of SBC cluster selection after bioinformatic analysis of the two producing *Verrucosispora* strains. The extensively noted similarity between congocidine and proximicin compounds laid the foundations for much of this potential route construction. Initial research supported this hypothesis with analogous mutations appearing in adenylation domains thought to be responsible for heterocycle incorporation in both systems, the theory we had unearthed a structural residue which confers the novel specificity began to form (Juguet et al., 2008). This was disproved with the revision of the congocidine pathway (Al-Mestarihi et al., 2015), shattering previously homology-based biosynthetic routes. It was shown that during congocidine synthesis, the previously identified heterocycle activating A-domain – Cgc18 - was instead responsible for incorporation of a linear molecule, and that another non-NRPS protein – Cgc3* - was responsible for the pyrrole incorporation (Fig. 82a); the perceived structural homology conferring heterocycle activity no longer held any weight. As previously outlined potential biosynthetic routes were in doubt, and due to the identification of the *ppb* cluster in the two producers, efforts were widened to focus on both the potential homologues, as well as novel, routes to heterocycle incorporation (Fig. 82b). In doing this, two opposing theories began to appear regarding the nature of furan group incorporation. The core of these discrepancies are based on the timing of furan synthesis: prior to activation by A-

domain, and hence peptide bond formation or after being activated as a linear molecule, then cyclised. As discussed in Chapter 3, the former theory is supported by previous work on congocidine, and the latter by the presence of cyclases enzymes in the putative cluster. It should be noted that the application of the findings described in the subsequent discussions do not rely on the natural action of these A-domain enzymes, but rather the exploitation of the apparent promiscuity to substrates Ppb120 possesses. Conversely, it should not be ignored that one initial goal of this research was to determine the proximicin biosynthetic route present in *Verrucosispora* spp., and this was not accomplished in its entirety. Although discussed in depth in prior chapters, two specific issues were encountered: i) successful production of A-domain recombinant proteins, consisting of many multifaceted issues and ii) inefficient CRISPR/Cas9 activity. As discussed in Chapter 3, rectification of the former issue would consist initially of addressing issues associated with high GC organism cloning and recombinant expression techniques. This is likely to be facilitated by increased research in these organisms, in combination with improved understanding of solubility issues surrounding adenylation domains. I predict innovation directed to solubility issues will centre on the presence of MbtH-like proteins, proposing that much is still to be uncovered regarding establishing the boundaries to roles these co-proteins play; this notion provides the foundation of discussed future research avenues. In respect to issues concerning inefficient gene editing of *ppb* genes, resolution will likely arise as the significance of microbial applications continues to be realised. It may be noted that CRISPR/Cas9 gene deletion and metabolite studies were not an area of intense investigation; this was due to another collaborating group having the sole research focus of random gene deletions in *Verrucosispora* spp., with the aim of disrupting any secondary metabolite synthesis. They were recently able to delete *ppb* genes outlined in Chapter 2, confirming our predictions by proving it is the gene cluster responsible for proximicin biosynthesis. This work will likely continue to shed light on the order of furan group incorporation, by deleting genes sequentially and analysing the metabolic trace to identify accumulating and absent compounds; unfortunately, this was not in the scope of this research.

As with any research, limitations were encountered, and largely these were overcome. Following a similar tendency to the accidental growth of *Peniciilum* spp.,

on Alexander Fleming's windowsill inadvertently leading to the revolution of medicinal capabilities, the area of this research which holds the most clinical application, was unintentional. The unforeseen ability of Ppb120 to adenylate an array of structurally diverse substrates, provides a potential path to novel compound synthesis. The unconventional domain structure present, also conveys important information regarding structure-activity relationships within these complex enzymes. The following sections will discuss the potential these discoveries possess, primarily in relation to its utilisation in the introduction of novel chemistry into other chemical scaffolds with the aim of novel antibiotic compound discovery.

5.3 Application of Findings - Novel heterocycle-containing compounds

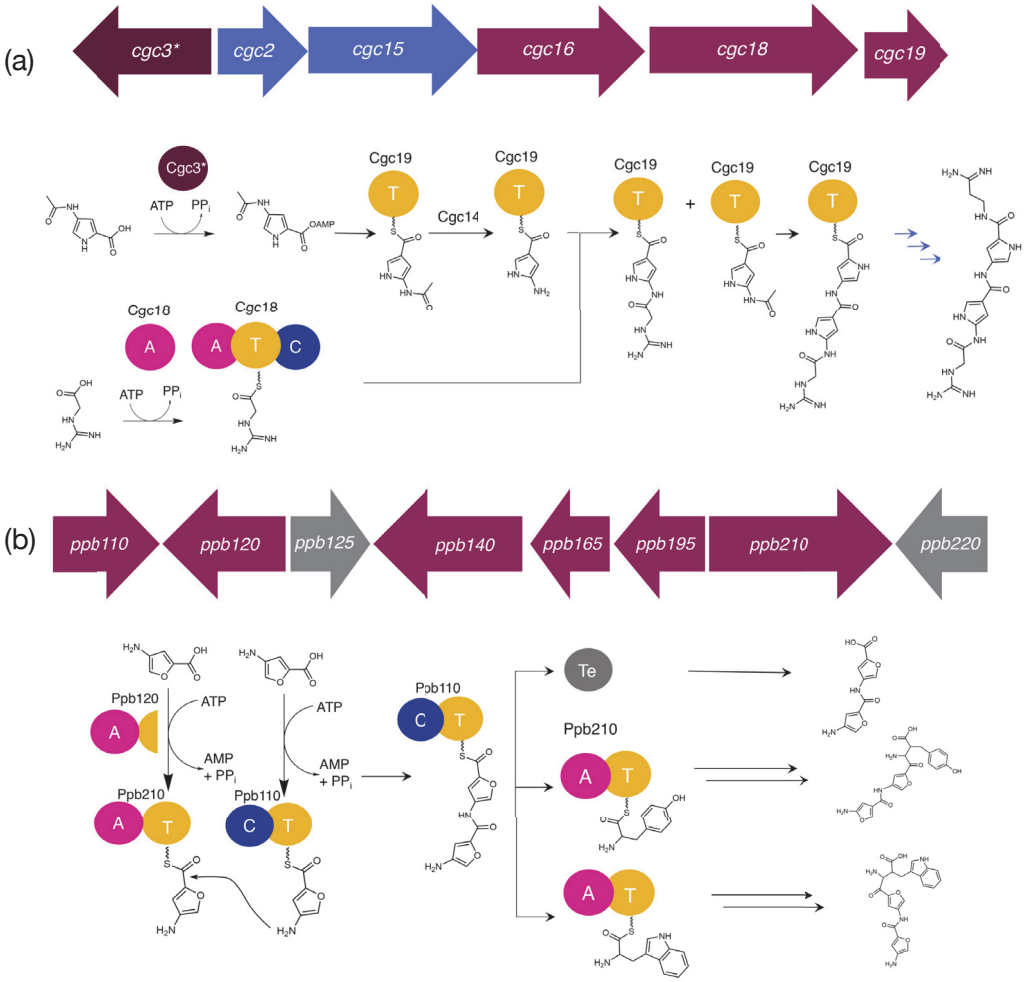
Heterocycles are a common feature in naturally occurring biologically active compounds, often shown to be essential and responsible for conferring the specific action. This ability has made them the target of much interest; however, the complexity associated with their synthesis has hindered the common application of some heterocycles in combinatorial chemistry approaches to compound production. An example of this is the 2,4-disubstituted furan group. The furan group is a common heterocycle found in natural products, however, the appearance of this specific substitution pattern is sparsely recorded. Molecules where it has been reported have diverse biological activities – from antifungal compounds such as flufuran produced by *Aspergillus flavus* to antiviral and anti-inflammatory metabolites from soft coral to scent components produced in butterflies; specific to *Verrucosispora* spp., it conveys DNA-binding independent cell death by disrupting the cell cycle. Interest in these biologically active molecules diminishing is likely owed to the complex synthesis protocols they necessitate; they are considered as poor clinical leads. Here, we outline the gene cluster responsible for 2,4-disubstituted furan synthesis in *Verrucosispora* spp., along with the adenylating enzyme – Ppb120 - responsible for its incorporation into the backbone of the natural peptide family: proximicins. The demonstrated unique activity of Ppb120 presents a facile research avenue for exploitation. All routes are based on utilising the novel enzymology identified here, departing from previously small scale combinatorial chemical synthesis approaches.

Exploiting NRPS systems for novel production is not a particularly recent advance, a large amount of research effort been expended on adapting precursor supply to result in modified products. However, limitations have been met due to the uncompromising substrate flexibility exhibited by native adenylating enzymes. This led to work being preoccupied with adding in new domains, rather than attempting the potentially laborious task of modifying existing ones. Most of the key work done in this area was done by Baltz et al., (2006) during the development of daptomycin, a clinically approved NRPS-synthesised cyclic antibiotic. Initial success for daptomycin derivative synthesis (Baltz et al., 2006; Coeffet-Le Gal et al., 2006; Miao et al., 2006), has led to investigation into other modifications; these are discussed below along with potential application to novel heterocycle-containing compound production.

5.3.1 Ppb120 promiscuous activity

Previous work centred on utilising NRPS systems for novel compound production, such as that on daptomycin by Baltz et al. (2006) are typically hindered by the strict activity of the A-domains; this high level of specificity towards substrates exhibited by NRPS adenylation domains is one of their characteristic features. In this respect, the identified Ppb120 adenylation domain capable of activating different heterocycle substrates is unusual. We demonstrated here the ability of Ppb120 to activate pyrrole containing heterocycles with the same substitution pattern exhibited by the predicted native furan containing precursor. This novel promiscuity will likely allow pyrrole/furan compounds to be incorporated into the peptide chain, without having to produce a recombinant gene cluster; by eliminating precursor-synthesising genes and feeding into the reaction the pyrrole containing analogue of the native precursor, it would likely result in the production of pyrrole/furan hybrids. This presents a potentially straightforward initial research avenue to assess the activity these two opposing heterocycles – furan and pyrrole – confer. Previous work by Wolter et al., (2009) focused on producing congocidine/proximicin hybrids, was only concerned with exchanging the terminal groups, not the nature of the heterocycle core. This preliminary approach would likely direct future research attempts by

Or have another function which would likely help us elucidate their function in NRPS systems uncovering the activity these specific groups confer, informing future research directions. This method would overcome issues regarding the production



(c)

Protein	Domains	Adenylating Activity
Cgc3*	Non-NRPS A-domain (A)	ATP PPi → Cgc19 (T)
Cgc18	A T C	ATP PPi → T
Ppb120	A T	ATP PPi → Ppb210 (T)
Ppb210	A T	ATP PPi → T

Figure 82. Overview of enzymes involved in proximicin and congocidine biosynthesis which pose attractive opportunities for exploitation. (a) Genome organisation of *cgc* cluster showing NRPS genes (pink), tailoring genes (blue) and non-NRPS adenylating enzyme (dark pink) and the proposed route to congocidine biosynthesis (Al- Mestarihi et al., 2015) **(b)** Genome organisation of *ppb* cluster showing NRPS proteins (pink) and associate proteins (grey) and proposed route for proximicin biosynthesis **(c)** a summary of the adenylating enzymes and the reactions they catalysed discussed in future applications.

of complex furan and pyrrole-containing compounds via organic chemistry synthesis routes. However, as pyrrole is not the native precursor involved in proximicin synthesis, this would likely interfere with application of this approach on a wide scale basis; the enzyme activity would be detrimentally altered resulting in low yields of desired compounds, this obstacle was demonstrated here by the slow activity rate of Ppb120 towards the pyrrole precursor. Because of this, more reliable and robust methods are outlined, including the exploitation of the protein responsible for pyrrole incorporation in congocidine production being incorporated into the proximicin pipeline, as discussed below. It should be noted that all the routes and compounds produced are the result of manipulation of those involved in only two NRPS biosynthetic pathways – proximicin and congocidine, a minute fraction of the enzymes available for exploitation (Fig. 82). These two SMC were chosen as (i) previous work has already been focused on proximicin/congocidine hybrids; (ii) heterocycle containing compounds typically have biological activity, and (iii) some enzyme in each cluster carry out similar activities but in very different ways, and so by comparing them it will inform A-domain structure/activity study. For sake of clarity, the adenylating enzymes responsible for core groups in both congocidine and proximicin are summarised along with the reactions they catalyse in Figure 82.

5.3.2 Exchanging NRPS subunits

The general workflow of this approach proceeds by deleting genes known for incorporation of a precursor and complemented *in trans* with another entire NRPS gene consisting of entire A-, T-, C-domains, under the control of a strong promoter; if the new addition requires a rare precursor, this is then fed in. Initial studies exchanging whole NRPS proteins entirely focused on the replacement with homologous subunits from similar pathways to help minimise any interference with subunit/subunit interactions within the multimeric structure. Baltz et al., (2006) carried out proof of concept experiments using the structural similarity between lipopeptide antibiotics. Concluding genes responsible for the additional of the final two amino acid units in daptomycin and A54145, were swapped resulting in the production of daptomycin derivatives. One reason denoting such success of this initial research was the presence of two similarly active adenylation domains – loading either Glu or 3mGlu - being chosen, theoretically resulting in minimal disruption to the protein subunit structure (Miao et al., 2006). The high level of structural similarity exhibited by congocidine and proximicin, in addition to the well characterised *cgc* gene cluster responsible for biosynthesis, provides an attractive starting point for similar combinatorial biosynthetic approaches which are outlined in Figures 83a & b., which could target the core heterocycle chemistry or tailoring groups. With the aim of further reducing the obstruction of cross-subunit connections, the last NRPS subunit involved in proximicin biosynthesis should be chosen, negating any downstream interactions that could affect activity (Figure 83b). In Chapter 3, the importance of the linker regions in the activity and solubility of Ppb A-domains became increasingly clear, and was a major hurdle encountered for A-domain activity studies. This demonstrated reliance on linker regions suggests this method, in which linker regions would be completely exchanged, would not be tolerated by the *ppb* cluster and complete abolishment in any production would ensue. Although presenting an exciting research opportunity, subtle revisions would likely avoid these issues, and hence be prioritised.

5.3.3 Module and domain exchanges

After demonstrating the initial lack of resilience of some NRPS proteins to tolerate subunit exchanges, more refined approaches began to emerge. Research focused on replacing either individual domains (e.g. A) or modules (e.g. C-A-T) from other

regions of the same cluster, followed by non-native exchanges in place of entire subunits/genes, to result in reduced impact on linker regions and downstream interactions. Initial studies conducted by Nguyen et al., (2006), again using daptomycin as a trial system, demonstrated the ability to replace C-A-T from one module with that of another, leading to easily predictable and detectable changes to the produced peptide. These smaller alterations were better tolerated and, although resulting a slight decrease in yield, permitted novel compound production (Nguyen et al., 2006). Domain exchanges could be used in proximicin studies; Fiedler et al., (2008) demonstrated that the anticancer and antimicrobial abilities of the proximicin family was not uniform, for example only B exhibits antimicrobial activity. This added capability exhibited by proximicin B must be due to the presence of an additional terminal tyrosine, hence it is logical to assume that alteration of this would have a substantial effect on its resultant biological activity (Fig. 83c). Further, novel chemistry incorporation could be assessed, such as groups from the *cgc* gene cluster (Fig. 83c). An interesting application would be the introduction of the apparently surplus adenylation domains found in the neighbouring cluster (Fig. 83d). This would be an interesting avenue of research; the resultant proximicin derivatives would likely resemble previously produced compounds if *ppb* does represent a relic of a previously larger ancient SBC. This partial NRPS system, regardless of origin, was naturally selected and hence, it is logical to assume that it did/does aid in a competitive advantage to its producer – a completely novel action could be revealed upon its reconstitution. This approach could also be utilised to build on previous work aimed at proximicin/congocidine derivative synthesis, circumventing the previously outlined laborious synthesis task associated with the combinatorial chemistry approach (Wolter et al., 2009). Domains responsible for the addition of terminal groups could be swapped within and between clusters, resulting in similar derivatives by an increasingly facile method (Fig. 83c). More extreme modifications could be envisioned, including the domains responsible for the core heterocycle incorporation which could be exchanged with that of congocidine, producing furan/pyrrole hybrids (Fig. 83c). Similar heterocycle core modification studies have been done successfully, resulting in the production of activity-altered analogues (Nguyen et al., 2006). This approach could be applied in the approach for the systematic altering of pyrrole groups in congocidine to furans, by replacing *cgc3** with *ppb120*, and testing biological activity. For equivalent studies in proximicins, the unique level of

promiscuity exhibited by Ppb120 could be exploited. It was shown here that, despite low efficiency, the Ppb120 adenylation domain can activate a 2,4-disubstitued pyrrole in place of the native furan. Hence, simply feeding of the pyrrole precursors would likely result in the production of furan/pyrrole hybrids, without any cloning steps.

5.3.4 Deleting or repeating modules

Alterations to yield novel activity are not restricted to inserting chemistry possessed by other clusters or organisms; an alternative route to be explored is the alteration of the peptide chain length. Deleting core substrate-incorporating module has been shown to result in reduction of resultant compound ring size (Mootz et al., 2002), and the opposite – addition of modules – has led to the increased size (Butz et al., 2006) (Further explored - Appendix X). As we predict Ppb120 acts iteratively to add heterocycle groups, deleting this enzyme would likely terminate proximicin production entirely, and so this is not an application which could be used. Using the same notion, due to the repetitive addition of substrates, it would likely be more successful to determine the controlling mechanism for heterocycle addition and then alter this to result in additional precursor incorporation. As many biologically active compounds rely on the presence of heterocycles for their function, it would be interesting to establish whether additional heterocycles increase this activity? It has been demonstrated that chain length in congocidine production is controlled by competition between acetylation of the N terminal amine (thus preventing further condensation and peptide bond formation) and the further condensation reactions. No homologous enzyme is present in the *ppb* cluster - there is no obvious acetylase, so what controls chain length? One hypothesis could be that the donor portion of the condensation domains has such high specificity, resulting in the termination of furan addition at a certain chain length/ As mentioned, altering the chemistry of terminal groups present in proximicin could yield novel activity, the degree of this could be further investigated by increasing the number of additional terminal groups. The adenylation domain protein Ppb210, predicted to be responsible for tyrosine or tryptophan addition in proximicin B and C targeted. Incurring issues similar to *ppb120*, *ppb210* if deleted would just result in proximicin A, and so not a new compound. However, it could be duplicated and the corresponding effect on activity measured (Fig. 83e).

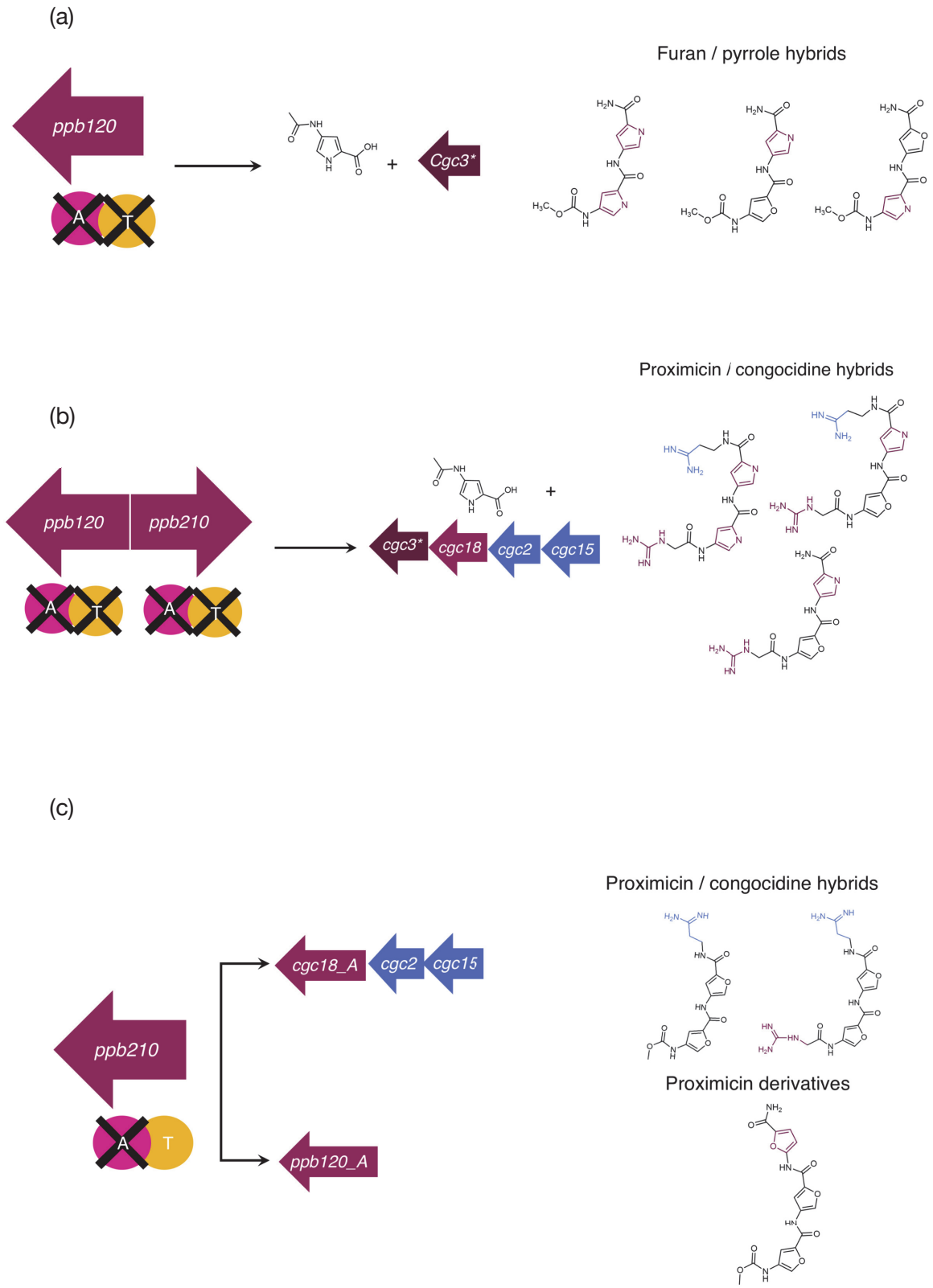


Figure 83. Continues on next page.

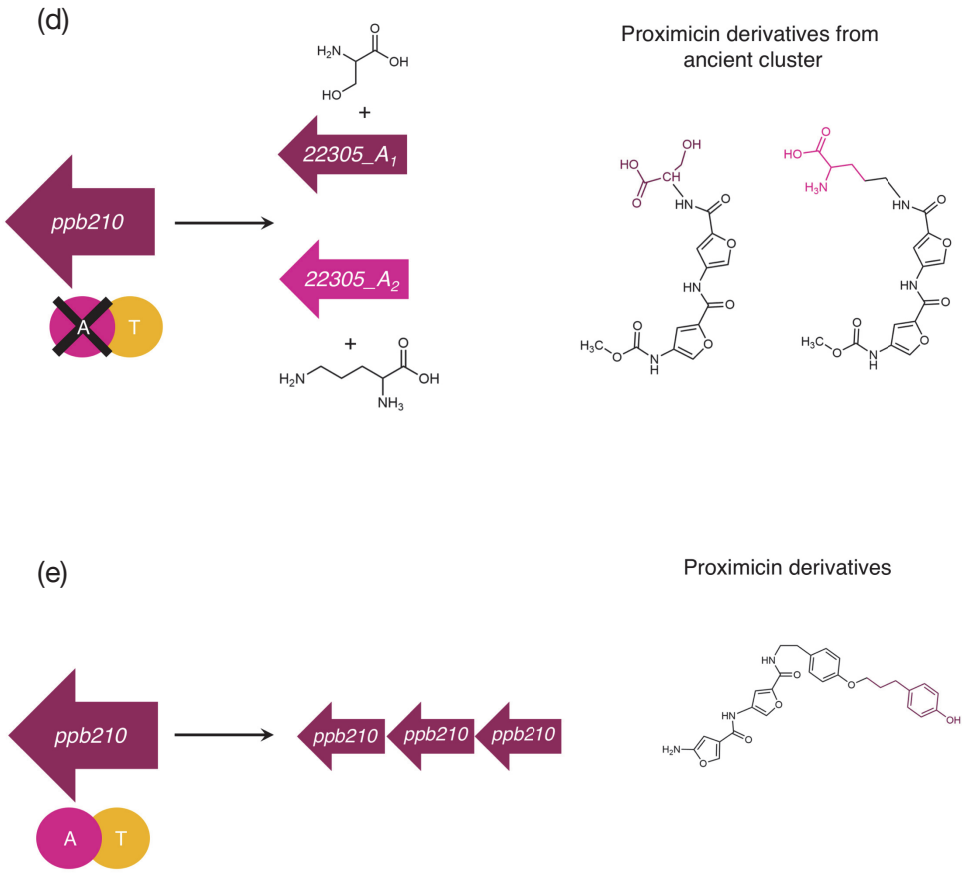


Figure 83. Potential routes to novel heterocycle-containing compound production based on the exploitation of NRPS enzymes involved in proximicin and congocidine biosynthesis. (a) proximicin derivatives containing pyrrole heterocycle core in place of the furan, by deleting entire furan incorporating enzyme – Ppb120 – and replacing with pyrrole activating Cgc3*. **(b)** congocidine/ proximicin hybrids by, in addition to that described in (a), also the deletion of other genes responsible for the addition of other peptides e.g. ppb210 – proposed to activated tyrosine and tryptophan in proximicin biosynthesis and replace with Cgc18 known to incorporate GA. **(c)** A increasingly subtle approach to core heterocycle alteration – deleting just the A-domain region of *ppb120* – not the entire protein as in (a). This would again produce furan/pyrrole hybrids; this is illustrated along with incorporation of other *cgc* genes involved tailoring of the congocidine molecule. **(d)** replacement of *ppb120* A-domain with the A-domains present in the downstream cluster. **(e)** repetition of A-domains in different areas of the cluster, resulting in multiple precursor incorporation and increased peptide chain lengths. All reactions include a control of the deleted gene to ensure normal biosynthesis is reinstated.

5.3.5 Altering specific residues for novel activity

The previously outlined methods to novel compound synthesis, despite steps taken in its reduction, are continually blighted by issues surrounding inactivity of pathways post modification. That reported here with issues regarding activity and solubility of just A-domain region of NRPS proteins, in combination with previous research, demonstrates the importance of surrounding regions to successfully create active native and hybrid NRPS assembly lines, respectively (Doekel et al., 2008; Yu et al., 2013; Beer et al., 2014). Module-module linker regions play an important role in NRPS systems and overlooking their role typically results in truncated peptides as well as products lacking cyclization and other modifications (Butz et al., 2006). Continual extensive research into the relationships and conformational roles these linker regions play, is slowly building a picture describing the full extent of their function, and hence, how they can be manipulated. However, routes to novel compound synthesis not dependent on altering these regions would be advantageous. One potential tactic would be to utilise the knowledge garnered concerning substrate-specificity residues to design A-domains with either relaxed or altered activity, exploiting the pathway to implement the carefully chosen changes introduced. This would likely overcome previously encountered issues as modification all arise internally, with no direct contact with other modules and linker regions occurs. Previous work has utilised a similar initiative (Crusemann et al., 2013; Kries et al., 2015); with Han et al., (2012) showing the directed evolution of a promiscuous enzyme towards a specific substrate by identifying likely successful mutations by comparing the A-domains active site with that of others which activate the desired substrate. Adenylation domain engineering via site directed mutagenesis has also been shown to alter the specificity of an A-domain to a completely different substrate; Kries et al., (2014) conducted an elegant experiment in which the core residues known to be responsible for adenylation domain specificity were mutated sequentially in the Phe-activating GrsA adenylation domain, resulting in complete alteration of its activity profile.

The novel adenylation domain discovered and characterised here – Ppb120 – contains mutations in motifs residues previously considered to be intolerable, which undoubtable contributes to its ability to adenylate a unique heterocycle-containing

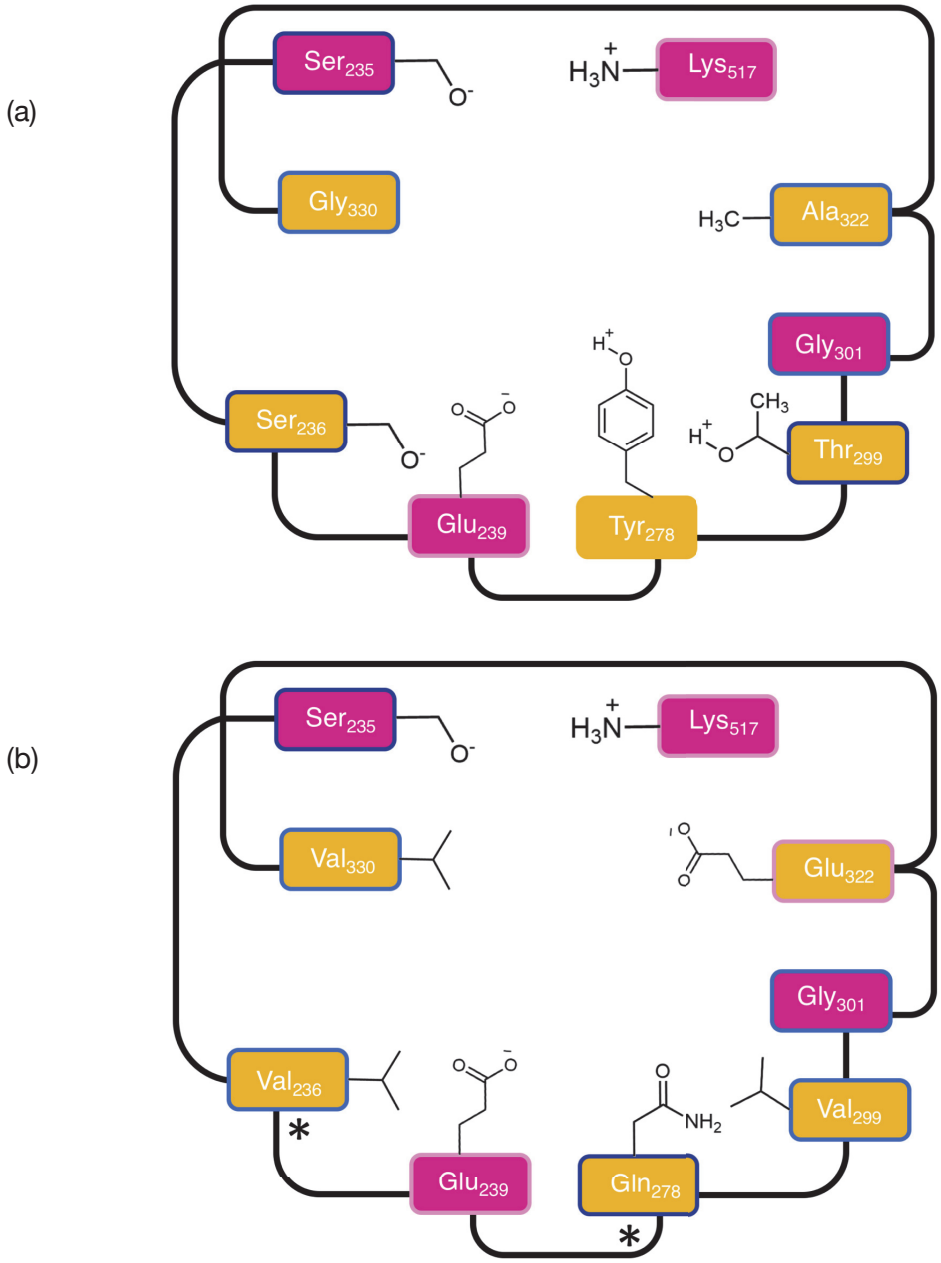


Figure 84. Potential routes for resolving binding residues responsible for novel heterocycle incorporation. (a) the active site of *ppb120* adenylation domain **(b)** the active site of *cgc18* adenylation domain. These two A-domains are very similar in two respects (i) the variant amino acid (yellow) residue structure – non-polar barrel (light blue border) ended with acidic (pink border) and polar groups (dark blue border) and (ii) the presence of a novel mutation in in-variant core residue Ser_{235}. They differ in that Cgc18 contains an addition polar and non-polar group at the barrel end, marked with an asterisk. Despite their apparent structural similarities, these two A-domains have very different substrate specificities. To better understand the residues which dictate this difference, these two residues make attractive starting points. The substrate conferring residues, and their amount of variance across A-domains, is based on work done by Stachelhaus et al., (1999); A-domain structure in the style of work done by Challis et al., (2000).

precursor as exhibited here (Fig. 84a). This newly discovered ability for A-domains to maintain activity while possessing these changes in apparently core regions widens the potential pool of alterations which can be attempted. Introducing this mutation, along with other unique aspects of the Ppb120 binding pocket, into other A-domains may alter their activity resulting in novel NRP production. Initial research and comparison of the Cgc18 and Ppb120 A-domain specificity conferring code highlighted a potential residue which strongly supported activity towards heterocycles; this theory was refuted as the congocidine pathway was revised. And so it remains, which features in Ppb120 enables this novel activity, if not this Asp_{235} mutation? Unfortunately, the similarly active Cgc3* AMP-binding protein does not belong to the NRPS sub family, and hence, the substrate specificity conferring core structure is not present, and so little can be garnered from the comparison of Cgc3* and Ppb120 due to the substantial deviation in homology. The in-depth analysis of the Ppb120 enzyme in chapter III highlighted two potential origins: an altered A5 domains, and the presence of the Asp_{235} mutation. One possibility is that the ability to adenylate heterocycles is dependent on a cumulative mutational effect, facilitated by both of these interacting; it follows, that if we successively alter Cgc18 binding pocket (Fig. 84b) towards Ppb120 mimicry, both within the active site initially and then in the distant A5 region we should see the correlated change in activity towards heterocycle containing substrates. When analysing the binding pockets within Cgc18 and Ppb120, in addition to the unique Asp_{235} mutation, they both share a similar overall structure: a non-polar barrel ending with polar and acidic groups. The main difference being Ppb120 contains more polar groups, in comparison to Cgc18; is this disparity enough to dictate such a difference in activities? Or does the distal A5 region contribute a larger proportion of specificity conference than previously established? Point mutations in Cgc18 in regions both within the binding pocket and A5, in combination with activity studies would allow confirmation of this. This would inform what we currently know about the adenylation domain specificity conferring residues, and structure-activity relationships present within A-domain proteins; progressing the fundamental goal of gaining the ability to tailor an A-domain to any substrate incorporation: an invaluable instrument in the ‘molecular toolbox’.

5.4 Application of findings - Large scale implementation

Substantial prospects of research described here build on work previously done, shedding new light on the once set boundaries of chemistry which can be introduced. Proximicin, and its furan containing derivatives, represent potentially limitless opportunities for novel antimicrobial and anticancer active compounds, not only in the ability to introduce heterocycle chemistry, but to inform mutational studies on A-domain specificity boundaries. However, it is important that we don't get hooked on the same snare as the 'combinatorial chemistry' approach of the late twentieth century which infamously yielded very little. We can gather from that futile era, that simple chemical novelty is not adequate, we need to utilise all the currently available scientific advances to produce intelligently designed compounds which we predict will have activity against a specific target. Implementing the previously discussed avenues for novelty, whilst incorporating information we now know about NRPS structure, residues dictating A-domain specificity, and quick activity assessment techniques, to create feedback loops to ensure effort is not exerted on research routes yielding novel but also biologically active and clinically applicable molecules. New scientific advancements need to be united into strategies specifically designed and integrated for this task. The novel enzymology uncovered here represents a potential source of one aspect of this, the novel chemistry by routes as outlined by the proposed route; however, this must be used in combined with other recent discoveries and advances to allow advancement to gain clinical applicability and relevance. Large scale antimicrobial compound searches are typically hindered in two ways: issues encountered in synthesising novel biologically active compounds, and the costs associated with screening many compounds. How the work described here can be integrated with cell free technology and intelligent design, into a system designed to uncover biologically active heterocycle compounds is outlined below (Fig. 85).

5.4.1 Cell free synthesis approach

Identifying compounds from bacteria with novel antimicrobial activity presents a paradox: moving biosynthesis into a heterozygous host for investigation into the enzymes and system responsible for production, results in host death, halting research. One approach to avoiding this issues, while optimally exploiting the

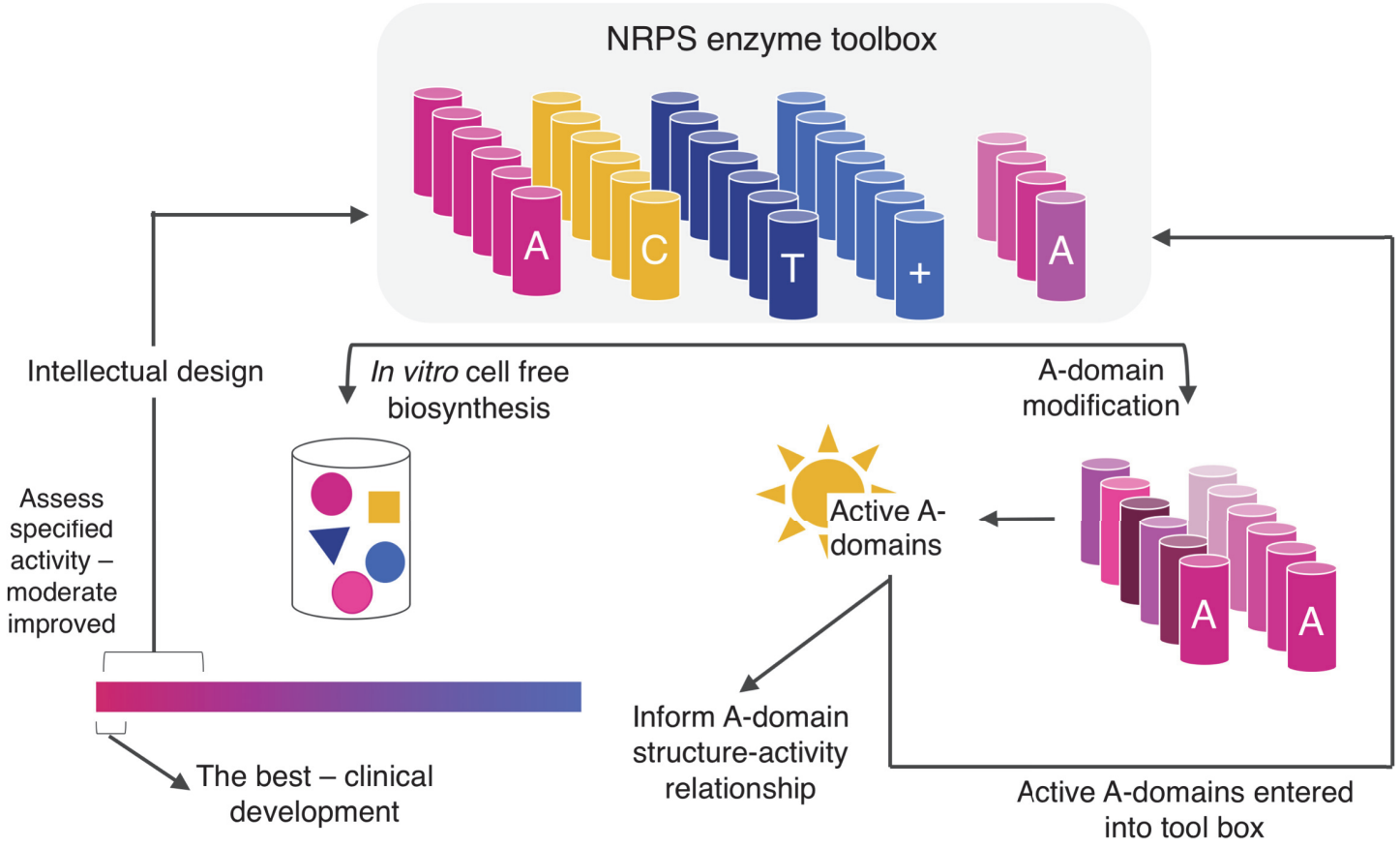


Figure 85. Large scale implementation of the research here to produce the 'NRPS-toolbox'. Characterised NRPS enzymes are purified to produce a cache of purified enzymes. These are then utilised by mixing different combinations together and the resultant compounds tested for biological activity; ones with moderate activity are re-cycled to be improved until high biological activity is achieved and passed on to clinical development. Adding additional novelty to this system, A-domains could be mutated in their specificity conferring residues, tested quickly for activity via the luciferase assay and then added to the toolbox for use in novel compound production.

modular nature of NRPS systems, would be to purify each – potentially specifically modified -individual NRPS, along with associated enzymes and substrates, and mix them together in a cell-free system as outlined in Figure 85. Reliance on cell free systems is increasing as the technology improves in its ability; it is a widely used system which negates issues encountered with the complex interactions found in a whole cell. These systems have demonstrated a surprising ability to tolerate toxic compounds (Wang et al., 2011) and work under a large range of conditions (Zhu et al., 2014). Once synthesised by the manipulated NRPS enzymes *in vitro*, the produced compounds would be quickly assayed for either DNA binding or cell cycle arresting activity, coupled with intelligent-design feedback loops ensuring only increasingly active chemical scaffolds are built up. Exploiting NRPS synthesis to this end would produce a facile and rapid approach to novel compound production; bringing the 'molecular toolbox' representation into reality, by creating libraries of well characterised NRPS as well as tailoring enzymes, checked for activity and ready for cell-free novel compound synthesis. The combination of NRPS enzymes and hence hybrid compounds would give an almost limitless ability for biologically active compound elucidation.

5.5 Learning from past mistakes: embracing novelty

In the process of envisaging applications of the novel enzymology discovered here, a bottleneck in innovation was anticipated: novel enzymes being overlooked due to the currently strict rules dictating A-domain structure. Here, we have shown that a functioning A-domain exists outside of the confines previously established to dictate A-domain activity. Assessment of the likelihood of a Ppb120 representing an operational A-domain, prior to the work done here, would have most likely suggested this enzyme was incapable of activity. No substrates were predicted by any bioinformatical program, it contained mutations in core specificity conferring residues and an essential motif was missing. These assumptions rely on previous extensive work into other A-domains, multi-sequence alignments and the identification of regions which are present in all active, and lacking in all inactive enzymes. For the most part, this approach works well; that is until enzyme novelty is being sought. The paradox of isolating new enzymes based on the structure and homology of other identified members of the same class, directly exerts restrictions on the level of

enzyme novelty which can be recognised. In attempts to halt the looming resistance crisis, all novelty created by nature must be embraced; we must implement approaches with this concept at the forefront. Here we identify a completely novel enzyme, and its huge range of potential exploitation for new antimicrobial drug discovery; however, a singular new group is not sufficient. We need to gather and exploit as many of these unusual, non-conforming enzymes to be used in this application. And hence, the catch twenty-two: how do we identify novel enzyme chemistry in the huge expanse of DNA yielded from sequencing experiments, if by its very nature, we don't know what we are looking for, or have any homologs to utilise?

5.5.1 Informing future NRPS cluster bioinformatic searches

Here, I propose a novel approach to NRPS identification: utilising the co-proteins – MbtH-like – to identify NRPS systems, in strict opposition to previous bioinformatical approaches utilising pre-established core domains and motifs. These previous methods implemented by modern bioinformatical programs, use homology to sequences established to be present in currently characterised NRPS enzymes, fundamentally preventing the identification of novel enzymes. MbtH-like proteins are proteins found to be present and essential in many NRPS systems, being involved in A-domain activity and solubility. Although the exact nature of their role has yet to be established, they all exhibit high homology with the first identified – MbtH – hence the name MbtH-like. I propose an A-domain identification search based on the presence of MbtH proteins within the cluster, rather than homology to pre-determined A-domain residue confines. This would allow the potential discovery of completely novel substrate activating A-domains, conferred by their novel structure, to be added to the NRPS molecular toolbox and exploited for novel compound discovery. The necessity for this change in focus for A-domain identification can be demonstrated in the work here; using typical constraints, Ppb120 would have been considered an inactive enzyme and excluded from further development. This is supported by the revision of many previously bioinformatically formed biosynthetic routes after experimental analysis of the A-domains, demonstrating the extent we still have to learn about these complex enzyme schemes. The revision of the predicted congocidine, and hence the proximicin, biosynthetic route by Al-Mestarihi et al., (2015), was the consequence of an unrelated research attempt, and hence largely

Table 21. MbtH guided NRPS cluster identification. NRPS-gene clusters identified by NRPS-related gene motifs, and potential clusters identified by MbtH-guided searching. Many MbtH proteins have been putatively identified as phosphoglycerate mutases (blue) or unknown.

NRPS by core domain identification		NRPS by MbtH presence	
Name	Current function via homology	Name	Current function via homology
NRPS - Arylopolyene	Kedarcidin	NRPS - Arylopolyene	Kedarcidin
NRPS-M	Meilingmycin	NRPS-M	Meilingmycin
NRPS	**Proximicin**	**NRPS**	**Proximicin**
NRPS - T1PKS	Sporolide	NRPS - T1PKS	Sporolide
T1PKS - NRPS	Maduropeptin	T1PKS - NRPS	Maduropeptin
		VM* 1	Uncharacterized part of NRPS-M
		VM 2	**Phosphoglycerate mutase**
		VM 3	**Phosphoglycerate mutase**
		VM 4	**Phosphoglycerate mutase**
		VM 5	No homology
		VM 6	No homology
		VM 7	Squaline cyclase

* *Verrucosispora maris* AB18-032

accidental. If not established, the previously outlined routes would still have been considered accurate, when in reality the apparently similar pathways to congocidine and proximicin, occur in very distinctive ways, exploiting completely different enzymatic machinery. This is just a single example of how the novelty of an enzyme was overlooked, an issue which will only be magnified during large scale implementation of traditional bioinformatical approaches to novel NRPS discovery.

Recent research, and that included here, is continually working towards understanding the role of these simple MbtH-like co-proteins; the core signatures they possess which determine and hence allow the prediction of the nature of their governance over NRPS systems are becoming increasingly apparent. The structure of an A-domains determines the chemistry of its potential activation pool, this uniquely strict reliance on structure-activity relationships makes it possible to assume that MbtH-like proteins will not possess the same extent of dependence on core domains. Simply put, A-domains must retain the ability to act on every potential substrate nature has to offer – a potentially limitless necessity for novelty; MbtH proteins have to only exert activity on this family of proteins, and so search homology

within MbtH proteins will likely not be as detrimental to uncovering novelty in comparison to the analogous approach in A-domains. To initially test the applicability of this approach, the *V. maris* AB18-032 genome was used as an example. The first hurdle to integration is the identification of core MbtH residues to allow for homology searches to be employed – a process which will likely take refinement and large scale multi-alignment analysis of MbtH proteins; for the preliminary work done here, regions of homology between currently identified *V. maris* AB18-032 MbtH proteins were assimilated. Previous annotation by traditional SBC techniques led to the identification of five NRPS systems; in comparison, the MbtH-focused approach uncovered 12, an additional 7 MbtH-core motif containing proteins (Table 21). These identified hypothetical MbtH-like proteins demonstrated a high level of homology with the MbtH-like protein shown to be essential in proximicin biosynthesis – Ppb125 – and the first identified MbtH protein, specifically in regions previously identified to be essential in MbtH function. These proteins must be associated with some biosynthetic pathway or have another function which would likely help us elucidate their function in NRPS systems, or their presence wouldn't be maintained. A cursory look at these highlighted, identified a region of *V. maris* AB18-032 genome which were surrounded by other genes typically associated with SMB clusters such as – transport, tailoring and transcriptional regulator proteins. To further confirm the applicability of this approach, these regions require further investigation for even minor similarity to NRPS or other biosynthetic enzymology. As no currently available integrated bioinformatic tools actively search for novelty in A-domains, oppositely searching for homology to allow substrate prediction, this would have to be done by hand. Initially, genes should be prioritised on their sequence similarity to the ANL super family of enzymes – as for an A-domain to be active, it has to retain its adenylating activity justifying this ranking. And then other NRPS components searched for i.e. potential T, Te and C domains. Although this may highlight enzymes which have MbtH similarity randomly, or are a remnant of an ancient and now obsolete pathway, its ability to identify potentially unique A-domains for novel compound production, outweighs these issues. Its application is not limited here, as further investigation into the MbtH-like proteins may give insights into the presence of multiple, or currently unidentified proteins. One example of this is the identification of an MbtH protein downstream of the NRPS-M cluster (Table 21)– if this cluster was to

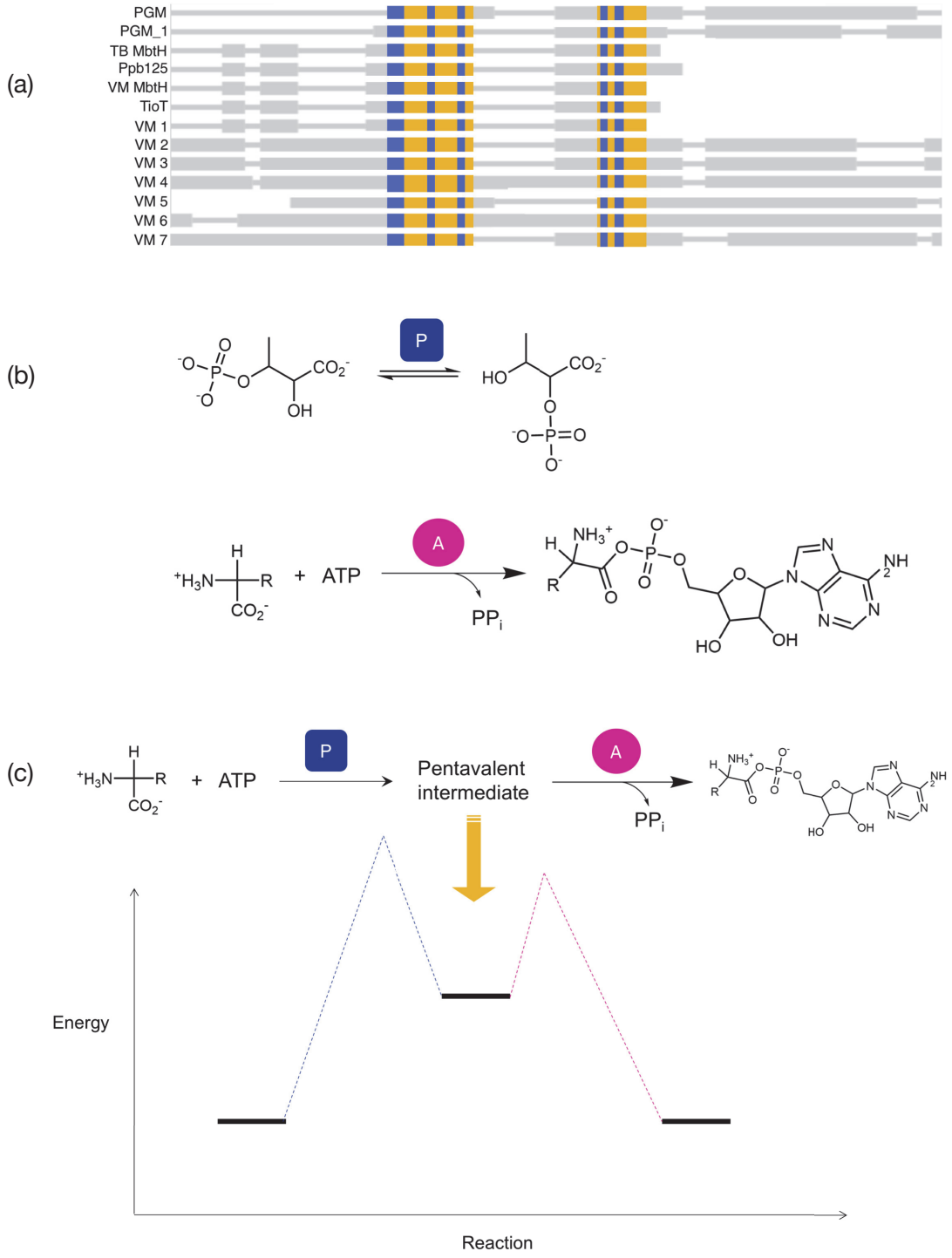


Figure 86. Potential homology between MbtH proteins and Phosphoglycerate mutases (PGM). (a) Multiple sequence alignment of characterised PGM enzymes and MbtH proteins, as well as those identified here: moderate similarity (yellow) and low similarity (yellow). **(b)** reactions catalysed by PGM and adenylation domains know to be helped by MbtH-like proteins. **(c)** potential role of MbtH based on PGM similarities – potentially adding an additional phosphate group creating a lower energy intermediate – pushing the forward adenylation reaction.

be investigated, issues would likely ensue regarding A-domain activity and solubility due to the corresponding MbtH-like protein not being accounted for.

It is not yet established the exact role which is played by MbtH-like proteins, only their requirement for in some NRPS-catalysed synthesis is proven. It is hence interesting to note that of those MbtH-similar enzymes which have been previously given a putative function, they are typically phosphoglycerate mutases (PGM) (Table 21). MbtH-like proteins and this family share moderate sequence homology, across entire sequences this is typically ~20%, however, these the similarity is seen in distinct regions maintain > 60% homology, opposed to being across the entire protein – suggesting some homologous structures (Fig. 86a). Phosphoglycerate mutases are an extremely highly conserved enzyme family which catalyse the isomerization of phosphoglycerate substrates (Fig. 86b), a process essential for glucose metabolism. It is interesting to note that this reaction involves the movement of a phosphate group within a molecule by first adding a phosphate and then removing another, which occurs in the same way as adenylating activity – the addition of a phosphate group in the form of a phosphonate – ADP. Could MbtH proteins be involved in adding a second phosphate group, similar to phosphoglycerate mutases, to produce a more reactive, lower activation energy or more suitable substrate for adenylation? If you subscribe to the ‘addition-elimination’ mechanism of phosphoryl transfer, MbtH- like proteins could partake in a similar reaction to its similar PGM proteins. Working as a catalyst to produce a pentavalent intermediate during adenylation activity, acting as a nucleophile resulting in a species which occupies an energy valley – pushing the forward adenylation reaction (Fig. 86c). Although completely hypothetical, this could explain why adenylation domain activity is lowered/terminated when their MbtH counterparts are not present; it does not, however, explain their effect on A-domain solubility. This information, along with further homology searches may give further insight into the role these elusive proteins play. MbtH-like proteins may have been subjected to a similar bottle-neck affect as described here in NRPS systems – strict confines on what symbolises an NRPS protein has hindered novel enzyme discovery. The same could be said for MbtH-like proteins – using stringent homology searches to identify other MbtH proteins we may be inadvertently selecting against the discovery of those which represent early diverging proteins – those dissimilar to what we consider MbtH-like, but descriptive of a putative function, such as PGM. To

harness NRPS specific novelty, these essential enzymes and their function require increased investigation and their similarity to PGM enzymes represents a suitable starting point. It is widely accepted, since the genome sequencing revolution, that microbes possess the ability to produce infinity more SM and other biologically active compounds than once predicted; the approach outlined here presents an innovative means of identifying these unresolved clusters and hence, harnessing their activity.

5.6 Application of findings – Native Cas exploitation from editing target species

Gene editing using the CRISPR/Cas system was reported here for the first time in the Micromonosporaceae family of Actinomycetes; this proof of application study demonstrated the effective deletion of the pigment conferring genes present in *Verrucosispora* sp. str. MG37. Although successful for this particular phenotypically observable gene, problems regarding off-target Cas activity resulting in the widely reported 'cas toxicity', prevented further application of the technique in *Verrucosispora* spp., for more precise gene editing attempts. As previously discussed in Chapter 4, one way of rectifying this issue would be to exploit the CRISPR/cas9 system found in the organism which is the target for gene editing. This is based on the theory that this native system must have acquired a mechanism to avoid affecting host DNA. The potential development of this approach would allow the application of a CRISPR/cas system with increased fidelity for specific use in non-model, genetically complex organisms, such as *Verrucosispora* species.

5.6.1 Native CRISPR/Cas exploitation

The *Verrucosispora maris* AB18-032 genome presents only a single area displaying characteristic features of a CRISPR/Cas loci, denoted in Uniprot as VAB18032_16190 – VAB108032_16250 (UniProt, 2017). This region contains multiple Cas-related genes spanning Cas I-III families; previous native application of this system is represented by 41 foreign spacers. The Cas proteins present show no conserved identity with spCas9 (<8%), the Cas protein most widely exploited for CRISPR/Cas applications, this divergence suggests that *Verrucosispora* spp. Cas (vmCas) proteins have evolved a level of specialization to this high GC, complex

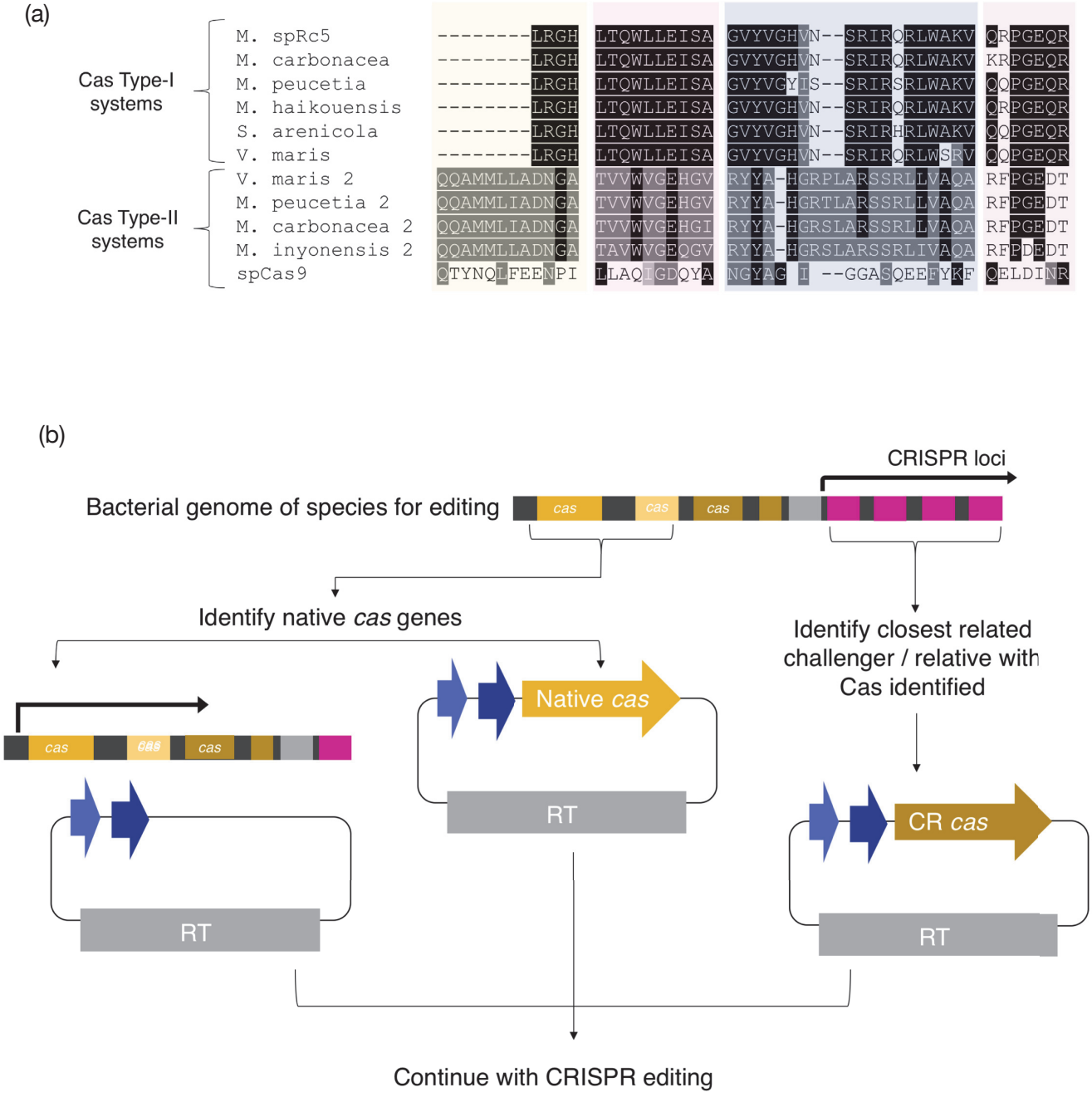


Figure 87. Potential exploitation of native CRISPR/Cas systems for more efficient gene editing. (a) Multiple sequence alignment of Cas genes from both type I and II systems, from Micromonosporaceae family in comparison to the model Cas protein – spCas9, showing the high variance. **(b)** Workflow of native *cas* genes for gene editing: they still rely on the synthesis of a synthetic guide RNA (blue arrows) but differ from other approaches as they (i) increase the expression of native cas genes or (ii) add the native cas genes on a plasmid under the control of a strong promoter. Alternatively, if native systems fail to work, (iii) the closest relative which is identified as an alien species should be used – this could be determined by identifying the species already present in the CRISPR loci.

species. This is supported by work shown here, as when sgCas9 is employed in *V. maris* AB18-032 it results in Cas9 induced toxicity. To further understand the evolution of novel *cas* genes, those present in *Verrucosispora* spp., and other Micromonosporaceae family members were compared to that of spCas (Fig. 87a); this demonstrated a very high level of sequence divergence in regions known to be involved in Cas activities. This was demonstrated in Cas genes even in the same subclass as spCas9. From this we can predict that the high level of cas toxicity exhibited in *Verrucosispora* spp., is due to the lack of native pathways which recognise the foreign spCas9 protein to allow prevention of off-target endonuclease activity. To circumvent this, one approach would be to lower the overall concentration of heterologous Cas proteins, however this would, although favourably lowering off-target Cas activity, also detrimentally affect targeted endonuclease action and hence, gene editing efficiency. Because of this paradox, I outline here a novel method in Figure 87b, of utilising the hosts own Cas proteins, in addition to synthetic components such as sgRNA, to allow efficient gene editing with the expectation that native Cas proteins will not act on host DNA to such an extent due to DNA protection pathways. This would require identifying the native *cas* genes within the host, then either increasing their expression by altering their promoter control or by placing it on a CRISPR/Cas designed plasmid in place of the spCas9 gene. One potential issue emerges when developing this application: native DNA protection pathways may be so effective that the native Cas protein will not act again its own DNA. To resolve this, there are two potential approaches: (i) identifying the closest relative included in the foreign DNA regions present in the CRISPR array as this demonstrates that the host recognises this as foreign and hence it must be dissimilar; or (ii) more simply, find the closest relative with a type-II *cas* gene characterised, and test for activity to confirm they have diverged enough. Although laborious to set the system up, once established this could be applied to all gene editing research in a given bacterial family or species, depending on the level of divergence. I propose this idea of finding ‘the sweet spot’ between reduced Cas toxicity as a result of off-target DNA degradation, and maintaining Cas endonuclease activity towards similar/native DNA. I think that this novel approach is worth investigation to further the somewhat lacking research into bacterial CRISPR/Cas application.

5.7 Conclusions

The requirement for biologically active novel chemistry compounds has never been greater; this exerts a pressure on scientific fields to begin yielding potential antibiotic drugs on a large scale. To accomplish this, we must pool all current technological innovations to ensure that the most efficient and effective approaches are being utilised. Microorganisms have developed the ability to defend themselves against competitors, utilising compounds refined over millennia of evolution; previously, they have been the source of much of what we consider modern medical compounds. Over time, attention on these compounds has dwindled due to apparently more financially lucrative research avenues which have yielded very little. Increased emphasis on returning to the search of novel compounds of microbial origins, has occurred, differing from previous approaches in the utilisation of modern techniques to develop the process.

Here, we demonstrate a facile pipeline starting from the potentially clinically interesting NRPS-secondary metabolite, to whole genome sequencing, assembly and annotation to allow the identification and characterisation of the responsible metabolic pathway. In doing this, novel enzymology was discovered, demonstrating a contradiction to the previously established confines dictating NRPS A-domain activity. To the best of our knowledge, no other A-domain has the ability to activate a 2,4-disubstituted heterocycle group as shown by Ppb120; likely owing to the non-conventional core residue, namely the absence of A5, present in know activity conferring domains. The significance of this research is two-fold: (i) it demonstrates the ability to sequence, assembly and annotate a microbial genome, along with novel SMC identification, in a small scale, academic environment effectively showing that these research applications can be put in the hands of small laboratories; in addition to outlining and refining a freely accessible bioinformatical pipeline for this application (ii) the identification and characterisation of a truly novel adenylation protein which has the ability to be involved in production of almost unlimited novel chemistry containing novel compounds. By assessing the applicability of NGS platforms, and their associated data handling shortcomings, in an academic setting and its ability to yield attractive findings, will hopefully inform other similar research in this area. The approach used here, as well as all the future application of these findings, present an

easily employable approach to the much required clinically important compounds discovery. It has already been established that microbes are likely the best source of this, requiring refinement of current sequencing and assembly methodology tailored to this application; here we outline such an approach. The subsequent research focusing on NRPS pathway resolution provides a framework for future similar work; these enzymatic clusters have already shown their ability to yield clinically important compounds and represent a suitable primary target when whittling down the many SMCs revealed following WGS, to yield a cache of likely the most interesting, to investigate. Approaches described here, including the luciferase assay for A-domain activity and MbtH-guided NRPS identification, could represent important developments to catalyse this approach to new antibiotic discovery, into large scale implementation.

Natures ability to produce compounds with completely novel chemistry and biological action is unparalleled; it will always greatly surpass man's power and capability to do the same. Combinatorial chemistry approaches have failed: we do not possess the ability to conjure up a solution to the looming antibiotic crisis, instead focus must begin to exploit these microbial foes which pose such a great risk to our medically competent civilizations. The research here represents the first step in this endeavour, a path which, optimistically, other research groups and the pharmaceutical giants will follow in, before antimicrobial resistance thrives to a level currently on the horizon, of no return.

5.8. References

Al-Mestarihi, A.H., Garzan, A., Kim, J.M. and Garneau-Tsodikova, S., 2015. Enzymatic evidence for a revised congocidine biosynthetic pathway. *ChemBioChem*, *16*(9), pp.1307-1313.

Baltz, R.H., Brian, P., Miao, V. and Wrigley, S.K., 2006. Combinatorial biosynthesis of lipopeptide antibiotics in Streptomyces roseosporus. *Journal of Industrial Microbiology and Biotechnology*, *33*(2), pp.66-74.

Beer, R., Herbst, K., Ignatiadis, N., Kats, I., Adlung, L., Meyer, H., Niopek, D., Christiansen, T., Georgi, F., Kurzawa, N. and Meichsner, J., 2014. Creating functional engineered variants of the single-module non-ribosomal peptide synthetase IndC by T domain exchange. *Molecular BioSystems*, *10*(7), pp.1709-1718.

Butz, D., Schmiederer, T., Hadatsch, B., Wohlleben, W., Weber, T. and Süssmuth, R.D., 2008. Module extension of a non-ribosomal peptide synthetase of the glycopeptide antibiotic balhimycin produced by Amycolatopsis balhimycina. *ChemBioChem*, *9*(8), pp.1195-1200.

Coeffet-Le Gal, M.F., Thurston, L., Rich, P., Miao, V. and Baltz, R.H., 2006. Complementation of daptomycin dptA and dptD deletion mutations in trans and production of hybrid lipopeptide antibiotics. *Microbiology*, *152*(10), pp.2993-3001.

Crüsemann, M., Kohlhaas, C. and Piel, J., 2013. Evolution-guided engineering of nonribosomal peptide synthetase adenylation domains. *Chemical Science*, *4*(3), pp.1041-1045.

Doekel, S., Coeffet-Le Gal, M.F., Gu, J.Q., Chu, M., Baltz, R.H. and Brian, P., 2008. Non-ribosomal peptide synthetase module fusions to produce derivatives of daptomycin in Streptomyces roseosporus. *Microbiology*, *154*(9), pp.2872-2880.

Han, J.W., Kim, E.Y., Lee, J.M., Kim, Y.S., Bang, E. and Kim, B.S., 2012. Site-directed modification of the adenylation domain of the fusaricidin nonribosomal peptide synthetase for enhanced production of fusaricidin analogs. *Biotechnology letters*, *34*(7), pp.1327-1334.

Juguet, M., Lautru, S., Francou, F.X., Nezbedová, Š., Leblond, P., Gondry, M. and Pernodet, J.L., 2009. An iterative nonribosomal peptide synthetase assembles the pyrrole-amide antibiotic congocidine in Streptomyces ambofaciens. *Chemistry & biology*, *16*(4), pp.421-431.

Kries, H., Wachtel, R., Pabst, A., Wanner, B., Niquille, D. and Hilvert, D., 2014. Reprogramming nonribosomal peptide synthetases for “clickable” amino acids. *Angewandte Chemie International Edition*, *53*(38), pp.10105-10108.

Kries, H., Niquille, D.L. and Hilvert, D., 2015. A subdomain swap strategy for reengineering nonribosomal peptides. *Chemistry & biology*, *22*(5), pp.640-648.

Meyer, S., Kehr, J.C., Mainz, A., Dehm, D., Petras, D., Süssmuth, R.D. and Dittmann, E., 2016. Biochemical dissection of the natural diversification of microcystin provides lessons for synthetic biology of NRPS. *Cell chemical biology*, *23*(4), pp.462-471.

Miao, V., Coëffet-Le Gal, M.F., Nguyen, K., Brian, P., Penn, J., Whiting, A., Steele, J., Kau, D., Martin, S., Ford, R. and Gibson, T., 2006. Genetic engineering in Streptomyces roseosporus to produce hybrid lipopeptide antibiotics. *Chemistry & biology*, *13*(3), pp.269-276.

Milne, C., Powell, A., Jim, J., Al Nakeeb, M., Smith, C.P. and Micklefield, J., 2006. Biosynthesis of the (2 S, 3 R)-3-Methyl Glutamate Residue of Nonribosomal Lipopeptides. *Journal of the American Chemical Society*, *128*(34), pp.11250-11259.

Mootz, H.D., Kessler, N., Linne, U., Eppelmann, K., Schwarzer, D. and Marahiel, M.A., 2002. Decreasing the ring size of a cyclic nonribosomal peptide antibiotic by in-frame module deletion in the biosynthetic genes. *Journal of the American Chemical Society*, *124*(37), pp.10980-10981.

Nguyen, K.T., Ritz, D., Gu, J.Q., Alexander, D., Chu, M., Miao, V., Brian, P. and Baltz, R.H., 2006. Combinatorial biosynthesis of novel antibiotics related to daptomycin. *Proceedings of the National Academy of Sciences*, *103*(46), pp.17462-17467.
H. D. Mootz, N. Kessler, U. Linne, K. Eppelmann, D. Schwarzer and M. A. Marahiel, *J. Am. Chem. Soc.*, 2002, **124**, 10980–10981

UniProt, 2017 http://www.uniprot.org/uniprot/F4F6M3 Accessed 8/09/18

Wang, Y., Huang, W., Sathitsuksanoh, N., Zhu, Z. and Zhang, Y.H.P., 2011. Biohydrogenation from biomass sugar mediated by in vitro synthetic enzymatic pathways. *Chemistry & biology*, *18*(3), pp.372-380.

Yu, D., Xu, F., Gage, D. and Zhan, J., 2013. Functional dissection and module swapping of fungal cyclooligomer depsipeptide synthetases. *Chemical Communications*, *49*(55), pp.6176-6178.

Zhu, Z., Tam, T.K., Sun, F., You, C. and Zhang, Y.H.P., 2014. A high-energy-density sugar biobattery based on a synthetic enzymatic pathway. *Nature communications*, *5*, p.3026.

Appendix A. Primer Table

Table A1. Primer Table with sequences of all utilised primers.

#	Name	Primer sequence in 5'-3' orientation
1	22190F	ATGCAGCAGCCCCCGATC
2	22190R	TCGGTCTCCGGACCCGTCGG
3	22110F	ATGGCCGACGACGACTG
4	22110R	CACCATGTTCAGGAAGTAGC
5	22120F	CCGACGACAGCGCGAGA
6	22120R	CACGTCGAGCACCTCGGC
7	22155F	ATGAGCCAGGTCTCGGTCT
8	22155R	GGTCCAGGGGTCGTAGAGC
9	22180F	ATCGGGCACACACCCCTGGT
10	22180R	GGACTCGACGAAGACCAACTC
11	POPINF_120F	AAGTTCTGTTTCAGGGCCCGATGGACCGGGTATCGACAGAG
12	POPINF_120R	TGGTCTAGAAAGCTTTAGCTCCGCGTGGCGATAC
13	POPINF_195F	AAGTTCTGTTTCAGGGCCCGATGGTCGCGGCGGCCT
14	POPINF_195R	TGGTCTAGAAAGCTTTAACGACGGGCTGCCCGTT
15	POPINF_210F	AAGTTCTGTTTCAGGGCCCGATGGTGGTGACCGCCGC
16	POPINF_210R	TGGTCTAGAAAGCTTTATCCACCGCACCGGGTACTAC
17	POPINF_220F	AAGTTCTGTTTCAGGGCCCGATGACGGCGGCCGGTAC
18	POPINF_220R	TGGTCTAGAAAGCTTTAGGGCAGACGGATCCGC
19	NCPBA4F	TCCCCAGGATAAATCTATTCATC
20	NCPBA4R	TAGGAGTGCGGTGTCTAAGTCTGG
21	PET28A_120F	CTATCC**CATATG**GACCGGGTATCGACAGAG
22	PET28A_120_185F	CAGGCG**CATATG**CACCCCAGGCGTCTCGCC
23	PET28A_120_672R	GCGGAG**AAGCTTTTA**GTCGGGTGCGGGCAGC
24	PET28A_120R	GACGAG**AAGCTT**TCAGCTCCGCGTGGCGATAC
25	PET28A_210F	TGGAGG**CATATG**GTGACTGAGGCGCTGACCTC

26	PET28A_210_271F	GTCGCC**CATATG**CACCCGGACGCCACCGCGG
27	PET28A_210_750R	GGCGTG**AAGCTTTTA** CGACGACTCCGGCAGG
28	PET28A_210R	CCGCAC**AAGCTT**CTACAGCCGGGCGACCGCTG
29	PET28A_195F	CCAACC**CATATG**GTGGTCGCGGCGGCCTGGG
30	PET28A_195_591R	CCGCAG**AAGCTTTTA**GACGAGTGCGGCCTTG
31	PET28A_195_690R	CTCCTG**AAGCTTTTA**CAGGGCCTGGAGCAGC
32	PET28A_220F	GACCGT**CATATG**ACGGCGGCCGGTACGAC
33	PET28A_220R	GGTCCT**AAGCTT**TCAGGGCAGACGGATCCGCG
34	PACYC_125F	CCGAGA**GAATTCG**ATGAGCGTGGCGGACG
34	PACYC_125R	CGACGG**AAGCTT**TCACCGCGTCGACGACC
35	PET30SYN195F	TAAGGA**CATATG** CCC CAT GAG AAG CAC TCG CTG
36	PET30SYN195P1R	TGGTAG**AAGCTTTTA** GCGTTCCTGACTTACTGTTGC
37	PET30SYN195FP2R	TGCATT**AAGCTTTTA**CACAGGTTCCTGAGGGGTTGC
38	PET30SYN195R	CCGCAGCGCCGCGGCGACGAGTGCGGCCTTGTC
39	PET30_210F	GGTATTGAGGGTCGCGTGACTGAGGCGCTGACCTCG
40	PET30_210R	AGAGGAGAGTTAGAGCCACTACAGCCGGGCGACCGC
41	CR_37_120_1F	ACGCCGGCGATCCGCTGTGCCGGA
42	CR_37_120_1R	AAACTCCGGCACAGCGGATCGCCG
43	CR_37_120_2F	ACGCACCCGCAGGGCGACCTGGCG
44	CR_37_120_2R	AAACCGCCAGGTCGCCCTGCGGGT
45	CR_37_120_3F	ACGCCGTCGTCCGACCGGACGGTA
47	CR_37_05470F	ACGCCCGGGTGTACGCCGCGAGGA
48	CR_37_05470F	AAACTCCTCGCGGCGTACACCCGG
49	CR_37_120_RT_LF	GGGTTTTTTGTCTAGATCGGTACTGGACGGATCGG
50	CR_37_120_RT_LR	ACATCCCGACCTGTCTAGCGTACACCCGG
51	CR_37_120_RT_RF	GTGGGCCACCGACCGGTGCAGACGGATCGG
52	CR_37_120_RT_RR	GATATCCTCTAGAGCTGGCGCGCCTCGGTGCCGTC
53	CR_37_0547_RT_LF	GGGTTTTTTGTCTAGATCCGAACACGTAGATGGCAC
54	CR_37_0547_RT_LR	GACGAGATCCCGGGTGTAATCGGTTCATCTCCACCGAC
55	CR_37_0547_RT_RF	AACCGATTACACCCGGGATCTCGTCA
56	CR_37_0547_RT_RR	GGGATCCTCTAGAGATCGTGCGCCCATCATG

Appendix B

Table A2. Explanation of terminology relating to genome assembly. Some terminology is not used, however, the meaning of all commonly used terminology is explained.

Output	Description
# contigs (≥ x bp)	The total number of contigs of length (≥ x bp). Not affected by the minimum contig parameter
Total length (≥ x bp)	Total number of bases (≥ x bp). Not affected by the minimum contig parameter
# contigs	Total number of contigs in assembly
Largest contig	The length of the largest contig
Total length	Total number of bases in assembly
Reference length	Total number of bases in reference genome
GC (%)	The total number of G and C nucleotides in the assembly, divided by the length of the assembly
Reference GC (%)	Percentage of GC nucleotides in the reference genome
N50	The length for which the collection of all contigs is the length or longer covers at least half of an assembly
NG50	Is the length for which the collection of all contigs of that length covers at least half a reference genome
N75 and NG75	Are define similarly as above, but 75 instead of 50
L50 (L75, LG50, LG75)	Is the number of contigs as long as N50 (N75, NG50, NG75)
# mis-assemblies	Is the number of positions in the contig that satisfy one of the following criteria: the left flanking sequence aligns over 1kbp away from the right flanking sequence; the flanking sequence overlap more than 1kbp or the flanking sequence align to different strands or chromosomes
# misassembled contigs	The number of contigs that contain miss-assembly events
Misassembled contig length	The total number of contigs in miss-assembled contigs
# local miss assemblies	The number of breakpoints that satisfy the following conditions: two or more distinct alignments cover the breakpoint; the gap between the left and right flanking sequence is less than 1kbp and the left and right flanking sequences both are on the same strand of the same chromosome of the reference genome
# unaligned contigs	The number of contigs that have no alignment to the reference genome. The value X + Y means X is totally unaligned contigs plus Y partially unaligned contigs.
Genome fraction (%)	The percentage of aligned bases in the reference. A base is in the reference is aligned if there is at least one contig with at least one alignment to this base. Contigs from repetitive regions may map to multiple regions, and thus may be counted multiple times
Duplication ratio	The total number of aligned bases in the assembly divided by the total number of aligned bases in the reference. If the assembly contains many contigs that cover the same region the duplication ratio will be much higher than 1
# N's per 100Kbp	Average number of uncalled bases (N's) per 100,000 assembly bases
# mismatches per 100Kbp	Average number of mismatches per 100,000 aligned bases. Sequencing errors and true SNP are counted similarly
Largest alignment	Largest continual alignment in assembly.
NA50m NGA50, NA75, NGA75, LA50, LA75, LGA50, LGA75	A stands for aligned. Similar to corresponding metrics without 'A' but in this case aligned blocks instead of contigs are considered. Blocks are obtained by breaking contigs involved in miss-assembly events and removing all unaligned bases

Appendix C

Table A3. Overview of integrated steps used during trimming of reads by Trimmomatic. Trimmomatic integrates a selection of programs for efficient trimming, this is an overview of each and how each program functions.

Trimming step used	Purpose
ILLUINACLIP	Cuts adapters and any other Illumina specific sequences from the read. Used in palindrome mode, this allows the removal of any contaminate partial adaptor sequences found at the 3' end of the read, due to adaptor read through.
SLIDINGWINDOW	Scans starting at the 5' end of the read and clips when the the quality drops below a threshold, using a sliding window approach. This stops the read being clipped when a single base of low quality is encountered, rather taking an average of the bases around it so that a single low quality base will not result in the loss of high quality data later in the read.
MAXINFO	An adaptive trimmer which cuts low quality sequences, but balancing read length and error rate to allow maximal information to be retained from each read.
LEADING	Cuts bases off the start of the read, if below a quality threshold.
TAILING	Cuts bases off the end of the read, if below a quality threshold.
CROP	Cuts bases off the end of the read, regardless of quality to ensure each read is below the specified maximal length.
HEADCROP	Cuts bases off the start of the read, regardless of quality to ensure each read is below the specified maximal length.
MINLEN	Removes reads that are below the specified minimal length.
AVGUAL	Removes the reads if the quality is below a specified minimal length.

Appendix D

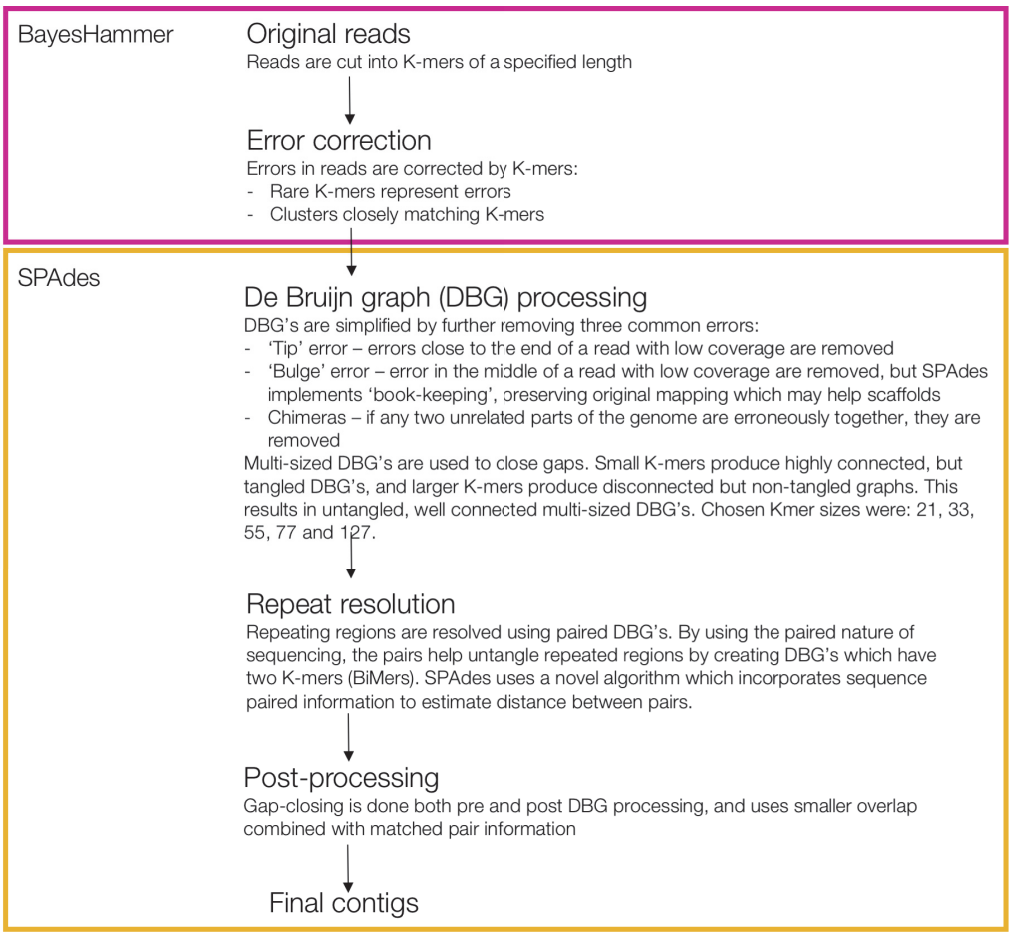


Figure A1. Overview of integrated steps used during assembly of reads by SPAdes. SPAdes is a combination of two programs – BayeHammer and SPAdes. BayesHammer is integrated to allow further error-correction and quality assessment of reads. Other QC and trimming programs were used separately, for sake of inclusivity the processing steps undertaken by this integrated program is stated.

Appendix E

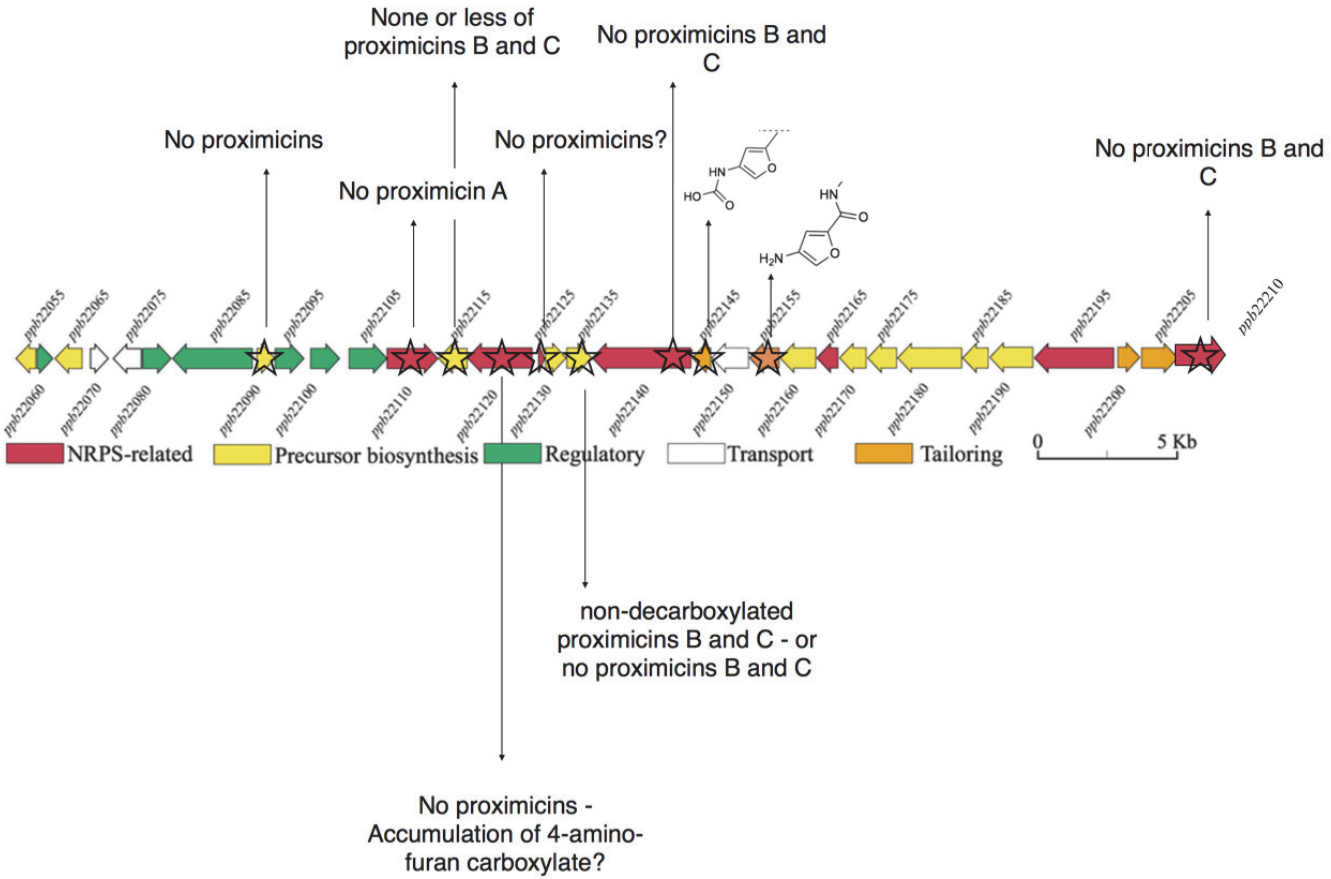
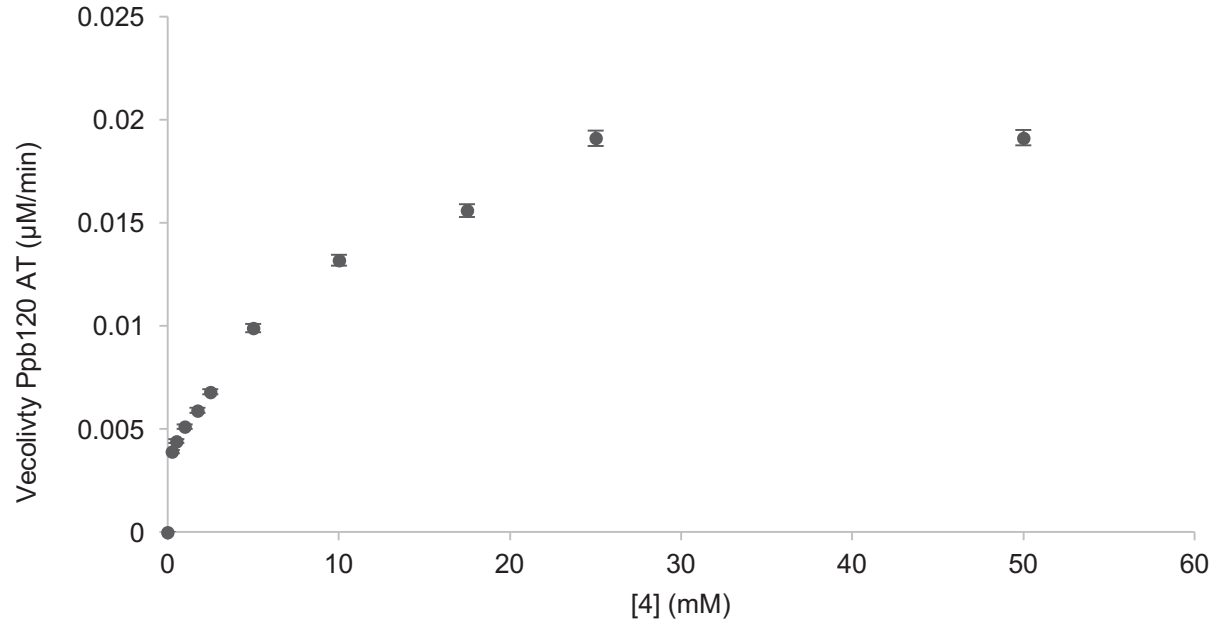


Figure A2. Overview of expected metabolites predicted to accumulate post *ppb* gene cluster deletion studies. CRISPR/Cas technology was successfully implemented in *Verrucosispora*, unfortunately it could not be fully explored in the scope of this research. Here is an overview of what we would expect to see accumulating in the media, assuming over proximicin biosynthetic route is correct, following specific gene deletions. This would have been explored, had time allowed.

Appendix F



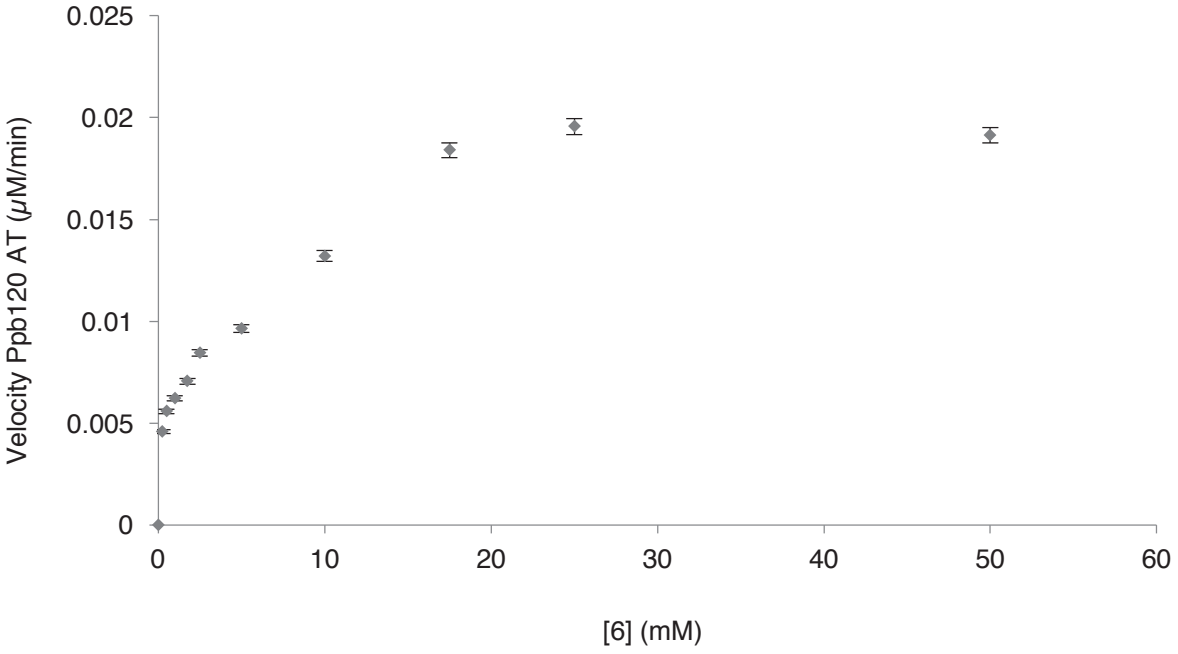


Figure F1. Michaelis-Menten kinetic parameters of Ppb120_AT catalysed adenylation of pyrrole derivatives – compound 4 and 6. Error bars represent SE.