Reprogramming Versus Subdomain Substitution: An Exploration of Strategies for NRPS Engineering Using the Unique Model System BpsA

By

Vincent Michael Kerr Collins

A thesis submitted to the Victoria University of Wellington in fulfilment of the requirements for the degree of Master of Science in Biotechnology

Victoria University of Wellington

(2020)

Abstract

Non-ribosomal peptide synthetases (NRPSs) are large enzymes that generate a plethora of important natural products, from antibiotics to immunosuppressants. These modular enzymes function like an assembly line, selecting and incorporating specific (and frequently non-proteinogenic) amino acids into a growing peptide chain. This modular structure offers promise for re-engineering NRPS units to generate new useful products, but progress has to date been limited by the complex and dynamic nature of key domains, and a failure to define generally applicable "rules" to guide engineering efforts. Early efforts to engineer NRPS enzymes relied of the substitution of entire NRPS modules or domains, but product yields were often very low. However, these studies did highlight the promise of targeting the adenylation domain, the part of each NRPS modules that is responsible for selecting each amino acid substrate.

Two particularly promising strategies for NRPS engineering aim to manipulate the adenylation domain in ways that minimise steric disruption to the assembly line. The first of these, reprogramming, makes the fewest possible changes to the NRPS primary sequence, but is dependent on those precise changes conforming to the existing structure of the adenylation domain binding pocket. More recently a second technique has been developed, subdomain substitution, which recombines a larger region of the adenylation domain to avoid perturbation of the binding pocket. The research described in this thesis examined and compared both approaches using the unique NRPS BpsA as a model system. BpsA is a single-module NRPS that generates a vivid blue pigment product, making for a reductionist system that offers a robust visual reporter capacity. Experiments with the reprogramming technique showed that small changes to the protein sequence had potential to exert major impacts on enzyme function, even when no change to function was intended. In contrast, experiments with subdomain substitution were generally more effective, showing that NRPS enzymes are very sensitive to the precise boundaries of the substituted region, but that activity can be restored to otherwise non-functional subdomain substitutions by modulation of the regional boundaries.

Acknowledgements

Thanks to my supervisors, Prof. David Ackerley and Dr Alistair Brown. Dave, thank you for being such an inspiring lecturer throughout my undergrad, and being such a supportive supervisor during my master's degree. Thank you for editing walls of text, for being relentlessly positive about my project, and for assembling such great people to make up your lab group. Alistair, thank you for constantly being there to answer questions and explain ideas, your patience is as remarkable as your enthusiasm!

Thanks to Hannah, for helping me and putting up with me right from that very first BTEC 101 lecture – you are the best. Thanks to Janine for going through master's right there along with me! Thanks to all my lab mates! Sarah, Mitch, Kelsi, Abby, Mark, Vincent (NZ), Manu, Jenni, and everyone I have inevitably forgotten to mention. Thanks to ex-lab mates too, Luke, Katherine and Michelle. Thanks to the wider SBS crew for your friendship and, when necessary, sympathy.

Thanks to Brooke, for supporting me and keeping me sane for the more than two years this has taken. Thanks to mum and dad for making me who I am, and for endless encouragement.

Contents

List of Figures

List of Tables

List of Abbreviations

His_6	six-histidine expression tag
A_{590}	absorbance at 590 nm
A-domain	adenylation domain
AMP	adenosine monophosphate
ATP	adenosine triphosphate
BLAST	basic local alignment search tool
bp	base pair
BpsA	blue pigment synthetase A
BpsA_SL	BpsA from *Streptomyces lavendulae*
C-domain	condensation domain
CoA	coenzyme A
DMSO	dimethylsulfoxide
E-domain	epimerisation domain
EDTA	ethylenediaminetetraacetic acid
EM	BpsA homologue from *Erwinia mallotivora*
EMD	EM downstream subdomain border variant
EMU	EM upstream subdomain border variant
FMN	flavin mononucleotide
IP	inorganic pyrophosphatase
IPTG	isopropyl β-D-1-thiogalactoside
kb	kilobase
LB	lysogeny broth
M-domain	methylation domain
min	minute
MQ	Milli-Q® ultrapure water
NRPS	non-ribosomal peptide synthetase
OD_{590}	optical density at 590 nm
OD_{600}	optical density at 600 nm
OD_{800}	optical density at 800 nm
Ox-domain	oxidation domain
PCP-domain	peptidyl carrier protein domain

Pi	PO_4^{3-} (inorganic phosphate)
PPi	$P_2O_7^{4-}$ (pyrophosphate)
Ppant	4’-phosphopantetheine
PPTase	4’-phosphopantetheinyl transferase
RO	reverse osmosis
RPM	revolutions per minute
S.D.	standard deviation
TE-domain	thioesterase domain
UV	ultraviolet
V	volts
v/v	volume/volume
WS	BpsA homologue from *Williamsia sterculiae*
WSD	WS downstream subdomain border variant
WSU	WS upstream subdomain border variant
w/v	weight/volume

1 Introduction

1.1 NRPSs

Non-Ribosomal Peptide Synthetases (NRPSs) synthesise peptides with diverse structures and bioactivities. NRPSs are large enzymes comprised of modules that sequentially select and incorporate monomers (usually amino acids) into a growing peptide product. NRPSs are found primarily, though not exclusively, in bacteria and fungi (Wang et al., 2015). Ribosomes are restricted to processing only the 22 proteinogenic amino acids (Finking & Marahiel, 2004), with additional tailoring only occurring through post-translational modification. In contrast, NRPSs are known to incorporate upwards of 500 unique monomers (Caboche et al., 2010), and through the action of unique domains can introduce structural complexity within the biosynthetic pathway, while still retaining a capacity for post-translational modification (Süssmuth & Mainz, 2017). The resultant NRPS products are variable and complex, and often of great biotechnological relevance (Felnagle et al., 2008).

NRPSs give rise to a plethora of medically useful compounds, including among others the β-lactam and daptomycin classes of antibiotics. Furthermore, a substantial number of anti-inflammatory and immunosuppressive drugs, as well as potential anti-cancer and anti-HIV compounds, are also NRPS-synthesised (Felnagle et al., 2008). Research into NRPSs has improved our understanding of the synthesis of extant compounds and opened avenues for the generation of new useful products. For example, researchers at Cubist Pharmaceuticals manipulated the assembly line structure of the NRPS responsible for the biosynthesis of daptomycin, an important antibiotic for the treatment of skin infections (Nguyen et al., 2006). In doing so, they generated several antibiotic peptides with a modified daptomycin cyclic core, including one with superior activity over standard daptomycin against an *Escherichia coli* model. This research was the first attempt at a true biotechnological application of NRPS manipulation, and remains the gold-standard example, with the field making only incremental progress in the ensuing decade.

In addition to the prospect of generating novel useful peptides, NRPS research offers an opportunity to optimise the synthesis of existing medically relevant compounds. Modern medicines derived from natural products often require additional synthesis steps after the natural precursor has been isolated. Genetically engineering NRPS enzymes, through

manipulation or addition of modules, has potential to enable direct in vivo production of the final useful compounds (Calcott & Ackerley, 2014).

1.2 Components of an NRPS Assembly Line

NRPSs synthesise peptides products in a manner analogous to an assembly line. The growing peptide is passed along the enzymatic template, with each module contributing to the final product – incorporating an amino acid or related monomer, or manipulating the structure in some way (Calcott & Ackerley, 2014). Synthesis of the final peptide product can depend on the action of multiple NRPS proteins, which aggregate to form a start-to-finish assembly line (Süssmuth & Mainz, 2017).

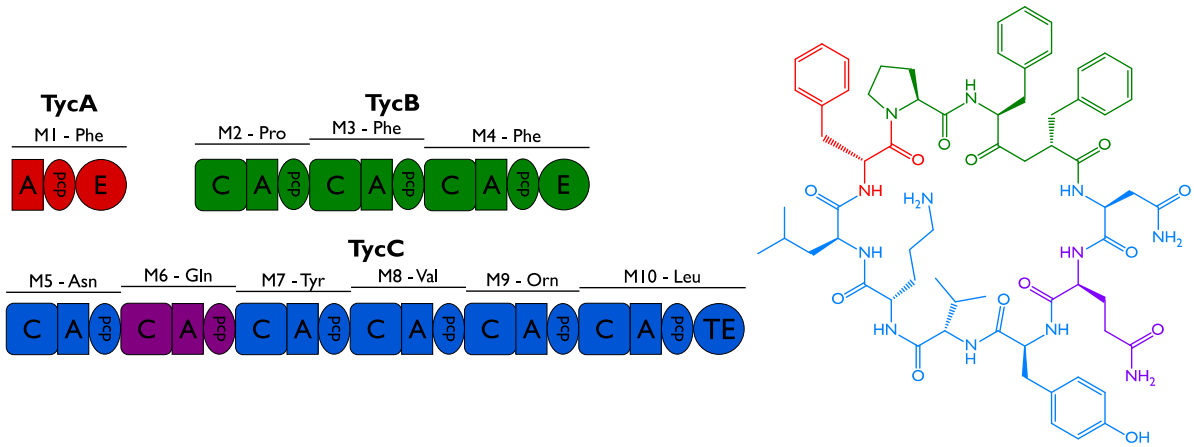


Figure 1.1 Schematic of the tyrocidine NRPS assembly line and its product. Individual peptide synthetases and corresponding monomers in the final product are colour-coordinated. Module 6 and the glutamine it incorporates (of particular relevance to this thesis) are shown in purple.

Figure 1.1 is a schematic diagram showing the NRPS assembly line responsible for production of the antibiotic compound tyrocidine A. Tyrocidine was the first NRPS product to be recognised as non-ribosomal in origin (Felnagle et al., 2008; Mach et al., 1963), and is the product of three NRPS enzymes acting *in trans* with 10 modules incorporating 10 monomers into a final product. Module 9 incorporates ornithine, a non-proteinogenic amino acid, and the final product is head-to-tail cyclised during biosynthesis; these two features exemplify key functional distinctions between ribosomal and non-ribosomal peptide synthesis.

1.2.1 The Adenylation Domain

For an amino acid or other monomer to be incorporated into a non-ribosomal peptide (NRP), it must first be identified and activated by the NRPS. This function is performed by the adenylation domain (A-domain). The A-domain determines the monomer to be added by a particular module and consequently each individual component of the peptide sequence. The

~60 kDa A-domain consists of a ~50 kDa N-terminal core domain linked by a 5-residue hinge region to an ~10 kDa minor domain (Süssmuth & Mainz, 2017). The solved crystal structure of the phenylalanine-activating A-domain in GrsA (part of the gramicidin synthesis pathway; PDB:1AMU) indicates that the amino-acid binding pocket of the A-domain is present on the concave face of the ~50 kDa core domain. The ~10 kDa minor domain is broadly sandwiched against this binding pocket, shown in Figure 1.2 (Conti et al., 1997).

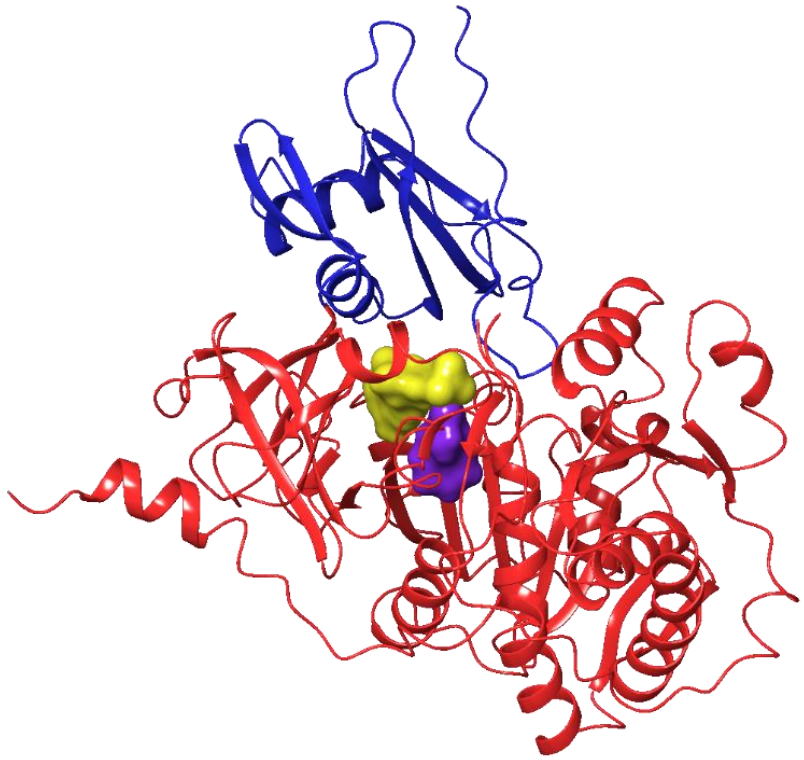


Figure 1.2 Cartoon ribbon structure of the GrsA A-domain. The ~50 kDa core domain is coloured red, and the ~10 kDa minor domain is coloured blue. Space-filling representations of phenylalanine (purple) and AMP (yellow) are shown in the binding pocket.

Adenylation domains contain 10 stretches of highly conserved sequence, called the A-domain core motifs (Takahashi et al., 2007). Between four and twenty amino acids long, the motifs can vary to a degree between A-domains, but they are far more conserved than other areas of the primary sequence. The location of these motifs (A1-A10) is essentially invariant across A-domains, and they occur sequentially from the protein N-terminus to the C-terminus (e.g. A1 is at the N-terminal end of the A-domain, and A10 is at the C-terminal end). The ~50 kDa core domain contains motifs A1-A8, whereas motifs A9-A10 are found in the ~10 kDa minor domain (Takahashi et al., 2007). The binding pocket is invariably present between motifs A3 and A6.

Within the binding site, 10 core residues were identified by Stachelhaus et al. (1999) as having conserved locations and being involved in recognition of the sidechain of the target amino acid. The identities of these residues, now known as the non-ribosomal or 'Stachelhaus' code, can be used to predict the acceptor substrate of the A-domain (Stachelhaus et al., 1999; Süssmuth & Mainz, 2017). Experimental data has allowed for the correlation of each residue combination with the associated substrate (Challis et al., 2000). The Stachelhaus residues are not the only points of variation between A-domains of differing selectivity. Changes in a broader 'subdomain' around the active site (encompassing the Stachelhaus residues) can also dictate the function of an A-domain (Kries et al., 2015).

Occasionally, A-domains are interrupted by auxiliary tailoring domains, most often between the A8 and A9 motif (Labby et al., 2015). In many cases where this has been observed, both the auxiliary domain and the interrupted A-domain remain functional. Among the tailoring domains found to interrupt A-domains are epimerisation (E-), oxidation (Ox-), and methyltransferase (M-)domains, described in Section 1.2.5. The Garneau-Tsodikova Lab recently presented the first crystal structure of an interrupted adenylation domain – a nonspecific A-domain interrupted by an M-domain. They found that the M-domain projected away from the A-domain, such that the typical three dimensional structure of the A-domain was preserved (i.e., was consistent with GrsA) (Lundy et al., 2020; Mori et al., 2018).

Following recognition of the acceptor substrate, the A-domain catalyses the adenylation of the amino acid at its carboxyl end. In a second reaction catalysed by the same A-domain, this activated monomer (in the form of an aminoacyl-AMP) is passed to the thiolation domain (Marahiel et al., 1997; Süssmuth & Mainz, 2017). In this reaction, the monomer is attached to the 'arm' of the peptidyl carrier protein domain, described below.

1.2.2 The Peptidyl Carrier Protein Domain

The peptidyl carrier protein (PCP) domain, also called the thiolation domain (T-domain), is responsible for the transport of substrates and peptide intermediates throughout an NRPS assembly line. It is small, at about 10 kDa (~100) residues and possesses a four-helix bundle structure (Marahiel et al., 1997). Inactive *apo* PCP-domains are activated by a partner phosphopantetheinyl transferase (PPTase), which installs a 4'-phosphopantetheine (Ppant) prosthesis derived from co-enzyme A to a conserved serine residue of the PCP-domain (Marahiel et al., 1997). The prosthesis on an activated *holo* PCP-domain functions like a robotic

arm, moving the substrates between domains and modules, and positioning them for catalysis, including the ultimate release of the nascent peptide from the NRPS in the termination module.

1.2.3 The Condensation Domain

The condensation domain (C-domain) catalyses peptide bond formation between the acceptor substrate monomer and the growing peptide chain/product (the 'donor substrate') that is proffered by the PCP-domain of the module immediately upstream in the assembly line (Stachelhaus et al., 1998). The ~50 kDa C-domain consists of two roughly equally-sized subdomains that form a V-shape, with the active sites located in the cleft of the V. Ppant arms of PCP-domains holding the donor and acceptor substrates approach the cleft from opposite sides of the V to reach their respective active sites (Bloudoff & Schmeing, 2017; Süssmuth & Mainz, 2017).

Like the A-domain, the C-domain shows a degree of specificity for incoming substrates, improving the accuracy of non-ribosomal peptide synthesis (Calcott & Ackerley, 2014). The C-domain 'proofreads' the acceptor substrate from the intramodular A-domain, and is generally believed to be selective both for the chirality and sidechain at the acceptor active site. In proofreading the donor substrate, it is believed to show L- versus D-enantiomer selectivity and general size selectivity, confirming the identity of the growing end of the peptide (Belshaw et al., 1999; Ehmann et al., 2000). These factors present a potential barrier to NRPS engineering efforts, as a non-native acceptor substrate from a re-engineered A-domain may not be accepted by an unchanged C-domain. More recently, however, research in our lab has indicated that C-domain proof-reading may not be a major hurdle to efficient NRPS re-engineering, with the demonstration that new peptides can be generated in diverse NRPS systems through A-domain substitutions alone (Calcott et al., 2020). The experimental work has been backed up by extensive phylogenetic analyses indicating that complete or partial A-domain substitution is also a primary mechanism for non-ribosomal peptide diversification in nature.

1.2.4 The Thioesterase Domain

Thioesterase (TE) domains are found in the final C-terminal (termination) module of an NRPS assembly line. Around 30 kDa in size, they catalyse the cleavage of the thioester bond holding the peptide product to the Ppant arm of the final PCP-domain. Cleavage by hydrolysis or aminolysis results in the release of a linear peptide product (Süssmuth & Mainz, 2017). However, by excluding water from the active site via a hydrophobic cavity and constraining the confirmation of the peptide, TE-domains commonly cause cyclisation of the peptide

products (Hur et al., 2012). Both head-to-tail and side chain-to-tail cyclisation can occur (Samel et al., 2006). Some TE-domains can also stockpile identical non-ribosomal peptide precursors for collective condensation or cyclisation, as in gramicidin S synthetase (Keating et al., 2001).

1.2.5 Tailoring Domains

As mentioned in Section 1.1, NRPSs are capable of the incorporation of non-proteinogenic substrates, often through the activities of auxiliary tailoring domains (Caboche et al., 2010). This section summarises those tailoring domains most relevant to the present study.

The most common tailoring domain is the E-domain. A majority of NRPs include D-amino acids in their structure. While some of these D-amino acids are recognised directly by A-domains, more commonly the L-enantiomer is selected by the A-domain, loaded onto the PCP-domain arm and subsequently racemised by an E-domain within the module (Süssmuth & Mainz, 2017; Walsh et al., 2001). E-domains are present exclusively in modules incorporating a D-monomer into the peptide product (Walsh et al., 2001). Typically ~50 kDa, they are descended from the C-domain, sharing sequence motifs and a base protein fold (Samel et al., 2014).

The Ox-domain catalyses the oxidation of an acceptor substrate. At 30 kDa in size, it is commonly found immediately downstream of a module's A-domain, or occasionally (as in BpsA) embedded between the A8 and A9 motif of the A-domain (Du et al., 2000). To oxidise the substrate, Ox-domains are dependent on an associated cofactor, flavin mononucleotide (FMN) (Du et al., 2000).

Many of the amino acids in NRPs are methylated. While this can be achieved by independent trans-acting methyltransferases, it is most often accomplished while the amino acid is at the end of a Ppant arm, where it is methylated by an embedded M-domain (Hur et al., 2012; Walsh et al., 2001). M-domains can be situated upstream of the module's A-domain or more frequently integrated into the A-domain, usually between the A8 and A9 motifs (Labby et al., 2015; Mori et al., 2018; Walsh et al., 2001).

1.3 Previous NRPS Manipulations

The modular and specific nature of NRPSs appears to lend itself to possibilities of enzyme design. Synthesis of novel compounds through the recombination and re-coding of modules is a key goal of NRPS researchers. Nature has clearly generated a plethora of module combinations and natural products over the course of evolution. Thus far however, artificial

NRPS redesign experiments have met with limited success (Ackerley, 2016; Calcott & Ackerley, 2014).

1.3.1 Module Substitution and Exchange Units

Cubist Pharmaceuticals' work on daptomycin, as mentioned in Section 1.1, is widely regarded as a best-in-class example of applied module substitution (Brown et al., 2018; Kries, 2016). In combination with other engineering techniques, Cubist researchers were able to generate a number of potentially useful daptomycin derivatives (Nguyen et al., 2006). The majority of their enzymes, however, were non-functional, and those that were functional often only produced a peptide product in markedly reduced volumes. This pattern of success and caveats is mirrored in the work of Zobel et al. who generated hybrid two-module NRPSs for the production of cyclodepsipeptide derivatives (Zobel et al., 2016). Again, product was generated, but production was greatly reduced in yield. With such significant activity losses seen with even the closely-related modules that were used in the above studies, whole-module substitution appears to be a non-optimal method of re-engineering NRPS enzymes (Kries, 2016).

A nascent similar technique involves the substitution of exchange units (A – PCP – C) rather than modules proper (C – A – PCP). This technique generated a number of new peptides, but adds complexity to the re-engineering task and is not immune to drops in product titre (Bozhüyük et al., 2018).

1.3.2 CA- and A-domain Substitutions

As the A-domain determines the monomer to be incorporated into the peptide, large-scale manipulation of non-ribosomal peptide structure necessarily begins with changes to A-domain specificity. The first successful swapping of an A-domain was performed by Stachelhaus, Schneider and Marahiel (1995), who took A- and T-domains from 5 different NRPSs and substituted them in the place of a leucine-specifying A-T-domain in a surfactin NRPS template. Each of the substituted A-domains specified a different amino acid substrate (i.e., was non-synonymous), and all were found to generate a detectable amount of novel product. However, in a subsequent publication it was revealed that the yield had been heavily compromised relative to the levels of unmodified surfactin produced by the wild-type assembly line (Schneider et al., 1998).

The reduction in yield may have been due to the specificity of the partner C-domain for the wild-type substrate. The specificity exhibited by both the A- and C-domains presents a potential barrier to the isolated manipulation of the A-domain (Calcott & Ackerley, 2014). Even synonymous substitutions can run into trouble at the C-domain – it has been suggested that the 'catalytic platform', where the C- and A-domains contact each other in space, is dependent upon an intact C-A interface for activity (Tanovic et al., 2008), and -A- and A-T-domain substitutions may disrupt this linker region and destroy activity.

This issue can be avoided, though not resolved, by targeting modules (like the initiation module) or NRPSs (like BpsA) that lack a C-domain entirely (Calcott & Ackerley, 2014). Substituting entire A-domains within the initiation module has been successful, with the first amino acid of surfactin altered from glutamate to glutamine by the substitution of a non-synonymous A-domain (Yakimov et al., 2000).

A more manifest solution to the issue of C-A compatibility might be to substitute seemingly inseparable pairs together – i.e., perform C-A-domain substitutions. This technique was the approach that was successfully adopted in the Cubist Pharmaceuticals daptomycin study (Nguyen et al., 2006). An example from our own lab found that CA-substitution could be used to rationally introduce non-native residues into pyoverdine while retaining high yields (Calcott et al., 2014). From seven constructs, two generated product at near-wild-type levels – a moderate success rate, and better than the zero percent success rate of the corresponding A-domain-only swaps (Calcott et al., 2014).

A further potential solution is to focus efforts on manipulation of the native A-domain already present in the NRPS of interest, thereby avoiding C-interface disruption issues.

1.3.3 Specificity Code Reprogramming

The Stachelhaus residues in the active site of the A-domain are thought to be responsible for determining substrate identity. Databases and algorithms have been developed which allow researchers to predict, and therefore design, the substrate specificity of an A-domain based on those eight residues (Rausch et al., 2005; Röttig et al., 2011). This does not address the issue of C-domain acceptor specificity, but it should leave the C-A interface, and hopefully the wider A-domain structure, unadulterated. The first studies into coding residue manipulation involved swapping in the Stachelhaus code of other known A-domains (Stachelhaus et al., 1999). These experiments saw successful change in the specificity of their A-domains, although changes were usually quite conserved (e.g., asparagine to aspartate) and often to the detriment of overall

levels of catalytic activity (Calcott & Ackerley, 2014). Other studies pursued rational redesign or random mutagenesis of the binding pocket residues to modify module specificity. While these strategies were sometimes successful, overall activity levels again suffered (Calcott & Ackerley, 2014). With efficiency in mind, one study sought to take an A-domain of broad specificity and reprogram it to select for a specific substrate (Han et al., 2012). Using site-directed mutagenesis, they were able to dramatically increase selectivity for their chosen substrate, Phe, from 9.4 to 26.4%. The best mutant generated a threefold higher titre of the target product than the wild-type enzyme. This demonstrates the potential for the 'tweaking' of A-domains, without causing widespread change to domain structure.

Recent studies have assessed the importance and variability of each specific coding residue in determining the specificity of the A-domain (Ishikawa et al., 2020; Throckmorton et al., 2019). Throckmorton et al. (2019) applied a site-saturation mutagenesis and directed evolution strategy to the A-domain coding residues of EntF, an *E. coli* NRPS involved in enterobactin biosynthesis. This strategy revealed upwards of 150 new codes that activated L-Ser in EntF but are not seen in any known A-domains. Throckmorton et al. (2019) found that some residues, mostly towards the top of the binding pocket, were far more conserved among functional specificity codes, whereas others, mostly towards the base of the pocket, were more variable. Serine is a small amino acid, and it is possible that because it does not protrude as far into the binding pocket as, say, phenylalanine, the interaction-potential of coding residues at the base of the pocket is reduced. This research suggests that the functional sequence space of specificity codes for any individual substrates could be far larger than has previously been observed in nature. Furthermore, it suggests that reprogramming strategies could be focused on key specificity-determining residues, while leaving other residues alone to minimise disruption of the native binding pocket (Throckmorton et al., 2019).

1.3.4 Subdomain Substitution

A nascent technique for the manipulation of A-domain specificity is subdomain swapping (Crüsemann et al., 2013; Kries et al., 2015). Rather than just substitution or reprogramming of the coding residues themselves, a broader stretch of sequence is transplanted, which includes the coding residues and avoids domain interfaces. To distinguish it from the physical subdomains of the A-domain (the ~50 kDa core domain and ~10 kDa minor domain), this functional subdomain is sometimes referred to as the Recognition Subdomain (Throckmorton et al., 2019). Essentially encapsulating the protein sequence of the binding pocket, it sits

roughly between the A3 and A6 motifs. This technique may bypass issues encountered in specificity code reprogramming where non-native coding residues interact poorly with the surrounding subdomain structure. Simultaneously, it avoids the detrimental structural changes introduced at domain borders in full-domain substitution strategies (Kries et al., 2015).

The Piel lab first identified the potential of subdomain swap strategies. Analysis of the hormaomycin gene cluster revealed that the A-domains in the cluster shared almost 90% identity across their length, with the exception of a ~400 bp region containing the majority of the binding pocket (Crüsemann et al., 2013). It appeared that exchange of only the recognition subdomain was responsible for the diversity of A-domain specificity found in the gene cluster. Guided by these evolutionary boundaries, they successfully generated A-domains with altered specificity *in vitro* (Crüsemann et al., 2013).

A different approach was taken by the Hilvert lab group in 2015. Informed by the evolutionarily-guided approach of the Piel lab, they took a structurally-guided approach to subdomain substitution (Kries et al., 2015). Through assessment of the crystal structure of the GrsA A-domain, they identified subdomain boundaries according to several structural criteria (e.g. include all coding residues, avoid domain interfaces). Compared to the subdomain borders of Crüsemann et al. (2013), the borders chosen by Kries et al. delineate a slightly larger subdomain, shifted slightly downstream (Fig. 4.1). The Kries et al. strategy yielded functional transgenic A-domains, efficient such that adenylation was no longer the kinetic bottleneck in *in vitro* synthesis (Kries et al., 2015). However, as yet there are no reports of modified peptide products generated in vivo via these types of subdomain exchange strategies.

1.4 Blue Pigment Synthetase A

Blue pigment synthetase A (BpsA) is a NRPS first identified in a strain of *Streptomyces lavendulae* that produces the blue pigment indigoidine (Fig. 1.3). Consisting of only a single module, it synthesises indigoidine from two L-glutamine monomers (Takahashi et al., 2007).

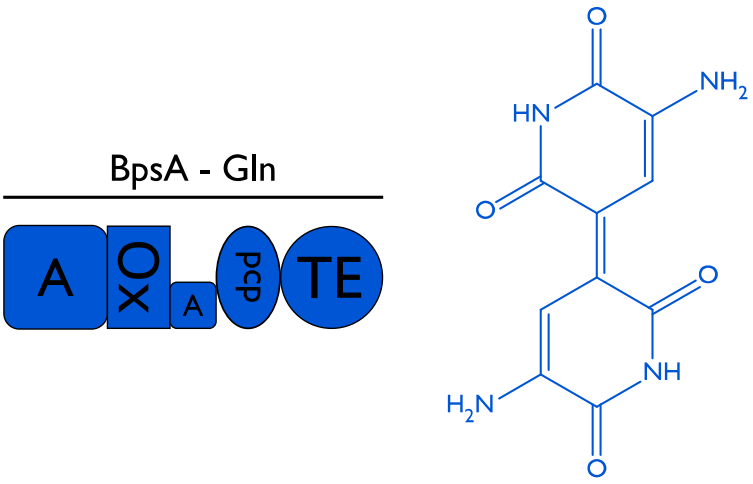


Figure 1.3 Simplified schematic of BpsA and its product, indigoidine. The core (A1-A8) and minor (A9-A10) subdomains of the A-domain are separated by the Ox-domain

As a small, simple NRPS that can be easily translated by *E. coli* into a functional and purifiable form, BpsA is a prime candidate for NRPS research and manipulation. BpsA also lacks a C-domain, bypassing issues of acceptor substrate specificity covered above (Takahashi et al., 2007).

BpsA itself has extensive utility as a biosensor. Recombinant BpsAs with non-native PCP-domains have been used to identify new PPTase enzymes from metagenomic libraries (Owen et al., 2012). These PPTases activate the PCP-domain (*apo* to *holo*), thus activating BpsA and producing indigoidine as a reporter pigment. Importantly, because PPTases are often linked to NRPS genes (and related polyketide synthase genes) within biosynthetic gene clusters, this provides a means of screening collections of cloned metagenomics DNA to recover natural product biosynthetic gene clusters (Owen et al., 2012). Purified BpsA can also be used to quantify levels of the medically and industrially important amino acid L-Gln, or required cofactors such as ATP, in unknown samples including blood, urine, or culture media (Brown et al., 2017).

1.4.1 Structure of BpsA

BpsA consists of a single module, with four core domains: An A-domain, Ox-domain, PCP-domain, and TE-domain. The A-domain harbours an interrupting Ox-domain between the A8 and A9 motifs (Fig. 1.3). Having only a single A-domain, BpsA synthesises indigoidine from only one kind of monomer, L-Gln (Takahashi et al., 2007).

The A-domain adenylates an L-Gln monomer and catalyses the thiolation reaction, attaching the amino acyl-AMP intermediate to the Ppant prosthesis of the PCP-domain. The PCP-domain

shuttles this L-Gln monomer from the A-domain to downstream domains. The Ox-domain dehydrogenates the carbon backbone of the amino acid, generating a C2-C3 carbon-carbon double bond (Kong et al., 2019). The PCP-domain delivers the monomer to the TE-domain, where it is cyclised and released. The released cyclic compounds are thought to spontaneously dimerise, in the presence of oxygen, to form the final indigoidine product (Kong et al., 2019; Walsh & Wencewicz, 2013). The timing of Ox-domain double-bond installation has not been experimentally confirmed. It may occur while the monomer is attached to the Ppant prosthesis of the PCP-domain, prior to cyclisation and release, as described above. Alternatively, it may occur after release of the cyclic compound by the TE-domain (Takahashi et al., 2007; Walsh & Wencewicz, 2013). The structure of BpsA has not been solved, and uncertainty around the effects and positioning of the Ox-domain preclude informed modelling approaches.

When BpsA is expressed in *E. coli*, it must be co-expressed with an appropriate PPTase, e.g. that from *Streptomyces verticillus*, *svp* (Takahashi et al., 2007) or PcpS from *Pseudomonas aeruginosa* (Owen et al., 2011). This PPTase is required to attach the Ppant arm to the PCP-domain, generating the *holo*-form of the NRPS.

1.4.2 Previous BpsA Manipulations

Directed evolution has become a principal technique in the field of protein engineering (Kries, 2016). The random or targeted mutagenesis of an enzyme, coupled with selection for desired activity, allows researchers to effectively recover or develop enzyme functionality without complete understanding of the underlying structure. By happenstance and throughput, directed evolution can reveal minor changes with unexpected positive consequences that could not have been discovered with a rational engineering approach (Kries, 2016).

BpsA has been used as a platform for the successful directed evolution of NRPS domains (Owen et al., 2016). BpsA has been shown to be amenable to PCP-domain substitution (Owen et al., 2012). It is the only currently-known single-module NRPS to autonomously generate a pigmented product, and this property lends itself to the screening of BpsA for functionality – the bluer the colony, the more functional the BpsA variant (Owen et al., 2016). Owen et al. used BpsA as a surrogate for the expression of non-native domains, namely the thiolation domain from the first module of PvdD (a bi-modular NRPS required for synthesis of pyoverdine, a yellow-green and UV fluorescent siderophore of *P. aeruginosa*). When moved to the termination module of PvdD, this first-module PCP-domain could not interact with the TE-domain to release the product. The PvdD PCP-domain was substituted into BpsA, and a

directed evolution campaign was performed, following which the introduced PCP-domain could interact successfully with the TE-domain of BpsA. Upon introduction of this evolved domain back into the termination module of PvdD, it was now functional in concert with the native TE-domain (Owen et al., 2016). This not only provided insight into key residues within PCP-domains, it also established BpsA as a suitable model for NRPS recombination and directed evolution studies.

1.5 This Study

The previous sections have detailed several different approaches to re-engineering of NRPS templates, while also making it clear that this is generally a very difficult proposition. Whereas nature has clearly exploited the modular nature and conserved sequences of NRPS enzymes to generate novel assembly lines and products (Calcott et al., 2020), recapitulating this in the laboratory setting has proven difficult, with transgenic NRPSs often being non- or barely functional. Moreover, most published studies have focused exclusively on a single re-engineering approach, meaning that there is a lack of quantitative data around which strategy might prove most effective for a particular scenario. To address this, we sought to directly compare two strategies in terms of their outright effectiveness, and the success of subsequent optimisation efforts. This study is interested in those strategies that make changes within the A-domain. Coding residue reprogramming has met with some success in past research. While it was argued that such focused changes of just the few coding residues should have minimal effect on the macromolecular structure (Challis et al., 2000), this may not be the case in practice. Indeed, recognising this, algorithms have been developed to expand on the basic eight amino acid code defined by Stachelhaus et al. (1999) to predict substrate specificity based on a weighted contribution of all residues located within 8 Å of the binding pocket (Rausch et al., 2005). This suggested that the local environment of the substrate binding residues can potentially exert a substantial influence on their activity. Thereafter that ‘local environment’ was named the recognition subdomain, and early studies show that preserving this local environment by substituting the subdomain in its entirety can generate functional products.

To date, no direct comparison has been made between coding residue reprogramming and subdomain substitution techniques. In this work we sought to employ BpsA for parallel reprogramming and subdomain substitution studies, to empirically assess which approach might be more effective.

1.5.1 Aims

This study aims to assess the amenability of the BpsA A-domain to reprogramming wherein the native coding residues are replaced with an alternative set of L-Gln coding residues, to judge the limits of the reprogramming technique. To do this, we will generate a curation of experimentally confirmed L-Gln specificity codes and introduce them into BpsA. Non-functional reprogramming efforts will be investigated for ways to recover activity, allowing us to develop a rationale for why certain coding residues or specificity codes are not tolerated.

We also aim to compare these reprogramming results to synonymous subdomain substitutions. Subdomain substitution, we hypothesise, may only be a viable strategy if the borders of the subdomain are tailored to minimise steric disturbance of the recipient A-domain. To this end we aim to make adjustments – and improvements – to the subdomain substitution strategy pioneered by Kries et al. (2015).

2 Materials and Methods

2.1 General Reagents and Materials

Unless otherwise stated, all chemical stocks were obtained from Sigma-Aldrich or Thermo Fisher Scientific. Lysogeny broth and microagar were sourced in pre-formulated powders from Duchefa Biochemie. L-glutamine was purchased from iHerb. Agarose LE and ethidium bromide were purchased from Hydragene. Restriction enzymes and FastAP alkaline phosphatase were from Thermo Fisher Scientific. Further restriction enzymes, NEBuilder (Gibson Assembly) reagents and Q5 Polymerase were sourced from New England Biolabs. Bioline was the supplier of isopropyl β-D-1-thiogalactoside (IPTG) and BioMix™ Red Polymerase mastermix. T4 DNA ligase was sourced from Invitrogen. Materials were sterilised by autoclaving at 121 ℃ for 15 minutes minimum.

2.2 Bacterial Strains and Plasmids

2.2.1 Bacterial strains

Table 2.1 E. coli bacterial strains used in this study.

Strain	Relevant Characteristics	Source
BL21	F^- *ompT gal dcm lon hsdS*$_B$(r_B^- m_B^-) λ(DE3)	Novagen
BL21 Δ*entD*	Chromosomal *entD* PPTase gene deletion	Owen et al., 2012
DH5α	*supE*44 D*lacU*169 (Ø80 *lac*Z DM5) *hsdR*17	Invitrogen

2.2.2 Plasmids

Table 2.2 Plasmids used in this study

Plasmid	Relevant Characteristics	Source
pET-28a(+)	*kan*^{R}, expression vector for His-tagged enzyme purification, T7 promoter, ColE1 origin of replication	Novagen
pET-pcpS	pET-28a(+) + *pcpS*	Owen et al., 2011
pNOHISPET-pcps	pET-28a(+) backbone with N-terminal His-tag removed, + *pcpS*	Brown, 2017

pET-licAM1AD	pET-28a(+) + *licA* module 1 A-domain	This study
pCDFDuet-1	*lacI*^{q} ,T7prom, *spec*^{R}, CDFori	Novagen
pCDFDuet-bpsA	*lacI*^{q} ,T7prom, *spec*^{R}, CDFori; wt *bpsA* insert (XhoI - BamHI)	Owen et al., 2012
pCDFDuet-OxTe	pCDFDuet1 + *bpsA* Ox,PCP,TE domains (HinDIII – NdeI)	This study
pCDFDuet-wtbpsA_OxTe	pCDFDuet1 + wt *bpsA* (BamHI-[HindIII]-NdeI)	This study
pCDFDuet-licA_res_OxTe	pCDFDuet1 + wt *bpsA* with *licA* AD coding residues	This study
pCDFDuet-tycC_res_OxTe	pCDFDuet1 + wt *bpsA* with *tycC* AD coding residues	This study
pCDFDuet-FCR4V_OxTe	pCDFDuet-wtbpsA_OxTe with one changed residue (F-CR4-V)	This study
pCDFDuet-VCR6L_OxTe	pCDFDuet-wtbpsA_OxTe with one changed residue (V-CR6-L)	This study
pCDFDuet-ICR7V_OxTe	pCDFDuet-wtbpsA_OxTe with one changed residue (I-CR7-V)	This study
pCDFDuet-licA_sub_OxTe	pCDFDuet1 + wt *bpsA* with *licA* A-subdomain	This study
pCDFDuet-tycC_sub_OxTe	pCDFDuet1 + wt *bpsA* with *tycC* A-subdomain	This study
pCDFDuet-WS_sub_OxTe	pCDFDuet1 + wt *bpsA* with *WS* A-subdomain	This study
pCDFDuet-VN_sub_OxTe	pCDFDuet1 + wt *bpsA* with *VN* A-subdomain	This study
pCDFDuet-EM_sub_OxTe	pCDFDuet1 + wt *bpsA* with *EM* A-subdomain	This study
pCDFDuet-licR-wt_CR2_OxTe	pCDFDuet-licA_res_OxTe with CR2 restored to *bpsA* wild type	This study
pCDFDuet-licR-wt_CR3_OxTe	pCDFDuet-licA_res_OxTe with CR3 restored to *bpsA* wild type	This study

pCDFDuet-licR-wt_CR4_OxTe	pCDFDuet-licA_res_OxTe with CR4 restored to *bpsA* wild type	This study
pCDFDuet-licR-wt_CR7_OxTe	pCDFDuet-licA_res_OxTe with CR7 restored to *bpsA* wild type	This study
pCDFDuet-licR-wt_CR8_OxTe	pCDFDuet-licA_res_OxTe with CR8 restored to *bpsA* wild type	This study
pCDFDuet-licR-wt(3+2)	pCDFDuet-licA_res_OxTe with CR3 and CR2 restored to *bpsA* wild type	This study
pCDFDuet-licR-wt(3+4)	pCDFDuet-licA_res_OxTe with CR3 and CR4 restored to *bpsA* wild type	This study
pCDFDuet-licR-wt(3+7)	pCDFDuet-licA_res_OxTe with CR3 and CR7 restored to *bpsA* wild type	This study
pCDFDuet-licR-wt(3+8)	pCDFDuet-licA_res_OxTe with CR3 and CR8 restored to *bpsA* wild type	This study
pCDFDuet-bpsA+licA_R2	pCDFDuet-wtbpsA_OxTe with CR2 changed to the *licA* codon	This study
pCDFDuet-bpsA+licA_R3	pCDFDuet-wtbpsA_OxTe with CR3 changed to the *licA* codon	This study
pCDFDuet-bpsA+licA_R4	pCDFDuet-wtbpsA_OxTe with CR4 changed to the *licA* codon	This study
pCDFDuet-bpsA+licA_R7	pCDFDuet-wtbpsA_OxTe with CR7 changed to the *licA* codon	This study
pCDFDuet-bpsA+licA_R8	pCDFDuet-wtbpsA_OxTe with CR8 changed to the *licA* codon	This study
pCDFDuet-NEB_bpsA_sub_OxTe	Identical to pCDFDuet-wtbpsA_OxTe. Acted as a positive control NEBuilder construction	This study
pCDFDuet-NEB_EM_sub_OxTe	pCDFDuet-wtbpsA_OxTe with EM sequence from start codon – A8 motif	This study
pCDFDuet-NEB_WS_sub_OxTe	pCDFDuet-wtbpsA_OxTe with WS sequence from start codon – A8 motif	This study
pCDFDuet-EMD_x	pCDFDuet-1-based construct for the expression of EMD variants 1-15 and 18	This study

pCDFDuet-WSD_x	pCDFDuet-1-based construct for the expression of WSD variants 1, 2, 3, 6, 7 and 9	This study
pCDFDuet-EMU_x	pCDFDuet-1-based construct for the expression of EMU variants 1-13	This study
pCDFDuet-WSU	pCDFDuet-1-based construct for expression of WSU	This study
pCDFDuet-EM_K_OxTe	pCDFDuet-1 + wt *bpsA* with EM subdomain according to the Kries et a. boundaries	This study

2.3 Oligonucleotide Primers and Synthesised Gene Fragments

Primers were designed using Geneious® software and ordered from Integrated DNA Technologies (IDT) custom nucleotide service, or from Macrogen. Primers were reconstituted to a concentration of 100 µM using 1 x TE buffer, and stored at -20 ℃. Working stocks at 10 µM were made by dilution in Sigma-Aldrich Milli-Q® H_2O.

Table 2.3 Primers used in this study

Primer Name	**Sequence (5' – 3')**
Introducing restriction sites into *bpsA*	
bpsA_IP_Hind3_fwd	GAAGAAGAGCAAGCTTCAGGTCAA
bpsA_IP_Hind3_rev	TATATCACTCGCCGAGCAGGTAGCGGAT
bpsA_fwd_BamHI	TATAGGATCCAATGACTCTTCAGGAGACCAGCGT
bpsA_rev_NdeI	CTTGACCTGAAGCTTGCTCTTCTTC
Correcting residue changes in *licA/tycC*	
Tyc_IP_fwd	ACACTCCAGTGCGTTCCGACC
Tyc_IP_rev	GGTCGGAACGCACTGGAGTGT
Lic_IP_fwd_SH	ACCACCCTGGACTGTGTCCCGACC
Lic_IP_rev_SH	GGTCGGGACACAGTCCAGGGTGGT
Single BpsA Homologue AA changes	

CR4_F2V_IP_fwd	CAGCAGATCGTTAGCGGTGGC
CR4_F2V_IP_rev	GCCACCGCTAACGATCTGCTG
CR6_V2L_IP_fwd	CTCATCAACCTCTACGGGCCG
CR6_V2L_IP_rev	CGGCCCGTAGAGGTTGATGAG
CR7_I2V_IP_fwd	GAGACGACGGTCAACTCGTCC
CR7_I2V_IP_rev	GGACGAGTTGACCGTCGTCTC
Constructing subdomain-swap vectors by NEBuilder	
BpsA_SmallFrag_rev	ACAGCTGCGCCTTGACCTGAAGCTTGCTCTTCTTC GACGCGTG
BpsA_SmallFrag_fwd	CGAGATATCCCTGGCGATCGAGAACCAC
BpsA_BigFrag_rev	CGATCGCCAGGGATATCTCGTCGAGCTC
BpsA_BigFrag_fwd	ACCATCATCACCACAGCCAGGATCCAATGACTCTT CAGGAGAC
EM_SmallFrag_rev	ACAGCTGCGCCTTGACCTGAAGCTTGCTCTTCTTC GACGCGTG
EM_SmallFrag_fwd	TGAAATCCGGCTGGCGATCGAGAACCAC
EM_BigFrag_rev	CGATCGCCAGCCGGATTTCATCCAGTTC
EM_BigFrag_fwd	ACCATCATCACCACAGCCAGGATCCAATGACTTTC AAAAGCCC
WS_SmallFrag_rev	ACAGCTGCGCCTTGACCTGAAGCTTGCTCTTCTTC GACGCGTG
WS_SmallFrag_fwd	CGAAGTGCGCCTGGCGATCGAGAACCAC
WS_BigFrag_rev	CGATCGCCAGGCGCACTTCGTTCGTTTC
WS_BigFrag_fwd	ACCATCATCACCACAGCCAGGATCCAATGACTGT GGACCGTCC
EM downstream border sliding	
EM_F2	ATTAATGCCTCTTCCTTCCCCGTCGACCCC

EM_F3	TCTTCCTTTACCCTCGACCCCGCCGACCTG
EM_F4	ACCCTCGAACCTGAAGACCTGGACGAGGGA
EM_F5	CCTGAATCCGTTGAAGAGGGACCGCAGTCC
EM_F6	GTTGAAGGTTATCCGCAGTCCATCTCCATC
EM_F7	TATCCGGATGCGATTTCCATCGGCTCCCCG
EM_F8	GCGATTGCGATAGGCTCCCCGGTGCACGGC
EM_F9	ATAGGCAGGCCGGTCCACGGCACCACGTAC
EM_F10	CCGGTCGCTGATACCACGTACCACATCCTT
EM_F11	GATACCCTTTATCATATCCTTGACAAGGAG
EM_F12	TATCATCTGCTGGACAAGGAGACCCTCAAG
EM_F13	CTGGACAGCGCAGGCCTCAAGCCGGTCGGC
EM_F14	AGCGCAGGCGCGCCGGTCGGCGTCGGTGAG
EM_F15	GCGCCGGTTAAGCCCGGTGAGATCGGCGAG
EM_F16	AAGCCCGGTGAGACGGGCGAGCTGTACATC
EM_F17	GAGACGGGGGAACTCTACATCGGCGGCATC
EM_F18	GAACTCTACATTTCAGGCATCCAGCTGGCC
EM_F19	ATTTCAGGTCTACAGCTGGCCCGCGGCTAC
EM_F20	CTACAGGTGGCGCAAGGCTACCTGCACCGC
EM_F21	GCGCAAGGCTACTGGCACCGCGACGACCTG
EM_R1	GGGGAACGAGGACGAATTAATCGTGCATTC
EM_R2	GGGGTCGACGGGGAAGGAAGAGGCATTAAT
EM_R3	CAGGTCGGCGGGGTCGAGGGTAAAGGAAGA
EM_R4	TCCCTCGTCCAGGTCTTCAGGTTCGAGGGT
EM_R5	GGACTGCGGTCCCTCTTCAACGGATTCAGG
EM_R6	GATGGAGATGGACTGCGGATAACCTTCAAC
EM_R7	CGGGGAGCCGATGGAAATCGCATCCGGATA
EM_R8	GCCGTGCACCGGGGAGCCTATCGCAATCGC
EM_R9	GTACGTGGTGCCGTGGACCGGCCTGCCTAT
EM_R10	AAGGATGTGGTACGTGGTATCAGCGACCGG
EM_R11	CTCCTTGTCAAGGATATGATAAAGGGTATC
EM_R12	CTTGAGGGTCTCCTTGTCCAGCAGATGATA
EM_R13	GCCGACCGGCTTGAGGCCTGCGCTGTCCAG
EM_R14	CTCACCGACGCCGACCGGCGCGCCTGCGCT

EM_R15	CTCGCCGATCTCACCGGGCTTAACCGGCGC
EM_R16	GATGTACAGCTCGCCCGTCTCACCGGGCTT
EM_R17	GATGCCGCCGATGTAGAGTTCCCCCGTCTC
EM_R18	GGCCAGCTGGATGCCTGAAATGTAGAGTTC
EM_R19	GTAGCCGCGGGCCAGCTGTAGACCTGAAAT
EM_R20	GCGGTGCAGGTAGCCTTGCGCCACCTGTAG
EM_R21	CAGGTCGTCGCGGTGCCAGTAGCCTTGCGC
WS downstream border sliding	
WS_F1	GAAACAACCATTAACTCGTCCTCGTTCCCC
WS_F2	ATTAACGCGACTTGGTTCCCCGTCGACCCC
WS_F3	ACTTGGTTTGATTTTGACCCCGCCGACCTG
WS_F4	GATTTTACTGATGTGGACCTGGACGAGGGA
WS_F5	GATGTGGATTTAACTGAGGGACCGCAGTCC
WS_F6	GATTTAACTGGGACTCAGTCCATCTCCATC
WS_F7	GGGACTGCAGTGGTTTCCATCGGCTCCCCG
WS_F8	GTGGTTCCAGTTGGTTCCCCGGTGCACGGC
WS_F9	GTTGGTACGCCAGTGCACGGCACCACGTAC
WS_F10	CCAGTGACGGGCTGTACGTACCACATCCTT
WS_F11	GGCTGTGCTACCGTCATCCTTGACAAGGAG
WS_F12	ACCGTCGTCGTGGATAAGGAGACCCTCAAG
WS_F13	GTGGATGCCTCGGATCTCAAGCCGGTCGGC
WS_F14	GCCTCGGATACCGTCGTCGGCGTCGGTGAG
WS_F15	ACCGTCGTAGGTCCAGGTGAGATCGGCGAG
WS_F16	GGTCCAGGCTCTACTGGCGAGCTGTACATC
WS_F17	TCTACTGGTGAACTGTACATCGGCGGCATC
WS_F18	GAACTGCTGATTAGCGGCATCCAGCTGGCC
WS_F19	ATTAGCGGTGCACAACTGGCCCGCGGCTAC
WS_F20	GCACAACTGGCGTCGGGCTACCTGCACCGC
WS_F21	GCGTCGGGATATCGCCACCGCGACGACCTG
WS_R1	GGGGAACGAGGACGAGTTAATGGTTGTTTC
WS_R2	GGGGTCGACGGGGAACCAAGTCGCGTTAAT
WS_R3	CAGGTCGGCGGGGTCAAAATCAAACCAAGT
WS_R4	TCCCTCGTCCAGGTCCACATCAGTAAAATC

WS_R5	GGACTGCGGTCCCTCAGTTAAATCCACATC
WS_R6	GATGGAGATGGACTGAGTCCCAGTTAAATC
WS_R7	CGGGGAGCCGATGGAAACCACTGCAGTCCC
WS_R8	GCCGTGCACCGGGGAACCAACTGGAACCAC
WS_R9	GTACGTGGTGCCGTGCACTGGCGTACCAAC
WS_R10	AAGGATGTGGTACGTACAGCCCGTCACTGG
WS_R11	CTCCTTGTCAAGGATGACGGTAGCACAGCC
WS_R12	TTGAGGGTCTCCTTATCCACGACGACGGT
WS_R13	GCCGACCGGCTTGAGATCCGAGGCATCCAC
WS_R14	CTCACCGACGCCGACGACGGTATCCGAGGC
WS_R15	CTCGCCGATCTCACCTGGACCTACGACGGT
WS_R16	GATGTACAGCTCGCCAGTAGAGCCTGGACC
WS_R17	GATGCCGCCGATGTACAGTTCACCAGTAGA
WS_R18	GGCCAGCTGGATGCCGCTAATCAGCAGTTC
WS_R19	GTAGCCGCGGGCCAGTTGTGCACCGCTAAT
WS_R20	GCGGTGCAGGTAGCCCGACGCCAGTTGTGC
WS_R21	CAGGTCGTCGCGGTGGCGATATCCCGACGC
EM upstream border sliding	
EMU_1_fwd	GCCGAAGGGTGTGGTGGTGACCGGCGCT
EMU_1_rev	TCACCACCACACCCTTCGGCTTGCCCGT
EMU_2_fwd	TGTGATGATCACCGGCGCTAATATTTCTC
EMU_2_rev	TAGCGCCGGTGATCATCACACCCTTCGG
EMU_3_fwd	CGAGCACCGCAATATTTCTCATCAGATGGCCTG
EMU_3_rev	GAGAAATATTGCGGTGCTCGATCATCAC
EMU_4_fwd	CAGCATCGTCCATCAGATGGCCTGGTTG
EMU_4_rev	CCATCTGATGGACGATGCTGCGGTGCTC
EMU_5_fwd	CAACCAGCTCGCCTGGTTGAAAAAAGAGTTTG
EMU_5_rev	TCAACCAGGCGAGCTGGTTGACGATGCTG
EMU_6_fwd	CGGCTGGCTGAAAAAAGAGTTTGGTTTTGATCATC
EMU_6_rev	ACTCTTTTTTCAGCCAGCCGAGCTGGTT
EMU_7_fwd	GCGCGAGACCTTTGGTTTTGATCATCACGACC
EMU_7_rev	CAAAACCAAAGGTCTCGCGCAGCCAGCC
EMU_8_fwd	CTACGCGATCGATCATCACGACCGGATTTTAC

EMU_8_rev	CGTGATGATCGATCGCGTAGGTCTCGCG
EMU_9_fwd	CGACCGCAGCGACCGGATTTTACAAAAAACACC
EMU_9_rev	AAATCCGGTCGCTGCGGTCGATCGCGTA
EMU_10_fwd	CAAGGTCATCTTACAAAAAACACCGGTAAGTTTTG
EMU_10_rev	TTTTTTGTAAGATGACCTTGCTGCGGTC
EMU_11_fwd	CCTCCAGAAGACACCGGTAAGTTTTGATGC
EMU_11_rev	TTACCGGTGTCTTCTGGAGGATGACCTTG
EMU_12_fwd	GACCCCGATGAGTTTTGATGCGGCGCAG
EMU_12_rev	CATCAAAACTCATCGGGGTCTTCTGGAG
EMU_13_fwd	GAGCTTCGACGCGGCGCAGTGGGAAATAC
EMU_13_rev	ACTGCGCCGCGTCGAAGCTCATCGGGGTC
licR to wild-type single AA reversions	
LicR-wt_CR2_Q-W_fwd	GATGCGGCCCAGTGGGAAATCCTGTCT
LicR-wt_CR2_Q-W_rev	AGACAGGATTTCCCACTGGGCCGCATC
LicR-wt_CR3_D-Q_fwd	GTTACCACCCTGCAGTGTGTCCCGACC
LicR-wt_CR3_D-Q_rev	GGTCGGGACACACTGCAGGGTGGTAAC
LicR-wt_CR4_L-F_fwd	CTGCAGCAGATTTTCTCCGGCGGTGAA
LicR-wt_CR4_L-F_rev	TTCACCGCCGGAGAAAATCTGCTGCAG
LicR-wt_CR7_V-I_fwd	CCTACCGAGACCACTATTGACTCTTCT
LicR-wt_CR7_V-I_rev	AGAAGAGTCAATAGTGGTCTCGGTAGG
LicR-wt_CR8_D-N_fwd	CCTACCGAGACCACTGTTAACTCTTCT
LicR-wt_CR8_D-N_rev	AGAAGAGTTAACAGTGGTCTCGGTAGG
licR epPCR/overlapPCR	
LicR_A3_fwd	ATCTATACCAGCGGTTCGACC
LicR_A8_rev	ACGGTAACCGCGCAGTTTAAC
licR epPCR/NEBuilder	
L_frag_fwd	ACCATCATCACCACAGCCAGGATCCAATGACTCTT CAGGAGACCAGC
L_frag_rev	CGATCATAACGCCTTTCGGTTTGCCGGTC
R_frag_fwd	GCAATTTGCAGGTCGCGCCGATAACCAG

R_frag_rev	GACAGCTGCGCCTTGACCTGAAGCTTGCTCTTCTTCGACG
sub_frag_fwd	ACCGAAAGGCGTTATGATCGAACATCGCAGC
sub_frag_rev	CGGCGCGACCTGCAAATTGCACGGTGCC
Amplifying *licA* module 1 A-domain	
Lic_fwd	CTGGTGCCGCGCGGCAGCCAGATTGTTCCCGCTTTTGAAAAAG
Lic_rev	CGAGTGCGGCCGCAAGCTTGATATTCATTTCTCGGTGCC
Sequencing primers	
T7_fwd	TAATACGACTCACTATAGGG
T7_rev	GCTAGTTATTGCTCAGCGG
BpsA_a_rev	AGGCCCGGGTTGGACAGCTGCGC
BpsA_Aseq_rev	AGGATCTGGCCGAGTTCG

Gene fragments were synthesised by IDT or Twist Bioscience, depending on the GC-content of the required fragment.

2.4 Bacterial Growth and Maintenance

2.4.1 Media

All media was autoclaved at 121 °C for 30 minutes. Media supplements were added before or (if filter sterilised) after autoclaving, depending on heat stability of the supplement.

Lysogeny Broth (LB)

LB powder was made up in reverse osmosis (RO)-purified H_2O to a concentration of 20 $g.L^{-1}$. Component concentrations were consequently: tryptone at 10 $g.L^{-1}$, yeast extract at 5 $g.L^{-1}$, and NaCl at 5 $g.L^{-1}$.

LB Agar

For solid media, microagar powder is dissolved in LB to a concentration of 15 $g.L^{-1}$. Where the 'lift and scoop' method of IPTG induction would be used (see Section 2.4.2), a microagar concentration of 16 $g.L^{-1}$ was used.

2.4.2 Media Supplements

Antibiotic stock solutions were made up to 1000x final concentration in the appropriate solvent and filter sterilised using a 0.22 µM filter before being stored at -20 °C. Antibiotics were generally made in RO H_2O to a working stock concentration of either 50 $mg.ml^{-1}$ or 100 $mg.ml^{-1}$. Kanamycin at a final concentration of 50 $\mu g.ml^{-1}$ was used for maintenance of pET-28a(+) based plasmids and spectinomycin was used for maintenance of pCDFDuet-1 based plasmids at a final concentration of 100 $\mu g.ml^{-1}$. IPTG stocks were prepared to a final concentration of 100 $mg.ml^{-1}$. For pigment-producing media, L-glutamine was added to LB agar at a concentration of 14.6 $g.L^{-1}$, along with appropriate concentrations of antibiotics. IPTG induction was achieved with the 'lift and scoop' method – lifting the agar from the dish and pipetting ~120 µL of IPTG evenly across the bottom of the dish. The agar was then laid down again and the IPTG allowed to diffuse through it.

2.4.3 Bacterial Maintenance

Cultures were incubated at 37 ℃ unless otherwise stated. Liquid cultures were aerated at 200 rpm. For short-term storage, agar plates were stored at 4 ℃. For long term storage, overnight liquid cultures were pelleted and mixed with 80% glycerol in a 1:1 ratio, then snap-frozen and stored at -80 ℃.

2.5 Routine Molecular Biology

2.5.1 Polymerase Chain Reaction (PCR)

PCR amplification reactions for cloning were performed using Q5® high-fidelity polymerase. GC buffer was used, to account for the high GC content of the *Streptomyces lavendulae* genome and thus *bpsA.* Alternatively, Q5® High-Fidelity 2X Master Mix was used, negating the need for a GC buffer. For other applications including colony screening, Biomix Red™ was employed. PCR reactions were set up according to the manufacturer's protocols. Parameters and additives were optimised as required for individual genes. Unless otherwise stated in the text, the following parameters were used: Initial template melting at 95 °C for 5 minutes. 35 cycles of [1. Melting at 95 °C for 1 min 30 sec. 2. Primer annealing at ~55 °C for 20 sec. 3. Polymerase extension at 72 °C for 1 min 30 sec]. For colony screening a small amount of each colony was picked from an agar plate to be used as a template. Negative (no-template) controls were routinely run simultaneously.

2.5.2 Overlap PCR

For overlap PCR a method based on the protocols described in Williams et al (2014) was used. Briefly, the two fragments to be combined were amplified by PCR, and then purified using the methods described below. The fragments were then combined in an equimolar ratio and a primerless PCR reaction run for 10 cycles. One µL of the forward and 1 µL of the reverse primer at a concentration of 10 µM were then added to the reaction, which was run for a further 25 cycles to allow the amplification of the full-length fragment.

2.5.3 NEBuilder

NEBuilder reactions using NEBuilder® HiFi DNA Assembly Master Mix were conducted according to the manufacturer's protocols with one exception: Reaction volumes were typically reduced from 20 µL to 7 µL, and DNA concentrations were adjusted accordingly.

2.5.4 Agarose Gel Electrophoresis

To assess the size and quality of DNA products including those from PCR reactions, DNA was electrophoresed on a 1% (w/v) agarose gel, containing 1 $\mu g.ml^{-1}$ ethidium bromide. Gels were electrophoresed submerged in 1x TAE buffer (40 mM Tris-acetate pH 8.0, 1 mM EDTA) at 100-140 V for 30-70 min depending on the purpose of the gel and product size. Hyperladder I (Bioline) was run alongside the samples for size comparison and bands were visualised under ultra-violet (UV) light.

2.5.5 PCR Product Purification

PCR products were purified using a Zymo-Spin™ column (Zymo Research) according to the manufacturer's instructions. Products were eluted in Sigma Milli-Q® H_2O (MQ). In cases where it was necessary to maximise yield, the elution MQ and/or column was pre-heated to 60 °C. Eluted products were stored at -20 °C.

For PCR purification using a gel purification method the product was electrophoresed on a 1% (w/v) agarose gel with SYBR® Safe (Thermo Fisher Scientific) at a concentration of 7.5 µL per 100 mL gel volume. This was then visualised on a blue light imager. The gel band was excised and purified using a Zymo Zymoclean Gel DNA Recovery Kit (Zymo Research) according to the manufacturer's instructions, with the following exceptions: Before addition to the column, dissolved agarose was quickly cooled on ice. An initial 200 µL wash with Zymo bind buffer was conducted before the prescribed 200 µL wash with Zymo wash buffer. The

wash buffer wash step was also repeated, to improve purity. DNA was eluted using MQ heated to 65 °C.

2.5.6 Restriction Enzyme Digests

Restriction digests were performed according to the manufacturer's instructions and heat inactivated at 65 °C for 20 min prior to use in further ligation reactions. Where enzymes were unable to be heat inactivated, digests were immediately cleaned and concentrated using a Zymo-Spin™ column and eluted in 10 – 20 µL of MQ. Digests that were not used immediately were stored at -20 °C. Alkaline phosphatase was used in vector digestions to avoid self-ligation.

2.5.7 Ligation

Ligations were performed according to the manufacturer's instructions, using a 1:3 or 1:6 molar ratio of vector to insert, with each reaction containing no more than 10 ng/µL of DNA. Reactions were incubated overnight at 16 °C. Ligations were then used to transform competent cells or else were stored at -20 °C.

2.6 Bacterial Transformation Techniques

2.6.1 Preparation of Electrocompetent Cells

The *E. coli* strain to be made competent was inoculated from a glycerol stock and grown in 10 ml LB supplemented with any relevant antibiotics at 37 °C, 200 rpm overnight. This overnight culture was then used to inoculate 400 ml LB containing any relevant antibiotics with a starting OD_{600} of 0.1. The culture was grown until an OD_{600} of 0.3 -0.7 was reached, at which point it was transferred to 8 sterile 50 ml tubes and cooled on ice for 30 min. Cells were then centrifuged for 15 min at 2700 g, 4 °C. For the first wash step, pellets were washed with ice cold sterile RO H_2O by gentle pipetting. The pellets were resuspended in 200 ml ice cold sterile 10% glycerol. Cells were pelleted again through centrifugation, and this time the pellet was resuspended in 100 ml ice cold sterile 10% glycerol. Cells were then finally collected through centrifugation and the pellet was resuspended in 300-500 µl of ice cold sterile 10% glycerol. The OD_{600} of a 1/100 dilution was then measured and the final concentration of cells was adjusted to 2-3 x 10^{10} cells/ml by the addition of further ice cold sterile 10% glycerol (OD_{600} of 1.0 = ~2.5 x 10^8 cells/ml). Aliquots of 200 µL were transferred to pre-chilled 1.5 ml microcentrifuge tubes on ice. Cell aliquots were then snap frozen in a metal tube block which had been pre-chilled to -80 °C. Aliquots of competent cells were stored at -80 °C until needed.

2.6.2 Transformation of Electrocompetent Cells

Frozen cell aliquots were removed from -80 °C storage and thawed on ice. Plasmid DNA was then added to the cells, typically 10- 50 ng in 20 µL of MQ. The contents of the tube were then gently mixed and transferred to an ice cold sterile 2 mm gap electroporation cuvette. Cells were then electroporated at 2.5 kV, 25 µF, 100 Ω, following which 1 mL of room temperature LB was immediately added to the cuvette and gently inverted to mix. Cells were incubated at 37 °C, 200 rpm for 1 hour for recovery. Finally, cells were plated on LB agar plates containing the appropriate antibiotics. One hundred µl of undiluted, as well as 1/10, 1/100 and (in some cases) 1/1000 dilutions, of cells were plated initially to determine the optimal dilution for single colonies to be obtained on agar plates. The remainder of the transformation was mixed in a 1:1 ratio with sterile 80% glycerol and stored at -80 °C until needed.

2.7 DNA Analysis

2.7.1 Isolation of Plasmid DNA

Isolation of plasmid DNA was achieved using the High-Speed Plasmid Mini Kit (Geneaid Biotech Ltd) according to the manufacturer's instructions.

2.7.2 DNA Quantification and Sequencing

Measurement of the concentration and purity of DNA was achieved using a Nanodrop ND-1000 spectrophotometer (Thermo Fisher Scientific), according to the manufacturer's instructions. All DNA sequencing was carried out by Macrogen Inc. using Sanger sequencing technology.

2.7.3 Identification of Recombinant Clones by Colony PCR

For initial identification of clones containing the desired insert, colony PCR was used. ~20 discrete colonies were labelled on each transformation plate. Approximately half of each colony was liberated from the plate surface onto a sterile pipette tip and inoculated into an 8 µL Biomix Red™ PCR reaction. Primers used in these colony PCR reactions were typically T7_Fwd and bpsA_IP_Hind3_rev, amplifying the multiple cloning site (MCS) into which inserts had been cloned; thus, any inserts were also amplified. After amplification, 3.5-5 µL aliquots of each PCR reaction were analysed by agarose gel electrophoresis and insert-containing colonies identified. These 'hits' were re-sampled from the transformant plates and cultured overnight, in antibiotic media where appropriate

2.8 Protein Expression and Purification

All proteins to be expressed were cloned into His_6-tag vectors and transformed into the expression strain BL21 Δ*entD* for purification via nickel affinity chromatography.

2.8.1 Expression in LB

BpsA variants were routinely expressed at low temperatures to aid with protein purification. A single colony or glycerol stock of the expression strain was used to inoculate 10 ml LB supplemented with appropriate antibiotics and grown overnight at 37 °C, 200 rpm. The following day this overnight culture was used to inoculate a baffled 2.5 L flask with 400 ml of LB containing appropriate antibiotics, giving a starting OD_{600} of 0.1. This was incubated at 37 °C, 200 rpm until an OD_{600} of 0.6-0.8 was reached. Cultures were then transferred to an ice bath for 30 min before addition of 0.5 mM IPTG. Cultures were then grown for an additional 24 hours at 18 °C 200 rpm before cells were harvested through centrifugation at 2700 g, 20 min, 4 °C. The supernatant was discarded, and cell pellets were stored at -80 °C until required.

2.8.2 Cell Lysis and Soluble Fraction Separation

For protein purification, cell pellets were thawed and resuspended in 35 ml of modified1x His•bind™ binding buffer, which contained no imidazole. For BpsA purification this binding buffer was also supplemented with 12.5% glycerol (v/v). Cells were lysed by sonication at 25% amplitude, 3 pulses of 30 seconds over 3 minutes. Following lysis, the soluble and insoluble fractions were separated by centrifugation at 18,000 g for 25 min at 4 °C. Where downstream applications called for *holo*-form protein, cell lysate was incubated on ice for 10 minutes with an equal volume of lysate from cells expressing non-His-tagged *pcpS* PPTase before centrifugation.

2.8.3 Purification of Proteins by Ni-NTA Affinity Chromatography

2.8.3.1 Purification of BpsA

Purification of His_6-tagged BpsA and BpsA-variant proteins was achieved using Novagen's His•bind™ Ni-NTA chromatography kit. The following changes to the manufacturer's instructions were made for the purification of BpsA (due to the weak binding of BpsA to the resin). All buffers were supplemented with 12.5% glycerol (v/v) and all reagents were kept at ice cold temperatures for the duration of the purification. A large volume, approximately 6-7 ml, of resin was used. Columns were washed using 40 ml of binding buffer. The wash buffer

step that is usually employed at this stage of purification (according to the manufacturer's instructions) was omitted for BpsA. For the elution, the resin was resuspended in 8 ml elution buffer, and protein was eluted in an 8-10 ml total volume. Buffer exchange and concentration of the eluted protein was performed in 100 kDa molecular mass cut-off Millipore Amicon® ultra-15 centrifugal filter units (Merck Millipore), into 50 mM sodium phosphate buffer, pH 7.8, 12.5% (v/v) glycerol. For phosphate-sensitive downstream applications, proteins were suspended in a buffer of 50 mM Tris-HCl pH 7.8, 10% (v/v) glycerol. Final buffer composition was then adjusted to 40% (v/v) glycerol and aliquots were stored at -20 °C.

2.8.3.2 Purification of LicA A-domain

Purification of the LicA A-domain was also achieved using Novagen's His•bind™ Ni-NTA chromatography kit. Buffers were supplemented with 10% glycerol (v/v) and again kept ice cold throughout purification. Buffer exchange and concentration of the eluted protein was performed in 30 kDa molecular mass cut-off Millipore Amicon® ultra-15 centrifugal filter units (Merck Millipore), into 50 mM Tris-HCl buffer, pH 7.8, 10% (v/v) glycerol.

2.8.4 Protein Concentration Quantification

Protein concentrations were determined using absorbance at 280 nm using a Nanodrop ND-1000 spectrophotometer (Thermo Fisher Scientific), according to the manufacturer's instructions.

2.9 SDS-PAGE Analysis

For qualitative assessment of proteins, 12.5% or 15% SDS-Polyacrylamide Gel Electrophoresis gels were run to visualise proteins depending on the protein size. These were cast and electrophoresed on a 1 mM thickness Bio-Rad Protean II™ apparatus. For each gel, 5 ml of separating gel was loaded into the apparatus before being overlaid with isopropanol and left to set for 45 min. The isopropanol was then removed, and 1.5 ml of stacking gel was poured on top. A multi- well comb was inserted, and this was left to set for at least 45 min. If not used immediately, gels were stored for up to three days at 4 °C, wrapped in a paper towel soaked in RO H2O and tin foil. Samples were added to 3x loading buffer in a 2:1 ratio, and incubated at 95 °C for 10 min. Gels were electrophoresed in 1x SDS run buffer, at a constant voltage (180 V) for 60 - 75 min, until the pre-stained ladder had reached proper separation. Gels were then stained in Coomassie blue stain and gentle shaking was applied for 30-60 min. Gels were then

rinsed using tap water and left in destain solution with gentle rocking until the desired level of destaining was reached.

2.10 Generating *bpsA* Variant Screening Constructs

Most *bpsA* variants were generated by some combination of fragment amplification and/or overlap PCR, described in Section 2.5.

Directed evolution of the LicA residue construct subdomain was achieved using the GeneMorph® Mutazyme® II kit, according to the manufacturer's instructions. Template was prepared by amplification of the subdomain using Q5® high-fidelity polymerase. A high error rate was chosen, which was achieved through use of 1 ng purified PCR product as template in a 50 µl reaction and 35 amplification cycles. Prior to thermocycling, reactions were divided into equal amounts (12.5 µl) across 4 tubes to reduce likelihood of clonal mutations in the final library. The size and quality of the amplicons was determined by running a 3 µl aliquot of total pooled reactions on an agarose gel. The subdomain was then assembled to the upstream and downstream fragments by overlap PCR (Section 2.5.2) or NEBuilder (Section 2.5.3).

2.10.1 Insert Preparation

PCR-amplified fragments were purified using the Zymo DNA Clean & Concentrator™-5 ng protocol and eluted in MQ. The concentration of each purified reaction was determined by analysing a 1.2 µl aliquot on a Nanodrop ND-1000 spectrophotometer. One µg (or all eluent in the case of epPCR constructs) was then digested with 10 U of each restriction enzyme (BamHI and HinDIII) overnight, according to the manufacturer's protocol. Digested fragments were again purified using the Zymo DNA Clean & Concentrator™ -5.

2.10.2 Vector Preparation

The pCDFDuet-OxTe plasmid was isolated using a Geneaid® High-Speed Plasmid Mini Kit according to the manufacturer's instructions. For each column, 2.5 mL of overnight culture was used for a yield of 2.5 – 5 µg. For digestion, each µg of vector was digested with 10 U of each restriction enzyme (BamHI and HindIII), following the manufacturer's instructions for buffer composition and temperature. Following an overnight incubation, alkaline phosphatase was added to the reaction and incubated according to the manufacturer's instructions. Digests were then heat inactivated at 80 °C for 20 min and purified using a 5 µg capacity Zymo-Spin™ clean up column. Digested vector was eluted in 15 µL of MQ and then gel purified using the method outlined in Section 2.5.5. Following this the concentration of the vector was quantified

by analysing a 1.2 µL aliquot on a Nanodrop ND-1000 spectrophotometer. Vector was then aliquoted and stored at -20 °C, or for short periods (less than 1 week) at 4 °C.

2.10.3 Assessment of Vector Quality

It was important to assess vector quality before use in directed evolution library generation experiments as the intention was to visually screen for colonies that produced a blue pigment rather than using a selection-based method, meaning it was important to have a high insert rate. The wild-type *bpsA* core domain was amplified using Q5® high-fidelity polymerase and prepared for ligation as described above (Section 2.10.1). This ligation was then used to transform electrocompetent competent BL21 Δ*entD* cells harbouring a pET-pcpS plasmid. The transformation was plated onto LB agar plates containing 100 mM L-glutamine, kanamycin and spectinomycin. Plates were incubated at 37 °C overnight and BpsA expression was induced as described in Section 2.11.1. Plates were incubated at 18 °C. Transformation and ligation efficiency were determined by comparing numbers of blue colonies vs. white colonies after 8-12 h. Blue colonies indicated a successful ligation, while white indicated a self-ligation or uncut vector. A successful ligation rate of 70% or higher was deemed sufficient for use in further experiments.

2.11 Screening for Functional Constructs by Indigoidine Production

2.11.1 Solid Media *In Vivo* Screening

This screening took place on solid agar plates. For directed evolution library variants, colony density was carefully controlled to achieve 4000-8000 colonies per plate.

After 16 hours incubation at 37 °C, expression was induced in colonies on the agar plates. Induction was achieved by lifting the agar out of the plate and spreading 120 µL of 3% IPTG on the bottom of the agar plate and then replacing the agar back into the plate which allowed the IPTG diffuse through the agar. Plates were then incubated at 18 °C and monitored for the production of pigment. Clones were then picked from the plate and grown overnight in LB medium supplemented with 0.4% glucose and appropriate antibiotics. Glycerol stocks of clones were then prepared.

2.11.2 Liquid Media *In Vivo* Screening

Glycerol stocks of appropriate strains were used to inoculate 200 µl of LB containing appropriate antibiotics and then incubated overnight at 37 °C, 200 rpm in a 96 well plate. Twenty µl of each culture were used to inoculate a fresh 96-well microplate, with each well

containing 130 µl LB amended with 100 mM L-glutamine, 0.6 mM IPTG, and appropriate antibiotics. Plates were wrapped in foil and incubated at 18 °C, 200 rpm for 24-48 h. At appropriate time points based on visual inspection, OD_{590} and OD_{800} values were recorded using a PerkinElmer Enspire plate reader. After a reading was taken the plate was again incubated at 18 °C, 200 rpm.

2.12 Quantification and Normalisation of Indigoidine Production

A normalised value for indigoidine absorbance was calculated using the method described by Beer et al., (2014), as adapted from Myers et al., (2013). The OD values recorded represent the indigoidine sensitive wavelength (OD^S) of 590 nm; as well as the robust wavelength (OD^R) of 800 nm, which accounts for the cellular components present. $OD^S_{B,+I}$ is the OD at 590 nm which contains absorption from both bacterial cells and the indigoidine pigment, while $OD^S_{B,-I}$ accounts for the scattering of just the cellular components at OD^S (i.e. at 590 nm). $OD^S_{B,-I}$ is calculated as:

$$OD^S_{B,-I} = \delta.OD^R$$

where δ is the correction factor and is calculated by measuring the ratio of pure cell culture at both OD^S (590 nm) and OD^R (800 nm). The normalised calculation for indigoidine is then calculated as:

$$\text{relative indigoidine} = OD^S_{B,+I} - \delta.OD^R$$

Background correction was achieved by subtracting the mean of "culture medium alone" replicates from every value first.

2.13 Pyrophosphate (PP_i) Release Assay – Malachite Green Assay

This assay is designed to measure the adenylation activity of an A-domain by quantifying the PP_i released from ATP when adenylation occurs. PPi is metabolised by inorganic pyrophosphatase (IP) into Pi, which forms a green complex with the detection reagents.

This assay is based on the RA2 assay developed by Dr Jeremy Owen (Owen, 2010), which is itself based on an assay developed by the Garneau-Tsodikova Lab (McQuade et al., 2009).

PPgreen reagent A: 4.2 % w/v sodium molybdate in 4 M sulfuric acid

PPgreen reagent B: 0.135 % w/v brilliant green dye in sterile RO H2O

Triplicate reactions for each substrate were established in a 96 well plate and in a final volume of 100 µL contained: 50 mM Tris-Cl (pH 7.8), 10 mM $MgCl_2$, 1 mM ATP, 5-20 µg enzyme, 0.03 U IP and 1.25 mM amino acid substrate. Plates were then incubated for 30 min at 30 °C, after which reactions were terminated by addition of 20 µL PPgreen reagent A per well. Plates were then mixed at 500 rpm, and 20 µL PPgreen reagent B was added to each well. Plates were again mixed at 500 rpm and A_{650} measured immediately using a microplate reader.

2.14 In Vitro BpsA Kinetics

Kinetic analysis of BpsA was conducted using an adaption of the method established by Owen et al. (2011). Active *holo*-form BpsA variants were suspended in 80 µL of a reaction mix containing 50 mM Tris-HCl pH 8.5, 10 mM $MgCl_2$, 5 mM ATP, and 0.1 µM protein. Reactions were initiated by addition of 20 µL of 25 mM L-Gln, initiating indigoidine production. Indigoidine was then measured at 590 nm every 10 seconds using a PerkinElmer Enspire plate reader. Determining the rate of change of indigoidine concentration using the slope function in Microsoft Office Excel gave a measure of maximum enzyme velocity in each well.

3 Exploration of Coding Residue Diversity and Function

3.1 Introduction

The modular structure of NRPSs, and the utility of their products, make them prime candidates for enzyme engineering. Coding residue reprogramming of the A-domain once appeared to be a promising means to this end, but experiments over the past two decades have shown that poorly understood complexities abound that limit the efficacy of this approach (Brown et al., 2018).

An early study by Eppelmann et al. (2002) successfully reprogrammed an L-glutamate A-domain to activate L-glutamine with no loss of catalytic activity. However, L-Glu and L-Gln are very similar amino acids, and it was necessary to change only a single coding residue to achieve the switch in activity (a lysine at the base of the binding pocket was changed to glutamine; Table 3.1). While prior research has shown that drastic changes in substrate specificity are observed in nature with only a few point mutations (Christiansen et al., 2011), drastic changes engineered via reprogramming have been accompanied by significant reductions in catalytic activity. For instance, informed by improved specificity-code-predicting algorithms, Chen and colleagues successfully reprogrammed an L-Phe-specifying A-domain to activate L-Leu, but achieved only around 15% the activity of wild-type (Chen et al., 2009).

Table 3.1 Initial and reprogrammed specificity codes from the studies of Eppelmann et al (2002) and Chen et al. (2009). Asterisks denote changed coding residues

Coding Residue	1	2	3	4	5	6	7	8
Eppelmann et al. (2002) Glu to Gln		*						
L-Glutamate	Ala	Lys	Asp	Leu	Gly	Val	Val	Asp
L-Glutamine	Ala	Gln	Asp	Leu	Gly	Val	Val	Asp
Chen et al. (2009) Phe to Leu			*		*			
L-Phenylalanine	Ala	Trp	Thr	Ile	Ala	Ala	Ile	Cys
L-Leucine	Ala	Trp	Leu	Ile	Gly	Ala	Ile	Cys

It seems success can be limited not just by dissimilarity between the native and target substrates, but by the extent of change required to the coding residues themselves. In reprogramming an L-Asp A-domain for L-Asn recognition, Eppelmann et al. (2002) found that every additional change they introduced to the coding residues incurred additional losses in activity, despite their predictions that those changes should improve substrate recognition.

A more complete understanding of the role each coding residue plays in substrate recognition, and the way those residues interact with the surrounding protein structure, is required. With that over-arching goal in mind, the research described in this chapter investigated variation in the specificity codes of L-glutamine A-domains present in nature. Following examination of the existing variation in coding residues, experiments were performed to assess how well a particular host A-domain, that of BpsA, could tolerate that variation. It was reasoned that naturally occurring variations should be the most accessible changes, and that if these don't work in BpsA, it will be evidence that the active site context is influenced by second- and third-shell residues relative to the active site (e.g. the potential of these more distant residues to impose steric constraints or provide stabilising hydrogen bonds).

The use of BpsA as a host enzyme means that activity can be measured in terms of relative indigoidine production. This has advantages over the ATP exchange assays commonly used to assess NRPS activity. While the ATP assays (e.g., the malachite green assay; see Methods section 2.13) provides a measure of adenylation activity, this is only one aspect of the enzymatic activity of the A-domain, much less the NRPS as a whole. In contrast, measurement of the final product, indigoidine, confirms that the A-domain can not only adenylate, but also thiolate the target substrate, and that downstream domains are unimpeded.

Reprogramming experiments were restricted to glutamine specificity codes; no change to A-domain specificity was attempted. Specifically, we were interested in developing a comprehensive picture of the transferability of a single amino acid sequence space. The exclusive focus on L-Gln as a substrate was also a pragmatic consideration, as BpsA will only produce indigoidine from L-Gln, and we were reliant upon indigoidine production as a screen for functionality.

3.2 Identification of Coding Residues in the BpsA A-domain

The first step in examining the coding residues in BpsA and other NRPS enzymes was to develop a consistent method for identifying these residues and placing them in their structural context. The amino acid sequence and crystal structure of GrsA (PDB:1AMU), solved by Conti et al. (1997), established the basis for amino acid numbering of all A-domain coding residues. While individual A-domains vary in length and sequence identity, they can be aligned against each other by the 10 highly conserved A-domain motifs (A1- A10). When aligned according to the motifs, the coding residues of any known A-domain can be determined by comparison to the primary sequence of GrsA. For instance, Coding Residue 4 (CR4) in GrsA is the

isoleucine at position 299. This corresponds to position 284 in BpsA, where there is a phenylalanine - thus, CR4 in BpsA is Phe (Fig. 3.1). It is important to note that the coding residues are named according to their location in the primary sequence of NRPS A-domains, and their number does not reflect their relative location in 3D space within the binding pocket. For example, in the tertiary structure, CR1 is physically adjacent to both CR7 and CR2.

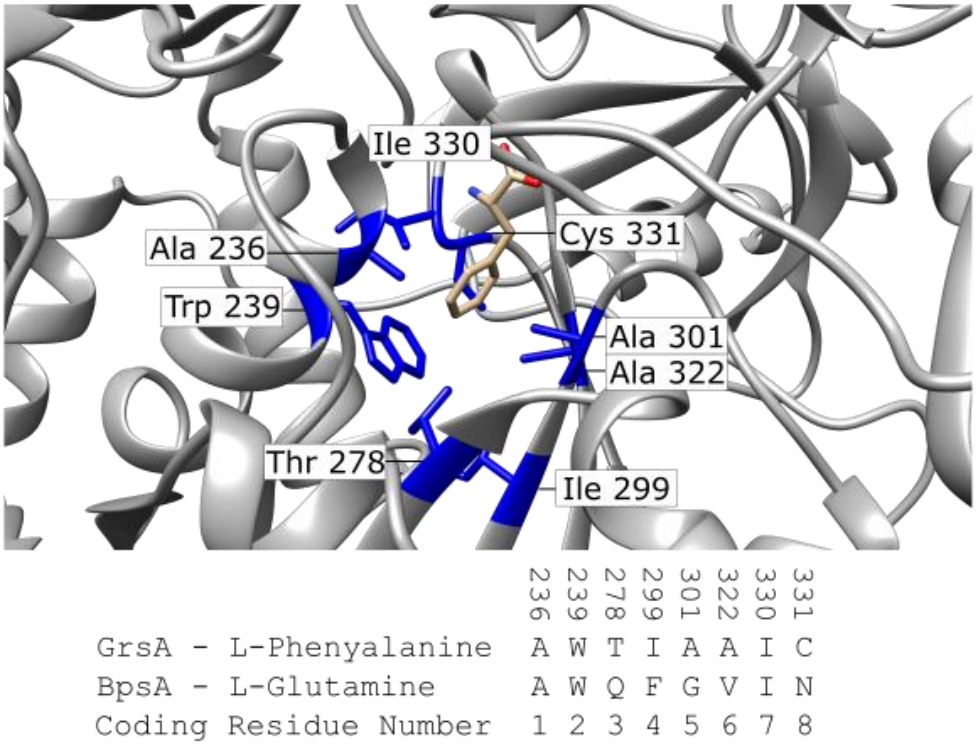


Figure 3.1 Crystal Structure of the binding pocket of GrsA (PDB:1AMU). Coding residues investigated in this study are shaded blue and labelled according to the GrsA protein sequence. The GrsA coding residues and corresponding residues in BpsA are aligned at the bottom of the figure, with coding residue (CR1 – CR8) denoted below.

3.2.1 Choosing the Coding Residues to be Assessed

Past studies examining A-domain specificity have differed in their assessment of how many amino acids comprise the 'coding residues' of the A-domain binding pocket. A seminal paper by Stachelhaus et al. (1999) identified 10 coding residues in the GrsA crystal structure that contribute to substrate specificity, beginning with the Asp residue at position 235. The subsequent study of Challis et al. (2000) implicated 9 of these 10 residues. Ultimately, we chose to focus on 8 residues in BpsA (CR1-CR8), these being the equivalents of residues 2-9 in the Stachelhaus paper, corresponding to Ala236 through Cys331 in GrsA. Two of the residues identified by Stachelhaus et al. and Challis et al. were omitted from our study; these were residues 1 and 10 as defined by Stachelhaus et al. The former of these was omitted because we found it to be universally conserved across L-glutamine-activating A-domains, which are

the primary focus of this study. Furthermore, it is almost universally conserved across A-domains as a whole, deviating in only 3% of known A-domains (Stachelhaus et al., 1999). Likewise, the 10th residue defined by Stachelhaus et al. was omitted because it is also universally conserved, with lysine present in all known A-domains.

The residue we selected as CR8 was not considered a binding pocket residue in the study of Challis et al. (2000). The researchers concluded that because the amino acid sidechain pointed away from the binding pocket in GrsA, it would not exert a selective effect on the pocket. They did, however, recognise that there was potential for this residue to play a role in substrate definition in other A-domains. Consistent with this, Stachelhaus et al. (1999), found that the residue at this position varied between A-domains that exhibit different substrate specificities – at 26% variability, it was the 3rd most variable of the 10 residues they selected. Based on the substrate-specific variability of this residue, and its prevalent and ongoing consideration in contemporary studies e.g. Throckmorton et al. (2019), we concluded that this residue is likely to play a role in selecting glutamine as a preferred substrate.

3.3 Coding Residue Substitution Testing Strategy

To test and compare the levels of activity resulting from different coding residue substitutions in BpsA, we built upon previous systems used in our lab for expression of BpsA in *E. coli*. We designed a vector containing silent restriction sites which allowed for the seamless integration of *bpsA* variants, after which it could be used to transform *E. coli* strains for indigoidine assays. We implemented a multi-tier strategy to allow us to assess activity in various contexts. The highest throughput but lowest accuracy test (described in Section 2.11.1) was a solid-media multi-day visual assay where colonies on agar plates were observed for the production of indigoidine. Following this assay, colonies could be picked and grown in liquid culture for a further liquid-culture assay (Section 2.11.2), which allowed for more accurate comparison of variant activity. This was usually sufficient as a measure of relative activity, but a third assay, *in vitro* analysis using purified protein, was used in cases where sufficiently few variants were being tested as to make it practical, and where distinctions could not confidently be made on the basis of *in vivo* assays alone.

3.3.1 Design and Construction of an A-domain Substitution Vector

To test the activities of A-domain substitutions in BpsA, a substitution vector was developed, into which alternative (e.g., reprogrammed) A-domains (up to the A8 motif) could be inserted and expressed as a full NRPS (Fig. 3.2). Incorporating the downstream region of *bpsA* into the

vector itself meant that manipulation and directed evolution efforts could be focused upon the A-domain region of interest alone, without the risk of introducing errors into the unchanging Ox – PCP – TE region.

3.3.1.1 Design

To ensure consistency of gene expression, the same plasmid vector was designed to be used in both the coding residue substitution experiments in this chapter, and the subdomain substitution experiments in the next chapter. The vector was designed to contain the downstream region of *bpsA* – encoding everything except the core structural subdomain of the A-domain, N-terminal to the Ox-domain. Into this vector the A-domain of *bpsA*, or the coding residue- or subdomain-variants, could be inserted. The BpsA enzyme characterised by Takahashi et al. (2007) was identified in *Streptomyces lavendulae*, and the corresponding gene was amplified from genomic DNA.

3.3.1.2 Choice of Plasmid Backbone

The plasmid backbone chosen for development of the substitution vector needed to encode a His-tag at the beginning of the open reading frame so that proteins could be purified by affinity chromatography, to assess protein kinetics *in vitro*. In order to function, NRPS enzymes require the Ppant prosthesis to be attached to their PCP-domain by the activity of a PPTase enzyme. As such, the A-domain substitution vector needed to be compatible with pET-28a(+) based plasmids like pNOHISPET-pcpS, the system typically used in the Ackerley lab for co-expression of the PPTase PcpS from *P. aeruginosa* (Owen et al., 2011) Furthermore, the plasmid backbone needed to include restriction sites that allowed separate and adjacent cloning of the A-domain and downstream Ox-PCP-TE-portion of BpsA. The plasmid backbone pCDFDuet-1 was found to fulfil all these requirements.

3.3.1.3 Choice of Insert Borders

To determine the downstream portion of *bpsA* to incorporate into the staging vector, the border between the A-domain region to be substituted and the downstream *bpsA* sequence was established.

The upstream border of the A-domain insert was necessarily the beginning of the *bpsA* gene. Due to DNA synthesis constraints, it was ideal for the downstream border of inserts to be within

1.8 kb of that upstream border. The downstream border was also selected to avoid the boundary with the Ox-domain, where changes to the primary sequence could disrupt function.

The precise location of the insert-Ox-PCP-TE border was dictated by locations where an appropriate (i.e. present in pCDFDuet-1) restriction site could be silently introduced into *bpsA*. The tool 're site finder' http://resitefinder.appspot.com/ was used to identify a point where a HindIII restriction site could be introduced with a single silent nucleotide change. The HindIII border site cuts at K497/L498 in *bpsA*.

The restriction sites at the beginning and end of *bpsA* as a whole were chosen as BamHI and NdeI respectively, as shown in Figure 3.2. Between the HindIII and NdeI sites in empty pCDFDuet-1 there is a second T7 promoter region. This is designed to allow two distinct, independently expressed genes to be incorporated into the plasmid. When the plasmid was cut and the Ox-PCP-TE region was ligated in, this redundant second promoter was excised. It was also necessary to introduce a single-nucleotide spacer to the beginning of each A-domain substitution, to keep the *bpsA* gene-variants in-frame with the His-tag upstream of the BamHI restriction site.

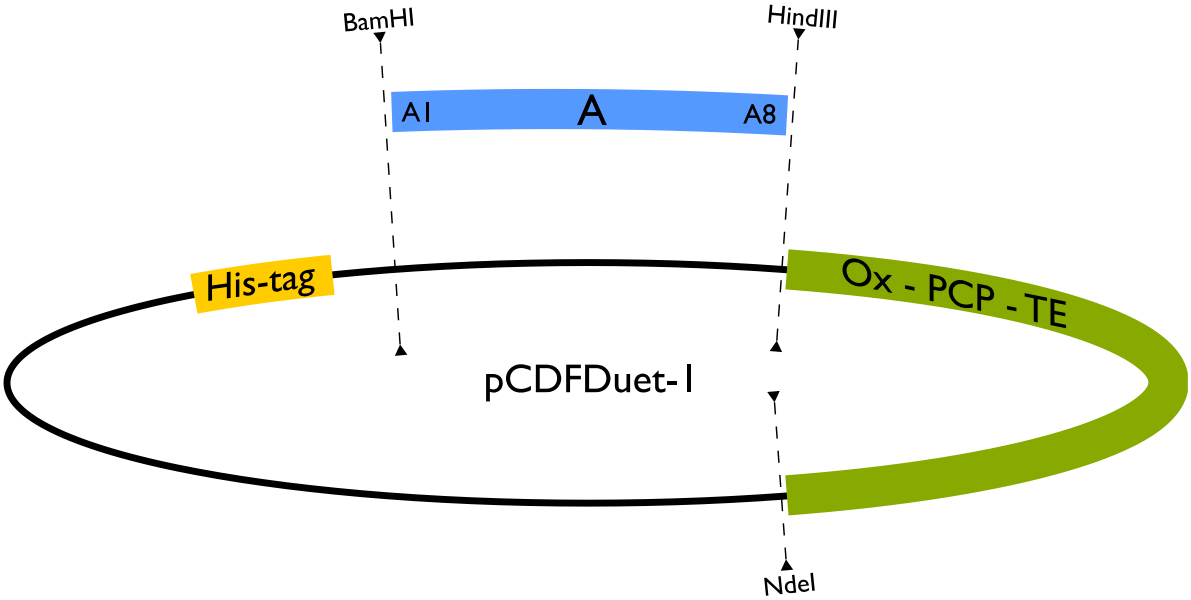


Figure 3.2 Schematic of the A-domain substitution vector pCDFDuet-OxTe, constructed from pCDFDuet-1. In green is the bpsA_OxTe region introduced during vector construction. The blue segment represents insertion of any reprogrammed or recombined bpsA A-domain sequence up to the A8 motif. The His-tag (required for nickel-affinity chromatography) is represented by the yellow segment. The plasmid backbone is pCDFDuet-1. Restriction enzyme recognition sites are labelled, and their locations indicated by dashed lines.

3.3.1.4 Construction of the Vector

The Ox-PCP-TE region of BpsA was amplified from pCDFDuet-bpsA, using primers bpsA_IP_Hind3_fwd and bpsA_rev_NdeI that introduced HindIII and NdeI restriction sites to the start and end of the sequence respectively. Digesting the pCDFDuet-1 plasmid with these restriction enzymes excised the second T7 promoter region between the two multiple cloning cassettes of the plasmid. The Ox-PCP-TE amplicon was ligated into the digested plasmid and the resultant vector was used to transform *E. coli* DH5α, an *E. coli* strain with recombinase and endonuclease knocked out, making it an ideal strain for plasmid storage. Transformed bacteria were plated on LB agar containing spectinomycin (pCDFDuet-1 contains the $spec^R$ spectinomycin resistance gene) and colony PCR was performed to identify colonies with the correctly inserted fragment. This A-domain substitution vector was named pCDFDuet-OxTe.

3.3.1.5 Validation

The integrity of pCDFDuet-OxTe was first validated by sequencing (See Section 2.2 for primer details), then further validated by testing incorporation of the wild-type *bpsA* A-domain into the plasmid. Much the same process occurred for inserting the A-domain as for incorporating the Ox-PCP-TE fragment in Section 3.2.2: PCR primers bpsA_fwd_BamHI and bpsA_IP_Hind3_rev were used to introduce the restriction sites (BamHI and HindIII) to the ends of the fragment amplified from pCDFDuet-bpsA, and the fragment was ligated into pCDFDuet-OxTe. Upon transformation into *E. coli* BL21 co-expressing the broad-range PPTase PcpS (hereafter called BL21 *pcpS*), functional wild-type BpsA was produced, as evidenced by colonies turning blue due to the presence of indigoidine (Fig. 3.3).

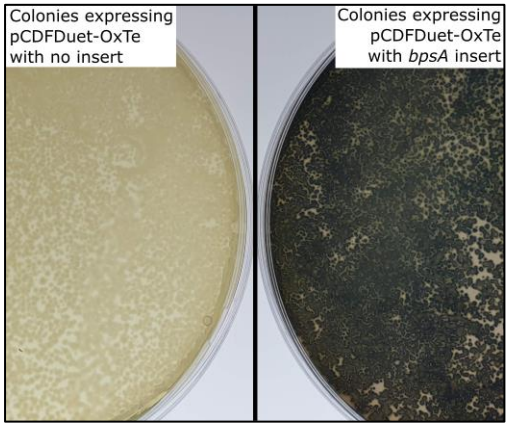


Figure 3.3 Agar plates validating the A-domain substitution vector. Blue colonies are indicative of a functional variant, leading to indigoidine production.

3.4 Coding Residue Permutations among Indigoidine Synthetases/BpsA Homologues

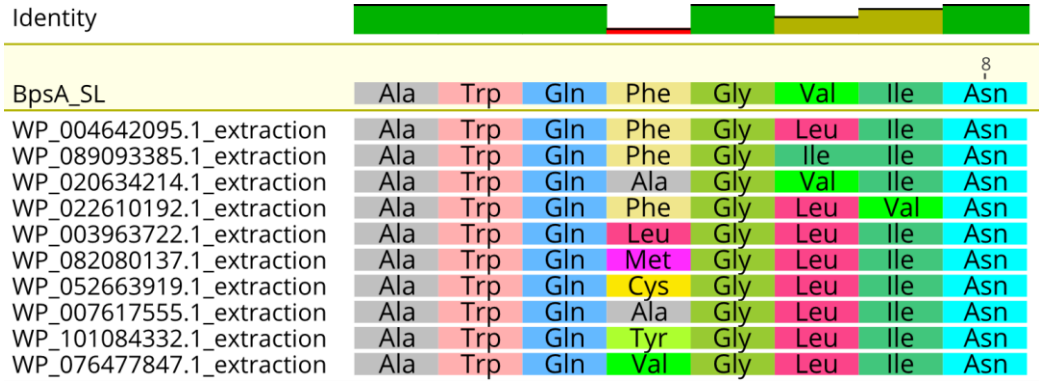
3.4.1 Identifying BpsA Homologue Specificity Codes

We next examined the degree of variability found in the specificity codes of A-domains most closely related to BpsA. A BLAST search was conducted against the GenBank accession number of the BpsA protein sequence uploaded by Takahashi et al. (2007), BAE93896, and the top 1000 hits were downloaded and then filtered using scripts written in Python by my co-supervisor Dr Alistair Brown. The first filtering criterion was size, removing any NRPSs that were too large (i.e. multi-module) or any incomplete fragments, after which 537 results remained. These were then filtered for the presence of an Ox-domain, i.e. sequences annotated with "mbc-like_oxidoreductase". Some more distantly related homologues occasionally had Ox-domains annotated as "Nitro_FMN_reductase", so these were also included in the filter, generating a list of 407 sequences. Following a manual curation and removal of multiple records from the same species, a final set of 236 unique sequences was collated – this curation is referred to as the "BpsA homologues". Across the recognition subdomain, the sequences shared a median 57.6% pairwise identity.

To distinguish 'wild-type' BpsA (BAE93896 from *S. lavendulae*) from the BpsA homologues, it is denoted as BpsA_SL for the remainder of this thesis.

A further script was used to extract the coding residues from an alignment of the A-domains of the 236 homologues. In total, we identified 11 unique specificity codes among the BpsA homologues, including that of BpsA_SL itself (Table 3.2).

Table 3.2 Geneious extraction showing the coding residues of each unique specificity code from the BpsA homologue curation. BpsA_SL is the wild-type BpsA coding residues, against which the homologue sequences were aligned and extracted. BpsA homologue specificity codes are labelled with the NCBI reference sequence of a representative protein, e.g. WP_076477847 is a BpsA homologue from Williamsia sterculiae. The "Identity" row summarises the degree of variability within columns.



3.4.2 Selection of BpsA Homologue Coding Residues for Assessment in BpsA

Analysis of the BpsA homologue curation revealed variation at three points in the specificity code – at CR4, CR6, and CR7. Table 3.3 shows the frequency of each residue among the homologues.

Table 3.3 Comparison of BpsA_SL coding residues and variations identified among BpsA homologues. The most common residue at each position is shaded in green.

	CR1	CR2	CR3	CR4	CR5	CR6	CR7	CR8
BpsA_SL	Ala	Trp	Gln	Phe (209)	Gly	Val (16)	Ile (227)	Asn
Homologues				Tyr (16)		Leu (219)	Val (9)	
				Cys (4)		Ile (1)		
				Val (3)				
				Ala (2)				
				Leu (1)				
				Met (1)				

At CR4, where BpsA_SL contains a Phe, several other amino acids are represented across the homologue curation. This suggests that within L-Gln coding A-domains, the CR4 position is flexible. In GrsA, CR4 is positioned at the base of the A-domain binding pocket (Fig. 3.1). In this position in the curated BpsA variants there was substantial variation in residue size, ranging from large aromatic Phe and Tyr to relatively small Val and Ala. With the exception of four cysteine-containing variants, all residues in this position were hydrophobic, suggesting that

there has been selection for hydrophobic residues, with little emphasis placed on size. To investigate this, we replaced the CR4 position in BpsA_SL with Val, to examine the effects of a very differently sized residue that possessed the same hydrophobicity characteristics (construct F-CR4-V).

CR6 is of particular interest for our BpsA screening system because for this residue, BpsA_SL is in the minority. The majority of similar A-domains encode a leucine at CR6, whereas BpsA_SL encodes a valine. Structurally, these two amino acids are similar, and are similarly hydrophobic. With such a clear majority of variants opting for leucine, the question arises whether a CR6-Leu imparts an increase in catalytic efficiency that BpsA_SL is yet to happen upon, or if the structure of BpsA_SL is such that Val is catalytically preferable, or if the two residues are indistinguishable. The construct V-CR6-L was generated to address this question, introducing the CR6-leucine to otherwise-unchanged BpsA_SL.

At the CR7 position, two amino acid possibilities exist among the BpsA homologues. BpsA_SL contains the majority option, isoleucine. As was the case with CR6, the alternative residue (valine) is very similar in both structure and charge. A construct was generated to investigate the effects of a substitution of this residue in BpsA_SL also, with construct I-CR7-V replacing the isoleucine with valine.

3.4.3 Construction of BpsA Homologue Coding Residue Substitutions

The three homologous-residue gene constructs (F-CR4-V, V-CR6-L and I-CR7-V) were generated, using overlap PCR, from a wild-type *bpsA* gene template in pCDFDuet-bpsA. The A-domains were amplified in two halves, each altering the target residue. Slight mismatch between the internal primers and wild-type sequence at the coding residue introduced the change in both half-domains (primers are listed in Table 2.3). Subsequently, overlap PCR was performed (as per the protocol outlined in Section 2.5.2) to generate a full-length gene encoding the desired substitution.

All CR-variant gene amplicons were digested with BamHI and HindIII, ligated into pCDFDuet-OxTe and used to transform BL21 *ΔentD* (an *E. coli* expression strain in which the endogenous PPTase gene *entD* has been deleted, to prevent conversion of BpsA into the *holo*-form prior to purification and thereby evade issues associated with mild indigoidine toxicity to *E. coli* cells). A colony PCR was run to identify colonies containing plasmid inserts of the correct size, which were then purified and sent for Sanger sequencing. Sequence data confirmed the correct plasmids had been generated for all 3 constructs. The V-CR6-L sequence

was found to contain an A→C point mutation at base 1371, however this was silent (residue 435 GTC Val → GTA Val).

When transformed into BL21 *pcpS* and grown on plates with appropriate antibiotics, all three single-residue substitution constructs gave rise to blue colonies (Fig. 3.4). This acted as a secondary confirmation of sequence, and also gave a first impression of the catalytic activity of reprogrammed variants, addressed in more detail below.

3.4.4 BpsA Homologue Coding Residue Substitution Functionality

With coding residue-substituted *bpsA* variants established in the pCDFDuet-OxTe vector, indigoidine production assays, both *in vivo* and *in* vitro, were conducted to assess their activities. As these were the first re-engineered constructs to be generated in this project, we sought to gauge the accuracy of the various assays intended to screen variant activity.

3.4.4.1 In Vivo Assessments of Enzyme Activity

The first screen for BpsA variant activity is a test on solid media. Variants were plated and induced using the plate-scooping technique described in Section 2.11.2. Relative activity, in comparison to each other or to BpsA_SL, could not be determined by visual inspection as all variants started producing blue pigment in the same timeframe (~12 hours), however it was clear the single-residue substitutions were functional.

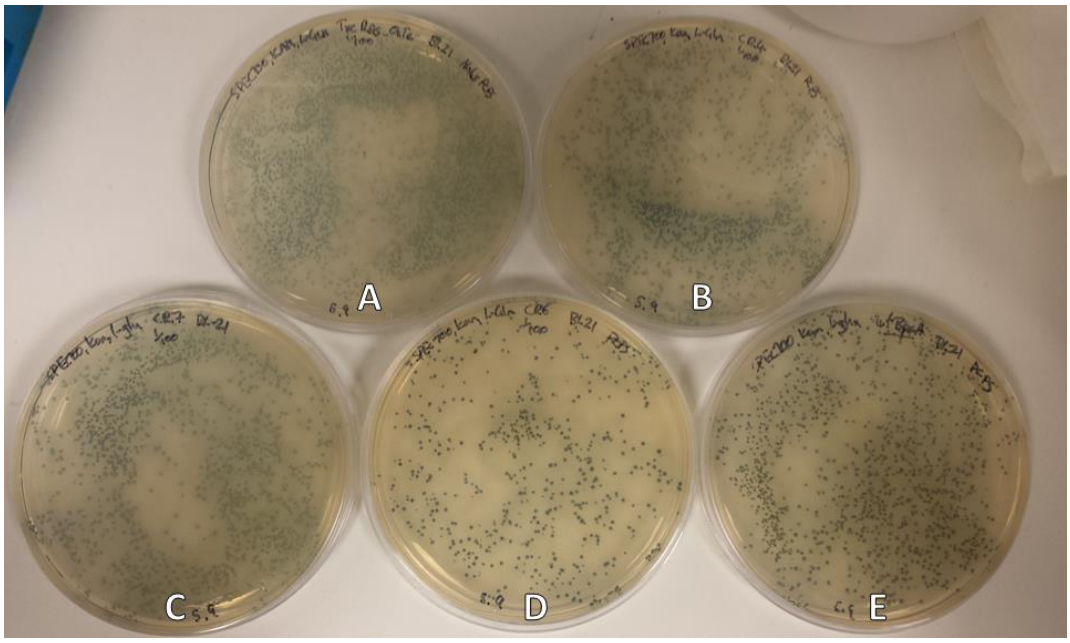


Figure 3.4 IPTG-induced LB agar plates supporting growth of coding residue variants. A = BpsATycC. B = F-CR4-V. C = I-CR7-V. D = V-CR6-L. E = BpsA_SL. After overnight growth of colonies on LB agar supplemented with antibiotics, the 'scoop' method (Section 2.4.2) of IPTG-introduction was used to induce indigoidine production.

Following this a liquid-media *in vivo* assay was conducted on all variants (as outlined in Section 2.11.2). This liquid-culture assay is more accurate than the plate-based assay in which cell density, agar thickness and colony location can all affect indigoidine synthesis. In the liquid-culture assay cells were incubated for 3 hours before the culture turbidity was measured, at an indigoidine-sensitive wavelength (A_{590}) and a cell-sensitive wavelength (A_{800}). Relative indigoidine production (which manifests as a linear relationship between the total indigoidine produced relative to the cell density of a culture; Owen et al., 2011) was quantified according to the formula described in Section 2.12.

Initial runs of the *in vivo* liquid-culture assay were unsatisfactory. Substantial inter-repeat variation made it impossible to judge the relative activities of the variants, or even determine if they differed at all. Additionally, visual inspection of the wells during the assay was inconsistent with the final calculated indigoidine synthesis values. We ascribed these issues to the rapid rate of indigoidine production seen in these variants. To maximise our ability to distinguish between variants, we had timed the reading of absorbance values to coincide with the time that BpsA_SL wells began to approach indigoidine saturation; around 3 hours post-initiation. However, due to the rapid activity of the BpsA variants, they would occasionally reach saturation before or around this point. When indigoidine reaches saturation in the well, it precipitates onto the base of the well, where physical occlusion can lead to inaccurate absorbance readings being reported. With the assay in this form we suspected well-performing variants had their activities under-reported due to high A_{800} values causing an over-correction for cellular component absorption (See section 2.12 for a description of the protocol for indigoidine normalisation).

The high activity of some variants provided another possible reason for the imprecise assay results. As the assays described above were seeded from overnight cultures with no glucose in the media, small amounts of leaky expression from the *lac* operator controlling the BpsA variants was probably occurring. This meant some quantity of indigoidine was likely being produced in the overnight culture, proportionate to the activity of the enzyme responsible. As indigoidine is slightly toxic to *E. coli*, its production exerts a negative selection pressure (Owen et al., 2011). We suspect bacteria producing more indigoidine (as a consequence of expressing a more active BpsA variant) would be more likely to recombine out or truncate their variant *bpsA* gene to eliminate the negative selection pressure. When these bacteria were next-day introduced to the liquid-culture assay, they would produce less indigoidine and therefore under-report the activity of their harboured BpsA variant.

The addition of 0.4% glucose to overnight cultures, for suppression of the *lac* operator, solved both of the problems described above. By suppressing overnight expression of BpsA variants, no negative selection pressure was applied before assay initiation. The assay endpoint was also delayed, improving the extent to which variants could be differentiated. We found this correction was only necessary when assayed variants synthesised indigoidine at rates comparable to wild type – assays described later in this thesis were conducted without glucose in overnight cultures unless stated otherwise.

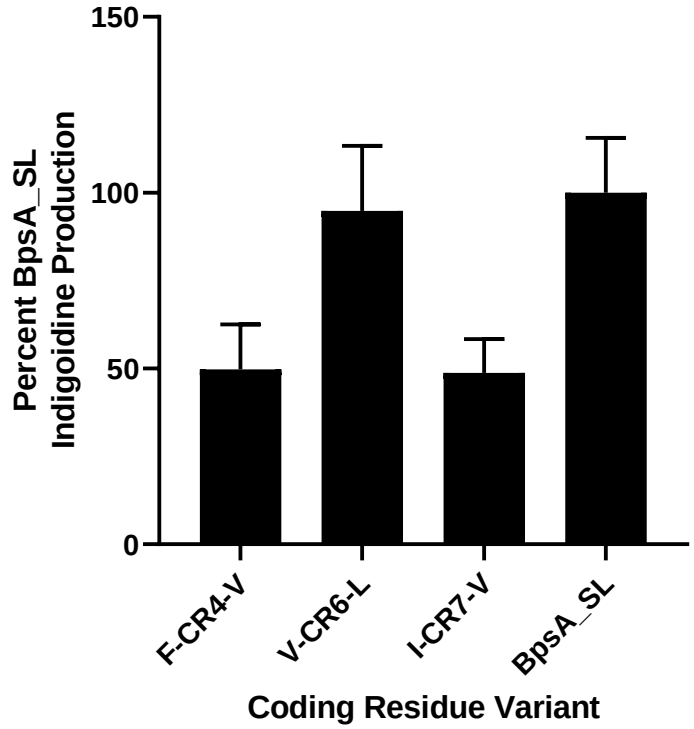


Figure 3.5 Graph of indigoidine production by E. coli cells expressing bpsA residue-reprogrammed variants as a percentage of wild-type. A_{590} and A_{800} were measured 9 hours after assay initiation and used to calculate the percentage indigoidine production relative to cells expressing wild type bpsA ("BpsA_SL"). Data shown in this graph are from assays in which glucose was added to the overnight cultures used to seed the assay. Data are from 3 biological repeats, each conducted in triplicate. Error bars represent 1 S.D.

Figure 3.5 shows the outcome of the assay after glucose-overnight adjustments were made. The most fundamental result from the *in vivo* analysis was that all three residue substitutions generated functional enzymes capable of producing indigoidine. The pattern of relative activity between variants is in line with expectations informed by bioinformatic analysis. At coding residue 4, where BpsA homologues appear to be selective for hydrophobic residues, introduction of a differently-sized but similarly hydrophobic residue resulted in a ~50% reduction in activity relative to BpsA_SL. At CR6, introduction of the overwhelmingly more prevalent coding residue among the BpsA homologues (V-CR6-L) gave rise to the best enzyme

from among the variants, not significantly different from BpsA_SL ($p = 0.45$, Welch's unequal variances *t*-test). Finally, the I-CR7-V variant was tolerated in the BpsA_SL A-domain (with a ~50% reduction in activity), not unexpected given the similarity of Ile and Val in both size and hydrophobicity.

3.4.4.2 *In Vitro Assessments of Enzyme Activity*

An *in vitro* assay of protein kinetics was conducted to further assess the relative activities of the substitution variants (Fig. 3.6). Proteins were purified in accordance with the method described in Section 2.8. Using the method established by Owen et al. (2011), maximum enzyme velocities were established. Active *holo*-form coding residue variants in a 96 well plate were supplemented with L-Gln substrate, initiating indigoidine production. Indigoidine was then measured at 590 nm every 10 seconds. Determining the rate of change of indigoidine concentration using the slope function in Microsoft Office Excel gave a measure of maximum enzyme velocity in each well.

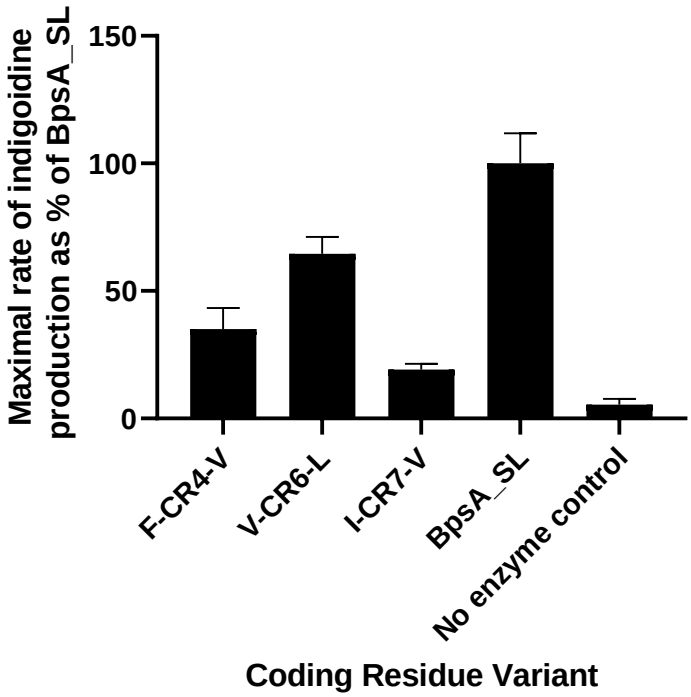


Figure 3.6 Graph of maximal rate of indigoidine production by reprogrammed BpsA variants in vitro. Maximal rate of indigoidine production (ΔA_{590}/second) was calculated and expressed as a percentage of BpsA_SL maximal rate. Data are derived from one biological repeat conducted in triplicate. Error bars represent 1 S.D.

3.5 Other L-Gln Specificity Codes

Having investigated the functional sequence space of known BpsA homologues, we then expanded our search to L-Gln-specifying A-domains unrelated to indigoidine synthesis.

We generated a comprehensive database of A-domains with experimentally validated substrate specificities. Data were derived from two major sources, both of which were databases compiled for the testing of A-domain specificity prediction software. The most significant source was the database that powered the online tool NRPS-PKS-substrate-predictor (Khayatt et al., 2013). A later database used for the validation of the SANDPUMA prediction software was also co-opted (Chevrette et al., 2017).

Table 3.4 L-glutamine specificity codes identified in our database of experimentally validated A-domains. Coding residues identical to BpsA_SL are shown in blue. Coding residues that differ from BpsA_SL are shown in orange.

Coding Residue	CR1	CR2	CR3	CR4	CR5	CR6	CR7	CR8
BpsA_SL	A	W	Q	F	G	V	I	N
BpsA Homologues	A	W	Q	F	G	L	I	N
	A	W	Q	F	G	I	I	N
	A	W	Q	A	G	V	I	N
	A	W	Q	F	G	L	V	N
	A	W	Q	L	G	L	I	N
	A	W	Q	M	G	L	I	N
	A	W	Q	C	G	L	I	N
	A	W	Q	A	G	L	I	N
	A	W	Q	Y	G	L	I	N
	A	W	Q	V	G	L	I	N
Other L-Gln A-domains	A	W	Q	F	G	L	I	D
	V	W	H	F	G	R	I	N
	A	W	H	F	G	G	V	D
	A	Q	D	L	G	V	V	D

Extracting the unique L-Gln specificity codes from both databases resulted in 15 unique L-Gln specificity codes (Table 3.4). For testing purposes, we prioritized the non-BpsA-homologue specificity codes from A-domains that were described, and the structure of their binding pockets predicted, in the study of Challis et al. (2000). These chosen specificity codes are shown in Table 3.5, aligned against BpsA_SL. Note that some of the deviations from BpsA_SL found in these two codes can be found elsewhere among the BpsA homologues.

Table 3.5 Alignment of Coding Residues from BpsA, LicA, and TycC, showing differences from BpsA_SL and the BpsA homologue curation. Plus symbols denote residues that differ from BpsA, but are found within other BpsA homologues. Asterisks indicate residues not found in BpsA_SL or the BpsA homologues.

Coding Res.	CR1	CR2	CR3	CR4	CR5	CR6	CR7	CR8
BpsA_SL	Ala	Trp	Gln	Phe	Gly	Val	Ile	Asn
TycC (M2)	Ala	Trp	Gln	Phe	Gly	Leu+	Ile	Asp*
LicA (M1)	Ala	Gln*	Asp*	Leu+	Gly	Val	Val+	Asp*

The more-similar specificity code was identified in TycC, an NRPS in the tyrocidine biosynthetic pathway, containing an L-Gln-activating A-domain in its second module. TycC differs from BpsA at only two coding residues (Table 3.5). The TycC substitution (relative to our BpsA coding residues) N-CR8-D is not represented anywhere among the BpsA homologues. This represents the only major departure from BpsA, as the V-CR6-L present in TycC is more common among the BpsA homologues than the valine at CR6 of BpsA_SL and did not affect the activity of BpsA in earlier experiments (Fig. 3.5).

The second specificity code, from the first-module A-domain of LicA, an NRPS in the lichenysin biosynthetic pathway, represents a significant departure from the coding residues of BpsA, and from other L-Gln specificity codes more generally. At CR2, CR3 and CR8, LicA displays entirely different amino acids to those found among the BpsA homologues. At CR4 and CR7, LicA differs from BpsA_SL again, though here the alternative amino acids can also be found in other BpsA homologues. The substantial level of difference (5 of 8 coding residues changed) suggests LicA has evolved a different solution to binding L-Gln within its substrate binding pocket. Looking specifically at certain coding residue positions is also informative. At CR2, a position totally conserved among the 236 BpsA homologues, instead of the very large and hydrophobic tryptophan, LicA instead contains a smaller and polar-uncharged glutamine. CR2 is also present at the bottom of the binding pocket, where it likely interacts with the distal end of the substrate sidechain (Fig. 3.1). We reasoned that introduction of the LicA residues into BpsA would provide an avenue to probe the limits of the apparent residue-flexibility of L-Gln specifying A-domains. Marked differences at CR3 also prompted investigation. The polar uncharged Gln of BpsA_SL contrasts with the smaller, negatively charged Asp of LicA. At the three coding residues where variation is seen among the BpsA homologues (CR4, 6 & 7), LicA contains a non-majority option in each case (note that while LicA and BpsA match at CR6, this is the position where BpsA_SL is an outlier relative to other BpsA homologues).

3.5.1 Construction of TycC and LicA Coding Residue Substitutions

The TycC residue substitution construct was generated in the same way as the single-residue substitutions described in Section 3.4.3, by overlap PCR. The two coding residues to be changed were changed sequentially. This generated the construct TycC_Res. Despite repeat efforts, the five LicA-informed residue changes were not able to be successfully incorporated by overlap PCR. This was likely due in part to the more drastic codon changes necessary, and the proximity of coding residues that needed to be changed simultaneously (e.g. CR7 and CR8 are neighbouring codons). Rather than attempting longer primers, it was decided that simply ordering the LicA residues in *bpsA* as a gene fragment (construct LicA_Res) would be a better time/value proposition. Following restriction cloning into pCDFDuet-OxTe, sequence validation, and transformation of BL21 *pcpS*, variants were subject to *in vivo* testing.

3.5.2 TycC and LicA Function

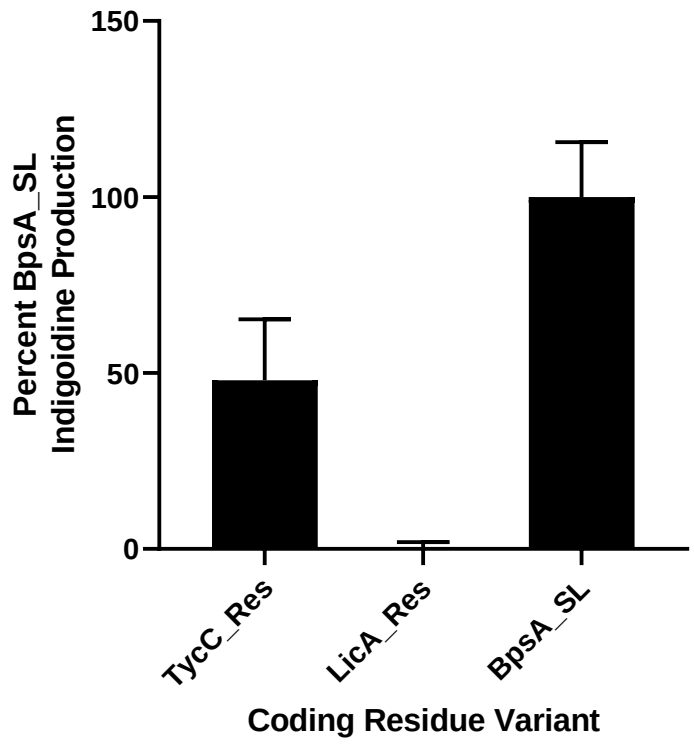


Figure 3.7 Graph of indigoidine production of TycC_Res and LicA_Res as a percentage of wild-type BpsA (BpsA_SL). A_{590} and A_{800} were measured 9 hours after assay initiation. Data shown in this graph are from assays in which glucose was added to the overnight cultures used to seed the assay. Data are from 3 biological repeats, each conducted in triplicate. Error bars represent 1 S.D.

In vivo liquid-culture assays (Fig. 3.7) showed that TycC_Res was functional, with around 50% activity in comparison to BpsA_SL (the reduced activity is likely attributable to the D-CR8-N coding residue change). In contrast, LicA_Res was completely non-functional. This finding

was reinforced by solid-media assays which found that LicA_Res colonies remained white (not pictured), whereas TycC_Res was indistinguishable on solid media from BpsA homologue single-residue substitutions (Fig. 3.4).

3.5.3 Visualisation and Analysis of L-Glutamine Stachelhaus Codes

Dr Alistair Brown generated a series of network diagrams using the BLOSUM matrix to score the similarity of any respective pair of specificity codes. Network diagrams showed that specificity codes tended to cluster according to the monomer they activated. In particular, we were interested in the glutamine and glutamate specificity codes within this network. A sub-network was built containing all the glutamate and glutamine coding residues from the dataset compiled in Section 3.5. The sub-network was then displayed using the network tool Cytoscape (cytoscape.org) using an edge-weighted layout where edges distances are based on the BLOSUM score (Fig. 3.8).

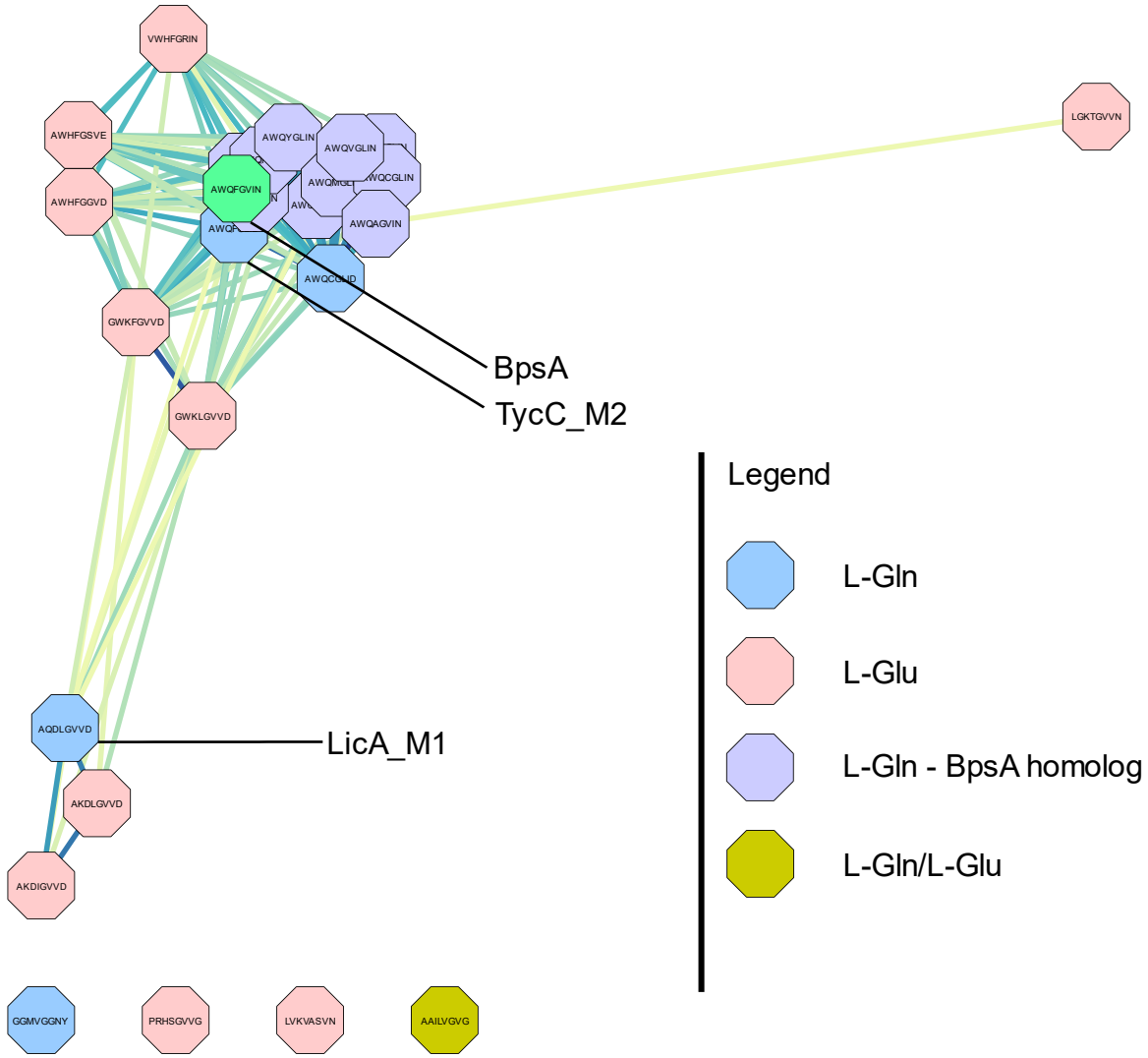


Figure 3.8 Edge-weighted network of glutamine and glutamine specificity codes. Key nodes are labelled. BpsA is highlighted in light green. Yellow edges represent low levels of similarity while blue edges represent high levels of similarity.

This targeted network shows that most glutamine specificity codes cluster together, including TycC. However, LicA appears to cluster with a group of glutamate specificity codes. This analysis suggests that the LicA module-1 A-domain may have evolved relatively recently from a glutamate-coding A-domain. Alternatively, it could suggest that LicA does not in fact recognise glutamine, but instead glutamate, and has been misclassified. While previous studies have reported that LicA module 1 does incorporate L-Gln in its native context, we decided to replicate these experiments to ensure the LicA specificity code had not been improperly classified in prior studies.

3.6 Experimental Confirmation of LicA Specificity for L-Glutamine

The malachite green assay is a way to measure adenylation activity independent of downstream reactions. When an A-domain adenylates a substrate, pyrophosphate (PPi) is released from ATP. In the malachite green assay, pyrophosphatase in the reaction mixture metabolises PPi into free phosphate, which forms a green complex with the detection reagents (McQuade et al., 2009).

The LicA A-domain, the source of the coding residues used in reprogramming studies, was ordered in full as a gene fragment from Twist Bioscience. The borders of the A-domain were those used by the seminal LicA characterisation study (Konz et al., 1999), with the fragment consisting of bases 1304-2931 of the LicA gene (U95370.1). A vector backbone, pET-28a(+), was linearized using the restriction enzymes NdeI and SalI, and the A-domain was cloned into the vector using NEBuilder. After colony PCR and validation by sequencing, plasmid isolated from validated clones was used to transform BL21 *ΔentD* for protein expression. Protein was expressed and purified by the method outlined in Section 2.8.

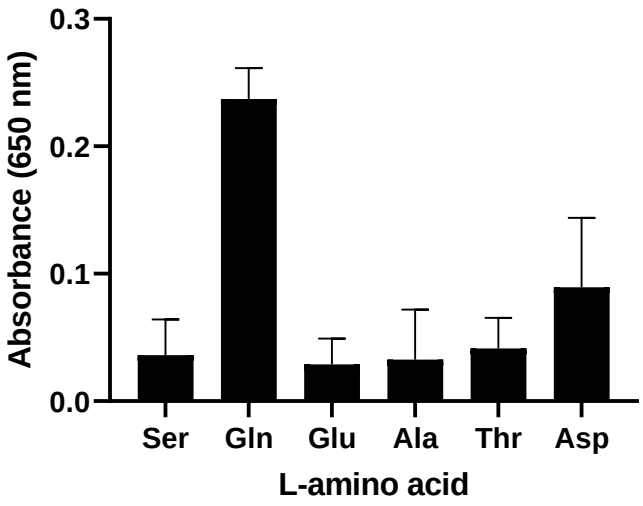


Figure 3.9 Results of the malachite green assay for the LicA A-domain. The A_{650} wavelength is sensitive to the green complex formed in the presence of adenylation activity. Values are shown after subtracting the A_{650} of well with no amino acid. Data are derived from 3 repeats. Error bars represent 1 S.D.

The malachite green assay (Fig. 3.9) confirmed that LicA is a glutamine-activating A-domain. Absorbance values depicted in Figure 3.9 have had the value of a negative control well (with protein, but no amino acid substrate) subtracted from them. The small positive values for other amino acids are attributable to the spontaneous breakdown of ATP in the highly acidic environment of the well after detection reagents are added and are not suggestive of low-level

substrate promiscuity in LicA. This includes L-aspartic acid in the final column of Figure 3.9 – as Asp was consistently measured last in the plate reader, there was more time for ATP to break down spontaneously. LicA was shown to be active with, and highly specific for, L-glutamine, and showed no activity with L-glutamate.

We also considered that the malachite green assay might prove useful as a further means of assessing BpsA variant activity (in addition to the solid agar and liquid media indigoidine synthesis assays). For example, this assay might provide insight into variants where the residue substitutions yielded active L-Gln adenylating substrate pockets that for some reason (e.g., steric constraints introduced by the new residues) were unable to subsequently interact effectively with other BpsA domains. However, difficulties were encountered in applying the malachite green assay to BpsA variants, as detailed here for BpsA_SL.

Purified BpsA protein was stored in a Tris-based buffer, as the use of a phosphate buffer would result in positive signals from the phosphate reacting with the malachite green dye, regardless of which amino acid was present with the protein. Unfortunately, assays gave negative results irrespective of the amino acid substrate tested. As an assay control, to ensure assay reagents were functional, comparison against EntF, an *E. coli* NRPS in the enterobactin biosynthesis pathway (Miller et al., 2016), showed the positive and expected signal for L-Ser, while the expected signal for L-Gln for BpsA was not observed (Fig. 3.10).

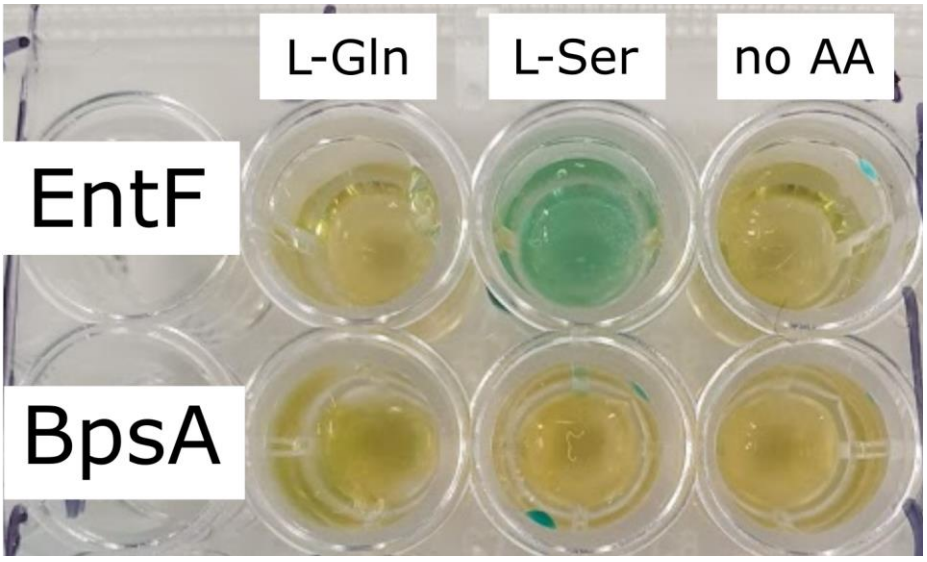


Figure 3.10 Picture of malachite green assay wells after addition of detection reagents. Wells where adenylation has occurred are green due to a complex of green dye and inorganic phosphate. The assay shows a strong signal for serine in EntF, validating the assay, but no expected signal for glutamine in BpsA.

Attempts at optimisation included adjustments to the amount of protein in the assay, the length of incubation in the presence of the amino acid, and a change in the brand of inorganic

pyrophosphatase used in the reaction. To ensure the BpsA protein was functional and intact, a sample was phosphopantetheinylated, and this *holo* BpsA was investigated for indigoidine production. The formation of indigoidine confirmed that adenylation was indeed occurring, but could not be detected in the malachite green assay (Fig. 3.11).

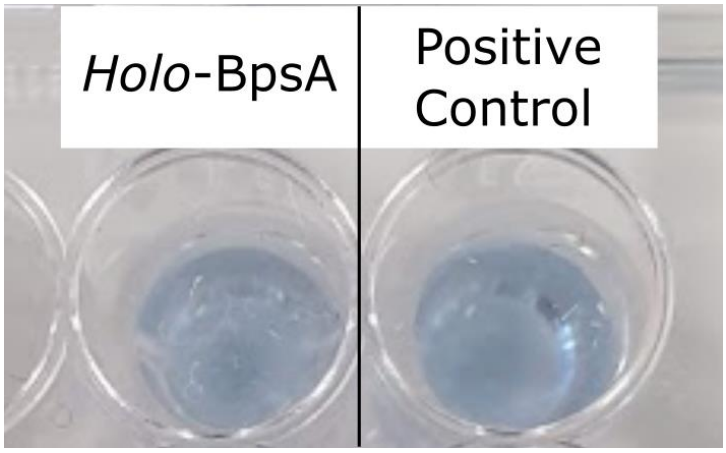


Figure 3.11 Picture of purified BpsA that was subsequently phosphopantetheinylated and incubated with L-Gln and ATP, compared to a positive control of holoform BpsA purified by a colleague. Both proteins are generating indigoidine, confirming the functionality of my purified BpsA.

Although the reason for BpsA not behaving as expected in malachite green assays is as yet unknown, we posit that the interrupting Ox-domain in BpsA may significantly hamper the rate of adenylation activity relative to the uninterrupted A-domains of LicA and EntF, and consequently BpsA gave no signal under conditions which permitted signal detection among the other A-domains.

3.7 Experiments to Recover Activity in LicA_Res

Having verified that the LicA specificity code does indeed recognise L-Gln in its native context, and that those residues cannot be directly substituted into the binding pocket of BpsA without destroying activity, we sought to establish if any specific residue substitutions were responsible for the disruption of the binding pocket.

3.7.1 Single Coding Residue Reversion

To assess if any individual residue substitution, introduced to BpsA from among the five differences present in LicA, was responsible for the inactivity of LicA_Res, five constructs were initially designed. Each consisted of LicA_Res in which a single LicA coding residue had been reverted to the wild-type residue. For instance, LicR_wtQ-CR2-W is LicA_Res with the LicA residue glutamine reverted to wild-type BpsA tryptophan.

Reversion constructs were generated using the NEBuilder strategy detailed in section 2.5.3. After transformation of DH5α, the constructs were validated by sequencing and used to transform BL21 *pcpS*.

The resulting cells were plated on solid media supplemented with IPTG. After 96 hours of incubation at 18 °C, the colonies remained white, indicating that no indigoidine was being produced by the residue-reversion constructs. This indicated that no single residue in the LicA_Res construct was responsible for the impaired indigoidine synthesis activity.

We next sought to determine whether individual substitution of LicA coding residues into BpsA might enable multiple activity-impairing substitutions to be identified. To this end we developed 5 constructs wherein BpsA was reprogrammed with a single LicA coding residue (i.e., each of the varying residues summarised in Table 3.5).

It was anticipated that reprogramming CR2 (W→Q) or CR3 (Q→D) would be most likely to have an adverse effect of BpsA function. These residues are not found among the BpsA homologues, had not been tested individually prior to this experiment, and represent the most significant changes in terms of amino acid size and charge. In contrast, the deviations in specificity code found at CR4 and CR7 are also found among the BpsA homologues, so we reasoned that these would be least likely to critically hinder activity. Indeed, the change at CR7 had already been made, and found to be functional, in Section 3.5. The change at CR8, while not seen among the BpsA homologues, was present in TycC_Res, which did function, so CR8 was also not expected to significantly impact activity.

3.7.2 Introduction of Individual LicA Residues into BpsA

All constructs were ordered as gene fragments from Twist Bioscience. These fragments were amplified with primers introducing BamHI and HindIII sites at the 5’ and 3’ end respectively, introduced into pCDFDuet-OxTe by restriction cloning, and used to transform the cloning strain DH5α. Plasmids were isolated, validated by sequencing and then used to transform BL21 *pcpS*.

3.7.2.1 *In Vivo Assay of Single LicA Residue BpsA Constructs*

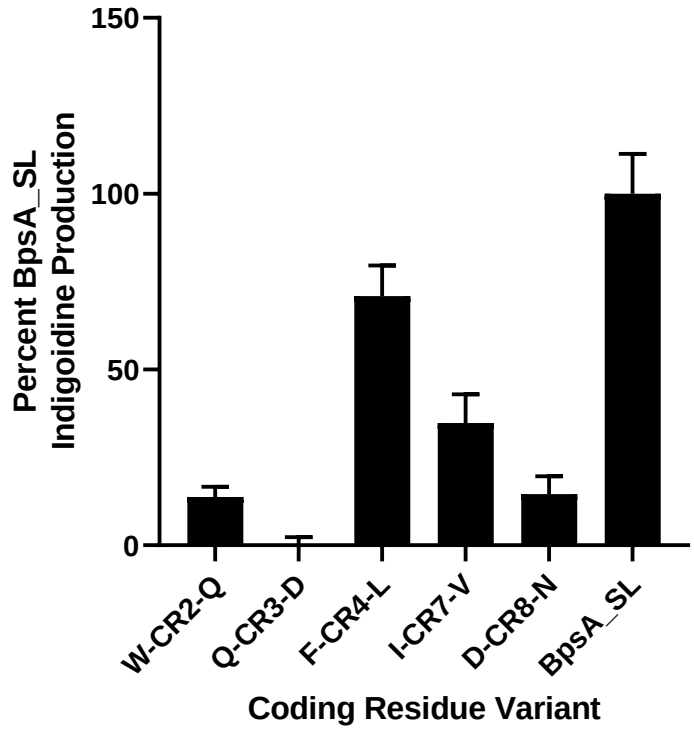


Figure 3.12 Graph of indigoidine production across single LicA-coding residue variants as a percentage of BpsA_SL. Variants were incubated for 6 hours before measurement. Data shown in this graph are from assays in which glucose was added to the overnight cultures used to seed the assay. Data are from 3 biological repeats, each conducted in triplicate. Error bars represent 1 S.D.

In vivo assays revealed that the inclusion of Q-CR3-D in BpsA was the only substitution that, in isolation, renders the enzyme non-functional. However, LicA_Res with this critical residue restored to wild-type remained entirely non-functional (Section 3.7.1), suggesting the existence of additional negative epistatic interactions between the coding residues and the surrounding structure (and/or each other).

Discrepancies in reported activity between Figure 3.12 and Figure 3.5 are due to differences in assay runtimes, as each assay was optimised to the specific variants being assayed. The I-CR7-V variant had a lower apparent activity in Figure 3.12 relative to Figure 3.5 because measurements shown in Figure 3.12 were taken at 6 hours, whereas measurements in Figure 3.5 were taken at 9 hours. When replicates shown in Figure 3.12 were measured again at 9 hours, the mean indigoidine production of I-CR7-V was 53.7% that of the BpsA_SL control, consistent with the 48.7% result in Figure 3.5. Together these results indicate that the I-CR7-V residue substitution does not critically impair BpsA.

F-CR4-L also did not critically impair BpsA, with ~70% indigoidine production compared to BpsA_SL; more active than the similar F-CR4-V residue substitution conducted in Section 3.4. The most pertinent conclusion from this experiment was that F-CR4-L is tolerated in BpsA, but the degree of difference in activity between F-CR4-L and F-CR4-V shows that despite CR4 being a 'wobble' position in L-Gln specificity codes (Table 3.3), the choice of residue at this position can have a substantial impact on the rate of indigoidine synthesis.

W-CR2-Q and N-CR8-D, the other two residue substitutions not found among the BpsA homologues, were significantly impaired relative to wild-type. In the case of W-CR2-Q this was anticipated, as the new coding residue is not seen in any other L-Gln specificity code identified in our database. The low activity of N-CR8-D was unexpected. This same residue substitution in TycC_Res (which also included a V-CR6-L substitution) produced 47.9% as much indigoidine as the BpsA_SL control after 9 hours of incubation. N-CR8-D alone, under the same conditions, was significantly slower with 25.4% BpsA_SL activity ($p < .05$, Welch's unequal variances *t*-test).

3.7.3 Second Degree Coding Residue Reversion – Restoration of Two Wild-type Coding Residues in the LicA Residue Construct

The above experiment demonstrated that BpsA cannot tolerate a glutamine residue in place of its native tryptophan at CR3. However, in the LicA_Res construct, simply restoring the wild-type residue at CR3 was not sufficient to restore enzyme activity.

We hypothesised that an additional combination of poorly tolerated LicA residues (beyond just CR3) was responsible for the inactivity of LicA_Res. Having confirmed that CR3 is critical, we then explored every second-degree reversion from the LicR_wtQ-CR3-W construct. That is, each possible combination of LicA_Res with both CR3 and one additional coding residue restored to wild-type.

Constructs were generated either using gene fragments from Twist Bioscience or as described in Section 2.5.

Constructs were generated by overlap PCR of pCDFDuet-licR-wt_CR3_OxTe using mismatched primers to individually restore each additional wild-type residue. Fragments were assembled into the expression vector using NEBuilder. These plasmids were used to transform DH5α, colony PCR was performed, and colony hits were picked and grown overnight. Plasmids were purified from these colonies and sent for sequencing. Construct L-w(3+2), i.e.

LicA_Res with coding residues 2 and 3 reverted to wild-type, failed sequence validation, so was instead ordered as a gene fragment, and inserted into an expression vector via restriction cloning.

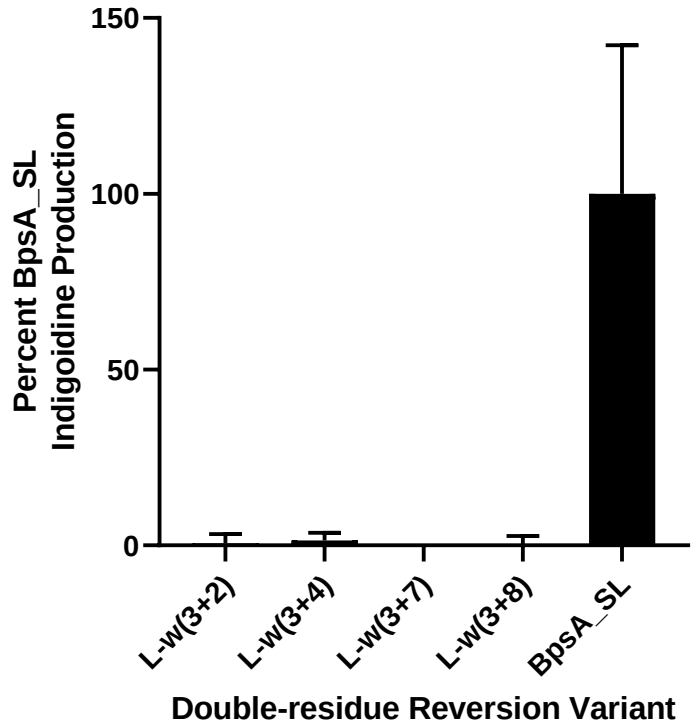


Figure 3.13 Graph of indigoidine production of second-degree coding residue reversion constructs expressed as a percentage of BpsA_SL activity. Measurements were made after 9 hours of incubation. Data are from three biological repeats, each conducted in triplicate. Error bars represent 1 S.D.

Liquid *in vivo* assays indicated that no double-reversion of the LicA residues to wild type was sufficient to restore indigoidine production in the LicA_Res construct (Figure 3.13). To ensure that not even trace amounts of indigoidine were being generated, variants were plated on solid media with IPTG, and incubated at 18 °C for a week. Past experience has indicated that in this amount of time, any amount of indigoidine production would cause an *E. coli* colony to turn blue (Dr Alistair Brown, *personal communication*).

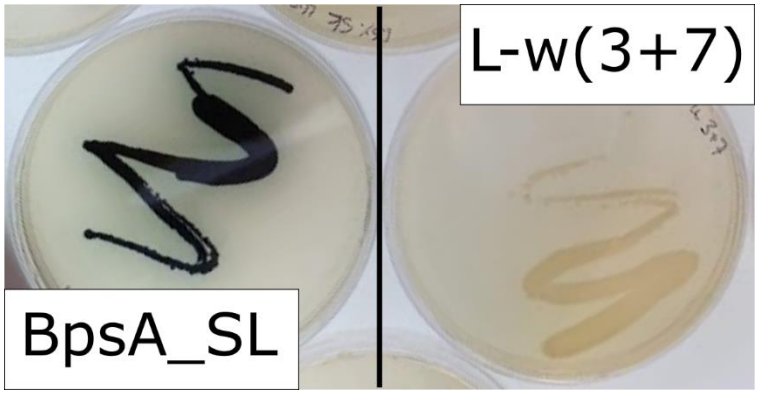


Figure 3.14 Picture of BpsA_SL and L-w(3+7) constructs expressed by BL21 pcpS plated on LB agar supplemented for indigoidine production, after one week of incubation at 18 °C. Figure is representative of all second-degree residue reversion constructs.

Solid media assays (Fig. 3.14) were consistent with the findings of the liquid culture assays, i.e. no variant was functional.

Taken together, these experiments demonstrated that BpsA_SL is critically impaired by substitution of CR3 from glutamine to aspartate, such that the enzyme does not produce indigoidine. The introduction of any other LicA residue to the binding pocket of BpsA_SL did not critically impair activity – these variants were still able to make indigoidine, though at a slower rate. When introduced all together however, these 4 subcritical residue substitutions resulted in a non-functional enzyme. Furthermore, introduction of any set of three of the four subcritical residues consistently eliminated enzyme activity.

3.8 Discussion of Chapter Results

Reprogramming experiments, replacing the coding residues of BpsA_SL with alternative L-Gln codes, showed that alternative specificity codes from BpsA homologues were broadly tolerated. The specificity code of TycC, an L-Gln A-domain with middling similarity to BpsA, was also tolerated. However, a more distantly related specificity code from LicA did not function when used to reprogram BpsA_SL. This indicated that reprogramming experiments can fail not just when attempting to engineer changed substrate specificity, but also when a synonymous code is substituted and no change in function is expected of the engineered enzyme.

In addition to reinforcing our suspicions that coding residues interact with their local environment, we found preliminary evidence of synergistic interactions between the coding residues themselves. The TycC_Res construct generated in our study had two changed coding residues relative to BpsA_SL: V-CR6-L and D-CR8-N (Table 3.5). While TycC_Res produced

indigoidine at a slower rate than the V-CR6-L substitution alone, it was much more active than the D-CR8-N substitution, suggesting that the presence of a leucine at the CR6 position was able to compensate for the otherwise deleterious introduction of an asparagine at CR8. This is not unprecedented; in a substantial EntF reprogramming study recently published by Throckmorton et al (2019), a single-residue substitution performing worse than a multi-residue substitution was found to be atypical but not uncommon. This suggests that, independent of interactions between the coding residues and the surrounding environment, interactions between the coding residues themselves in an otherwise unchanged binding pocket must be considered when reprogramming NRPSs. Similar to our study, Throckmorton et al. (2019) generated a large collection of experimentally confirmed L-Ser specificity codes in EntF (which naturally accepts L-Ser). They identified a pattern of covariation at certain coding residue sites, e.g. the identity of a particular amino acid at CR1 was predictive of a particular amino acid at CR2, suggesting that the covaried residues give rise to a better enzyme than when those residues are varied individually. Our equivalent TycC_Res result can be rationalised structurally, as CR6 and CR8 are proximal and both protrude into the binding pocket (Figure 3.1). To use a metaphor of Lego, there may be more than one way to build a particular three-dimensional shape, but not each piece will be interchangeable in the different models.

While inter-residue interactions likely modulated the function of our enzyme variants in this study, efforts to recover LicA_Res activity showed that the interaction between a specificity code and the surrounding protein structure was critical for any degree of enzyme function. Introducing the five alternative coding residues of LicA into BpsA_SL individually showed that only one residue change, Q-CR3-D, completely eliminated indigoidine production in BpsA. This result could be rationalised as a consequence of inflexibility at CR3 in the BpsA binding pocket. Throckmorton et al (2019) also observed inflexibility at the CR3 position of EntF, finding that while there was substantial diversity at the CR2 position, the CR3 residue was almost universally conserved across codes. If this conservation at CR3 is a consequence of restrictions that also apply in the BpsA binding pocket (as seems likely to be the case, due to the complete conservation of a Gln at the CR3 position in all known BpsA homologues), this rationalises our finding that CR3 in LicA_Res is inhibiting enzyme function.

CR3 is not the sole problem however, as a BpsA_SL variant containing all LicA coding residues except Q-CR3-D remained completely non-functional. This demonstrated epistatic or additive negative effects in at least some of the remaining four altered coding residues, as none was individually sufficient to eliminate function, but collectively there was a complete loss of

function. The effect persisted when any three non-critical changes were introduced. Without time constraints, one priority would be to examine all possible combinations of LicA versus BpsA_SL coding residues to identify precisely which combinations are not tolerated. Interpreting the structural effect of coding residue changes is complicated by the lack of a crystal structure of BpsA (several attempts to obtain one have been made by collaborating groups, but to date there has been no success; David Ackerley, *personal communication*). We suspect that a third-degree reversion of LicA_Res, generating a BpsA_SL construct with only two foreign LicA residues, would generate at least one functional enzyme. For instance, reversion of W-CR2-Q, Q-CR3-D and N-CR8-D simultaneously would eliminate the three worst-performing residues in single-residue substitution experiments (3.7.3). However this would also eliminate all variation unique to LicA – the remaining changes to BpsA_SL (F-CR4-L and I-CR7-V) are both present among the BpsA homologues (Table 3.4) – so this construct would not be informative with regard to permitted codes in BpsA_SL.

Questions pertaining to the LicA coding residues are further complicated by our discovery in Section 3.5.3 that LicA is much more closely related to L-glutamate A-domains than to L-glutamine ones. Eppelmann et al. (2002) successfully altered the substrate specificity of an L-glutamate A-domain to recognise L-glutamine by reprogramming it with the LicA coding residues. In contrast our experiments show the LicA specificity code is non-functional even in an A-domain of synonymous specificity, BpsA_SL. It is likely that Eppelman et al. were unaware of the anomalous nature of LicA amongst glutamine-specifying A-domains, as at that time LicA was regarded as the archetypical L-glutamine specificity code by Stachelhaus et al. (1999). Thus, they may have concluded reprogramming of A-domains to be easier than it generally is, based on their selection of a non-representative model system.

Our results suggest that reprogramming requires not just knowledge of a consensus substrate specificity code, but rather a specific one that is able to synergise with the surrounding protein structure. This leads into the question of whether reprogramming is an efficient means of altering A-domain specificity when compared to, for instance, domain substitution. Fewer et al. (2007) identified recurrent instances of full A-domain replacement in the microcystin synthetase gene cluster across modules and between species. A tendency in natural systems to rely on the substitution of large areas of protein structure rather than mutation or recombination of the coding residues specifically, coupled with increasing experimental evidence of the potential pitfalls of reprogramming strategies, leads to the conclusion that wider-scale

adjustments like whole- or subdomain substitution may be a more promising engineering technique than coding residue reprogramming.

4 Investigation into Subdomain Substitution

4.1 Introduction

The coding residue reprogramming experiments reported in chapter 3 of this thesis added to a growing body of evidence that specificity codes are optimised for their native local environment, and frequently do not function in the context of a non-native A-domain. By introducing the synonymous code from a distantly related NRPS domain, LicA, we also demonstrated that even the transplant of synonymous specificity codes can critically impair enzyme function, with no clear route to, or obvious rules for, engineering recovery.

The alternative strategy of subdomain substitution presents a potential solution to the problem of specificity code incompatibility with non-native subdomains. Transplanting an entire subdomain ensures the coding residues remain in their native context, where they might be more likely to retain their native function.

As in the previous chapter, subdomains substitution experiments in this project were restricted to synonymous subdomains – those with the same substrate specificity as the host A-domains – for three reasons. First, we sought to more precisely define permissive and non-permissive factors that influence the success of subdomain substitutions within a closely related (as well as less-closely related) set of NRPS variants. Second, we were thereby able to make direct comparisons to the equivalent reprogramming experiments in Chapter 3 – to our knowledge the first time the techniques have been directly compared. Finally, production of indigoidine remained vital to our functionality screen, and indigoidine is only produced when BpsA activates L-glutamine.

4.2 Preliminary Subdomain Substitutions using the Kries Boundaries

In Chapter 3, the specificity codes from two L-Gln A-domains, TycC and LicA, were substituted into BpsA. The first undertaking of this chapter was the substitution into BpsA of the full TycC and LicA subdomains (TycC_Sub and LicA_Sub respectively). The subdomain borders chosen for this experiment were initially those determined by Kries et al. (2015). As these were derived from the solved crystal structure of GrsA (PDB:1AMU), the equivalent boundaries could be readily identified within BpsA by alignment of the two amino acid sequences. The alternative borders proposed by the other team to have documented subdomain substitution efforts, Crüsemann et al. (2013), were deemed unsuitable as the downstream

boundary excluded the final two coding residues (CR7 and CR8), thereby precluding the possibility of a direct comparison to our reprogramming experiments.

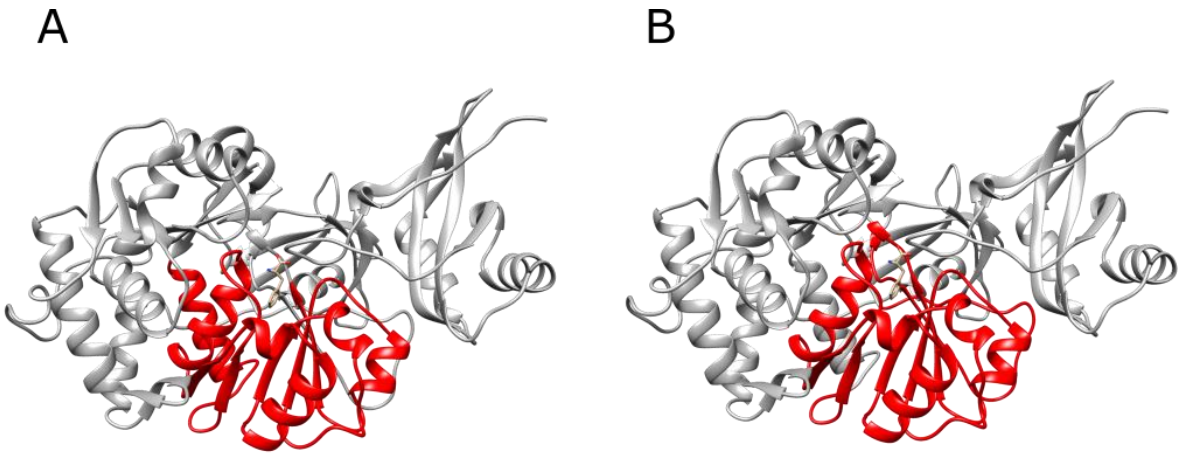


Figure 4.1 GrsA A-domain (PDB:1AMU), with ball-and-stick representation of phenylalanine in the binding pocket. Panel A shows the subdomain borders of Kries et al. (2015) in red. Panel B shows the subdomain borders of Crüsemann et al. (2013) in red. Adapted from Brown et al. (2018) with permission from the Royal Society of Chemistry.

Subdomains (together with the surrounding wild-type BpsA A-domain regions) were synthesised as gene fragments, amplified, and ligated into pCDFDuet-OxTe (described in Section 3.3.1). After transformation of DH5α and subsequent sequence validation, the resulting constructs were used to transform BL21 *pcpS*.

Neither LicA_Sub nor TycC_Sub were capable of synthesising indigoidine. This result was particularly interesting with regard to the TycC_Sub construct, as the TycC coding residues alone were tolerated in the BpsA binding pocket. It was expected that co-substitution with native surrounding activity would improve activity, rather than result in a non-functional enzyme.

One potential reason for this outcome was that the A-domain of BpsA has evolved to have a unique subdomain topography that is compatible with the inserted Ox-domain, whereas the A-subdomain regions from TycC or LicA were incompatible, e.g. due to wider-scale steric disruptions. We therefore tested whether BpsA_SL was more tolerant of subdomain substitutions from other BpsA homologues.

Two BpsA homologue subdomains were substituted into BpsA_SL in the same manner as employed for TycC_Sub and LicA_Sub. These homologue subdomain substitutions were WP_076477847 from *Williamsia sterculiae* (WS) and WP_084276134 from *Erwinia mallotivora* (EM). The WS construct was found to be non-functional, producing no indigoidine. EM showed trace levels of activity.

4.3 Subdomain Border-Sliding

As crystal structure of the BpsA A-domain has not been solved, the subdomain borders used for the substitutions in this section were determined by alignment of BpsA against the GrsA amino acid sequence. Taking into consideration the impact of the interrupting Ox-domain, it is likely that alignments of BpsA against GrsA are not entirely structurally accurate, and for this reason the borders, when transposed onto the BpsA sequence, may fall at inopportune locations. We suspected that steric clashes also may arise around the subdomain borders where incompatible native and non-native amino acids interact deleteriously. To avoid this, it would be necessary to delineate the subdomain borders carefully.

We reasoned that we would be able to improve upon the level of activity seen in the EM subdomain substitution using the Kries et al. (2015) borders (EM_Kries), by identifying more optimal boundaries for subdomain substitution. In the absence of both structural and phylogenetic information to rationally guide selection of the subdomain borders, we chose a more empirical approach of sequentially 'sliding' the borders of the subdomain though plausible regions, relying upon indigoidine synthesis as a rapid and high-throughput activity reporter.

The experiment was designed so that each border, the N-terminal upstream border and C-terminal downstream border (Fig. 4.2), was shifted independently. Having the two borders vary at the same time would make it impossible to assign changes in activity to one change or the other. We therefore first attempted to optimise the downstream border and then, having selected an optimal point, shift to optimising the upstream border. In order to select for activity, we wanted a level of baseline activity upon which to improve, i.e. we wanted a subdomain-substituted BpsA variant capable of producing some amount of indigoidine, which we had found in EM.

The EM subdomain therefore became the focus of subdomain border optimisation experiments. It was intended that optimal borders identified as improving activity of the EM construct could then be transposed onto other L-Gln subdomains that exhibited no activity with the Kries et al. borders, to see if the optimised borders restored activity.

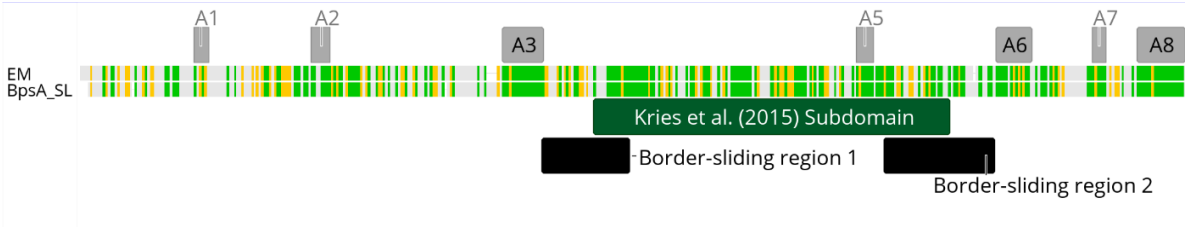


Figure 4.2 Gene-level alignment of BpsA_SL and EM, with intended regions of border-sliding annotated in black. A1 – A8 motifs are annotated in grey, with A4 excluded for clarity. The region corresponding to the subdomain of Kries et al. (2015) is annotated in dark green. Amino acid colour represents similarity across the alignment. Green = Identical. Yellow = Similar, with a BLOSUM score matrix value of +1 or higher. White = Dissimilar (BLOSUM value less than +1).

4.3.1.1 Downstream Border-Sliding

To exclude the possibility that critical steric clashes were occurring at the upstream border while the downstream border was being varied, the first set of test constructs lacked an upstream border altogether (Fig. 4.3). This was achieved by simply beginning the non-native sequence at the beginning of the protein, so that the only non-native/wild-type interface was at the downstream border itself. The key assumption was that the BpsA variant from *E. mallotivora* would be capable of forming a functional chimeric fusion with the downstream regions of our wild type (*S. lavendulae*) BpsA, provided a tolerant recombination boundary could be identified.

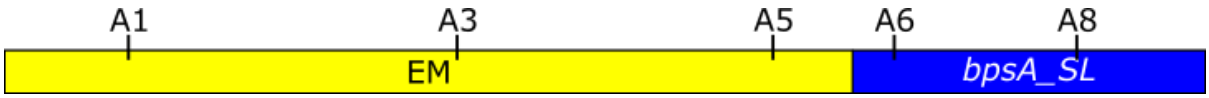


Figure 4.3 Abstract schematic of a downstream border-sliding construct. EM sequence (yellow) runs from the beginning of the gene to the downstream subdomain border. BpsA_SL sequence (blue) runs from the subdomain border for the remainder of the gene. Approximate locations of key A-domain motifs (A1, A3, A5, A6 and A8) are labelled above the construct.

4.4 Construction of Border-Sliding Vector/Starter Construct

4.4.1 Construction

The border-sliding vector took much the same form as the coding residue substitution vector, utilising the same pCDFDuet-OxTe backbone, digested at the same BamHI and HindIII restriction sites. The initial construct was designed with EM sequence from the beginning of the gene through to the area encoding the A8 peptide motif (Fig. 4.2). This region extended well beyond the region we expected to manipulate in border-sliding experiments, to ensure our experiments were not artificially limited by the limits of our ordered gene fragment. The EM sequence was ordered as a gene fragment from Twist Bioscience. The remaining span of gene sequence, from the A8 motif through to the resumption of the vector backbone at the HindIII site (~150 nucleotides) was *bpsA_SL* sequence amplified from pCDFDuet-wtBpsA_OxTe.

These two fragments were assembled into the plasmid backbone using NEBuilder. Through overlap PCR and later NEBbuilder, we retained the ability to reposition the border between the EM and BpsA_SL sequence with full discrimination over the region we intended to examine.

4.4.2 Validation

Sequence confirmation of the border-sliding vector confirmed no errors had been introduced to the A-domain region during construction. In parallel to the border-sliding vector containing EM sequence, a wild-type *bpsA* vector was constructed, with corresponding component fragments, as a positive control. *E. coli* BL21 *pcpS* cells expressing *bpsA_SL* from this vector turned blue, confirming the vectors were expressing *bpsA* variants in frame. Additionally, BL21 *pcpS* expressing the border-sliding vector turned blue. This meant the BpsA variant comprised of EM sequence from the N-terminus to the A8 motif was able to produce indigoidine. This was not unexpected; the A8 motif is a highly conserved region, so no steric clashes were anticipated between adjacent EM and wild-type regions. Nevertheless, this was a reassuring outcome as it confirmed there were at least some points where creation of chimeric BpsA constructs was tolerated.

4.5 EM Downstream Border Sliding

4.5.1 Design

We decided to vary the downstream subdomain border across a 60-amino acid region that was bounded upstream by CR8, and downstream by the A6 motif. This region is annotated as "Border-sliding region 2" in Figure 4.1 and examined at a sequence level in Figure 4.4. CR8 is necessarily as far upstream as the downstream border can reside, to retain the ability to directly compare subdomain substitution and specificity code reprogramming. The A6 motif, as a region of high homology and structural importance, was deemed an appropriate downstream boundary.

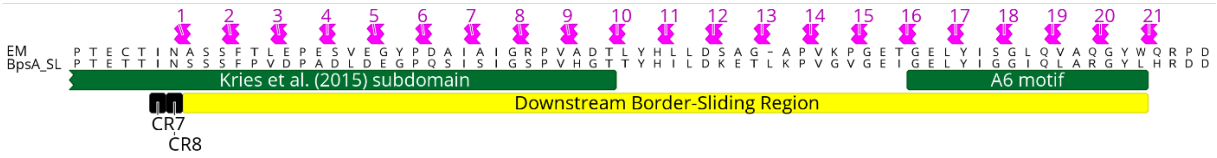


Figure 4.4 Sequence-level alignment of EM and BpsA_SL showing the downstream border-sliding region. The final two coding residues (CR7 and CR8) are denoted by black boxes. The Kries et al. (2015) subdomain and A6 motif are shown by labelled green annotations. Magenta annotations above the alignment show the 21 downstream subdomain border points.

Initially, 21 constructs were designed, having border points at 3-amino acid intervals across the border sliding region (Fig. 4.4). Constructs were named EMD_1 through to EMD_21.

EMD_1 switched from EM-sequence to BpsA-sequence immediately after CR8. EMD_10 was the border exactly equivalent to the Kries et al. (2015) downstream border.

4.5.2 Overlap PCR construction of EMD constructs

At the outset, construction for all border-sliding variants was attempted by overlap PCR. In this process, primers were designed and used to amplify the upstream EM sequence, and downstream *bpsA* sequence. Overlap PCR (Section 2.5.2) of the two fragments generated a full-length A-domain, which could be restriction cloned into the expression vector. We attempted to generate all 21 EMD constructs in this fashion. This proved more difficult than anticipated, with inefficiencies at various points in the protocol complicating construction. Ultimately, five constructs out of 21 were successfully generated using the overlap PCR method, namely EMD_1, 8, 11, 14, and 18.

4.5.3 Initial EMD Construct Results and Decisions on Next Steps

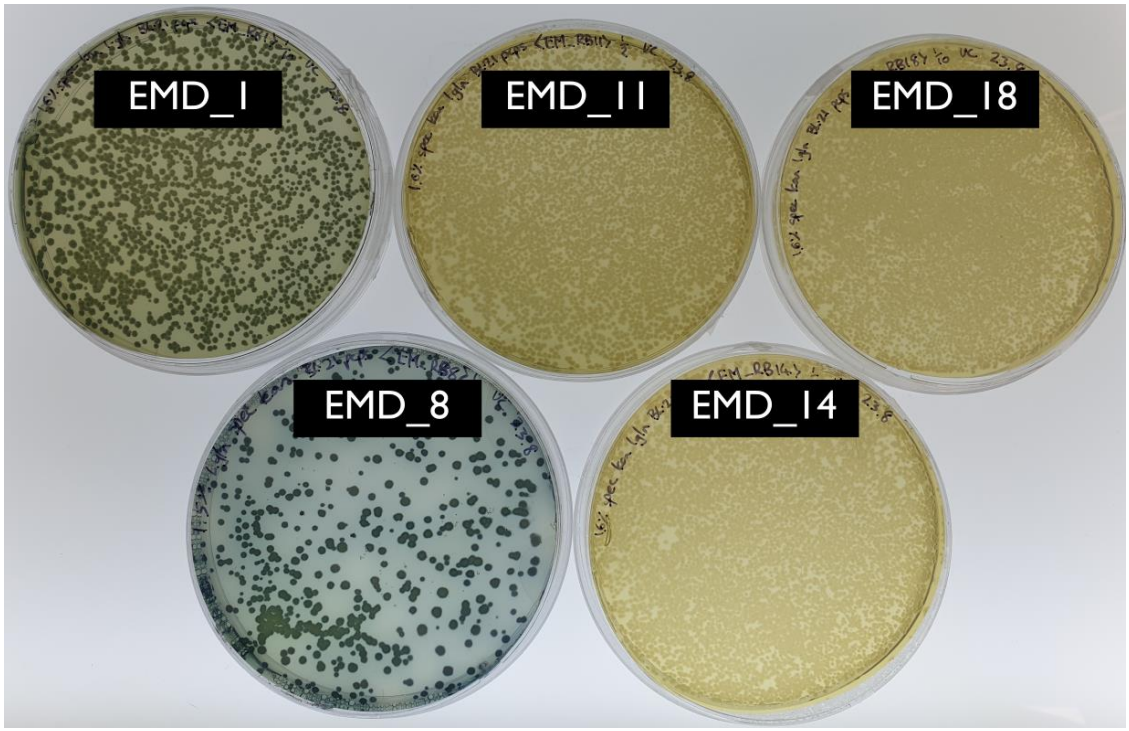


Figure 4.5 Picture of BL21 pcpS expressing the five preliminary EMD constructs on agar plates, 18 hours after IPTG induction using the 'lift and scoop' method.

The five preliminary EMD constructs were used to transform BL21 *pcpS*, which were then plated on LB agar supplemented for indigoidine production and screened for function using the method described in Section 2.11.1. Results of preliminary EMD construct expression indicated that there was substantial differential activity between different subdomain borders (Fig. 4.5). This provided tangible evidence that the precise location of the subdomain border

can have marked impacts on the activity of constructs, and therefore the success of subdomain-substitution experiments.

Having observed the results with these first five variants and experienced complications in attempting high-throughput overlap PCR, we decided to focus our experimental region more narrowly. Seeing a pattern whereby activity seemed strongest in the upstream area of the sliding zone, we decided to build all of the variants from EMD_1 through to EMD_15, abandoning attempts to generate EMD_16 and above (apart from the already-generated EMD_18). Further justification for this decision was that the omitted borders (16,17, 19, 20 and 21) fell entirely within the conserved A6 motif where there was likely to be little activity difference between neighbouring borders (Fig. 4.4). EMD_18 therefore acted as a representative border, falling right in the middle of the A6 motif.

4.5.4 Completing the EMD Panel

In an effort to improve construct generation throughput, it was decided to assemble fragments using NEBuilder. The comparatively higher cost of this method, particularly when utilised at scale, was a reason NEBuilder was not attempted initially. Serendipitously, the primers designed and synthesised for overlap PCR generated a sufficiently large region of homology between fragments as to permit their use in NEBuilder. Assembly by NEBuilder proved far more efficient, and this method was used exclusively from this point onward.

4.5.5 EM Downstream In Vivo Assays

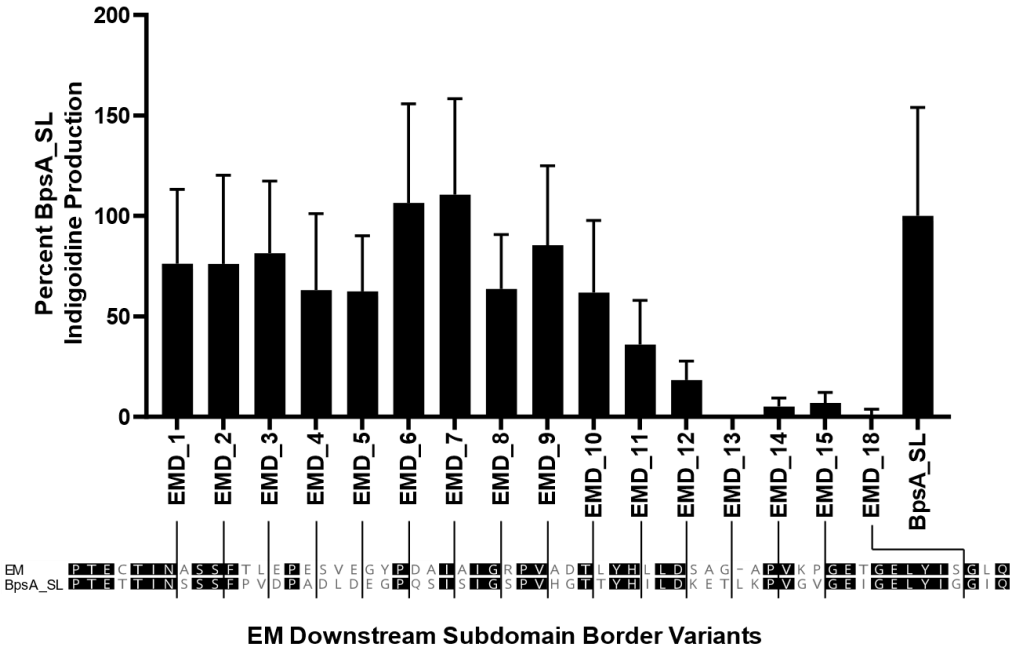


Figure 4.6 Indigoidine production by EMD variants expressed as a percentage of BpsA_SL production. An alignment of EM and BpsA_SL is presented beneath the bar graph, with lines drawn showing the respective locations of the downstream subdomain borders for each variant. Where EM and BpsA_SL sequences are identical, amino acids are highlighted in black. Data were collected after 6 hours of incubation. Data are derived from four biological repeats, each conducted in triplicate. Error bars represent 1 S.D. A consistent trend was seen across all repeats.

Figure 4.6 shows the relative activity of the EMD variants in liquid culture (assay described in Section 2.11.2). Measurements shown in Figure 4.6 were taken after 6 hours of incubation. Because BpsA_SL wells in this assay tend to reach indigoidine saturation at around 3 hours, the activity of the variants is not a true comparison *versus* BpsA_SL; all EMD variants are substantially slower. This endpoint of this assay was chosen to coincide with the time that the best EMD variant approached indigoidine saturation – this allowed for the most accurate distinctions to be drawn between the relative activity of the EMD variants. A positive aspect of the slower activity of the EMD variants is that none of them reach saturation before wild-type, so assays are less likely to encounter ceiling effects. We note also that the fairly substantial error bars in Figure 4.6 are a consequence of inter-replicate variation, rather than inherent variability between different constructs. Owing to equipment limitations in our lab, multiple shaking incubators were used across different replicates, one of which had difficulty in maintaining a constant temperature. While the quantitative values differed between replicates, the mean values depicted in Figure 4.6 accurately reflect the consistent pattern of activity seen across replicates.

The distribution of activities portrayed in Figure 4.6 suggests that there was significant variation in activity across the border-sliding region, and that activity did not cluster around one particular favourable transition-zone. However, it was notable that the Kries et al. (2015) border-equivalent variant, EMD_10, was not the best performing variant, nor did it even appear to be a 'local maximum', with the proximal EMD_9 being the best of its own nearest neighbours. Overall, the two most prominent peaks were EMD_6 and EMD_7.

4.5.6 Selection of Optimal Downstream Subdomain Boundary

In order to begin upstream border-sliding experiments in our EM construct, we needed to select a fixed downstream subdomain border from our panel. In terms of performance in EMD *in vivo* assays, EMD_6 and EMD_7 were essentially indistinguishable as the best performing variants. Anticipating that we would subsequently be attempting to recover activity in a WS subdomain substitution construct using our preferred border, the selection of the "top" border variant included consideration of which border was *most likely* to function in WS on a protein sequence basis.

We selected EMD_7 on this basis, as within 5 amino acids of the border point WS and BpsA are 30% pairwise identical, and 60% pairwise positive using the BLOSUM62 substitution matrix. In contrast, within 5 amino acids of the EMD_6 border, WS and BpsA share only 20% pairwise identity and are 40% BLOSUM62 pairwise positive. This contrast is represented visually in Figure 4.7.

Figure 4.7 Extraction of alignment of WS and BpsA sequence at the downstream border. CR7 and CR8 are visible at the upstream edge of the extraction. EMD_6 and EMD_7 are marked in purple. Amino acid colour represents similarity across the alignment. Green = Identical. Yellow = Similar, with a BLOSUM score matrix value of +1 or higher. White = Dissimilar (BLOSUM value less than +1).

4.6 EM Upstream Border Sliding

4.6.1 Design and Construction

Having identified a preferred downstream border, we then sought to identify a preferred upstream border that was compatible with this and that would enable a functional substitution of the EM subdomain into BpsA_SL. To do this, we maintained the EMD7 downstream border and varied upstream border across 36-amino acid region bounded upstream by the A3 structural

motif, and downstream by the first coding residue investigated in this study (Fig. 4.8). Thirteen constructs were designed at 3-amino acid intervals across this 36-amino acid range.

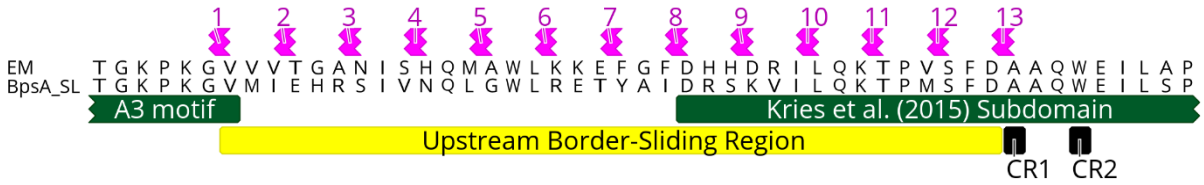


Figure 4.8 Sequence-level alignment of EM and BpsA_SL showing the upstream border-sliding region. The first two coding residues (CR1 and CR2) are denoted by black boxes. The Kries et al. (2015) subdomain and A3 motif are shown by labelled green annotations. Magenta annotations above the alignment show the 13 upstream subdomain border points.

The thirteen constructs, named EMU_1 through EMU_13, were assembled using NEBuilder in the same manner as described for the downstream border-sliding constructs.

4.6.2 EM Upstream In Vivo Assays

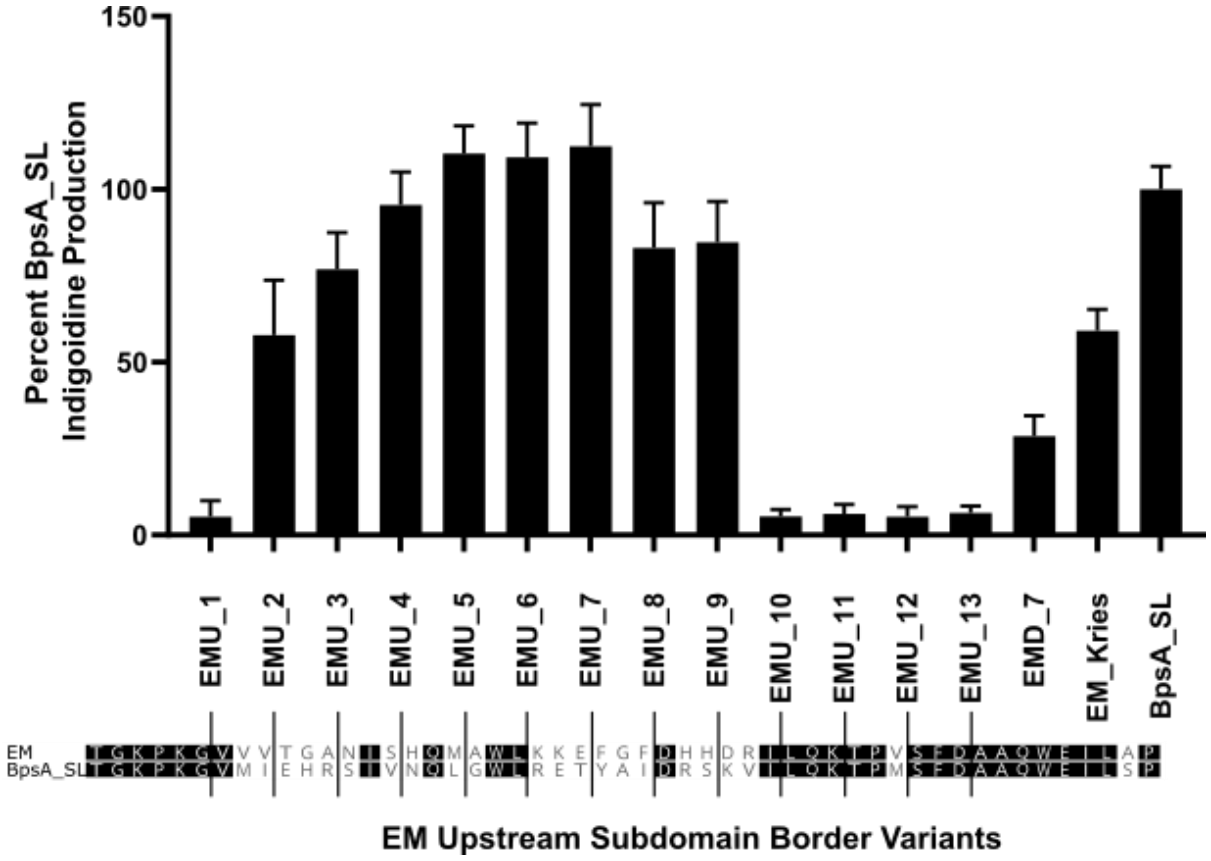


Figure 4.9 Indigoidine production by EMU variants expressed as a percentage of BpsA_SL production. An alignment of EM and BpsA_SL is presented beneath the bar graph, with lines drawn showing the respective locations of the upstream subdomain borders for each variant. Where EM and BpsA_SL sequences are identical, amino acids are highlighted in black. Also included in the assay are EMD_7, the best construct from the downstream-border panel, and the EM subdomain substituted into BpsA at the Kries boundaries (EM_Kries). Data were collected after 3 hours of incubation. Data are derived from 3 biological repeats, each conducted in triplicate. Error bars represent 1 S.D.

The 13 upstream border-sliding variants were then tested in *in vivo* liquid culture assays (Section 2.11.2). Figure 4.9 demonstrates that the sequential border-sliding strategy was indeed successful in identifying superior EM subdomain substitution constructs; the best EMU variants showed markedly improved activity over construct EM_Kries which utilised the borders of Kries et al. (2015). Moreover, the best EMU variants showed better activity than the best of what were essentially partial A-domain substitutions, the downstream-border-only variants, represented by construct EMD_7 in Figure 4.9. This indicated that a genuine/canonical subdomain substitution yielded better activity than constructs with a single subdomain border alone. This feature is further exemplified by the shortened length of incubation required in this EMU assay; whereas EMD variant *in vivo* assays were read after 6 hours of incubation, EMU variants began to approach saturation after only 3 hours.

The construct EMU_8, which consisted of our optimised downstream border (EMD_7) and the Kries et al. (2015) upstream border, was not the best of the EMU variants. Coupled with the observation that EMU_8 was more active than EM_Kries, this indicated additive improvement from both optimised borders of the subdomain. That is, optimising the downstream border (over those employed by Kries et al.) improved activity, and subsequently optimising the upstream border improved it further still.

Also of interest is the relationship between upstream border location and relative activity. Activity seems to 'peak' slightly before the midpoint of the border-sliding region, with a sharp drop-off in activity at the upstream and downstream ends. EMU_1, the variant with an upstream subdomain border situated within the A3 motif, had very low activity, which was somewhat unexpected considering the motifs are highly conserved. This may suggest that steric clashes are particularly deleterious just downstream of the A3 motif; constructs with border points further downstream (EMU_2 through EMU_9) had much higher activity despite their borders falling in regions of lower homology.

4.7 Introducing Optimised Subdomain Borders to WS

The optimal subdomain borders identified in EM border sliding experiments (i.e., upstream corresponding to EMU_5 and downstream to EMD_7) were introduced to the WS construct, with the expectation that optimised borders may give rise to a function enzyme, in contrast to the non-functional enzyme generated using the Kries et al. (2015) boundaries. In addition to a full subdomain substitution using the preferred boundaries, six chimeric BpsA constructs were generated that recombined at the downstream optimised border position only, analogous to the

EMD constructs generated in Section 4.6. This was to ensure that continuing experiments using the EMD_7 downstream border only would not preclude discovery of potential routes to recovery of WS activity; we wanted to be sure that EMD_7 was the best downstream recombination point in WS constructs.

4.7.1 WS Downstream Border Optimisation

With a number of downstream subdomain borders showing promise in EM experiments, and limited rationale for how they might perform in a WS context, the top 6 subdomain borders from EM downstream border-sliding experiments were chosen for introduction into the WS vehicle. These 6 points; 1, 2, 3, 6, 7, and 9, maintained a breadth of probing across the subdomain, maximising our chances of finding a transferrable border.

Constructs were assembled using NEBuilder and named WSD_1, 2, 3, 6, 7, and 9.

In vivo assays of the WSD variants, both on solid media and in liquid culture, returned negative results across the board (Fig. 4.10).

Figure 4.10 Picture of WSD_1 (left) and WSD_3 (right) in comparison to BpsA_SL (centre), incubated at 18 °C for 48 hours after IPTG induction. Picture is illustrative of results for all 6 WSD variants.

While no WS Downstream-border-only variant produced indigoidine, these constructs were essentially partial A-domain substitutions, from the N-terminus to the downstream border, rather than full subdomain substitutions. We remained hopeful that upon introduction of an optimised upstream subdomain border, some activity might be seen.

4.7.2 WS Upstream Border Optimisation

The top three EM upstream border-sliding variants, 5, 6, and 7, were essentially indistinguishable from each other in liquid culture assays (Fig. 4.9). Due to time constraints it was necessary to select only one upstream border point for introduction into the WS construct, so that a gene fragment could be ordered for synthesis. The border equivalent to that seen in EMU_5 was selected as the upstream border, by the same justification as the EM downstream

border in Section 4.5.6 – we predicted it to have the highest likely tolerability in WS. Across the closest 10 residues to this border WS shares 80% identity to BpsA_SL and is 90% BLOSUM62 pairwise positive (Fig. 4.11).

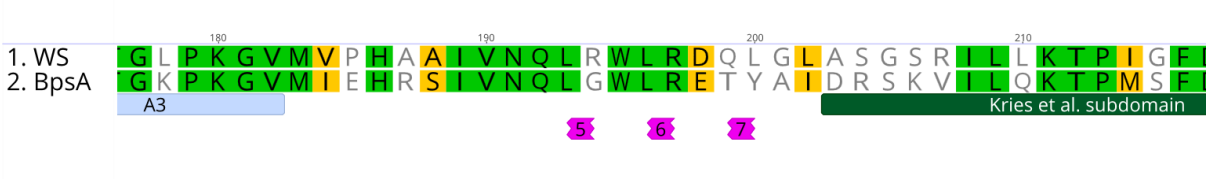


Figure 4.11 Extraction of alignment of WS and BpsA sequence at the upstream border. Subdomain border candidates 5, 6 and 7 are marked in purple. Amino acid colour represents similarity across the alignment. Green = Identical. Yellow = Similar, with a BLOSUM score matrix value of +1 or higher. White = Dissimilar (BLOSUM value less than +1).

The synthesised gene fragment was subsequently amplified and assembled into the vector backbone by restriction cloning. This resulted in a circular vector, named WSU, identical to an equivalent construct generated through the NEBuilder strategy applied in earlier experiments.

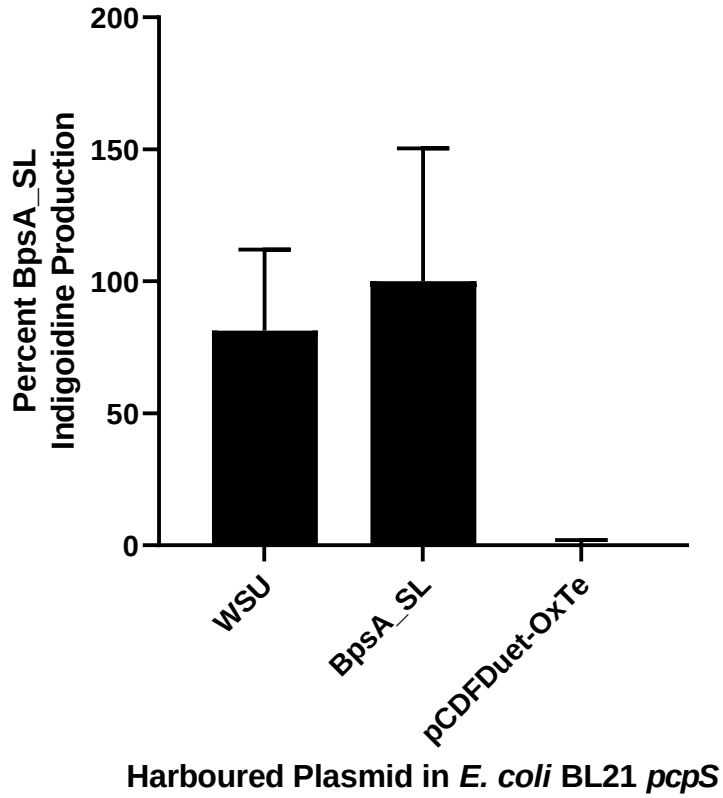


Figure 4.12 Indigoidine production of WSU variant (BpsA harbouring a WS subdomain using our optimised subdomain borders), expressed as a percentage of BpsA_SL production. Data for the negative control well, expressing a truncated BpsA protein from empty pCDFDuet-OxTe and against which indigoidine readings were normalised, are also shown, to illustrate the difference between the variants and background. Measurements were made 25 hours after induction. Data are derived from three biological repeats, each conducted in triplicate. Error bars represent 1 S.D.

Liquid culture assays (Fig. 4.12) demonstrated that the WSU construct was able to generate indigoidine, confirming that our optimised borders recovered activity in a previously non-

functional subdomain substitution. However, the rate of indigoidine synthesis by WSU was substantially reduced in comparison to BpsA_SL, a fact not well represented by Figure 4.12. Due to the low activity of the WSU enzyme, the assay was run for 25 hours, by which time BpsA_SL wells were indigoidine-saturated. An accurate comparison between WSU and BpsA_SL cannot be drawn from Figure 4.12 although a negative control column, BL21 *pcpS* harbouring an empty expression vector, is included to illustrate the degree to which WSU measurements differ from a non-indigoidine-producing cell line.

4.8 Introduction of Optimised Subdomain Borders to a LicA Construct

Having identified subdomain recombination boundaries that were superior in the BpsA context to those used by Kries et al. (2015) in GrsA, the final challenge was to see whether these would permit functional substitution of the most divergent L-Gln-activating subdomain in our study – that from LicA. When this construct (LicA_Opt_Sub) was generated via restriction cloning, used to transform BL21 *pcpS*, and assayed for indigoidine production, it was found to be non-functional. In the solid media *in vivo* assay (Section 2.11.1) after 96 hours no blue pigment could be observed in LicA_Opt_Sub colonies. The liquid culture assay (Section 2.12.1), which we found to be more sensitive to indigoidine production for A-domain variants of BpsA, confirmed this negative result.

4.9 Discussion of Chapter Results

Initial experiments in this chapter showed that subdomain substitution utilising the borders defined by Kries et al. (2015) produced an enzyme capable of synthesising indigoidine in only one out of four tested cases (i.e. when substituting the subdomain derived from the BpsA homologue of *Erwinia mallotivora*; Section 4.2). In contrast when the alternative strategy of coding residue reprogramming was used (Section 3.4 - 3.5), three out of four equivalent substitutions generated a functional enzyme.

Subdomain border-sliding experiments show that BpsA, and likely other NRPSs, are very sensitive to the precise boundaries of the substituted region. The difference between an inactive and near-wild-type-activity enzyme can be as little as a 3-AA shift in a single border location (e.g. EMU_9 vs EMU_10; Fig. 4.9). While subdomain substitution may, as originally theorised by Kries et al. (2015), avoid the issue of binding pocket perturbation, it can clearly introduce other issues if borders are not chosen carefully. Experiments conducted in this chapter were more thorough than either of the published subdomain substitution papers to date (Crüsemann et al., 2013; Kries et al., 2015); whereas both prior studies identified a single set of subdomain

borders on a rational basis, the use of BpsA as an effective reporter of activity allowed us to identify optimal borders experimentally. Our preferred pairs of borders generated functional enzymes where the Kries et al. (2015) borders did not, indicating that our borders may be a superior basis for further experimentation. However, it remains to be seen if our optimised borders are transferrable to other NRPS systems (beyond BpsA).

It is possible that the interrupting Ox-domain in BpsA will impact the transferability of our subdomain borders to other A-domains. With no crystal structure available for BpsA, it is unclear what structural effect the Ox-domain has on the A-domain binding pocket, however Takahashi et al. (2007) noted that when the Ox-domain of BpsA_SL is mutated to be non-functional, the adenylation activity of the A-domain drops by 50-70%. A strong interaction between the A-domain and Ox-domain in BpsA_SL may mean our findings do not translate directly into orthodox A-domains where the Ox-domain is absent. In the work completed in this thesis, only subdomain substitutions into BpsA_SL from other BpsA orthologs (i.e., also containing Ox-interrupted A-domains) were found to function, with the caveat that the only non-interrupted A-domain to be tested was LicA, a significant outlier to other L-Gln A-domains. Across the optimised subdomain region LicA shared 28.8% amino acid identity to BpsA_SL, in comparison to shared identity of 40.3% between WS and BpsA_SL, and 50.7% between EM and BpsA_SL. Had time permitted, it would have been desirable to test alternative L-Gln specifying subdomains from non-BpsA enzymes, however the constrained Masters research timeline did not allow this.

Further testing with other non-interrupted A-domains (ones more similar to BpsA than LicA), e.g. from TycC, is required to make a judgement about transferability of findings (TycC shares 35.9% identity to BpsA_SL across the optimised subdomain). Additionally, the TycC L-Gln A-domain that was investigated in Chapter 3 is from the second module – an elongation module – of the TycC NRPS, whereas BpsA homologues and LicA are taken from initiation modules. This means in its native context the TycC A-domain interacts with an upstream (and downstream) C-domain (Degen et al., 2019). If this characteristic also has an impact on the structure of the A-domain binding pocket, being able to achieve effective subdomain substitution of TycC with our optimised subdomain borders would be a strong endorsement of the general transferability of our borders into other NRPS modules. A smaller step in this same direction could be achieved by subdomain substitution of certain BpsA homologue A-domains which appear to have vestigial fragments of C-domains at their N-termini (Walsh & Wencewicz, 2013). If these substitutions do not work despite otherwise sharing a high level of

identity with BpsA_SL, it would imply that the vestigial C-domain has a steric impact on the binding pocket.

To our knowledge the experiments reported in this chapter represent the first time complete NRPS products have been synthesised *in vivo* using the subdomain substitution technique. This reflects the utility of the BpsA reporter system and highlights an advantage of assays which detect the final product, as opposed to detection of ATP consumption as an intermediate proxy for adenylation activity. Crüsemann et al. (2013) relied on ATP assays of adenylation activity as a reporter for engineered enzyme function. Kries et al. (2015) also relied primarily, though not exclusively, on ATP assays to determine the functionality of their engineered enzymes. The foremost advantage of measuring product directly is it confirms all steps in NRP synthesis are completed successfully.

5 Discussion

5.1 Summary of Key Findings

The over-arching goal of the research described in this thesis was to compare two promising A-domain manipulation strategies for NRPS engineering: coding residue reprogramming and subdomain substitution. Reprogramming experiments in the BpsA model system showed that even minor changes – as little as a single amino acid – could have critical impacts on the function of the enzyme, even when those changes were derived from A-domains with the same substrate specificity. The alternative subdomain substitution strategy was initially largely ineffectual, demonstrating either that this method is not directly transferable with any degree of efficiency, or that the single set of boundaries tested by Kries et al. (2015) were not optimal for effective subdomain substitution. Consistent with the latter supposition, Kries et al. found their best subdomain-substituted enzyme variant was 300 times slower than wild-type, suggesting that these boundaries are inefficient. We showed that with some adjustments to the borders of the subdomain region, inactive substitutions could have their activities recovered, and substitutions with initially poor activity could be greatly improved.

This research addressed a gap in understanding around the level of reprogramming that is sufficient to 'break' an A-domain, and found that the limit of tolerability can fall within the set of synonymous specificity codes. This finding is reflected in the relatively minor changes in specificity achieved in the literature thus far, and amplifies concerns in the field about the utility of the reprogramming strategy (Brown et al., 2018; Kries, 2016).

The seminal papers on subdomain substitution (Crüsemann et al., 2013; Kries et al., 2015) selected subdomain borders on a rational basis. However, the size and complexity of NRPS enzymes, coupled with the paucity of solved A-domain structures, means that these borders may not be transferrable to other systems, as our results indicated. Our findings have been reinforced in the recent study of Iacovelli et al. (2020), who attempted subdomain substitution in the first module of the bacterial NRPS ACVS, part of the β-lactam synthesis pathway. They successfully substituted a synonymous fungal subdomain into ACVS using the borders defined by Kries et al. (2015), but saw a significant drop in product titre relative to the wild-type enzyme, very similar to our findings with EM_Kries (Fig. 4.9). Our technique of subdomain border sliding, the most in-depth analysis of recognition subdomains to date, proved useful in identifying much superior subdomain borders – though it remains to be seen if these borders

are transferrable to other NRPS systems or unique to BpsA. Regardless of whether the precise borders are applicable to other NRPSs, we have developed a promising system for optimising subdomain substitution experiments in the form of this subdomain border sliding technique.

5.2 Critical Evaluation of BpsA as a Model System

The strength of BpsA as a model system lies in its relative simplicity and ability to generate a visually identifiable and readily-quantified product. The primary benefit of this is that it provides a holistic measure of enzyme function. Many NRPS products are only detectable by chromatography or mass spectrometry, and consequently assays of adenylation activity (via the proxy of ATP consumption), rather than assays of the final product, are often used as a primary measure of NRPS engineering success due to their comparatively lower cost and higher throughput. These two measures are not equivalent. For example, Zhang et al. (2013) developed a method of directed evolution to engineer alterations in substrate specificity in the A-domain of DhbE in *Bacillus subtilis*, measured by ATP/PPi release assay (Section 2.13). However, they found that a point mutation common among their evolved A-domains prevented the downstream thiolation reaction (attachment of the adenylated substrate to the PCP-domain 'arm') from occurring. Had they not conducted assays of downstream synthesis steps, they might have been left with the false impression that their evolved enzymes were functional without adjustment. Use of an end-product assay confirms up-front that all synthesis steps are working.

However, the production of indigoidine as an end product is also a limitation of our model system, due to complications it introduces to *in vivo* assays. Indigoidine may be utilised as an antimicrobial secondary metabolite in some species (Cude et al., 2012), or may be a shunt product of antibiotic synthesis (Kong et al., 2019) and is mildly toxic to *E. coli* (Owen et al., 2011). Consequently *E. coli* cells that express *bpsA* variants – particularly more active variants – are under a selection pressure to cease that expression, thereby under-reporting enzyme activity. This issue can be largely overcome by suppressing gene expression until the assays begin, and extending the length of time the assay runs. Nevertheless, selection pressure during the assay remains an extraneous variable. Indigoidine introduces an additional complication to liquid culture assays due to its propensity to drop out of solution over time and with increasing concentration (Brown et al., 2017). This can make it difficult to accurately compare enzymes of differing speeds, as wells containing faster enzymes may reach indigoidine saturation before slower enzymes have produced non-trivial quantities of indigoidine. This issue could also be

addressed by the introduction of additional steps to the assays used in this thesis (Section 2.11.2), such as solubilisation of precipitated indigoidine in a controlled volume of DMSO, a technique used successfully by Brown et al. (2017) (this was not attempted in this study as such an end-point assay requires re-optimisation of reaction duration for each new set of samples). Perhaps the biggest indigoidine-dependent limitation of our system is that it restricts us to L-glutamine as a target substrate. However, in our study this L-glutamine dependence was actually advantageous, as we were not investigating other substrate specificities and it ruled out accidental detection of off-target adenylation. This feature could also be useful in future studies with non-specific/promiscuous A-domains; in combination with ATP/PPi assays, it could act as a sensitive means of detecting trace levels of L-Gln adenylation. Finally, as noted in Section 4.11, it remains to be discovered whether the embedded Ox-domain on BpsA might render effective subdomain substitutions in this system less transferrable to other NRPS enzymes.

5.3 Future Directions

The most pressing future direction to emerge from the work described in this thesis is the need for more comprehensive investigation of the applicability and transferability of the optimised subdomain borders. Testing the optimised subdomain of TycC would be illuminating – the TycC coding residues functioned in the BpsA_SL binding pocket, while transplantation of the TycC subdomain according to the Kries et al. (2015) boundaries did not. Optimisation and analysis of a TycC subdomain construct could answer the question of whether the optimised borders work for subdomains from non-interrupted A-domains. In a similar vein, subdomain substitutions from additional BpsA homologues and other L-Gln A-domains would provide still greater clarity as to the transferability of our borders. From there, the next step would be to apply our optimised borders, and possibly the border-sliding strategy, to other model systems. A good candidate NRPS is PvdD, part of the pyoverdine synthesis pathway in *Pseudomonas aeruginosa.* This NRPS has been used as a model system in our own lab and found to be amenable to full domain substitutions (Calcott et al., 2014). Moreover, the product of PvdD, pyoverdine, is an iron-chelating siderophore. Consequently, functional engineered PvdD variants can be selected/identified *in vivo* by expressing the variant *pvdD* genes in a *pvdD* deletion strain of *P. aeruginosa* grown under iron-limited conditions. Furthermore, owing to their characteristic fluorescence, generation of pyoverdines, functional or otherwise, can be detected spectrophotometrically *in vivo* (M. J. Calcott et al., 2014). Another advantage of the PvdD model system is that the peptide chain of pyoverdine is variable, and across different

strains of pseudomonads many different monomers are incorporated into a varied family of pyoverdines (Meyer et al., 2008). The natural ability of the pyoverdine pathway to incorporate many different monomers would allow us to broaden the scope of our subdomain experiments to investigate numerous substrate specificities beyond L-Gln, and importantly investigate non-synonymous substitutions – a key goal of NRPS engineering (Brown et al., 2018). These potential experiments would in turn permit investigation of substrates with a greater number of known specificity codes. In contrast to the 15 unique L-Gln specificity codes identified in nature and collated in this thesis, Throckmorton et al. (2019) were able to experimentally confirm upwards of 150 novel L-Ser specificity codes in an EntF model, showing that the panel of L-Gln codes currently known is comparatively very small. A downside of working with PvdD as a model system is that *P. aeruginosa* is less amenable than *E. coli* to high-throughput transformation studies (Prof. David Ackerley, *personal communication*). The *E. coli* EntF model system is another promising NRPS for testing of our subdomain borders and sliding strategy. Although lacking the fluorescent chromophore of pyoverdine, the product of EntF, enterobactin, is also a siderophore and can similarly be selected *in vivo*.

Both EntF and PvdD models would also expose our strategies to the complications arising from the proofreading action of the C-domain (Ackerley et al., 2003; Gehring et al., 1998). The absence of a C-domain in BpsA was useful as it simplifies the model, but in systems where it is present it has complicated attempts to change NRPS substrate specificity (Bozhüyük et al., 2018, 2019; Calcott et al., 2020)

As this thesis approached completion, Iacovelli et al. (2020) reported a similar subdomain substitution experiment to our initial experiment (utilising the boundaries of Kries et al., 2015) in ACVS. Their synonymous subdomain substitution suffered significantly reduced yield relative to the wild-type enzyme, mirroring our findings with EM subdomain substitutions in BpsA. In our case, optimising the borders of the subdomain in a BpsA context dramatically improved activity of negligibly-active and non-functional subdomain substitutions – attempting these same optimised borders in the ACVS model of Iacovelli et al. would be an excellent measure of the applicability of our optimised borders. Additionally, Iacovelli et al. (2015) attempted to make non-synonymous subdomain substitutions, all of which failed to generate product. This could be a consequence of downstream C-domain gatekeeping activity, or it could be a consequence of a totally non-functional A-domain. Applying our subdomain borders, or border-sliding strategy, may elucidate the cause and potentially enable A-domain activity to be recovered.

More minor future steps involve expanding upon preliminary results obtained during this Masters project. Due to laboratory shutdown as a consequence of Covid-19, various experiments were not able to be completed to our satisfaction. For example, more repeats of the *in vitro* protein assays would provide a more complete picture of relative enzyme activity. An additional experiment we had intended to run was an SDS-PAGE analysis of subdomain substitution variants, to ensure engineered protein variants deemed non-functional were still soluble – this too becomes a future direction. The malachite green assay (Section 2.13), once optimised to detect adenylation in BpsA, would also provide useful information about the work conducted in this thesis and hypothetical future work. While indigoidine production is an excellent measure of overall enzyme activity, if enzyme variants fail to generate indigoidine, a malachite green assay could narrow down the cause of the dysfunction and establish if engineered changes have resulted in an enzyme unable to adenylate, or if downstream synthesis steps – most likely thiolation (as per Zhang et al., 2013) – are impaired. When work expands into attempts at non-synonymous subdomain substitution, for instance in PvdD, the relative signal in the malachite green assay for various substrates could be compared to the relative titre of pyoverdines incorporating the corresponding substrate.

Finally, more work is needed to ascertain the reasons for the non-functionality of BpsA variants that were reprogrammed with the LicA coding residues. Completing the panel of all possible permutations/combinations of the LicA coding residues will permit determination of exactly which residue combinations are not tolerated by BpsA. Past research into A-domain coding residues has found that certain coding residue locations ("wobble" positions) are more tolerant to change than others, with "wobbly" coding residues varying by substrate specificity (Chevrette et al., 2017; Eppelmann et al., 2002; Throckmorton et al., 2019). We suspect a full picture of LicA residue tolerability in BpsA would inform understanding of wobble positions in BpsA, or L-Gln codes more generally. Subsequent investigations could utilise a directed evolution technique, subjecting the LicA_Res construct from Section 3.5.1 to random mutagenesis and a subsequent screen to recover enzyme activity. A similar directed evolution technique has been utilised in BpsA for improvement of non-native PCP-domains to great effect (Owen et al., 2016). We suspect that if improved mutants were identified, they would show changes to the recognition subdomain which allow the BpsA A-domain to better accommodate the LicA coding residues (e.g., optimising sidechain orientation within the binding pocket). If this technique were successful it would provide detailed insight into key factors limiting the success of the coding residue reprogramming strategy.

5.4 Concluding Remarks

The research described in this thesis compared the A-domain engineering techniques of reprogramming and subdomain substitution, and found that both have individual benefits, though neither is a 'one-size-fits-all' solution. The technique of reprogramming can lead to perturbation of the binding pocket when changes to the coding residues are introduced alone. While changing only the coding residues can alter the specificity of an A-domain, the reprogramming technique does not aim to mimic nature – the probability of constructively mutating only 8 specific codons in a gene many thousands of bases long is infinitesimal. Perhaps early on in the evolution of NRPSs, small mutations in the binding pocket were an important phenomenon for changing substrate specificity. Now, with such a plethora of specific A-domains in existence, subdomain substitution seems perhaps more plausible as a means of acquiring novel activity. While both techniques are finicky and situational, they both offer promise for future development. In particular, where directed evolution can productively be applied, this offers excellent scope to adapt a native binding pocket to optimise the positioning of reprogrammed residues; while, as demonstrated here, a small-scale border-sliding campaign has potential to greatly enhance product yields following subdomain substitutions.

Bibliography

Ackerley, D. F. (2016). Cracking the nonribosomal code. *Cell Chemical Biology*, *23*(5), 535–537. https://doi.org/10.1016/j.chembiol.2016.05.001

Ackerley, D. F., Caradoc-Davies, T. T., & Lamont, I. L. (2003). Substrate specificity of the nonribosomal peptide synthetase PvdD from Pseudomonas aeruginosa. *Journal of Bacteriology*, *185*(9), 2848–2855. https://doi.org/10.1128/JB.185.9.2848-2855.2003

Beer, R., Herbst, K., Ignatiadis, N., Kats, I., Adlung, L., Meyer, H., Niopek, D., Christiansen, T., Georgi, F., Kurzawa, N., Meichsner, J., Rabe, S., Riedel, A., Sachs, J., Schessner, J., Schmidt, F., Walch, P., Niopek, K., Heinemann, T., … Di Ventura, B. (2014). Creating functional engineered variants of the single-module non-ribosomal peptide synthetase IndC by T domain exchange. *Mol. BioSyst.*, *10*(7), 1709–1718. https://doi.org/10.1039/C3MB70594C

Belshaw, P. J., Walsh, C. T., & Stachelhaus, T. (1999). Aminoacyl-CoAs as probes of condensation domain selectivity in nonribosomal peptide synthesis. *Science*, *284*(5413), 486–489. https://doi.org/10.1126/science.284.5413.486

Bloudoff, K., & Schmeing, T. M. (2017). Structural and functional aspects of the nonribosomal peptide synthetase condensation domain superfamily: discovery, dissection and diversity. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, *1865*(11), 1587–1604. https://doi.org/10.1016/j.bbapap.2017.05.010

Bozhüyük, K. A. J., Fleischhacker, F., Linck, A., Wesche, F., Tietze, A., Niesert, C. P., & Bode, H. B. (2018). De novo design and engineering of non-ribosomal peptide synthetases. *Nature Chemistry*, *10*(3), 275–281. https://doi.org/10.1038/NCHEM.2890

Bozhüyük, K. A. J., Linck, A., Tietze, A., Kranz, J., Wesche, F., Nowak, S., Fleischhacker, F., Shi, Y.-N., Grün, P., & Bode, H. B. (2019). Modification and de novo design of non-ribosomal peptide synthetases using specific assembly points within condensation domains. *Nature Chemistry*, *11*(7), 653–661. https://doi.org/10.1038/s41557-019-0276-z

Brown, A. S., Calcott, M. J., Owen, J. G., & Ackerley, D. F. (2018). Structural, functional and evolutionary perspectives on effective re-engineering of non-ribosomal peptide synthetase assembly lines. *Natural Product Reports*, *35*(11), 1210–1228. https://doi.org/10.1039/C8NP00036K

Brown, A. S., Robins, K. J., & Ackerley, D. F. (2017). A sensitive single-enzyme assay system using the non-ribosomal peptide synthetase BpsA for measurement of L-glutamine in biological samples. *Scientific Reports*, *7*(1), 41745. https://doi.org/10.1038/srep41745

Caboche, S., Leclère, V., Pupin, M., Kucherov, G., & Jacques, P. (2010). Diversity of monomers in nonribosomal peptides: towards the prediction of origin and biological activity. *Journal of Bacteriology*, *192*(19), 5143–5150. https://doi.org/10.1128/JB.00315-10

Calcott, M. J., & Ackerley, D. F. (2014). Genetic manipulation of non-ribosomal peptide synthetases to generate novel bioactive peptide products. *Biotechnology Letters*, *36*(12), 2407–2416. https://doi.org/10.1007/s10529-014-1642-y

Calcott, M. J., Owen, J. G., Lamont, I. L., & Ackerley, D. F. (2014). Biosynthesis of novel pyoverdines by domain substitution in a nonribosomal peptide synthetase of Pseudomonas aeruginosa. *Applied and Environmental Microbiology*, *80*(18), 5723–5731. https://doi.org/10.1128/AEM.01453-14

Calcott, M., Owen, J., & Ackerley, D. (2020). Efficient rational modification of non-ribosomal peptides by adenylation domain substitution. *BioRxiv*, 2020.02.28.970632. https://doi.org/10.1101/2020.02.28.970632

Challis, G. L., Ravel, J., & Townsend, C. A. (2000). Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chemistry and Biology*, *7*(3), 211–224. https://doi.org/10.1016/S1074-5521(00)00091-0

Chen, C.-Y., Georgiev, I., Anderson, A. C., & Donald, B. R. (2009). Computational structure-based redesign of enzyme activity. *Proceedings of the National Academy of Sciences*, *106*(10), 3764–3769. https://doi.org/10.1073/pnas.0900266106

Chevrette, M. G., Aicheler, F., Kohlbacher, O., Currie, C. R., & Medema, M. H. (2017). SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. *Bioinformatics*, *33*(20), 3202–3210. https://doi.org/10.1093/bioinformatics/btx400

Christiansen, G., Philmus, B., Hemscheidt, T., & Kurmayer, R. (2011). Genetic variation of adenylation domains of the anabaenopeptin synthesis operon and evolution of substrate

promiscuity. *Journal of Bacteriology*, *193*(15), 3822–3831. https://doi.org/10.1128/JB.00360-11

Conti, E., Stachelhaus, T., Marahiel, M. A., & Brick, P. (1997). Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. *EMBO Journal*, *16*(14), 4174–4183. https://doi.org/10.1093/emboj/16.14.4174

Crüsemann, M., Kohlhaas, C., & Piel, J. (2013). Evolution-guided engineering of nonribosomal peptide synthetase adenylation domains. *Chemical Science*, *4*(3), 1041–1045. https://doi.org/10.1039/C2SC21722H

Cude, W. N., Mooney, J., Tavanaei, A. A., Hadden, M. K., Frank, A. M., Gulvik, C. A., May, A. L., & Buchan, A. (2012). Production of the antimicrobial secondary metabolite indigoidine contributes to competitive surface colonization by the marine roseobacter Phaeobacter sp. strain Y4I. *Applied and Environmental Microbiology*, *78*(14), 4771–4780. https://doi.org/10.1128/AEM.00297-12

Degen, A., Mayerthaler, F., Mootz, H. D., & Di Ventura, B. (2019). Context-dependent activity of A domains in the tyrocidine synthetase. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-41492-8

Du, L., Chen, M., Sanchez, C., & Shen, B. (2000). An oxidation domain in the BlmIII non-ribosomal peptide synthetase probably catalyzing thiazole formation in the biosynthesis of the anti-tumor drug bleomycin in Streptomyces verticillus ATCC15003. *FEMS Microbiology Letters*, *189*(2), 171–175. https://doi.org/10.1111/j.1574-6968.2000.tb09225.x

Ehmann, D. E., Trauger, J. W., Stachelhaus, T., & Walsh, C. T. (2000). Aminoacyl-SNACs as small-molecule substrates for the condensation domains of nonribosomal peptide synthetases. *Chemistry and Biology*, *7*(10), 765–772. https://doi.org/10.1016/S1074-5521(00)00022-3

Eppelmann, K., Stachelhaus, T., & Marahiel, M. A. (2002). Exploitation of the selectivity-conferring code of nonribosomal peptide synthetases for the rational design of novel peptide antibiotics. *Biochemistry*, *41*(30), 9718–9726. https://doi.org/10.1021/bi0259406

Felnagle, E. A., Jackson, E. E., Chan, Y. A., Podevels, A. M., Berti, A. D., McMahon, M. D., & Thomas, M. G. (2008). Nonribosomal peptide synthetases involved in the production

of medically relevant natural products. *Molecular Pharmaceutics*, *5*(2), 191–211. https://doi.org/10.1021/mp700137g

Fewer, D. P., Rouhiainen, L., Jokela, J., Wahlsten, M., Laakso, K., Wang, H., & Sivonen, K. (2007). Recurrent adenylation domain replacement in the microcystin synthetase gene cluster. *BMC Evolutionary Biology*, *7*(1), 183. https://doi.org/10.1186/1471-2148-7-183

Finking, R., & Marahiel, M. A. (2004). Biosynthesis of nonribosomal peptides. *Annual Review of Microbiology*, *58*(1), 453–488. https://doi.org/10.1146/annurev.micro.58.030603.123615

Gehring, A. M., Mori, I., & Walsh, C. T. (1998). Reconstitution and characterization of the Escherichia coli enterobactin synthetase from EntB, EntE, and EntF. *Biochemistry*, *37*(8), 2648–2659. https://doi.org/10.1021/bi9726584

Han, J. W., Kim, E. Y., Lee, J. M., Kim, Y. S., Bang, E., & 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), 1327–1334. https://doi.org/10.1007/s10529-012-0913-8

Hur, G. H., Vickery, C. R., & Burkart, M. D. (2012). Explorations of catalytic domains in non-ribosomal peptide synthetase enzymology. *Natural Product Reports*, *29*(10), 1074. https://doi.org/10.1039/c2np20025b

Iacovelli, R., Zwahlen, R. D., Bovenberg, R. A. L., & Driessen, A. J. M. (2020). Biochemical characterization of the Nocardia lactamdurans ACV synthetase. *PLOS ONE*, *15*(4), e0231290. https://doi.org/10.1371/journal.pone.0231290

Ishikawa, F., Nohara, M., Nakamura, S., Nakanishi, I., & Tanabe, G. (2020). Precise probing of residue roles by nrps code swapping: mutation, enzymatic characterization, modeling, and substrate promiscuity of aryl acid adenylation domains. *Biochemistry*, *59*(4), 351–363. https://doi.org/10.1021/acs.biochem.9b00748

Keating, T. A., Ehmann, D. E., Kohli, R. M., Marshall, C. G., Trauger, J. W., & Walsh, C. T. (2001). Chain termination steps in nonribosomal peptide synthetase assembly lines: directed acyl-s-enzyme breakdown in antibiotic and siderophore biosynthesis. *ChemBioChem*, *2*(2), 99–107. https://doi.org/10.1002/1439-7633(20010202)2:2<99::AID-CBIC99>3.0.CO;2-3

Khayatt, B. I., Overmars, L., Siezen, R. J., & Francke, C. (2013). Classification of the adenylation and acyl-transferase activity of NRPS and PKS systems using ensembles of substrate specific hidden markov models. *PLoS ONE*, *8*(4), e62136. https://doi.org/10.1371/journal.pone.0062136

Kong, L., Xu, G., Liu, X., Wang, J., Tang, Z., Cai, Y.-S., Shen, K., Tao, W., Zheng, Y., Deng, Z., Price, N. P. J., & Chen, W. (2019). Divergent biosynthesis of C-nucleoside minimycin and indigoidine in bacteria. *IScience*, *22*, 430–440. https://doi.org/10.1016/j.isci.2019.11.037

Konz, D., Doekel, S., & Marahiel, M. A. (1999). Molecular and biochemical characterization of the protein template controlling biosynthesis of the lipopeptide lichenysin. *Journal of Bacteriology*, *181*(1), 133–140. http://www.ncbi.nlm.nih.gov/pubmed/9864322

Kries, H. (2016). Biosynthetic engineering of nonribosomal peptide synthetases. *Journal of Peptide Science*, *22*(9), 564–570. https://doi.org/10.1002/psc.2907

Kries, H., Niquille, D. L., & Hilvert, D. (2015). A subdomain swap strategy for reengineering nonribosomal peptides. *Chemistry and Biology*, *22*(5), 640–648. https://doi.org/10.1016/j.chembiol.2015.04.015

Labby, K. J., Watsula, S. G., & Garneau-Tsodikova, S. (2015). Interrupted adenylation domains: unique bifunctional enzymes involved in nonribosomal peptide biosynthesis. *Natural Product Reports*, *32*(5), 641–653. https://doi.org/10.1039/C4NP00120F

Lundy, T. A., Mori, S., Thamban Chandrika, N., & Garneau-Tsodikova, S. (2020). Characterization of a unique interrupted adenylation domain that can catalyze three reactions. *ACS Chemical Biology*. https://doi.org/10.1021/acschembio.9b00929

Mach, B., Reich, E., & 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), 175–181. https://doi.org/10.1073/pnas.50.1.175

Marahiel, M. A., Stachelhaus, T., & Mootz, H. D. (1997). Modular peptide synthetases involved in nonribosomal peptide synthesis. *Chemical Reviews*, *97*(7), 2651–2674. https://doi.org/10.1021/cr960029e

McQuade, T. J., Shallop, A. D., Sheoran, A., DelProposto, J. E., Tsodikov, O. V., & 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), 244–250. https://doi.org/10.1016/j.ab.2008.12.014

Meyer, J. M., Gruffaz, C., Raharinosy, V., Bezverbnaya, I., Schäfer, M., & Budzikiewicz, H. (2008). Siderotyping of fluorescent Pseudomonas: Molecular mass determination by mass spectrometry as a powerful pyoverdine siderotyping method. *BioMetals*, *21*(3), 259–271. https://doi.org/10.1007/s10534-007-9115-6

Miller, B. R., Drake, E. J., Shi, C., Aldrich, C. C., & 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), 22559–22571. https://doi.org/10.1074/jbc.M116.746297

Mori, S., Pang, A. H., Lundy, T. A., Garzan, A., Tsodikov, O. V., & Garneau-Tsodikova, S. (2018). Structural basis for backbone N-methylation by an interrupted adenylation domain. *Nature Chemical Biology*, *14*(5), 428–430. https://doi.org/10.1038/s41589-018-0014-7

Myers, J. A., Curtis, B. S., & Curtis, W. R. (2013). Improving accuracy of cell and chromophore concentration measurements using optical density. *BMC Biophysics*, *6*(1), 4. https://doi.org/10.1186/2046-1682-6-4

Nguyen, K. T., Ritz, D., Gu, J.-Q., Alexander, D., Chu, M., Miao, V., Brian, P., & Baltz, R. H. (2006). Combinatorial biosynthesis of novel antibiotics related to daptomycin. *Proceedings of the National Academy of Sciences*, *103*(46), 17462–17467. https://doi.org/10.1073/pnas.0608589103

Owen, J. G., Robins, K. J., Parachin, N. S., & Ackerley, D. F. (2012). A functional screen for recovery of 4′-phosphopantetheinyl transferase and associated natural product biosynthesis genes from metagenome libraries. *Environmental Microbiology*, *14*(5), 1198–1209. https://doi.org/10.1111/j.1462-2920.2012.02699.x

Owen, Jeremy G., Copp, J. N., & Ackerley, D. F. (2011). Rapid and flexible biochemical assays for evaluating 4′-phosphopantetheinyl transferase activity. *Biochemical Journal*, *436*(3), 709–717. https://doi.org/10.1042/BJ20110321

Owen, Jeremy G, Calcott, M. J., Robins, K. J., & Ackerley, D. F. (2016). Generating

functional recombinant nrps enzymes in the laboratory setting via peptidyl carrier protein engineering. *Cell Chemical Biology*, *23*(11), 1395–1406. https://doi.org/10.1016/j.chembiol.2016.09.014

Owen, J. G. (2010). *Characterisation, Manipulation and Directed Evolution of Non-Ribosomal Peptide Synthetase Enzymes* (Doctoral Dissertation, Victoria University of Wellington, Wellington, New Zealand). Retrived from http://restrictedarchive.vuw.ac.nz/handle/123456789/1959

Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W., & 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), 5799–5808. https://doi.org/10.1093/nar/gki885

Röttig, M., Medema, M. H., Blin, K., Weber, T., Rausch, C., & Kohlbacher, O. (2011). NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Research*, *39*(2), 362–367. https://doi.org/10.1093/nar/gkr323

Samel, S. A., Czodrowski, P., & Essen, L. O. (2014). Structure of the epimerization domain of tyrocidine synthetase A. *Acta Crystallographica Section D: Biological Crystallography*, *70*(5), 1442–1452. https://doi.org/10.1107/S1399004714004398

Samel, S. A., Wagner, B., Marahiel, M. A., & Essen, L. O. (2006). The thioesterase domain of the fengycin biosynthesis cluster: a structural base for the macrocyclization of a non-ribosomal lipopeptide. *Journal of Molecular Biology*, *359*(4), 876–889. https://doi.org/10.1016/j.jmb.2006.03.062

Schneider, A., Stachelhaus, T., & Marahiel, M. A. (1998). Targeted alteration of the substrate specificity of peptide synthetases by rational module swapping. *Molecular and General Genetics*, *257*(3), 308–318. https://doi.org/10.1007/s004380050652

Stachelhaus, T, Schneider, A., & Marahiel, M. (1995). Rational design of peptide antibiotics by targeted replacement of bacterial and fungal domains. *Science*, *269*(5220), 69–72. https://doi.org/10.1126/science.7604280

Stachelhaus, Torsten, Mootz, H. D., Bergendahl, V., & Marahiel, M. A. (1998). Peptide bond formation in nonribosomal peptide biosynthesis: Catalytic role of the condensation domain. *Journal of Biological Chemistry*, *273*(35), 22773–22781.

https://doi.org/10.1074/jbc.273.35.22773

Stachelhaus, Torsten, Mootz, H. D., & Marahiel, M. A. (1999). The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chemistry & Biology*, *6*(8), 493–505. https://doi.org/10.1016/S1074-5521(99)80082-9

Süssmuth, R. D., & Mainz, A. (2017). Nonribosomal peptide synthesis—principles and prospects. *Angewandte Chemie - International Edition*, *56*(14), 3770–3821. https://doi.org/10.1002/anie.201609079

Takahashi, H., Kumagai, T., Kitani, K., Mori, M., Matoba, Y., & Sugiyama, M. (2007). Cloning and characterization of a Streptomyces single module type non-ribosomal peptide synthetase catalyzing a blue pigment synthesis. *Journal of Biological Chemistry*, *282*(12), 9073–9081. https://doi.org/10.1074/jbc.M611319200

Tanovic, A., Samel, S. A., Essen, L. O., & Marahiel, M. A. (2008). Crystal structure of the termination module of a nonribosomal peptide synthetase. *Science*, *321*(5889), 659–663. https://doi.org/10.1126/science.1159850

Throckmorton, K., Vinnik, V., Chowdhury, R., Cook, T., Chevrette, M. G., Maranas, C., Pfleger, B., & Thomas, M. G. (2019). Directed evolution reveals the functional sequence space of an adenylation domain specificity code. *ACS Chemical Biology*, *14*(9), 2044–2054. https://doi.org/10.1021/acschembio.9b00532

Walsh, C. T., Chen, H., Keating, T. A., Hubbard, B. K., Losey, H. C., Luo, L., Marshall, C. G., Miller, D. A., & Patel, H. M. (2001). Tailoring enzymes that modify nonribosomal peptides during and after chain elongation on NRPS assembly lines. *Current Opinion in Chemical Biology*, *5*(5), 525–534. https://doi.org/10.1016/S1367-5931(00)00235-0

Walsh, C. T., & Wencewicz, T. A. (2013). Flavoenzymes: Versatile catalysts in biosynthetic pathways. *Natural Product Reports*, *30*(1), 175–200. https://doi.org/10.1039/C2NP20069D

Wang, H., Sivonen, K., & Fewer, D. P. (2015). Genomic insights into the distribution, genetic diversity and evolution of polyketide synthases and nonribosomal peptide synthetases. *Current Opinion in Genetics & Development*, *35*, 79–85. https://doi.org/10.1016/j.gde.2015.10.004

Williams, E. M., Copp, J. N., & Ackerley, D. F. (2014). *Site-Saturation Mutagenesis by*

Overlap Extension PCR (pp. 83–101). Springer, New York, NY. https://doi.org/10.1007/978-1-4939-1053-3_6

Yakimov, M. M., Giuliano, L., Timmis, K. N., & Golyshin, P. N. (2000). Recombinant acylheptapeptide lichenysin: high level of production by Bacillus subtilis cells. *Journal of Molecular Microbiology and Biotechnology*, *2*(2), 217–224. https://www.caister.com/jmmb/v/v2/v2n2/09.pdf

Zhang, K., Nelson, K. M., Bhuripanyo, K., Grimes, K. D., Zhao, B., Aldrich, C. C., & Yin, J. (2013). Engineering the substrate specificity of the dhbe adenylation domain by yeast cell surface display. *Chemistry and Biology*, *20*(1), 92–101. https://doi.org/10.1016/j.chembiol.2012.10.020

Zobel, S., Boecker, S., Kulke, D., Heimbach, D., Meyer, V., & Süssmuth, R. D. (2016). Reprogramming the biosynthesis of cyclodepsipeptide synthetases to obtain new enniatins and beauvericins. *ChemBioChem*, *17*(4), 283–287. https://doi.org/10.1002/cbic.201500649