PRe-ART (Predictive Reagent-Antibody Replacement Technology): Engineering and analysis of randomised DNA libraries encoding designed armadillo repeat proteins

Ben Phillip Gordon Wagstaffe

Doctor of Philosophy

Aston University

September 2021

Aston University

PRe-ART (Predictive Reagent-Antibody Replacement Technology): The engineering and analysis of randomised designed armadillo repeat protein DNA libraries

Ben Phillip Gordon Wagstaffe

Doctor of Philosophy

September 2021

Thesis Abstract

PRe-ART (Predictive Reagent- Antibody Replacement Technology) aims to replace reagent monoclonal antibodies with designed armadillo repeat proteins (dArmRPs), made from sequence-defined modular subunits capable of specific and conserved dipeptide recognition and binding. These modular units joined in a 'Lego-brick' fashion, will generate proteins capable of binding a user-defined target peptide, removing the costly and timely traditional immunisation process and the associated issue of unreproducible results.

This project contributed to PRe-ART by generating randomised DNA libraries targeting key binding resides of the two pockets within the armadillo repeat, aiming to alter the unit's specificity. Successful saturation of seven positions using MAX randomisation, produced a randomised DNA library of the dArmRP pocket that originally bound arginine. Computational designs provided by collaborators in the Höcker group (University of Bayreuth), directed the engineering of a specific DNA library, aiming to engineer an improved threonine-binder. Separately, to accommodate for saturating contiguous codons in the second, Lysine-binding pocket (not possible with MAX randomisation) a new saturation mutagenesis technology, ParaMAX randomisation, was invented. This MAX randomisation derivative was implemented on an adapted dArmRP sequence, generating a region of four contiguous randomised codons. Subsequently, novel Next Generation Sequencing (NGS) analysis techniques were developed to assess the success of positional saturation as a quality control stage before protein expression and screening by collaborators in the Plückthun group (University of Zurich).

Analyses of the DNA libraries engineered using MAX randomisation showed successful target saturation and were therefore used in protein production and screening. The proof of concept ParaMAX library analysis revealed further optimisation of the ParaMAX process was required to prevent deletions interfering with amino acid representation. This analysis also revealed limitations in existing alignment technologies when processing such unique DNA libraries. Alternative stratagems for ParaMAX and the processing of NGS data are considered in light of these results.

Keywords: repeat proteins, designed armadillo repeat protein, nondegenerate saturation mutagenesis, Next Generation Sequencing, protein engineering,

Acknowledgements

Firstly my thanks go to Professor Anna Hine and Dr Andrew Sutherland, for their guidance and generosity. Thank you for giving me such an amazing PhD experience; the opportunities to learn for myself while always being there to support me. Thank you to both Professor Andreas Plückthun, Professor Birte Höcker and their groups for their hospitality and expertise. Thanks also goes to FET Open: Horizon 2020 for funding my PhD.

My unending gratitude must also go to my friends who have been with me during this journey: Mo for your training and generosity, Anu your wisdom and experience, Marta your encouragement and mentorship, Sarah, John and Romez for the laughs and the venting.

Thank you to my family. Mom and Dad, you have always pushed me to do my best and without your continued support, I would not be where I am today. Thank you to Debbie and Tony, for your generosity and kindness.

And thank you Katie; my PhD journey has made me a better scientist, while you have made me a better person.

Contents

List of abbreviations

ATLIGATOR- Atlas-based LIGand binding site ediTOR)

At-Thr- ATLIGATOR- Threonine

Bp- Base pair

BLAST- Basic Local Alignment Search Tool

BSA- Bovine serum albumin

CAS/CSE1L- cellular apoptosis susceptibility protein

cDNA- Complementary DNA

CDR3- Complementarity-determining region 3

DARPins- Designed Ankyrin Repeat Proteins

DNA- deoxyribonucleic acid

dNTP- deoxyribonucleotide triphosphate

DTT- Dithiothreitol

ELISA- Enzyme-Linked Immunosorbent Assay

EDTA- Ethylenediaminetetraacetic acid

FACs- Fluorescence Activated Cell sorting

FRET- Fluorescence resonance energy transfer

hheC- Halohydrin dehalogenase

ITCHY- iterative truncation for the creation of hybrid enzymes

MHC- Major Histocompatibility Complex

mRNA- Messenger RNA

MW- Molecular weight

NGS- Next Generation Sequencing

NLS- Nuclear localisation signal

PCR- Polymerase Chain Reaction

PET- Positron Emission Tomography

RACHITT- Random Chimeragenesis on Transient Templates

RID- Random Insertion/Deletion

RNA- ribonucleic acid

S. cerevisiae- Saccharomyces cerevisiae

StEP- Staggered Extension Process

SDS-PAGE- sodium dodecyl sulphate–polyacrylamide gel electrophoresis

TCF/LEF- T cell factor/lymphoid enhancer factor

TRIM- trinucleotide-directed mutagenesis

TPs- trinucleotide phosphoramidites

Tris-HCL- trisaminomethane- hydrochloride

TAE- Tris base, acetic acid and EDTA

VHH- Single variable domain on a heavy chain

Wnt- Wingless and Int-1

ANS- 8-Anilinonaphthalene-1-sulfonic acid

List of Figures

List of Tables

List of Appendices

Chapter 1 Introduction

The importance of developing new protein binders is demonstrated in their ever-increasing popularity and number of applications. Protein binders are an invaluable therapeutic agent for a range of diseases (Dimitrov, 2012) and used extensively in laboratory-based diagnostics, for example serological tests using ELISA (enzyme-linked immunosorbent assay (Aihajj & Farhana, 2022). Developing new binders based on established scaffolds such as antibodies is the goal of many pharmaceutical companies, while new scaffolds such as Miniproteins are in their relatively early stages, but have significant therapeutic potential (Crook, Nairn and Olson, 2020; Service, 2022).

Outside of therapeutics, protein binders are essential tools in research. Tools such as ELISA are also a staple resource of a research laboratory, while other techniques such as flow cytometry, western blotting and immunohistochemistry would be impossible without engineered protein binders.

New binders from existing scaffolds, and perhaps more excitingly the development of novel binder systems, open the possibility of improved therapeutics and completely novel approaches to disease treatment as well as the possibility of adding to existing research or the ability to investigate new avenues not currently accessible. New resources allow new approaches and continued progression.

1.1 Protein engineering to develop novel binder scaffolds

Nature is a common starting point in the process of developing proteins with new and desirable properties. The most abundant naturally occurring protein binder is the antibody. The human immune system's naïve antibody repertoire, contains at least 10^{12} different antibody binders (Briney *et al*., 2019), each specific to their own epitope. This immense diversity, highlights the antibody's potential as an ideal scaffold, which could be artificially manipulated to bind user determined target epitopes.

1.1.1 Antibody based scaffolds

Antibodies are widely used reagents both therapeutically and in the laboratory setting, with target specificity determined by the complementarity-determining regions found in the antibody V domain (Tsuchiya & Mizuguchi, 2016). The specificity of these domains and

therefore the antibody, is directed by the immunogen used. Depending on the application of the end antibody product, either full length native protein, protein fragments or linear peptide targets can be used as an immunogen. All have their positive and negative aspects, for example using a full-length native protein immunogen increases the chances of cross-reactive antibodies against proteins sharing degrees of homology with the immunogen, while using peptide immunogens often prevents the antibody binding the native protein in its conformational shape (Forsström *et al*., 2015).

Antibodies have a wide range of therapeutic uses. Each of these are thoroughly tested, requiring approval at multiple stages before clinical use, by approving bodies such as the FDA and The Medicines and Healthcare products Regulatory Agency.

Unfortunately, these regulatory bodies are not involved with reagent antibodies. Bradbury and Plückthun (2015) encapsulated the existing issues with reagent antibodies, all relating to the problem of poor result reproducibility.

The lack of reproducibility can be rooted to the production and the lack of information provided regarding the reagent antibodies. Traditional immunisation where the antigen of interest is injected into an animal (Köhler & Milstein, 1975), results in a polyclonal antibody response that cannot be replicated even in the same animal. Once a hybridoma has been generated, there is no guarantee the antibodies produced can be indefinitely mined due to the fragility of the cell line. Even if the source of the monoclonal is still viable, purchasing the same reagent as one described in the literature is not always possible (Vasilevsky *et al*., 2013).

Reproducibility when using reagent monoclonal antibodies is also hindered by the lack of sequence data for the proteins used and the inherent un-conserved binding mode, where two reagents specific to the same target could be binding different motifs. Bradbury and Plückthun (2015) address the issue of reproducibility by calling for sequence defined, recombinant binders to replace the traditional reagent monoclonal antibody.

1.1.1.2 Nanobodies as binder scaffolds

Sometimes referred to as the third generation of antibodies, nanobodies are comprised of a heavy chain and a variable domain. As with traditional antibodies, the variable domain possesses three peptide loops called the complementarity determining regions, responsible for antigen specificity and binding. The variability in length, especially of the third loop, impacts significantly on a nanobody's antigen binding (Yang & Shah, 2020).

The simplistic structure and modularity of nanobodies make them ideal for conjugations, for example in diagnostic imaging where nanobodies are tagged with positron-emitting nuclides for PET (Positron Emission Tomography). Capitalising on their small size and low immunogenicity, nanobodies are ideal for tumour infiltration for identification and drug delivery (Verhaar, Woodham and Pleoegh, 2020). Other applications include crystalisation chaperones (Rivera-Calzada *et al*., 2013) or as biosensors (Vercruysse *et al*., 2011).

Nanobody generation can be divided into two categories, immune and synthetic libraries. Immune libraries are based on the traditional antibody library approach, but instead immunising bactrian camels, llama, dromedary or alpaca, who naturally generate single domain antibody fragments (Harmsen & Haard, 2007). The immunisation process takes approximately 2 months, during which the animal is exposed to the target antigen (immunogen) up to eight times. To increase the probability of successfully generating an immune derived nanobody, multiple animals are immunised with the same antigen target. After the course of immunisation, blood is taken from the animal with the mRNA present converted to cDNA. Synthetic nanobody libraries do not require the immunisation of an animal, but instead utilises existing scaffold information. The library focuses on diversity in the complementarity determining regions, with the amino acids selected dependant on the antigen target (Muyldermans, 2020).

Nanobodies possess many advantageous qualities over traditional antibodies and previous antibody derived binders. With the possibility of synthetic libraries eliminating the need for animal immunisation and the sequence defined nature of the recombinant protein, nanobodies are a promising reagent capable of meeting the criteria outlined by Bradbury and Plückthun (2015). Even though the issue of reproducibility would be solved, nanobodies do have limitations. The first being, each new nanobody binder, made using immune libraries or a synthetic approach, would still need to undergo the process of library generation and screening. The second limitation arises when considering the nanobody target, most likely fully folded proteins, as nanobodies have poor binding capabilities for linear antigen targets (Pardon *et al*., 2014), the primary targets for the PRe-ART project.

1.1.2 Non-antibody based scaffolds

Alongside the antibody and antibody derived binders exist a wide range of non-antibody-based scaffolds, ranging in size, composition, and origin (Vazquez-Lombardi *et al*., 2015). Some examples are: Affibodies, based on staphylococcal protein A's receptor domain Z (Nord *et al*., 1997). Afflins: a binder based on a β-sheet of human γ-B-crystallin (Ebersbach

et al., 2007) and DARPins (Designed Ankyrin Repeat Proteins): an optimised version of the naturally occurring ankyrin repeat domain (Li, Mahajan and Tsai, 2006).

Combined, these non-antibody derived scaffolds have extensive applications and potential (Vazquez-Lombardi *et al*., 2015), but a common limitation exists. The requirement for a target in its conformational state, means these binders are not engineered for linear peptide target recognition and binding.

With this project's end goal of generating a protein library made of modular subunits, that are sequence-defined and capable of binding linear peptide targets in a predictable manner, the most obvious candidates for further enquiry were naturally occurring proteins or domains capable of linear peptide recognition. Reichen, Hansen and Plückthun (2014) reviewed the potential of small naturally occurring binding domains in generating a designed binder capable of meeting the requirements established by Bradbury and Plückthun (2015), while also being able to target linear peptides. Included binding domains were MHC-I and MHC-II, unfortunately not a viable domain due to their low stability and yield. Small adaptor proteins were also considered as they use peptide recognition modules to form precise temporal and spatial complexes. As these complexes are transient, only low to medium affinities are achieved, also making them unsuitable (Borowicz *et al*., 2020). Alongside several other binding domains, Reichen, Hansen and Plückthun (2014) also reviewed the suitability of repeat proteins, proteins possessing a target binding region consisting of individual modular subunits. These subunits would have to be identical to efficiently establish a modular and predictable binding region, meaning repeat proteins such as Beta-propeller proteins which can be formed from multiple different repeat subunit types (Chen, Chan and Wang, 2011) were unsuitable. Ideally the protein would also have a conserved binding mode with sequence defined modules, to assist in data reproducibility. This meant repeat proteins such as tetratricopeptide repeat proteins (D'Andrea and Regan, 2003) with their unconserved binding modes and HEAT proteins who's subunits have highly divergent repeat sequences were not appropriate scaffolds. The repeat protein capable of accommodating all these requirements was the armadillo repeat protein.

1.2 Naturally occurring armadillo repeat proteins

Armadillo repeat proteins were first discovered in 1987, as segment polarity genes, initially observed in Drosophila (Perrimon and Mahowald, 1987; Wieschaus and Riggleman, 1987). Armadillo repeat proteins are characterised by a repeating 42 amino acid motif, possessing three individual α-helices (H1, H2 and H3), folding in tandem to form a right-handed super-helical structure. Each of the armadillo subunits bind a single dipeptide unit of the peptide target in its extended conformation. The armadillo domain formed by these stacked subunits are flanked by specialised N and C terminal caps, which protect the protein's hydrophobic core (Figure 1.2). The armadillo repeat domain is present in an array of different proteins across the eukaryotic kingdom (Coates, 2003) indicating the domain's diversity and importance.

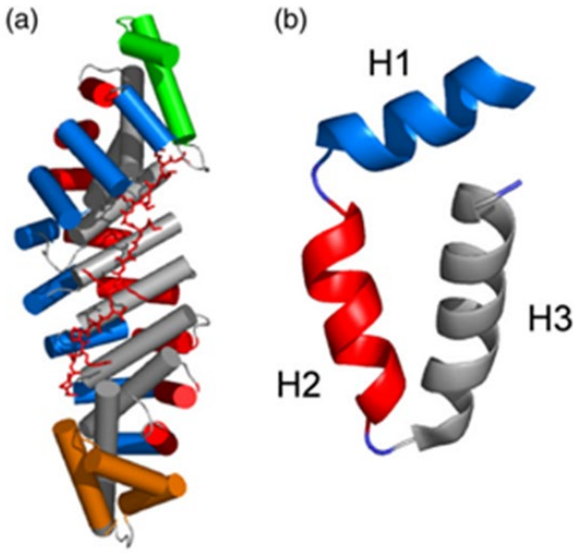

Figure 1.2 A) Structure of *S. cerevisiae* importin-α in complex with nucleoplasmin NLS (PDB 1EE5).

The α-helices of the protein are depicted as cylinders and the bound nucleoplasmin nuclear localisation signal as red sticks. Taken from (Conti & Kuriyan, 2000; Parmeggiani *et al*., 2008). The terminal caps are green and orange (N and C cap respectively), and the armadillo subunits coloured blue red and grey (H1, H2 and H3 respectively). **B) Isolated armadillo repeat subunit from importin-α (PDB 1EE5).** The three α-helices that form the armadillo repeat subunit are depicted as ribbons, with H1 coloured blue, H2 red and H3 grey. Taken from (Parmeggiani *et al*., 2008).

1.2.1 β-catenin

β-catenin possesses an armadillo repeat motif stretching over 12 individual armadillo subunits, capped by N and C terminal structures displaying the characteristic positively charged super helical structure (Figure 1.2.1) associated with armadillo repeat proteins (Huber, Nelson and Weis, 1997).

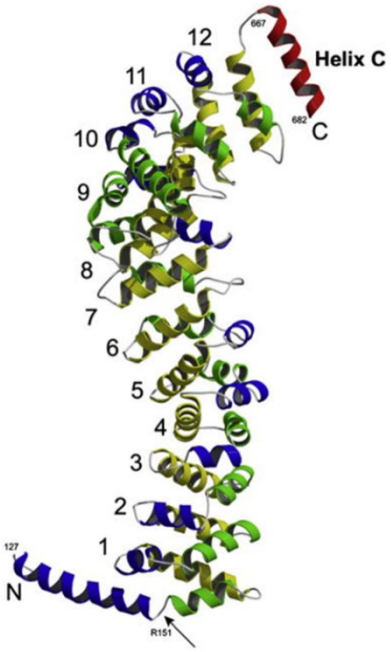

Figure 1.2.1 Crystal structure of Zebrafish β-Catenin.

Each of the 12 armadillo repeat subunits depicted as ribbons and numbered, with their constituent α-helices coloured: H1 blue, H2 green and H3 yellow. The C-terminal helix is coloured red and the N-terminal helix coloured blue and green. Taken from (Xing *et al.,* 2008).

Nuclear β-catenin is an essential component in Wnt signalling (MacDonald, Tamai and He, 2009). Its disassociation from the cytosolic destruction complex means that β-catenin is left un-ubiquitylated and can act alongside other nuclear factors as an activator of the TCF/LEF transcription repressor. This now activated transcription factor recruits an RNA polymerase, which produces mRNA for key cell cycle proteins. This central armadillo domain is the key region of the protein used to bind other Wnt signalling pathway proteins (Tolwinski and Wieschaus, 2004). Another nuclear role of β-catenin is its involvement in the intercentrosomal linker complex, and its vital role in mitotic centrosomal separation, with the armadillo domain identified as the key β-catenin motif responsible for β-catenin's centrosome localisation (Bahmanyar *et al*., 2008). The importance of β-catenin at the centrosome was demonstrated by Kaplan *et al* (2004) who showed that β-catenin depletion resulted in monopolar spindles, generated by failed centrosome separation. β-catenin's plethora of functions (Valenta, Hausmann and Basler, 2012), demonstrates its importance but also means it is often a factor in disease states (Shang, Hua and Hu, 2017; Kim *et al*., 2019).

1.2.2 Importin-α

Another important and well-documented armadillo repeat protein is importin-α. Importin-α possesses an armadillo repeat domain made of 10 individual armadillo subunits located in the middle of the protein flanked by an N-terminal importin-β binding site and a C-terminal region binding CAS/CSE1L (Figure 1.2.2). Each region of importin-α plays an important role in the protein's function.

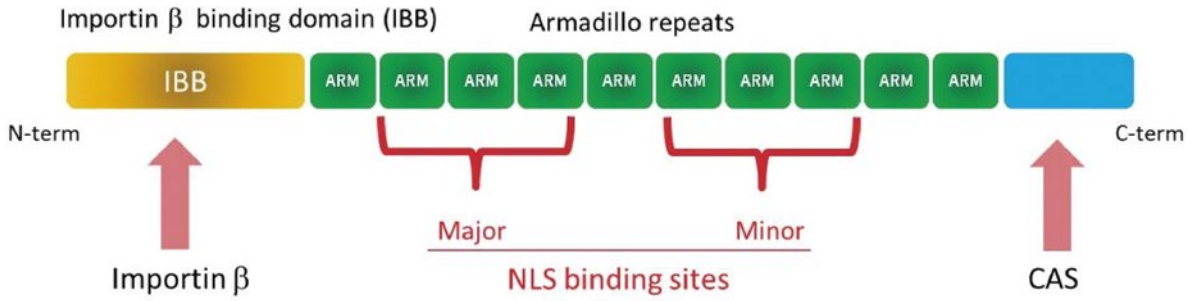

Figure 1.2.2 Diagrammatic representation of importin-α structure.

Taken from (Oka and Yoneda, 2018)

The central armadillo region of importin-α recognises the nuclear localisation signal (NLS) of nucleus bound proteins. Importin-α binds the NLS at two separate binding locations in the armadillo domain called the major (repeats 2-4) and minor (repeats 6-8) binding sites (Figure 1.2.2). Classical NLSs are typically 8-10 amino acids long with a high lysine and arginine content. These sequences recognised by the importin-α armadillo domain are either a contiguous amino acid stretch (monopartite) or organised into two separate clusters of lysine/arginine rich groups (bipartite) (Miyamoto, Yamada and Yoneda, 2016). Importin-β, another key carrier protein for nucleus bound proteins, is bound by the N-terminal importin-α domain (Oka and Yoneda, 2018) and once the nucleus bound protein has been successfully chaperoned into the nucleus, NLS-free importin-α is preferentially bound via its C-terminal region to CAS (Cellular Apoptosis Susceptibility gene) and is moved back to the cytoplasm (Kutay *et al*., 1997).

Just like β-catenin, importin-α has a range of functions (Oka and Yoneda, 2018). One such example occurs during mitosis, when the nuclear envelope is broken down making the typical chaperone protein function redundant. Work conducted by Ems-McClung, Zheng and Walczak (2004), showed importin-α can bind spindle assembly factors alongside importin-β, to inhibit their cytoplasmic function.

The in-depth sequence and structural knowledge of the β-catenin and importin-α families of armadillo repeat proteins were essential for the initial stages of engineering designed armadillo repeat proteins.

1.3 Engineering the Designed armadillo repeat protein

To successfully engineer a high quality peptide binder, required over a decade of work investigating the designed armadillo repeat protein, finding the balance between structure, binding and therefore functionality. The development of the protein could be subcategorised by structural components of the protein peptide complex; the terminal caps, the internal armadillo repeat subunits and peptide binding optimisation.

1.3.1 Engineering and optimisation of the terminal designed armadillo repeat protein caps

As with naturally occurring armadillo repeat proteins, specialised terminal regions called caps were necessary to protect the hydrophobic core composed of the internal armadillo repeat units. Parmeggiani *et al* (2008) used two methods to generate caps; the first modified the most suitable naturally occurring capping repeats, those belonging to the importin-α of *Saccharomyces cerevisiae*. The second approach used the consensus sequence generated for the internal repeat modules and engineered artificial caps from it by substituting the exposed hydrophobic residues within the internal repeat consensus sequence with hydrophilic ones. The yeast derived capping domains were coined Ny and Cy, while the artificial caps were named Na and Ca. By using multiple types of engineered internal repeats, they demonstrated that the highest levels of soluble protein expression were produced by proteins using a cap combination of Ca and Ny. Subsequently, while investigating the biophysical impacts of altering the number of internal repeats and attempting to produce crystal structures, Madhurantakam *et al* (2012) conducted molecular dynamic simulations that suggested five point mutations in the end cap repeats. These mutations resulted in second generation capping domains, Y_{II} (second-generation yeast N-terminal cap) and A_{II} (second-generation C terminal cap). Crystals of $Y_{II}M_3A_{II}$ (three internal repeats, flanked by second-generation N and C caps) revealed an N-terminus domain-swapping event (Figure 1.3.1 A). This was also found in $Y_{II}M_4A_{II}$ (four internal repeats, flanked by second-generation caps). Domain swapping is when two or more identical regions of a protein exchange part of their structure to generate a dimer or even larger oligomer (Rousseau, Schymkowitz and Itzhaki, 2013). Optimisation work used both $Y_{II}M_3A_{II}$ and

$Y_{II}M_4A_{II}$ to combat the protein dimerization, and therefore the domain swapping. The knowledge that solenoid protein caps have two distinct interfaces (one which faces the hydrophobic core and the other which interacts with the solvent) was implemented, with several mutations used to change solvent exposed hydrophobic residues to hydrophilic ones, generating the third generation Y_{III} and A_{III} caps. All second and third generation cap combinations were tested with three internal M-type repeats by ANS (anilino-napthalene-8-sulfonate) binding, thermal and guanidinium chloride denaturation. All four different cap combinations resulted in correctly folded, stable proteins, with differences occurring in thermal and guanidinium chloride denaturation transition midpoints. For both temperature and guanidiium chloride denaturation, Y_{III} increased stability while A_{III} caused a decrease in stability. As the Y_{III} cap was more beneficial to protein stability, crystal structures for both $Y_{III}M_3A_{II}$ and $Y_{III}M_3A_{III}$ were determined. Neither demonstrated domain-swapping events, meaning the redesign of the N-cap had been successful. Since A_{III} had reduced overall protein stability, A_{II} was preferred (Figure 1.3.1 B). Domain swapping was also investigated by Reichen *et al* (2014), using $Y_{III}M"_5A_{III}$ (the M-type described by Madhurantakam *et al* (2012), but with two mutations at position 34 and 36). Reichen *et al* (2014) showed the considerable impact a N-terminal His_6tag had upon domain swapping events. They showed His_6- $Y_{III}M"_5A_{III}/Ca^{2+}$ formed a stable dimer that prevented domain swapping, whereas in the absence of the His_6 tag, $Y_{III}M"_5A_{III}/Ca^{2+}$ still engaged in domain swapping. Since the super helical properties of the domain-swapped $Y_{III}M"_5A_{III}/Ca^{2+}$ and the dimeric His_6- $Y_{III}M"_5A_{III}/Ca^{2+}$ were identical, the huge impact that crystallisation conditions can have on designed armadillo repeat proteins was highlighted. Reichen *et al* (2014) similarly concluded that reverting to the A_{II} cap was beneficial for the progression of the designed armadillo repeat proteins as a way of improving interaction energies between the M-repeats and the cap.

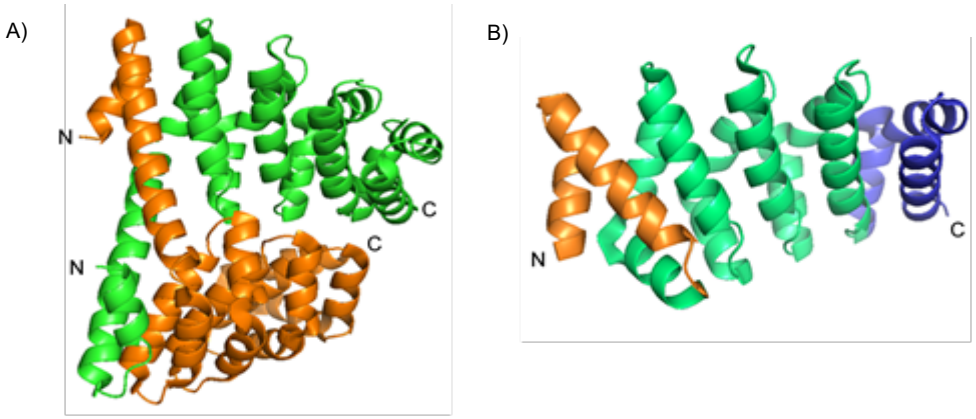

Figure 1.3.1 Crystal structures demonstrating N-cap optimisation

A) $Y_{II}M_3A_{II}$ dimer (PDB: 4DBA), with individual designed armadillo repeats coloured green or orange.

B) $Y_{III}M_3A_{II}$ structure (PDB: 4DB6), with the N-terminus Y_{III}-type cap (orange), three internal M-type repeats (green), and C-terminus A_{II}-type cap (blue).

Adapted from Madhurantakam *et al* (2012).

1.3.2 Engineering and optimisation of the internal repeats of the designed armadillo repeat protein

The internal repeats of an armadillo repeat protein play a significant role in the biophysical properties of the protein, as well as being the region responsible for peptide target recognition and binding. This meant during the development of the internal repeats, a fine balance was struck between overall structural requirements, while developing a scaffold whose target specificity could be changed.

1.3.2.1 Structural optimisation of the designed armadillo repeat protein by engineering internal repeats

The first designed armadillo repeat subunits were produced by Parmeggiani *et al* (2008) who started with the generation of consensus amino acid sequences by using the Swiss-Prot database. After removing unsuitable sequences, 319 hits remained. These 319 sequences gave an amino acid profile for 40 of the 42 amino acids in an armadillo repeat subunit, excluding the loop between H3 and the adjacent H1 helix. As the 319 sequences found using

the Swiss-Prot database originated from different sub-families of armadillo repeat protein, individual consensus sequences from each of the sub-family types were generated to avoid any sequence incompatibility. This method produced a consensus sequence for the armadillo repeat unit based on β-catenin (and plakoglobin) sequences (type T), importin-α sequences (type I) and a consensus sequence based on both families (type C). Protein expression characteristics, thermal and guanidinium chloride denaturation and ANS binding determined the internal repeat type-C (consensus based) had the most promising characteristics, even though hydrophobic core packing issues were indicated by strong ANS binding. This was addressed by including a range of mutations across sixteen different positions involved in hydrophobic core properties. One of the generated mutants, YM_4A (yeast C-terminus, four mutant internal repeats, artificial N-terminus), produced a monomeric protein capable of specific binding to the NLS target peptide (Parmeggiani *et al*., 2008). Optimisation work carried out by Madhurantakam *et al* (2012), investigated designed armadillo repeat proteins containing 3,4,5 or 6 M-type internal repeats, where the type M-internal repeat differed from the type-Cs used by Parmeggiani *et al* (2008) by three point mutations that were indicated by computational modelling to improve hydrophobic core properties. Each of the proteins were exposed to ANS which no longer caused an increase in fluorescence that would have been associated with an exposed hydrophobic core, allowing the inference that each protein had successfully folded into its native formation. Temperature and guanidinium chloride denaturation investigations indicated an increase in protein stability with increased numbers of internal repeats in a linear fashion, with temperature-induced folding reversible, meaning increasing the numbers of internal repeats was beneficial for protein stability.

1.3.2.2 Investigations into designed armadillo repeat target peptide binding

The only attempt at modifying the binding specificity of a designed armadillo repeat protein was conducted by Varadamsetty *et al* (2012). They identified a consensus sequence of ArmRP positions responsible for peptide binding by mapping the crystal structures of different ArmRPs bound to their respective target. All key binding residues belonged to helix three expect for position 4 (H1) and position 41 (no helix). The well-documented Asn at position 37, responsible for direct peptide backbone binding via two hydrogen bonds, was present in the consensus and therefore retained. The use of a consensus approach established positions 4, 30, 33, 36, 40 and 41 as randomisation targets. As position 4 was also involved in the formation of the hydrophobic core as well as peptide binding, the amino acids used to saturate position 4 were limited to Glu, His, lys, Arg, Ile, Gln and Thr. Varadamsetty *et al* (2012) also acknowledged the potential importance of positions 26, 29

and 30 as these are naturally occurring positions for binding across different ArmRP sub families. The decision to only include position 30 instead of 26 or 29 in the randomised library was based on position 30 identified most frequently as the residue out of 26 ,29 and 30 that formed target side chain interactions. The importance of positions 26 and 29 could not be ignored as previous work by Parmeggiani *et al* (2008) had identified an electrostatic repulsion between the consensus lysines at these positions. This meant these backbone positions were also investigated in the randomised library by either fixing both or either position for glutamine. The rationale for choosing each position as a saturation target was as follows: position 33 was of interest due to its spacing that could allow many different residues to take the consensus tryptophan's place, changing the amino acid at position 36 could affect the Asn37 backbone binding, position 40 was most frequently an amino acid with large side chains, so the impact of smaller amino acids would be of interest, position 41 was located between helix three and helix one of the adjacent repeat, interacting with the target peptide backbone via hydrogen bonds, and therefore also an interesting target. All of these positions of interest were saturated with all the natural amino acids except for glycine, proline and cysteine. Glycine and proline were excluded as most positions were situated in helix three and the exclusion of cysteine was to avoid any disulphide bridge formation. Randomisation of these positions had a detrimental impact on protein stability, which was countered by introducing flanking consensus modules to either side of the three repeats possessing saturation targets. Before the protein screening could occur, the optimal library had to be established. As the backbone changes at positions 26 and 29 resulted in four different variations and the incorporation of flanking consensus repeats had created another two library subsets, there were 8 different libraries to investigate. The deciding factor to determine the optimal library format was how the protein performed during purification.

The library containing the flanking consensus repeats that included glutamine at positions 26 and 29 had the highest proportion of monomeric soluble library members, so was used. The library was screened via Ribosome display with the 13mer peptide, neurotensin (QLYENKPRRPYIL), the designated target. The highest enrichment was achieved by alternating binder exposure to synthetic neurotensin and phage λ protein D-neurotensin fusion protein. Clones were made and those demonstrating the strongest neurotensin binding via ELISA analysis were sequenced. 29 of the 30 clones sequenced had an identical sequence with the anomalous clone possessing a single amino acid change from tyrosine to histidine at position 116, corresponding to position 30 of a helix 3 within the protein. Both protein types were purified using immobilized metal-ion affinity chromatography with peptide recognition tested by ELISA showing the protein without the 116 change was the superior binder. The binder specificity was then tested by exposing the designed armadillo repeat to 10 different peptide targets, to which the protein only bound neurotensin in a significant manner. As selection had prioritised specificity over affinity, only a moderate affinity of 7μM

at 4 °C was observed. The optimisation of the designed armadillo repeat protein in order to generate high affinity binders was conducted by Hansen *et al* (2016) who investigated the binding of $Y_{III}M_XA_{II}$ to a $(KR)_5$ peptide where X was a varying number of internal repeats. These experiments differed from those of Madhurantakam *et al* (2012), as the focus was now on the impact differing numbers of repeats had on target specificity and binding as opposed to the impacts on the designed armadillo's structural properties. The use of a $(KR)_5$ peptide target was due to the original type-C consensus sequence generated by Parmeggiani *et al* (2008). This biased the protein's binding preference towards a peptide target mimicking a NLS, which typically contains a significant amount of K and R residues (Lu *et al*., 2021). Hansen *et al* (2016) demonstrated an increased affinity between the protein and peptide target as the number of M repeats and number of KR peptide units increased. Designed armadillo repeat proteins with M repeats ranging from 3-7 were tested, but K_d values could not be determined for all, since the designed armadillo repeat proteins containing six and seven M repeats resulted in peptide binding so tight, that method sensitivities prevented the measurement of K_d values. Instead the K_d for these proteins was approximated by extrapolating the data obtained for M3-5 proteins (Figure 1.3.2.2.1).

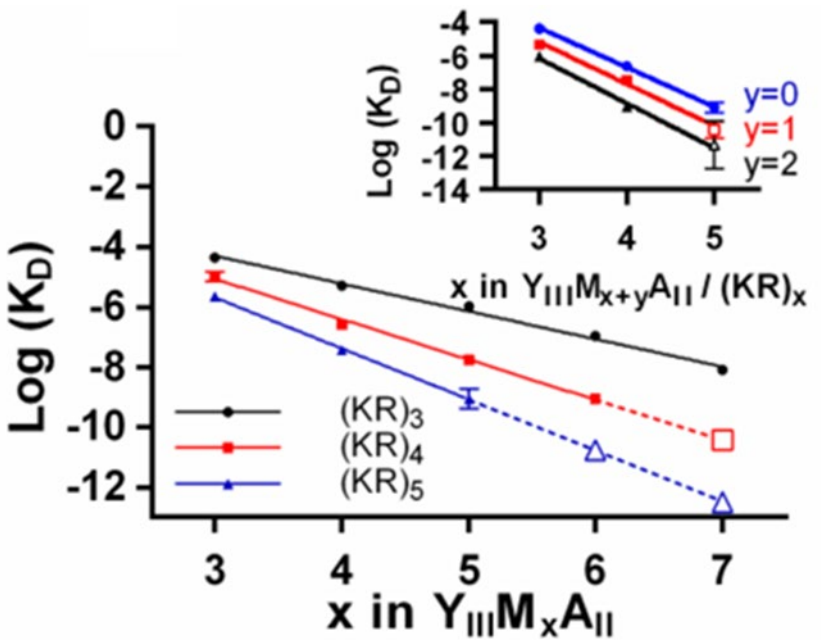

Figure 1.3.2.2.1 Impact of the number of armadillo repeats on K_d when binding different KR length peptides.

A linear relationship was observed between number of internal "M" repeats and length of target peptide with more repeats demonstrating a lower K_d. K_d values for M6 and M7 were extrapolated and shown with a dotted line. Inset: Impact on K_d with a step wise increase in the number of M repeats and peptide KR repeats, demonstrating a linear regression. Taken from (Hansen *et al*., 2016).

This meant the designed armadillo repeat protein $Y_{III}M_5A_{II}$ was the optimal binder as its K_d was accurately measured to be 1.1 ± 0.8nM when binding a $(KR)_5$ peptide. The linear relationship between the impact on K_d as the number of internal repeats and KR subunits changed, demonstrated a consistent and modular effect of adding a M repeat. The binding energy of an M repeat and a KR unit was calculated to be −14.4 ± 0.7 kJ/mol thanks to alanine scanning of the individual arginine and lysine binding pockets. The crystal structure of $Y_{III}M_5A_{II}$ was determined, with the bound KR_5 peptide in the typical antiparallel orientation characteristic of armadillo repeat proteins (Figure 1.3.2.2.2).

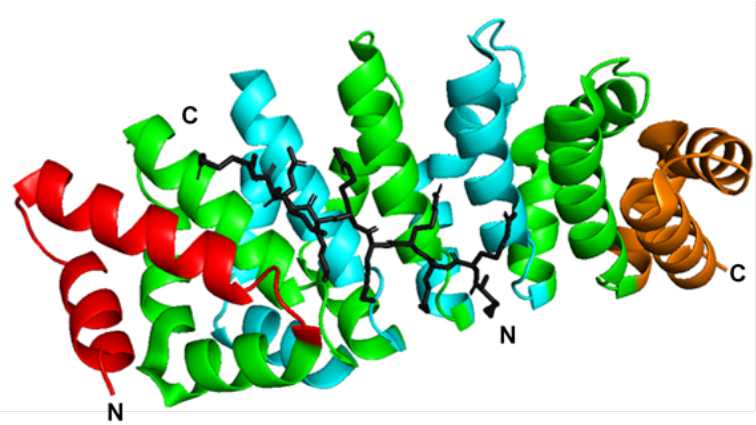

Figure 1.3.2.2.2 Structure of $Y_{III}M_5A_{II}$−$(KR)_5$ (PDB ID: 5aei).

The protein was depicted as ribbons with the N-terminal cap coloured red and the C-terminal cap orange. The five internal M repeats, each possessing three α-helices, were coloured alternating green and cyan. The $(KR)_5$ peptide, represented as sticks (black), was seen bound to $Y_{III}M_5A_{II}$ in an antiparallel fashion. Adapted from (Hansen *et al*., 2016).

With the successful engineering of the optimal designed armadillo, $Y_{III}M_5A_{II}$, Ernst *et al* (2020), investigated the binding register of the target peptide to the protein. An inherent problem of repeat proteins, especially when binding repetitive sequences, as with $Y_{III}M_5A_{II}$ and $(KR)_5$, is the ability of the peptide to bind in several different possible registers up and down the protein as described in Figure 1.3.2.2.3.

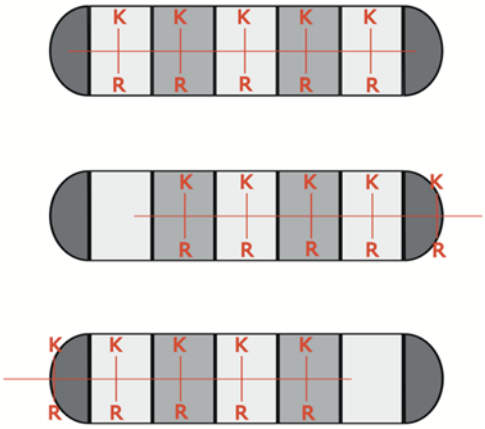

Figure 1.3.2.2.3 Diagrammatic representation of different potential binding registers of the KR5 peptide with $Y_{III}M_5A_{II}$.

The KR_5 peptide (orange) is shown to have the capability of binding to $Y_{III}M_5A_{II}$ in three different orientations. $Y_{III}M_5A_{II}$ is coloured in alternating grey tones to help visualise the different binding register, but all internal repeats are identical. Taken from (Ernst *et al.,* 2020)

The peptide having a range of binding options results in different amino acids occupying different binding pockets, making the selection for sequence specific amino acid binding incredibly difficult to accomplish. To combat the movement of the peptide, Ernst *et al* (2020) countered the repetitiveness of the protein-peptide interaction by grafting a hydrophobic binding pocket originating from β-catenin to the N-terminus of the protein. By introducing a hydrophobic binding domain into a protein associated with binding polar residues, and introducing a complementary hydrophobic region into the peptide target based on consensus ligand sequences to the β-catenin pocket (LSF), Ernst *et al* (2020) hypothesised the peptide would be forced into a single binding register. The N-terminal hydrophobic domain was coined Lock 1. Several rounds of hydrophobic domain optimisation resulted in a second generation domain coined Lock 2 which bound its target sequence (AKITW) with greater affinity than that of Lock 1 and LSF. To determine if the Lock 2 had forced a single binding register, FRET (Fluorescence Resonance Energy transfer) was used with the $Y_{III}M5A_{II}$:$(KR)_5$ complex compared with the Lock 2:KRKRKAKITW complex. The two histograms produced shared a similar peak, but the narrower distribution of the Lock 2:KRKRKAKITW complex meant a higher proportion of the Lock 2:KRKRKAKITW complexes were in the correct binding formation, meaning that Lock 2 had been successful.

1.4 Saturation mutagenesis techniques to generate randomised designed armadillo repeat protein libraries

With the stable $Y_{III}M_5A_{II}$ designed armadillo repeat protein established, possessing a LSF lock as of the project's start, meant the next stage was to introduce changes to binding specificities of the protein.

There now exists a multitude of techniques available in the molecular biology tool kit, for the randomisation of a target nucleic acid, and therefore the protein encoded. This directed evolution can be broadly defined into two categories, random and site-directed. Both approaches can be used in the now standard combinatorial approach where typically a DNA library of variants is made, the corresponding proteins are expressed and screened to identify new proteins of interest.

1.4.1 Introducing sequence changes at random

Random mutagenesis as the name suggests is a category of techniques that introduce changes to the original sequence in both random places and by introducing random changes. These changes can be introduced either by recombination, where there are no base changes but instead the sequence order is rearranged, or by the introduction of point mutations, insertions and or deletions.

1.4.1.1 Randomisation by recombination

The biggest positive of recombination induced randomisation, is the ability to join advantageous protein mutations, while avoiding undesirable ones.

Examples of recombination randomisation techniques are DNA shuffling, random chimeragenesis on transient templates (RACHITT), the staggered extension process and incremental truncation for the creation of hybrid enzymes (ITCHY) (Neylon, 2004). DNA shuffling takes the parental DNA sequences and cleaves randomly using a DNAse or endonuclease (Stemmer, 1993) creating small fragments which are then forced to undergo self-priming events, to reconstitute a full-length product (Figure 1.4.1.1). This product is then amplified via PCR. RACHITT can generate cross over events at an increased number of random locations in comparison to other recombinant techniques by fragmenting all but one of the parental DNA sequence strands. The fragments are then reassembled using the one

non-fragmented strand as the transient template. Fragment sections not involved in matches are removed and the remaining fragments are ligated to form a new template strand (Figure 1.4.1.1) (Coco, 2003).

Unlike the random placement of fragments seen in DNA shuffling and RACHITT, the staggered extension process (StEP) builds a full-length randomised product by addition of sequence sections at the terminal end of the growing sequence. Random binding to a parental gene sequence coupled with a limited extension, introduces a fragment in a sequential process elongating the recombinant sequence (Figure 1.4.1.1) (Zhao *et al*., 1998).

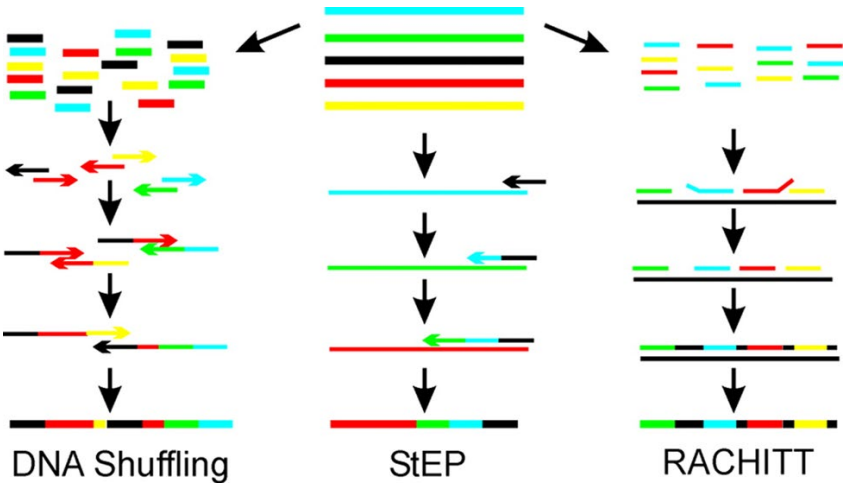

Figure 1.4.1.1 Diagrammatic representation of DNA shuffling, StEP and RACHITT.

All three methods use a pool of parental DNA. Both DNA shuffling and RACHITT start by fragmenting their parental DNA pool sequences; DNA shuffling fragments all parental sequences while RACHITT leaves one intact. DNA shuffling uses self-priming events between the fragments to generate a full length product, while RACHITT produces randomisation by using the intact parental sequence as an annealing platform for a mixture of fragments originating from different parental sequences. StEP uses random annealing of the growing cassette to a parental sequence acting as template, coupled with controlled cycling conditions to introduce sections of sequence in a step wise fashion.

Taken from (Neylon, 2004).

The above-mentioned methods of recombination induced randomisation, require a degree of homology between the parental genes, a requirement not necessary for ITCHY. Ostermeier (1999) originally generated novel sequences by restricting parental library DNA sequences using exonuclease III and SI restriction and ligating 5’ and 3’ restricted entries, removing the

necessity for homology at the cross over site, but at the same time increasing the possibility of nonsense sequences, due to the randomness of the connection. The iterative part of the truncation was originally performed by timed restriction reactions, but as these were difficult to control the methodology was optimised by the introduction of αS-dNTPs, coining THIO-ITCHY (Lutz, Ostermeier and Benkovic, 2001). Spiking low levels of αS-dNTPs, dNTPS capable of preventing restriction, meant a base-by-base restriction was theoretically possible, meaning upon ligation of the restriction products, all possible combinations would be achieved. This library could then undergo other types of recombinant based randomisation or be screened for viable proteins of interest.

Recombinant randomisation generates diversity without directly changing any sequence bases, but instead relies on the rearrangement of parental sequences to generate novel cassettes. This is complimented by directed evolution methods that induce randomisation by directly changing the base sequence via insertions/deletions and point mutations. This can be a targeted or random process.

1.4.1.2 Randomisation via insertions, deletions and point mutations

A popular technique for introducing random point mutations within a sequence is error-prone PCR. Significant amounts of time and effort are invested into making the process of PCR as consistent as possible, from the optimisation of reagent concentrations to the engineering of proof-reading polymerases. But the innate error rate of PCR can also be advantageous when increased. Typical error rates of a standard Taq DNA polymerase PCR reaction (the number of mismatched bases) are not sufficient for sequence randomisation, so artificially increasing the error rate by commercial kits or manually incorporating Mn2+ (Cadwell, 1994) or biasing specific dNTPs can make the error rate appropriate for use in randomisation.

The simplicity of error prone PCR makes it a valuable technique, but one which has an innate flaw. A key consideration when using error prone PCR is bias. Bias can be introduced by the natural characteristics of the polymerase having preference towards specific mutations at specific positions, the degeneracy of the genetic code and the last being PCR bias. The exponential growth of a sequence caused by PCR, means an early mutation during the PCR reaction will be amplified throughout the entire PCR reaction, meaning that mutation will be overrepresented within the library.

Regarding insertions and deletions, error prone PCR can introduce these types of mutations, but at a much lower frequency than point mutations. A simplistic method for introducing deletions was developed by Pikkemaat and Janssen (2002), where they used BAL-31 nuclease to delete DNA from the terminal ends of a template sequence, which was then re-

ligated together. This simplistic approach was prone to missense products due to ligated products being out of frame. On the other end of the complexity scale is Random Insertion/Deletion (RID) mutagenesis designed by Murakami *et al*, (2002), a multi-step process capable of deletions and then insertions of different lengths.

Randomisation based on untargeted directed evolution is the ideal starting point for protein studies where no previous knowledge of the protein is available. In protein studies where existing data is available, specific changes to relevant key positions can be introduced, allowing a direct investigation between randomisation of positions and phenotypic change in the protein. The most efficient randomisation technique for introducing relevant changes to investigate specific protein qualities, for example the binding of a designed armadillo repeat pocket, is saturation mutagenesis, the process of substitution-based randomisation. This category of randomisation techniques can be subdivided into degenerate and non-degenerate saturation. Depending on the library randomisation requirements, a mix of degenerate and non-degenerate techniques can be used.

As demonstrated by Varadamsetty *et al* (2012), successfully generating designed armadillo repeat protein binders with different specificities required the simultaneous saturation of several key binding residues in the same randomised library. Position 4 only required saturation with seven different amino acids which was achieved using three degenerate oligonucleotides, while the remaining five target positions required saturation with large amino acid subsets (17 amino acids per position) (Varadamsetty *et al* 2012), a magnitude of saturation not possible using degenerate saturation mutagenesis techniques.

1.4.2 DNA degeneracy in randomised DNA libraries

Fully degenerate saturation mutagenesis, uses all 64 different codons (via NNN) to represent the 20 natural amino acids. If a single position in the DNA sequence was being targeted, full degeneracy using NNN would result in 64 different sequence outputs. Increasing the number of positions saturated using NNN causes a rapid exponential growth of required sequences to encode all possibilities making fully degenerate saturation impractical for multiple positions. The degeneracy of NNN randomisation is reduced by changing the third base from N to K or S (K representing T or G and S representing G or C). This reduces the degeneracy encoding all 20 natural amino acids from 64 codons to 32. Kille *et al's* (2012) use of degenerate primers to generate libraries successfully reduced the genetic code redundancy to 22 codons. By using two degeneracy containing primers, possessing either NDT or VHG alongside a TGG specific primer, all 20 natural amino acids could be encoded, removing all nonsense codons and all redundancy except for two codons each for leucine and valine.

Being close to complete degeneracy, makes the 22c-trick significantly more efficient in comparison to positional randomisation using NNN or NNK (or variants of NNK), but still does not compare to fully non-degenerate saturation mutagenesis.

1.4.3 Non-degenerate saturation mutagenesis techniques

The need to saturate five positions, each with 17 different amino acids, within the first randomised designed armadillo repeat protein DNA library, lead Varadamsetty *et al* (2012) to use Trinucleotide phosphoramidites (Virnekäs *et al*., 1994), a nondegenerate saturation mutagenesis technique. Using non-degenerate saturation mutagenesis catered for the simultaneous randomisation of multiple key residues within the same DNA library, making the process of identifying a new binder more efficient. As demonstrated by Varadamsetty *et al* (2012) the choice of saturation mutagenesis technique is library-dependent with the power of non-degenerate saturation mutagenesis invaluable for efficient library production. Accordingly, the following sections will address alternative methodologies for achieving non-degenerate saturation mutagenesis.

1.4.3.1 Trinucleotide phosphoramidites (TRIM)

The first non-degenerate saturation mutagenesis was called TRIM. Extension during chemical DNA synthesis is normally achieved by adding a single base at a time in individual reactions, while TRIM is capable of adding three bases at once. This allowed the saturation of a codon by adding the predetermined trinucleotide phosphoramidites (TPs). The original method of generating the TPs can now be simplified by purchasing commercially produced codons (Virnekäs *et al*., 1994). Based on work by Holm, (1986) less common codons for each amino acid could be avoided meaning that with careful consideration, 20 trinucleotides, each representing one amino acid, could be generated from either T or G and one of seven different dinucleotides (AT, CT, GT, TT, AG, GG, TG). Conditions for coupling of the TPs were similar to those of mononucleotide couplings, with each of the 20 TPs used in individual automated oligonucleotide synthesis experiments originally using an Applied Biosystmes DNA synthesizer 380B. Coupling yields for the TPs were optimised by introducing increased coupling times and a second round of coupling for each TP, with yields of 96-98.5% achieved (mononucleotide yields being 98-99.5%) (Virnekäs *et al*., 1994). The application of TRIM in saturation mutagenesis relies on the ability of the trinucleotide phosphoramidites to be mixed in a user-defined pool. Virnekäs *et al* (1994) attempted this with a subset of amino acids encoding the eight hydrophobic residues. Each of the TPs, one for each amino acid were

combined in equimolar quantities and used to generate TP-incorporated oligonucleotides. These oligonucleotides were then used as primers to amplify a 2H-10 antibody fragment. Although all the expected trinucleotides had been incorporated into the primers and were observed via Sanger sequencing, they were not present in equal amounts, most likely due to differing coupling efficiencies of the TPs (Virnekäs *et al*., 1994). Though resulting in a biased library, TRIM provided the foundation for using specific codons to represent specific amino acids, and therefore the removal of DNA degeneracy.

1.4.3.2 MAX randomisation

MAX randomisation (Hughes *et al*., 2003) eliminates the degeneracy of the DNA code by using a one-to-one codon to amino acid ratio. MAX randomisation uses a traditionally randomised template strand, where positions of interest are randomised using NNN. This template strand acts as a docking station for the MAX selection oligonucleotides. MAX selection oligonucleotides are 9bp long sequences, containing a 6bp invariant region that acts as a localisation sequence to ensure accurate annealing to the template strand and the 3bp MAX codon, coined from the typical use of the maximally expressed codon for each amino acid for downstream protein expression. Figure 1.4.3.2 shows three positions of interest being saturated with three different MAX selection oligonucleotide pools (red, blue and green). The MAX selection oligonucleotides anneal to their complementary sequence found in the NNN template and are flanked by constant oligonucleotides which share 6bp of sequence complementarity with the 5’ and 3’ end of the template strand. The individual MAX oligonucleotides are ligated together to form a MAX-containing strand. The incorporation of flanking oligonucleotides cater for specific PCR amplification of the MAX-containing strand, meaning the NNN template is not amplified by PCR (Hughes *et al*., 2003). The inclusion of restriction sites in the flanking oligonucleotide sequence allow for the incorporation of the randomised cassette into parental genes.

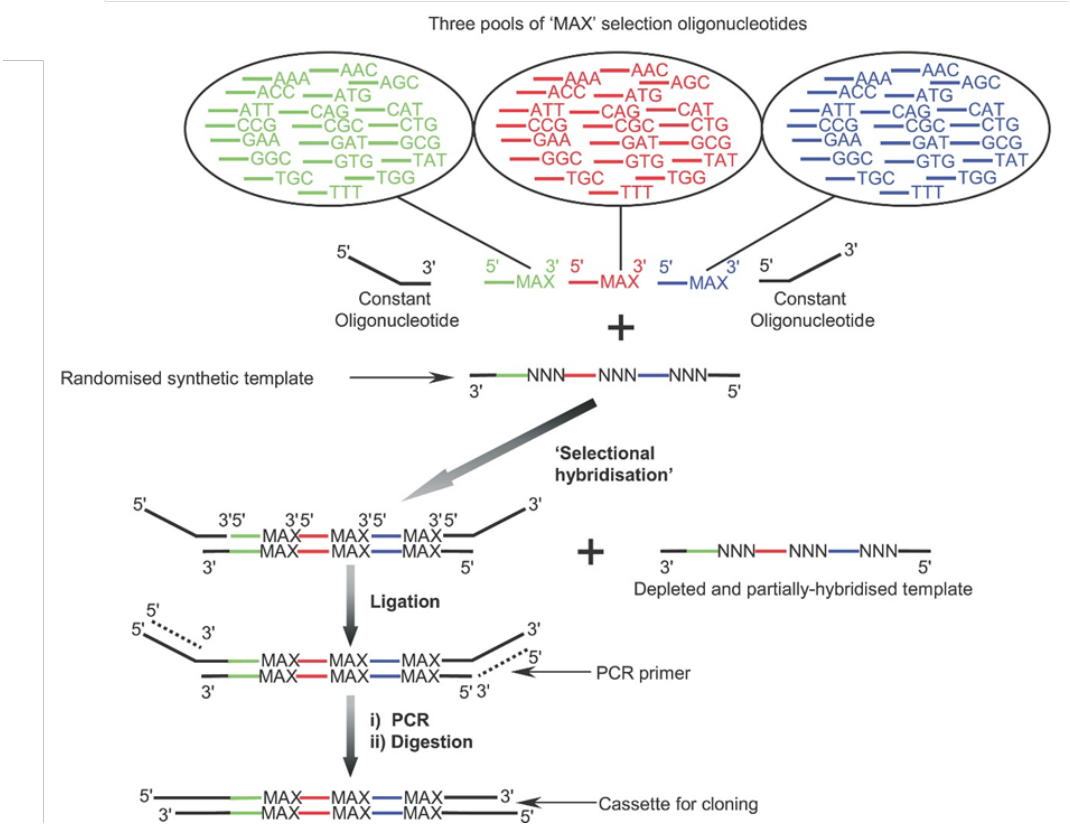

Figure 1.4.3.2 Schematic representation of MAX randomisation.

MAX selection oligonucleotide pools are made for each positional target by mixing the MAX selection oligonucleotides encoding the desired amino acids (green, red and blue). The MAX selection oligonucleotides bind to their complementary sequence found in the traditionally randomised NNN template strand. Non-randomised constant oligonucleotides also hybridise to the 5' and 3' ends of the template strand. Ligation of the MAX selection oligonucleotides and the flanking constant oligonucleotides generates a MAX codon containing DNA strand, which undergoes PCR amplification made possible by the constant oligonucleotides providing a location for primer annealing (dashed lines). The randomised cassette library can then be digested using restriction sites present in the constant oligonucleotide sequences, ready for incorporation into a vector and cloning. Taken from (Hughes *et al*., 2003).

MAX randomisation allows the non-degenerate saturation of non-contiguous targets. The requirement for the 6bp invariant region of the MAX selection oligonucleotide to ensure correct localisation of the MAX codon with the NNN of the template strand, means that saturating more than two adjacent positions is not possible using MAX randomisation.

1.4.3.3 ProxiMAX randomisation

To address the need for a non-degenerate saturation mutagenesis technique that was capable of saturating contiguous codons, Ashraf *et al* (2013) developed ProxiMAX randomisation. ProxiMAX randomisation uses two distinct types of double-stranded oligonucleotide, the donor and acceptor. A mixture of donor molecules each possessing a MAX codon at the 3' terminal is made by hybridisation, with the MAX codon being user-defined. Using a 1:1 ratio of donor molecules to amino acid removes DNA degeneracy.

The pool of donor molecules are mixed with the acceptor DNA and ligated to only the 5' end of the acceptor, ensured by the acceptor DNA's 5' phosphate. The donor-acceptor construct is then amplified via PCR and purified. The construct is then subjected to type IIS restriction via *MlyI* (5'…GAGTC$(N)_5$…3', 3'…CTCAG$(N)_5$…5'). The *MlyI* restriction occurs upstream of the MAX codon within the donor DNA, meaning it remains a part of the acceptor sequence after restriction. This elongated acceptor can then be used for the next round of ProxiMAX (Figure 1.4.3.3). Alternating between different donor sequence sets, allows specific primer annealing and PCR amplification during each ProxiMAX cycle, preventing the amplification of any donor-acceptor complexes generated using the previous round's donor. ProxiMAX has been documented to be successful in saturating 11 contiguous codons in a CDR3 domain of an antibody variable heavy chain, by engineering two smaller randomised constructs, and ligating the acceptors sequences together (Ashraf *et al*., 2013). ProxiMAX was licensed to Isogenica in 2010, marketing the technique as Colibra (Isogenica, 2015). With the technique automated, a region of 23 randomised codons was successfully engineered (Frigotto *et al*., 2015).

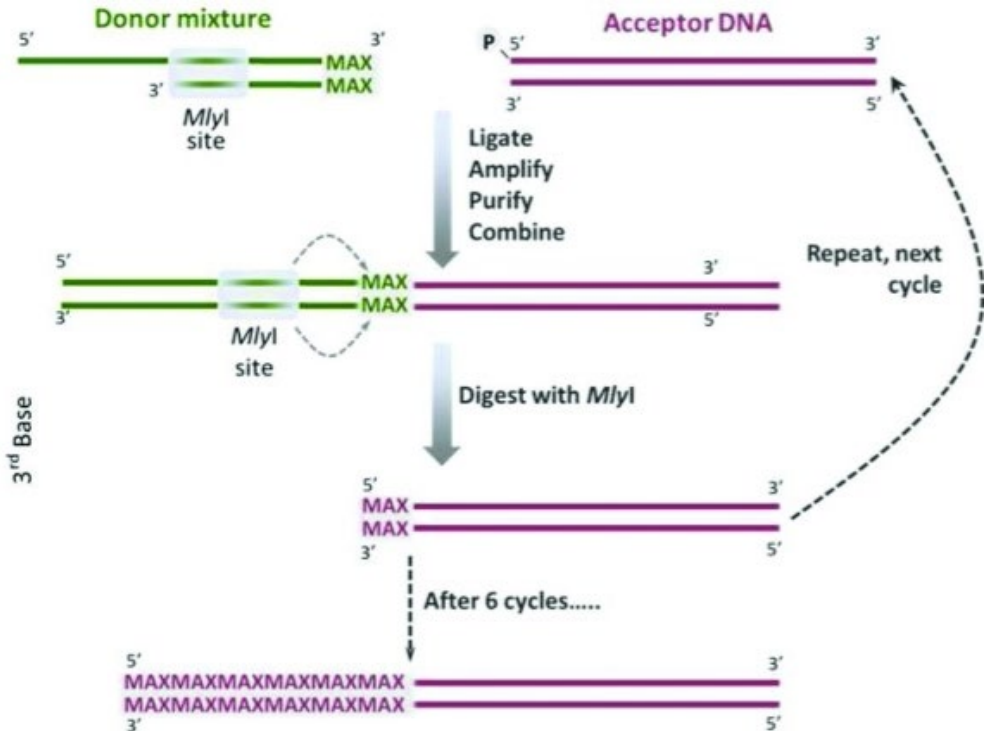

Figure 1.4.3.3 Diagrammatic representation of ProxiMAX randomisation.

ProxiMAX randomisation uses iterative rounds of ligation, PCR amplification, purification and *MlyI* restriction to generate stretches of contiguous randomised codons. A double stranded donor sequence possessing a MAX codon at its 3' termini, is ligated to a acceptor sequence, with the newly formed construct amplified via PCR. The PCR product is then purified and restricted using *MlyI*, exposing the MAX codon now belonging to the acceptor sequence. This elongated acceptor sequence can be used in the next round of ProxiMAX in an iterative cycle. Illustrated is a contiguous region containing six randomised codons produced from six rounds of ProxiMAX randomisation. Taken from (Ashraf *et al*., 2013).

1.4.3.4 Slonomics

The use of restriction enzymes to donate randomised codons to another sequence is not limited to ProxiMAX randomisation (Ashraf *et al*., 2013). A related method of transferring a randomised codon between sequences by type IIs restriction is also seen in Slonomics (Van den Brulle *et al*., 2008). Slonomics uses two different oligonucleotides coined splinkers and anchors, both forming DNA hairpins that provide structural stability to the molecules. There are several differences between these two molecules both structurally and within the DNA sequence. The splinker possesses a three base single stranded overhang, each representing a single codon, making 64 different splinker molecules. The anchor has a region of seven bases at the 5' end, three of which form a single strand overhang. This overhang is used for complementary annealing to one of the 64 splinkers. The triplet adjacent to the overhang in the anchor sequence, contains the NNN used for the splinker's next round of elongation, meaning 64^2 anchors are required (Van den Brulle *et al*., 2008). Each anchor possesses a biotin modification in its loop allowing for immobilisation to a streptavidin-coated 96 well plate. The anchor sequence contains the recognition site for the type IIs restriction enzyme *Eam1104I* (5'-…CTCTTC$(N)_1$….3', 3'…GAGAAG$(N)_4$…5'). Digestion with this enzyme releases the newly elongated splinker (the original complementary overhang triplet and the newly donated codon from the anchor cleavage which acts as the splinker's new 5' overhang). The plate is washed and the supernatant containing the elongated splinker is used in the next round (Figure 1.4.3.4). This process is repeated up to a length of 18bp, with this construction coined the elongation block (Van den Brulle *et al*., 2008).

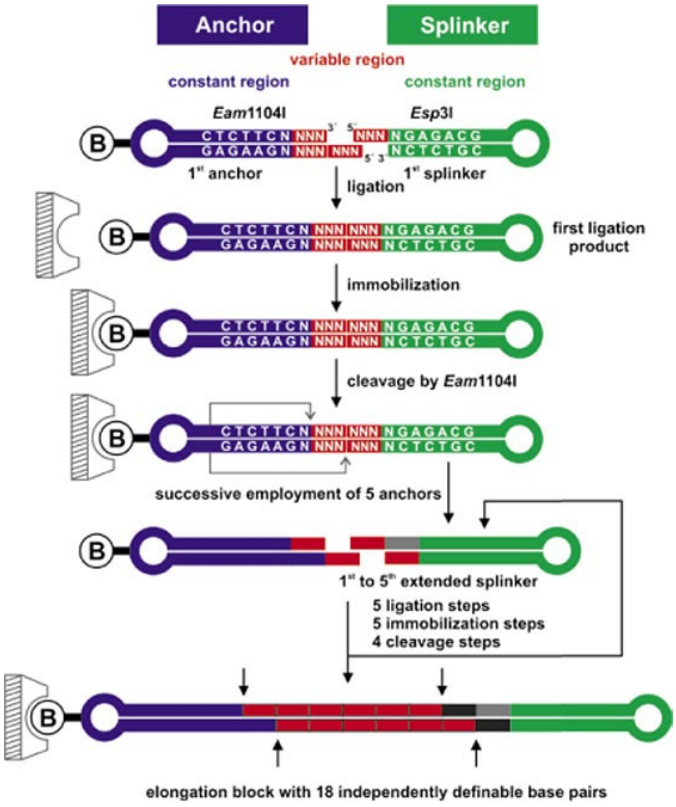

Figure 1.4.3.4 Diagrammatic representation of Slonomics

Complementary anchor and splinkers are ligated via sticky end ligation forming a single construct. A biotin moiety in the anchor accommodates attachment and immobilisation to a streptavidin-coated well. *Eam1104I* restriction releases the splinker and the newly incorporated codon into the supernatant, allowing its use in subsequent elongation steps. Taken from (Van den Brulle *et al*., 2008). Note that in this figure, NNN refers to a specific, rather than a degenerate codon.

Within the splinker sequence is the recognition site for *Esp3I* (5'…CGTCTC$(N)_1$…3', 3'…GCAGAG$(N)_5$…5') which generates a 4bp overhang. This means each elongation block can be restricted using *Eam1104I* and *Esp3I* to generate a 5' 3bp overhang and a 3' 4bp overhang. The difference in overhang length allows for highly selective assembly of elongation blocks in a pair wise fashion. This process can be repeated to assemble even longer regions of randomisation (Van den Brulle *et al*., 2008). Sloning then went on to develop SlonoMAX™ , a commercialised application of Slonomics ™ ,capable of encoding protein libraries with user defined amino acids at specific positions.

1.4.3.5 DC and MDC-Analyzers

The MDC-Analyzer is a computational approach to saturation mutagenesis, designed as an extension to the DC analyser (Tang *et al*., 2012), but now targeting contiguous sites, while the DC Analzyer was more suited to single codon mutagenesis.

The tools use optimised degenerate primer sets allowing for the encoding of contiguous amino acids with minimal redundancy. Computational primer predictions were conducted by Tang *et al* (2014) for the randomisation of three contiguous codons, T134,P135,F136 of A. *radiobacter* AD1 halohydrin dehalogenase (hheC), resulting in the primer sequences ATTACCTCTGCAXXXXXXXX XGGGCCT-3′, where XXXXXXXXX represented KGCWYGWYG, ATCWYGWYG, KGCKWCWYG, DKCCAGWYG, and ATCKWCWYG. These primer oligonucleotides were then synthesised and mixed to a ratio of 32:16:32:24:16 to accommodate for the frequency of each amino acid encoded. Mixing the degenerate primers simplified the experimental protocol, by only needing to perform a single PCR amplification, but increases the possibility of primer dimers or bigger oligomers forming. This could impact upon PCR amplification of specific sequences thus skewing the amino acid frequencies. The work demonstrated by Tang *et al* (2014) is based on the mutagenesis of three contiguous codons and required a primer length of 27bp. This included 9 degenerate bp flanked upstream by 12bp and downstream by 6bp. An increase in the number of codon targets would increase the length and complexity of the primers, increasing the potential complications when saturating a repetitive sequence for example, that of a designed armadillo repeat protein.

1.4.4 Key considerations when choosing a saturation mutagenesis technique

When deciding which saturation mutagenesis technique to use when engineering a randomised DNA library, there are several key considerations to deliberate. The majority of these considerations are biological, while others are financial (Acevedo-Rocha, Reetz and Nov, 2015). The biological considerations can be said to revolve around sequence space. A DNA library's sequence space is the encoding space required to provide a single DNA sequence for every possible randomisation outcome. As sequence space is finite, it can become a limiting factor for DNA library diversity (the number of unique DNA sequences present in the randomised library). This means a library's diversity is dependent on efficient use of the sequence space. Figure 1.4.4 A, B and C summarise the impact of changing the level of degeneracy on diversity, encoded bias and encoded truncation.

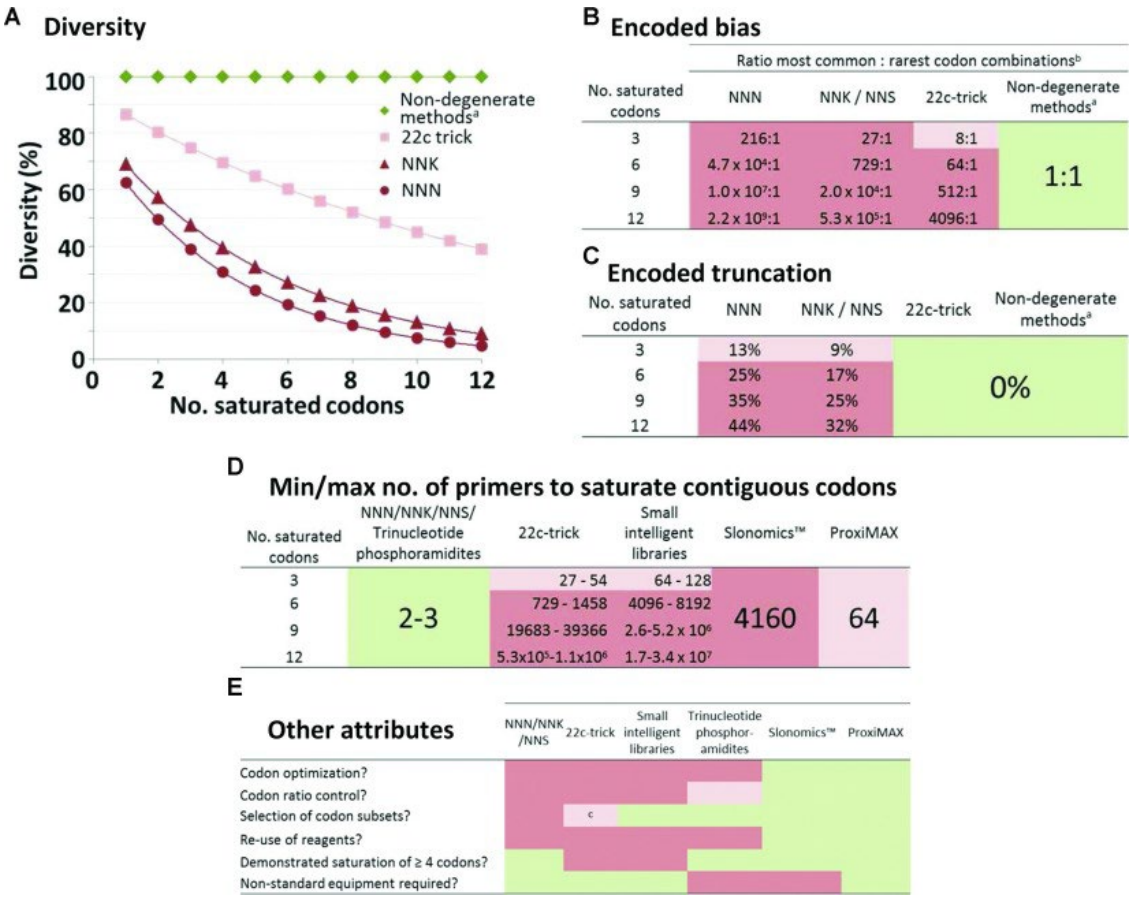

Figure 1.4.4 Performance comparison of different saturation mutagenesis techniques

(A) Graphical representation of potential diversity generated by different saturation mutagenesis techniques, calculated using the formula d=1/(NΣkpk2) (Makowski and Soares, 2003). (**B**) Comparison of theoretical gene ratios demonstrating codon bias between the most degenerate amino acids (leucine/arginine/serine having 6 codons each) and the least degenerate (methionine and tryptophan having 1 codon each) for different numbers of saturated targets. (**C**) The probability of downstream protein truncation due to the incorporation of 1 or more nonsense codons across different numbers of saturated positions. (**D**) Number of primer oligonucleotides required for each of the compared saturation mutagenesis techniques. (**E**) Other key considerations when comparing saturation mutagenesis techniques. Taken from (Ashraf *et al*., 2013)

Non-degenerate saturation mutagenesis is inherently more efficient in terms of sequence space, thus maximising library diversity (Figure 1.4.4A). The removal of degeneracy also eliminates amino acid representation bias (Figure 1.4.4B) and the incorporation of nonsense codons at saturated positions (Figure 1.4.4C). Figure 1.4.4D compares the number of primers required by each saturation mutagenesis technique to randomise different numbers

of contiguous targets, allowing an inference about the associated cost of each technique, while other key considerations are compared in Figure 1.4.4E.

Each saturation mutagenesis technique has both positive and negative aspects, meaning a balance of library quality against cost, time, workload and other considerations should be assessed when saturating any DNA library. With the randomised designed armadillo repeat protein DNA libraries most likely containing multiple positions, each being saturated to encode a significant number of new amino acid residues, non-degenerate saturation is deemed essential to ensure maximal library diversity, therefore increasing the chance of identifying high quality novel binders.

1.5 Determining the sequence of randomised designed armadillo repeat proteins

Upon completion of a randomised designed armadillo repeat protein DNA library, an assessment of its quality is essential before continuing with the downstream processes of expression and screening. To investigate the success of an engineered DNA library made of billions of different sequences, if not more (depending on the saturation), requires considerable sequencing power that is both cost and time effective. To direct downstream screening, an accurate measurement of encoded amino acid representation at each of the saturated positions is required. First generation sequencing techniques such as chain termination sequencing (Sanger sequencing) (Sanger, Nicklen & Coulson, 1977) are not capable of providing the accurate quantification of bases, codons and therefore the amino acid representation at saturated positions. So called second generation sequencing techniques such as Pyrosequencing (Nyrén & Lundin 1985), (Hyman, 1988), (Margulies *et al*., 2005) also encounter issues when attempting to sequence libraries of considerable sizes.

Sixty years of sequencing development has resulted in techniques capable of providing such niche data requirements, with Illumina becoming synonymous with next generation sequencing, as Sanger was with the first generation.

1.5.1 DNA library sequencing via Illumina

Illumina DNA sequencing was developed as a parallel technique to Pyrosequencing, using the detection of fluorescently labelled dNTPs like that of Sanger sequencing, implementing a

novel amplification processed called bridge amplification (Voelkerding, Dames and Durtschi, 2009). The Illumina sequencing process can be separated into two distinct stages: library preparation and sequencing (Figure 1.5.1).

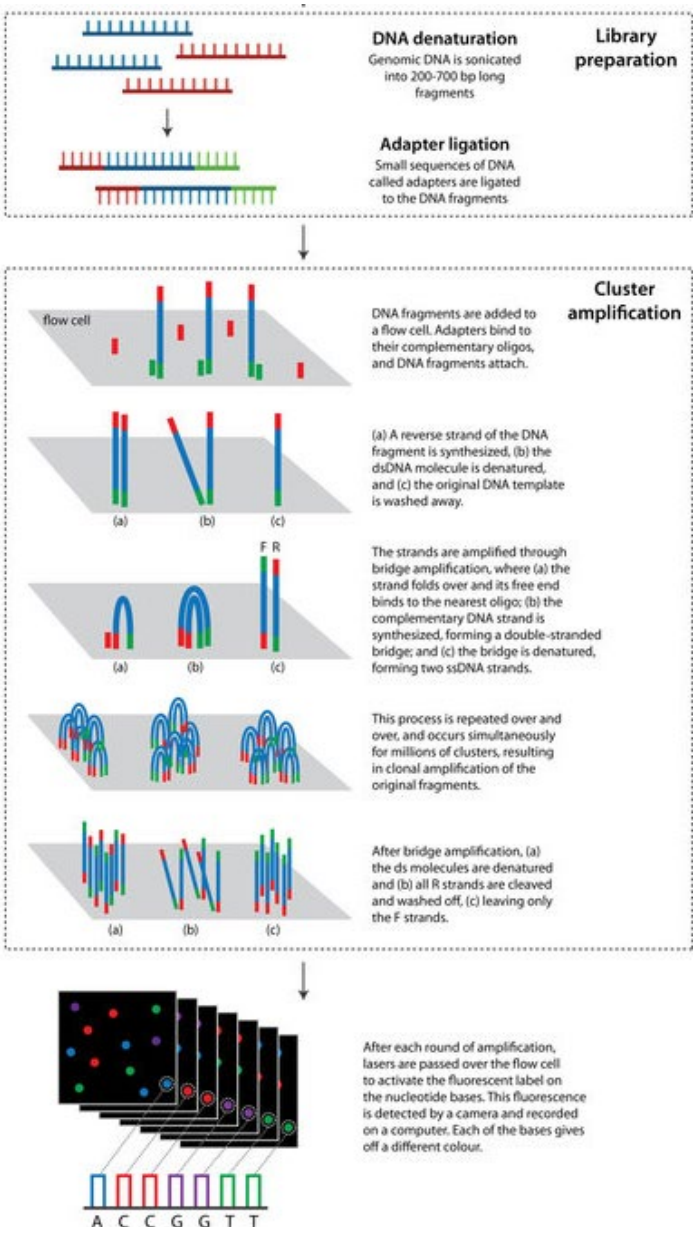

Figure 1.5.1 Process of Illumina next-generation sequencing.

The library sequences are fragmented and adapter sequences ligated to the 5' and 3' ends. Theses adapters facilitate binding to complimentary oligonucleotides found on the flow cell. Synthesis of the reverse strand begins, by using the forward reading strand as template. The original template strand is removed from the flow cell and via bridge amplification, the complementary DNA strand is synthesised incorporating fluorescently labelled dNTPS. The double stranded bridge is separated into single stranded DNA and bridge amplification is repeated. After each round of amplification and the incorporation of the fluorescently labelled dNTP, laser excitation reveals the base identity. Adapted from (Young and Gillung, 2019).

The DNA to be sequenced is first fragmented into smaller sections, with these fragments undergoing "tagmentation". During tagmentation two different transposomes break the DNA strands and add adaptor sequences to the 5' and 3' which are joined via ligation. These adaptor oligonucleotides possess motifs that allow for flow cell interactions and primer binding location for use during sequencing. The DNA fragment binds to a complementary oligonucleotide attached to the flow cell using the newly introduced motif and is used as a template strand for amplification via DNA polymerase. Denaturation of this double stranded oligonucleotide releases the original DNA strand leaving the newly synthesised strand to fold over and use the exposed terminal adaptor motif to bind to its complementary flow cell oligonucleotide. Amplification of this bridged oligonucleotide occurs and denaturation leaves two single strand oligonucleotides attached to the flow cell. This bridge amplification is repeated many times, and simultaneously with all the generated fragments from the library preparation stage, until clusters of newly synthesised oligonucleotides are formed. The reverse stand oligonucleotides are cleaved from the flow cell and the clustering step is completed. The sequencing primer now binds to its complementary site and the sequencing by synthesis stage commences. Each of the four nucleotides are fluorescently labelled, (each with its own colour identifier) with the base identity determined by laser excitation after each addition. This sequencing by synthesis occurs simultaneously across all of the clusters on the flow cell. Each incorporated dNTP is a non-permanent synthesis terminator, preventing the incorporation of more than one dNTP per cycle by blocking the 3' hydroxyl position with the fluorophore. When the fluorophore is cleaved the next cycle of dNTP incorporation and identification can occur (Turcatti *et al*., 2008). Sequencing by synthesis, is also performed in the opposite direction. Bridge amplification generates the reverse strand, with the forward strand cleaved from the flow cell, washed away and the sequence determined by identifying the incorporated fluorescently labelled dNTP each cycle.

With 454 Pyrosequencing discontinued, and newer next generation sequencing technologies aimed at sequencing longer nucleic acid sequences (Slatko, Gardner and Ausubel, 2018), Illumina sequencing is the most viable method of sequencing engineered designed armadillo repeat protein DNA libraries. Illumina sequencing is perfectly adapted to handle libraries of around 400bp, the expected sequence size estimations based on Varadamsetty *et al* (2012) randomising positions across three repeats (42 amino acids long so therefore 378bp in length). Illumina sequencing is also capable of producing the high quality read data essential for accurate amino acid distribution calculations (Illumina, 2011), while still being cost effective.

1.6 Protein screening from randomised designed armadillo repeat protein DNA libraries

With the library engineering and quality control performed, the screening of the randomised DNA libraries will then be performed by the Plückthun group, based on the well-established yeast display methodology developed by Boder and Wittrup (1997), using a-agglutinin in *S.cerevisiae*.

The vector pYD is engineered to contain the designed armadillo sequence, where the randomised cassette is inserted using homologous recombination, flanked by two detection tags, HA at the N-terminus and c-Myc at the C-terminus. This protein fusion is expressed as part of the naturally encoded Aga2 subunit of a-agglutinin (Figure 1.6), all present in the pYD vector. The counter subunit to Aga2, Aga1, is encoded within the yeast genome. Expression of both Aga1 and the Aga2-fusion protein, results in the surface display of the designed armadillo repeat protein library, by the formation of disulphide bonds between the Aga1 and Aga2 subunits (Figure 1.6). The flanking tags of the designed armadillo repeat protein can then be used to ensure the display of the protein has occurred correctly (detection of HA-tag) and to determine if the full length protein has been produced (detection of c-Myc tag) by flow cytometry. This insurance that the full-length protein has been displayed (detection of the c-Myc tag) can occur at the same time as investigating the designed armadillo repeat protein library binding. Using a fluorescently labelled linear peptide target, varying in one position from the consensus $(KR)_5$ binder, FACs (Fluorescence Activated Cell sorting) can distinguish between: cells not expressing full length protein, cells expressing full length protein but are not recognising the fluorescent peptide target and those cells expressing complete protein capable of binding the peptide target. This allows for all cells, except for full length proteins capable of binding, to be discarded. Multiple rounds of FACs including the use of competitive peptide targets (for example leucine when investigating an isoleucine binder) are used to enrich the binder pool. The pools undergo NGS sequencing, which after the completion of enrichment provide sequences for potential binders. The resulting protein is then investigated on an individual basis, by *Kd* investigations and crystallography (Y Stark 2020, personal communication, 9 March).

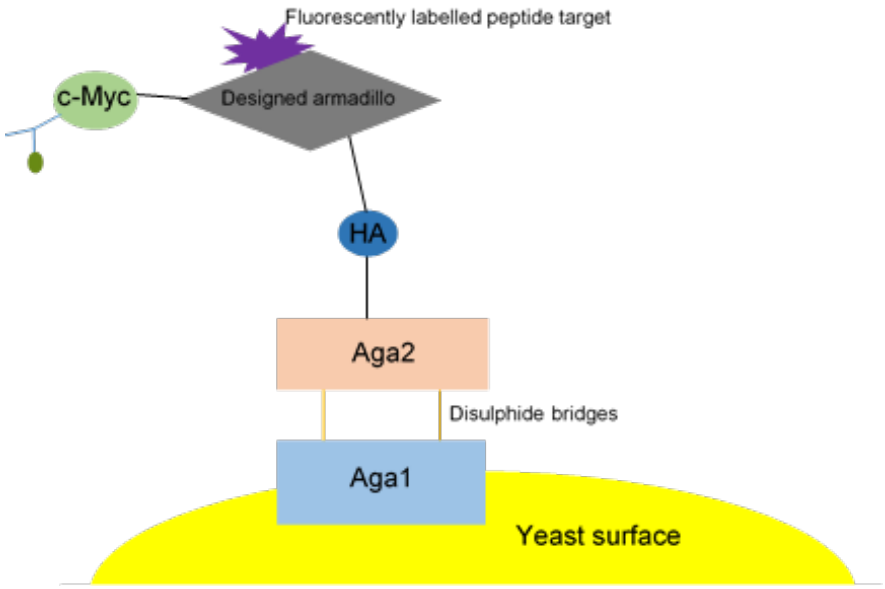

Figure 1.6 Diagrammatic representation of a-agglutinin based yeast display of designed armadillo repeat proteins.

Disulphide bonds (orange) between the Aga1 (pale blue) and Aga2 (peach) subunits allow for the display of the designed armadillo repeat protein (grey), flanked by two control tags (HA-blue and c-Myc- pale green). Binding of fluorescently labelled peptide targets (purple) by the designed armadillo repeat protein alongside or in absence of c-Myc (green) (recognised by fluorescently labelled antibodies), enables FACs.

1.7 The aims of PRe-ART (Predictive reagent antibody replacement technology)

The most abundantly used detection reagent in the research environment is the monoclonal antibody (Köhler and Milstein, 1975) with their involvement being pivotal to many discoveries over the last five decades. In order to generate monoclonal antibodies, immunisation against a specific epitope of interest occurs (often in mice), with the B-lymphocytes produced obtained by removal of the animal's spleen. The fusion product of the B-lymphocytes and immortalised myeloma cells, hybridomas, are then cultured. Each individual hybridoma clone is separated and the specific antibody portfolio of each clone screened for desirable activity (Liu, 2014). This means for each new epitope to be targeted, the timely process of generating and screening monoclonal antibodies is required. Unfortunately, after this extensive process the monoclonal antibody produced can suffer from batch variations (Voskuil, 2014), questioning the comparability of results, supported by Berglund *et al* (2008) reporting only 49% of commercial available antibodies were functional. In 2015 the call for sequence defined recombinant binding reagents was made (Bradbury and Plückthun, 2015),

highlighting the research community's desperation for high quality and reliable reagents, capable of reproducible results. Even with the movement to recombinant libraries, a selection process for each new target is still required meaning the quality of reagent has increased but the time and cost have not improved.

PRe-ART aims to revolutionise the reagent antibody field, by producing sequence-defined, modular regents that bind specifically and tightly to linear peptide targets. As the designed armadillo repeat proteins require a peptide target in an extended orientation, the potential application of these binders is huge, from binding denatured proteins in SDS-PAGE and western blots, to the specific recognition and binding of artificial protein tags. By generating individual armadillo repeat subunits that are dipeptide specific and sequence defined, a peptide specific, reproducible binder can be engineered by assembling the necessary modules (Figure 1.7).

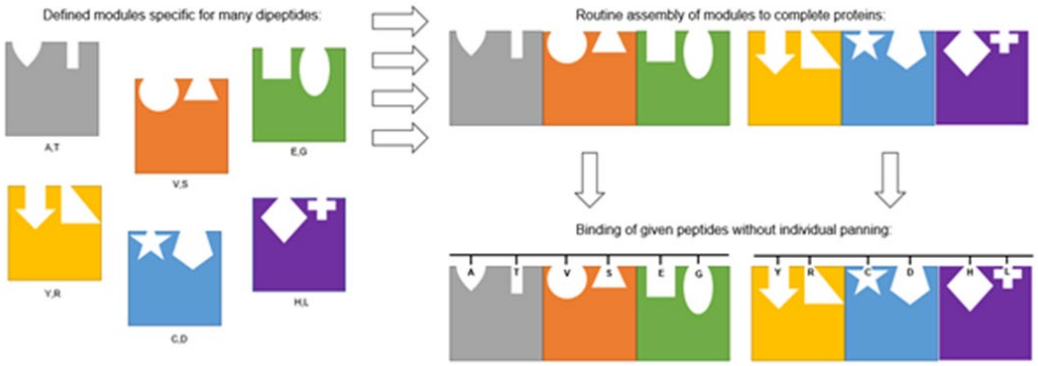

Figure 1.7 Overall aim of Pre-ART

A library of armadillo repeat subunits are generated, each with its own dipeptide specificity (a total of 400 subunits) capable of modular binding to generate a peptide specific designed armadillo repeat protein. Example shown is an armadillo repeat being produced to bind the peptide targets, THVESG & ATVRGY (Adapted from: A Plückthun 2017, Personal communication, 1st September 2017).

Having a library of modules that are fully-defined, eliminates the need for timely and therefore expensive binder selection processes and ensures that experiments using such reagents would be reproducible. The successful production of a stable designed armadillo repeat protein achieved by the Plückthun laboratory, means the engineering of repeat subunits with altered binding preference can begin. This will be attempted using a collaborative approach between the Plückthun, Höcker and Hine laboratories using both experimental and computational approaches to direct dArmRP engineering.

1.7.1 PhD Project Aims

This PhD project aims to contribute to the generation of the designed armadillo repeat protein module repertoire, via the following work streams:

- Creating generic, saturated libraries of the arginine pocket
- Creating focussed libraries of the arginine pocket specific to threonine binding
- Developing non-proprietary methodology to saturate the lysine pocket (ParaMAX randomisation)
- Developing methods to assess NGS data of saturated libraries containing contiguous and non-contiguous saturated codons.

Chapter 2 Materials and Methods

2.1 Materials

2.1.1 Buffer constituents

2.1.1.1 Pfu DNA polymerase buffer (10x)

200mM Tris-HCl (pH 8.8 at 25°C), 100mM KCl, 100mM (NH4)2SO4, 20mM MgSO4, 1.0% Triton® X-100 and 1mg/ml nuclease-free BSA. Purchased from Promega.

2.1.1.2 T4 DNA ligase buffer (10x)

300mM Tris-HCl (pH 7.8 at 25°C), 100mM MgCl2, 100mM DTT and 10mM ATP. Purchased from NEB.

2.1.1.3 CutSmart® Buffer

1X CutSmart® Buffer, 50 mM Potassium Acetate, 20 mM Tris-acetate, 10 mM Magnesium Acetate, 100 µg/ml BSA, (pH 7.9 @ 25°C). Purchased from NEB.

2.1.1.4 TAE (1x)

40 mM Tris, 20 mM acetic acid, 1 mM EDTA. Purchased from ThermoFisher Scientific

2.1.1.5 Blue/orange loading buffer (6x)

0.4% orange G, 0.03% bromophenol blue, 0.03% xylene cyanol FF, 15% Ficoll® 400, 10mM Tris-HCl (pH 7.5) and 50mM EDTA (pH 8.0). Purchased from Promega.

2.1.2 Other solutions

2.1.2.1 dNTP Mix

Sodium salts of dATP, dCTP, dGTP and dTTP, each at 10mM in water at pH 7.5; the total concentration of nucleotides is 40mM. Purchased from Promega.

2.1.2.2 Oligonucleotides

All oligonucleotides were purchased from Eurofins Genomics.

2.2 Methods

2.2.1 General methods

2.2.1.1 *Mly*I Restriction

*Mly*I restrictions were performed using a 50 µL reaction mix of 1x CutSmart® buffer, 1µg of DNA to be restricted, 1 µL *Mly*I (0.2 U/µL) and the remaining reaction volume millQ water.

The reaction mix was incubated at 37°C for 1 hour, followed by enzyme deactivation at 65°C for 20 minutes.

2.2.1.2 Blunt-end Ligation

Blunt-end ligations were performed at a volume of 20 µL using a reaction mix of 1x T4 DNA ligase buffer (NEB), 50% of the reaction volume being DNA to be joined, divided equally between all ligation counterparts, 1 µL T4 DNA ligase (100 U/µL) (NEB) with the remaining volume made by millQ water.

The reaction was incubated at 16°C overnight, followed by enzyme deactivation at 65°C for 10 minutes.

2.2.1.3 Agarose gel electrophoresis

3% Agarose gels were made by dissolving 3g LE Multi-Purpose Agarose (Geneflow) in 100ml 1x TAE buffer unless otherwise stated. As the solution was cooling, ethidium bromide solution was added to achieve a final concentration of 0.5 µg/ml. The solution was then poured into the casting tray of a Fisherbrand™ Sub-Gel Midi Horizontal Gel Electrophoresis Unit unless otherwise specified. Once the gel had solidified, the comb was removed, the casting tray placed into the electrophoresis tank, which was then filled with TAE 1x buffer. DNA was mixed with loading buffer and loaded into gel wells. Gels were electrophoresed using a constant voltage of 20 V/cm was used unless otherwise specified. Once the DNA had migrated sufficiently through the gel, it was visualised using a Syngene G:BOX.

2.2.1.4 PCR product purification

To purify PCR products, the Promega Wizard® SV Gel and PCR Clean-Up System was used according to manufacturer's instructions.

2.2.1.5 DNA quantification via NanoDrop

DNA was quantified using a Thermo Scientific NanoDrop™ 2000 Spectrophotometer, blanked using milliq water.

2.2.1.6 DNA sequencing

DNA libraries were sent to Genewiz using the amplicon EZ sequencing service, following Genewiz sample submission guidelines: DNA sample to be standardised to 20ng/µL with a minimum of 500ng submitted.

2.2.2 Saturation Mutagenesis

2.2.2.1 MAX Oligonucleotide selection pools

Each individual oligonucleotide used in the MAX randomisation process, was synthesised externally by Eurofins Genomics at a concentration of 100 µM, suspended in water unless otherwise specified. With all MAX oligonucleotides at the same concentration, to generate a MAX oligonucleotide selection pool, the MAX selection oligonucleotides encoding the desired amino acids were mixed in equal volumes and therefore equal concentrations. A mix of 20

different MAX selection oligonucleotides would containing each oligonucleotide at an individual concentration of 5 μM, in a selection pool of 100 μM. The codons selected to represent each amino acid based on *S. cerevisiae* preference using GenScript Codon Frequency Table(chart) Tool- Saccharomyces cerevisiae (gbpln).

2.2.2.2 MAX randomisation

The master mix made in the MAX randomisation process used 1x T4 DNA ligase buffer mixed with the NNN template oligonucleotide (final concentration 10 μM), a variable number of MAX oligonucleotide selection oligonucleotide pools (final concentration of 10 μM for each total selection oligonucleotide pool), the oligonucleotide upstream of the saturated region (final concentration 10 μM) and the oligonucleotide downstream of the saturated region (final concentration 10 μM), with the remaining volume (minus T4 DNA ligase) made of milliQ water unless otherwise stated in a reaction volume of 20 μL.

Reaction incubation started at 99°C with a decrease of 1°C/minute until 4°C was reached, at which point T4 DNA ligase was added to an end concentration of 200U/μL, unless otherwise stated.

2.2.2.3 ParaMAX randomisation

Multiple individual couplets were made simultaneously using the MAX randomisation process, amplified using PCR (2.2.4), restricted using *MlyI* ready for blunt end ligation to form a Quad. Repetition of process to generate larger contiguous randomised codon construct.

2.2.3 Control test

2.2.3.1 Self-priming

All possible different combinations of non-MAX selection oligonucleotide, oligonucleotides were made. To each of these reactions containing different mixes of the oligonucleotides (0.5 μM), 1x Pfu DNA polymerase buffer (Promega), 200mM dNTP mix (Promega), and 0.025U/μL Pfu DNA polymerase were added, with the remaining volume of each mix being milliQ water.

The PCR cycling conditions used were: denaturation at 95°C for 2 minutes, followed by 35 cycles of: 95.0°C 30 seconds, 52.0°C 30 seconds, 72.0°C 2 minutes. Final extension at 72°C 5 minutes and 4°C hold.

2.2.4 PCR

2.2.4.1 PCR amplification

PCR amplification used a master mix of 1x Pfu DNA polymerase buffer (Promega), 200μM dNTPs (Promega), forward primer 0.5 μM and reverse primer 0.5μM, Pfu DNA polymerase 0.025U/μL and MAX randomisation product as template at 2% of master mix volume, unless otherwise stated.

The PCR cycling conditions used were: denaturation at 95°C for 2 minutes, followed by 30 cycles of: 95.0°C 30 seconds, 65°C 30 seconds, 72.0°C 20 seconds with a final extension at 72°C 1 minute and 4°C hold, unless otherwise stated.

2.2.4.2 Overlap PCR

To join multiple sequences, an overlap PCR between the individual constructs/oligonucleotides was used. An equal volume of each construct (or if overlapping oligonucleotides equamolar amounts)of each oligonucleotide was mixed with 200μM dNTPs (Promega), 0.03U/μL Pfu DNA polymerase with the remaining master mix volume being made of milliQ water.

The PCR cycling conditions used were: 95°C 2 minutes followed by 30 cycles of: 95°C 30 seconds, 60°C 30 seconds, 72°C 1 minute with a final extension step of 72°C 2 minutes and 4°C hold, unless otherwise specified.

2.2.5 NGS data analysis

2.2.5.1 Alignment sequence design

In preparation for NGS data analysis a reference sequence was required, which was made using UGENE (http://ugene.net/). The full library sequence of interest (unless otherwise stated) was copied into the programme in a 5'-3' direction with MAX codons replaced with NNN and alphabet option Raw selected.

2.2.5.2 NGS data preparation using Galaxy

The processing of the raw NGS files was performed using Galaxy (https://usegalaxy.org/).

2.2.5.2.1 Fastq join function

Mate 1 and mate 2 Fastq files containing the raw sequencing data were uploaded to the Galaxy server. The two mate files were joined using the *FASTq Join* function (unless otherwise stated). The parameters selected were: Paired-End, Maximum percentage difference between matching segments: 5, Minimum length of matching segments: 9 unless otherwise stated.

2.2.5.2.2 Filter by quality function

The output of the joining function underwent quality testing via the Filter by quality function with the following parameters: Quality cut-off value: 30, Percent of bases in sequence that must have quality equal to / higher than cut-off value= 90 (unless otherwise stated).

2.2.5.2.3 Filter by Filter FASTQ function

Raw illumina sequence data (Fastq file format) was filtered using the following parameters: Minimum and maximum length: 216, Minimum quality score: 30 (no upper MAX quality score assigned), 12 bases allowed outside of quality score range (unless otherwise stated)

2.2.5.2.4 Alignment of reference sequence and NGS data using Bowtie2

Bowtie 2 alignment was performed between the output of the Filter by quality function and the uploaded UGENE alignment sequence. The following Bowtie 2 parameters were selected: Single end, Using default settings.

2.2.5.2.5 Alignment of reference sequence and NGS data using BWA-MEM

BWA-MEM alignment was performed between the output of the Filter by quality function and the uploaded UGENE alignment sequence. The following BWA-MEM parameters were selected: Single end, Using default settings.

2.2.5.2.6 File reformatting for Excel use

The now aligned output from the *Bowtie 2* function required reformatting so it could be accessed in Excel. The file was converted from its BAM format to SAM using the function *Convert BAM to SAM*, selecting the option to exclude the header as this does not contain any read data. The new SAM formatted file was converted to Interval using the *SAM to Interval* function. This Interval file was then downloaded from the history tab and opened in Excel. Columns were delimited using the default setting and all non-sequence containing columns deleted.

2.2.6.3 Codon Frequency counting in Excel

To determine codon frequency at saturated positions within the library, Excel's *countif* function was implemented using the MAX selection oligonucleotides as the count criteria in the following formula:.

=sum(countifs(range,{"*___*","*___*"…})), where 'range' describes the Excel column in which the NGS reads are displayed, and ___ indicates the MAX oligonucleotide being counted unless otherwise stated.

Chapter 3 Mutagenesis of a synthetic designed armadillo repeat protein arginine binding pocket

3.1 Introduction

Within the $Y_{II}M_5A_{II}$ designed armadillo repeat protein protein, are five internal 'M' repeats responsible for the recognition and binding of the target peptide sequence. The H3 helix, possessed by each of the M repeats, forms a binding groove with its adjacent H3 helix of the neighbouring repeat, capable of binding a single dipeptide of the target peptide. The target peptide sequence of the original $Y_{II}M_5A_{II}$ designed armadillo repeat protein protein was $(KR)_5$, with each H3 groove recognising a KR dipeptide. To achieve the overall aim of generating a library of M repeats capable of binding all dipeptide combinations, the binding specificities of the pockets within the H3 grooves, currently recognising either arginine or lysine, needed to be altered. This chapter focuses on the arginine binding pocket located in the binding groove formed by M repeats 4 and 5, with the key binding residues responsible for arginine recognition (Figure 3.1) saturated using MAX randomisation.

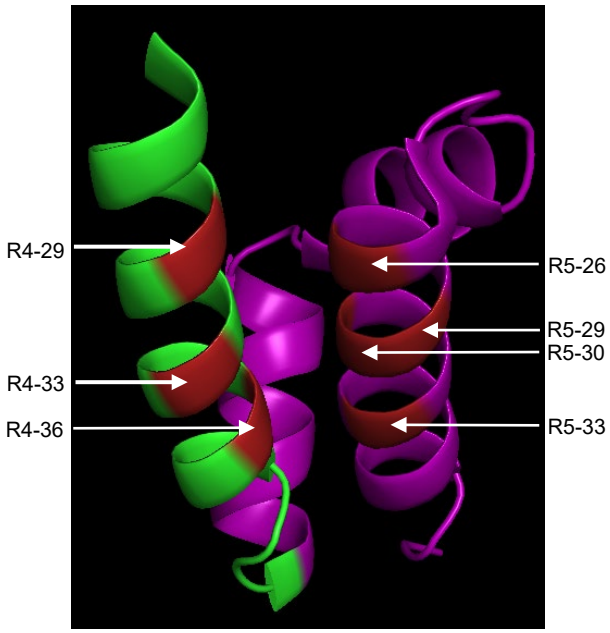

Figure 3.1: Visualisation of the binding groove formed from the H3 helices of repeat 4 and repeat 5.

Key residues for peptide binding coloured red in H3 helix of repeat 4 (green) (positions 29,33,36) and in the H3 helix of repeat 5 (purple) (positions 26, 29, 30, 33). Adapted from (Hansen *et al.*, 2016).

As this was the first designed armadillo repeat protein to be saturated using MAX randomisation, a broad-based library was most appropriate, where the key binding residues were saturated with the widest range of appropriate amino acids. Based on previous designed armadillo repeat protein work conducted by Varadamsetty *et al* (2012), R4-36 saturation was limited to alanine, serine, threonine and valine as naturally occurring armadillo repeat proteins favour smaller residues at this position. The exclusion of cysteine at all target positions was to prevent the occurrence of disulphide bond formation and proline exclusion was to prevent any binding groove configuration changes due to its proclivity to disrupt α-helices.

3.2 Arginine library design to facilitate MAX randomisation

The arginine pocket to be targeted belonged to the H3 binding groove generated by M repeats 4 and 5. A short section of M repeat 3 was included in the arginine library design as a conserved region (Figure 3.2 grey), to allow for downstream incorporation of the randomised cassette into the designed armadillo repeat protein via homologous recombination. Repeat 4 of the arginine library possessed three positions of interest while repeat 5 possessed four. As MAX randomisation requires a unique 6bp invariant region to allow for specific annealing of the MAX selection oligonucleotides (Figure 3.2), the invariant region neighbouring R4-36 was changed by silent mutation from its original sequence of GCACTG (identical to the invariant region neighbouring R4-33) to GCTTTA (Figure 3.2).

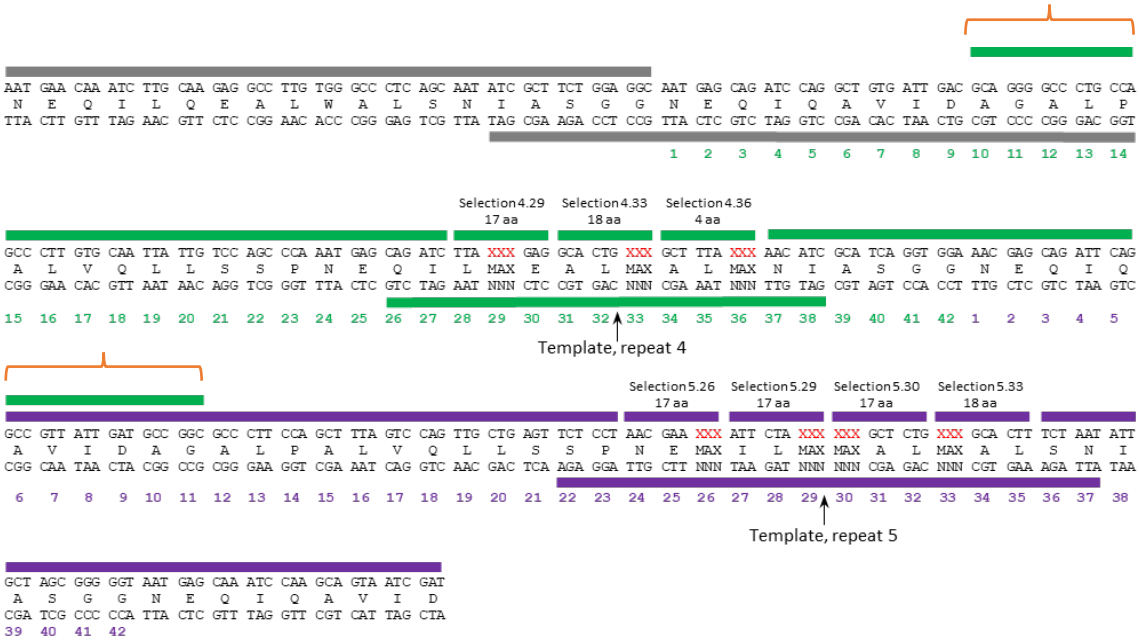

Figure 3.2: DNA sequence for internal repeat 3, 4 and 5 of the designed armadillo repeat protein, overlaid with the MAX randomisation library design.

The DNA sequence is split into three separate constructs, a conserved region (grey), repeat 4 (green) and repeat 5 (purple), with positions of saturation highlighted red. Overlapping regions between constructs are indicated by orange brackets.

The arginine library sequence was separated into three constructs which were engineered separately and ultimately combined to form the full length arginine library cassette.

3.3 Generation of MAX selection oligonucleotide pools for positional saturation of key binding residues in an arginine binding pocket

Before commencing library construction, pools of MAX selection oligonucleotides were needed for each targeted position. The codons selected to represent each amino acid were based on *S. cerevisiae* preference using GenScript Codon Frequency Table(chart) Tool-Saccharomyces cerevisiae (gbpln) (2.2.2.1), as this was the organism of choice for protein expression. Even though the lysine codon AAA, has a higher usage frequency the alternative codon, AAG, was chosen. This was in an attempt to alleviate any potential ligation bias during MAX randomisation that can be caused when DNA sequences have an excess of consecutive adenine bases. The required MAX selection oligonucleotides to represent the desired amino acids at each position (Table 3.3) were mixed to form seven different MAX selection oligonucleotide pools, one for each target position (2.2.2.1). (See Appendix 1 for

non-MAX selection oligonucleotide oligonucleotides used to engineer the randomised single arginine DNA library).

	Position	R4-29	R4-33	R4-36	R5-26	R5-29	R5-30	R5-33
	Invariant sequence showing MAX position	TTAMAXGAG	GCACTGMAX	GCTTTAMAX	AACGAAMAX	ATTCTAMAX	MAXGCTCTG	MAXGCACTT
Identity of MAX codon	Aspartic acid	GAT	GAT		GAT	GAT	GAT	GAT
	Glutamic acid	GAA	GAA		GAA	GAA	GAA	GAA
	Aspargine	AAT	AAT		AAT	AAT	AAT	AAT
	Glutamine	CAA	CAA		CAA	CAA	CAA	CAA
	Threonine	ACT	ACT	ACT	ACT	ACT	ACT	ACT
	Serine	TCT	TCT	TCT	TCT	TCT	TCT	TCT
	Methionine	ATG	ATG		ATG	ATG	ATG	ATG
	Leucine	TTG	TTG		TTG	TTG	TTG	TTG
	Isoleucine	ATT	ATT		ATT	ATT	ATT	ATT
	Valine	GTT	GTT	GTT	GTT	GTT	GTT	GTT
	Alanine	GCT	GCT	GCT	GCT	GCT	GCT	GCT
	Phenylalanine	TTT	TTT		TTT	TTT	TTT	TTT
	Tyrosine	TAT	TAT		TAT	TAT	TAT	TAT
	Tryptophan	TGG	TGG		TGG	TGG	TGG	TGG
	Histidine	CAT	CAT		CAT	CAT	CAT	CAT
	Arginine	AGA	AGA		AGA	AGA	AGA	AGA
	Lysine	AAG	AAG		AAG	AAG	AAG	AAG
	Glycine		GGT					GGT
	Cysteine							
	Proline							

Table 3.3: Table showing the encoded amino acids in the MAX selection oligonucleotide pools created for saturation of each target position in the randomised arginine library. Each cell highlighted indicates that amino acid's inclusion in the corresponding MAX selection oligonucleotide pool. Each MAX selection oligonucleotide was received at 50 µM in aqueous solution. To make the MAX selection oligonucleotide pool for each position, an equal volume and therefore concentration of each MAX selection oligonucleotide was mixed. The resulting end concentration of each selection pool was 50 µM, with individual MAX selection oligonucleotide concentration at 2.94 µM for positions: R4-29, R5-26, R5-29, R5-30, 2.78 µM for R4-33, and R5-33 and 12.5 µM for R4-36.

3.4 Production of the repeat 4 construct for incorporation into a randomised arginine pocket DNA library

As each of the three library components were to be made separately as shown in Figure 3.2, the repeat 4 construct (green, Figure 3.2) was focused upon first.

3.4.1 Determining the optimal repeat 4 MAX randomisation product template dilution and annealing temperature for PCR amplification

In order to generate a high quality PCR product, the optimal conditions for the repeat 4 PCR amplification had to be determined. This occurred in a two-pronged approach, ascertaining the optimal annealing temperature for the amplification while at the same time finding the optimal template dilution. The repeat 4 MAX randomisation product template was produced via MAX randomisation (2.2.2.2) (Figure 1.4.3.2) and then diluted in milliQ water to form a template dilution range of 1/10, 1/100, 1/1000 – fold dilutions. Undiluted template and each of the template dilutions was used in an individual annealing temperature gradient PCR amplification, allowing for the determination of the optimal annealing temperature and template dilution in one set of experiments (2.2.4.1). PCR products were visualised via agarose gel electrophoresis (2.2.1.3).

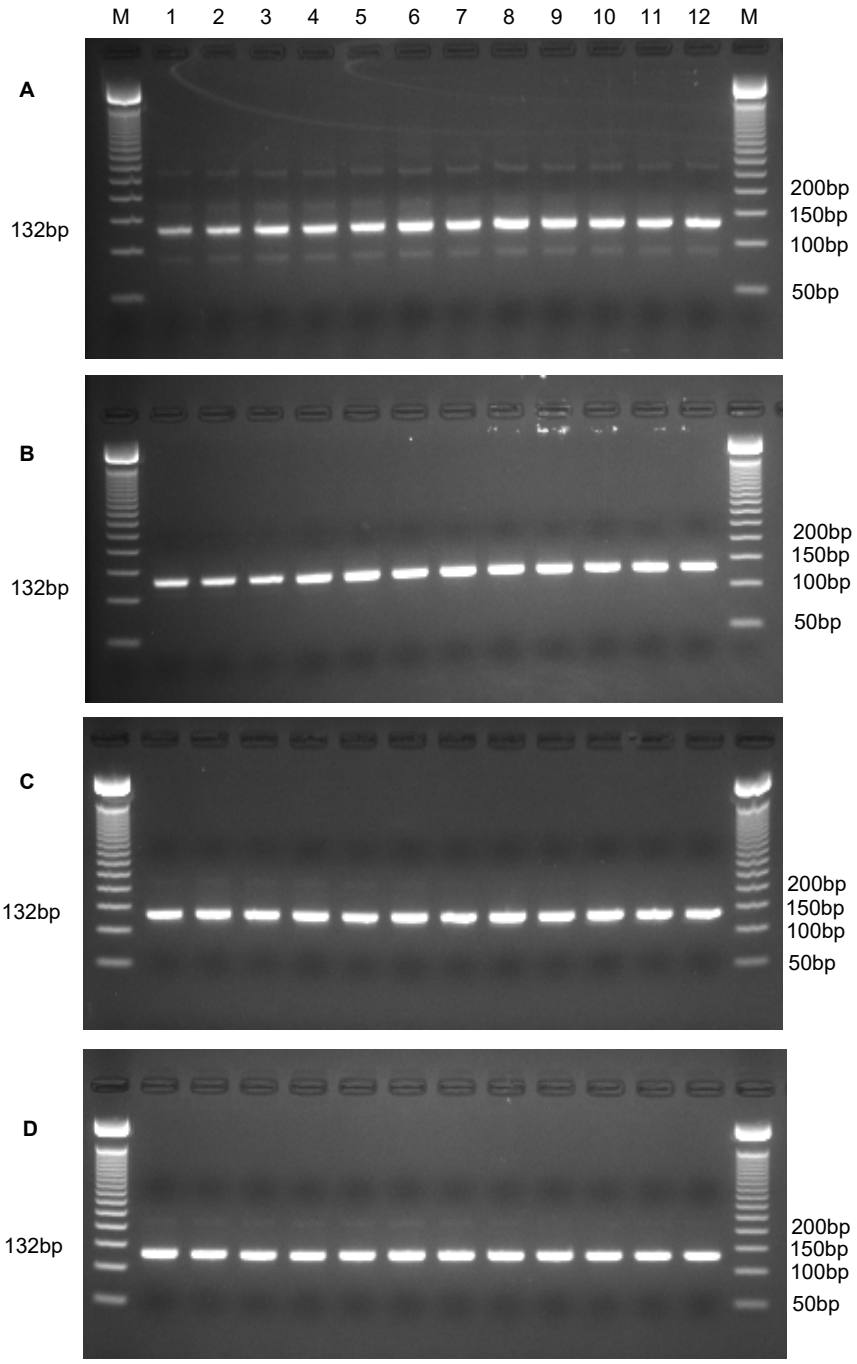

Figure 3.4.1: Four annealing temperature gradients using MAX randomisation product as template at varying dilutions to determine optimal annealing temperature and template dilution for repeat 4 PCR amplification.

Four individual PCR reactions were made using (**A**) neat MAX randomisation product, (**B**)1/10 diluted MAX randomisation product, (**C**) 1/100 diluted MAX randomisation product and (**D**) 1/1000 diluted positive MAX randomisation as template and divided into 12 equal reactions. Annealing occurred at the temperatures described below. Lanes: M= 50bp MW ladder, annealing temperatures; **1**.45.0, **2**. 45.4, **3**.46.5, **4**.48.5, **5**.51.1, **6**.53.7, **7**.56.1, **8**.58.7, **9**.61.2, **10**.63.2, **11**.64.4,**12**. 65.0 (°C).

Figure 3.4.1 shows a bright product band of 132bp across the entire annealing temperature gradient using the four different MAX randomisation product template dilutions. Figure 3.4.1A also shows two other products of ~90bp and 275bp for all annealing temperatures tested, with a third band of 175bp present for annealing temperatures 45.0 - 56.1°C, that becomes less defined when using higher annealing temperatures. Due to the presence of unwanted products, the use of neat MAX randomisation product as template was not possible. A faint product of ~225bp could be seen across the annealing temperature range when using 1/1000 MAX randomisation product as template (Figure 3.4.1D). This ~225bp product could also been seen in Figure 3.4.1C (1/100 template dilution) for temperatures 45.0 - 63.2°C, becoming less intense as the temperature increased. In comparison with the other template dilutions, 1/10 MAX randomisation product template dilution (Figure 3.4.1B) showed fainter unwanted products. This may have been a result of the non-product band intensity being blocked by the dye front causing them to appear fainter. As the 225bp band was seen across all the template dilution gradients, with a minimal difference in brightness (Figure 3.4.1B, C and D), it was concluded the increasing dilution from 1/10 to 1/1000 had no considerable impact on the 225bp band and therefore the template dilution of 1/10 was considered optimal as it was the lowest template dilution. Across Figure 3.4.1B, the 132bp band intensity increased from 48.5°C, but as the band also became denser (potentially hiding unrequired products), an annealing temperature of 45.0°C, which produced a lesser but still bright band was deemed most appropriate for use in downstream applications. This annealing temperature was chosen for future amplifications alongside the 1/10 template dilution. No-template controls were not performed for any of the annealing temperature and template dilution combinations seen in Figure 3.4.1. Instead, a no-template control was performed using the optimal conditions determined from Figure 3.4.1 (seen in Figure 3.5.4.3) demonstrating the product resulted from specific amplification of the repeat 4 template.

3.4.2 Testing asymmetry of MAX randomisation of repeat 4 under optimised conditions

To eliminate non-degeneracy and therefore produce unbiased DNA libraries, MAX randomisation relies upon the specific, asymmetric amplification of the MAX codon-containing DNA strand. This DNA strand is made from the ligations of the constituent MAX selection oligonucleotides and flanking oligonucleotides (Figure 3.2) and is not produced in the absence of ligase.

To determine if the 132bp product generated was a result of the required, specific asymmetric amplification, the optimal conditions of 45.0°C annealing using a 1/10 dilution of MAX randomisation product as template were then used to amplify the control non-ligase

repeat 4 MAX randomisation product (2.2.4.1). The PCR product was visualised via agarose gel electrophoresis (2.2.1.3).

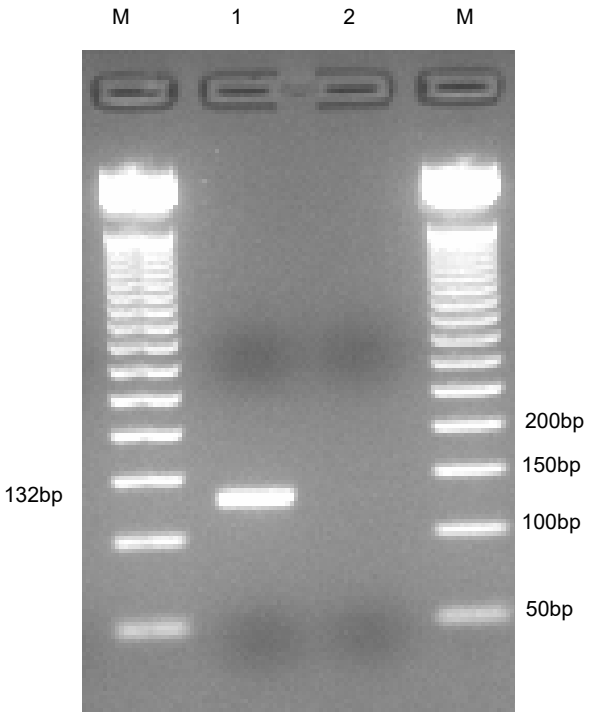

Figure 3.4.2: Repeat 4 negative controls to assess asymmetric amplification.

Repeat 4 was created as described (2.2.4.1.) under optimised conditions of 1/10 diluted template and an annealing temperature of 45°C, but with the omission of ligase during construction. Lanes: M=50bp MW ladder, **1**. Ligase-negative control, **2**. No template control.

A bright product band was seen at 132bp, meaning the product was formed by amplifying something other than the required, MAX codon-containing DNA strand.

3.4.3 Source of full length product in ligase negative control

The generation of a PCR product in the absence of ligase (Figure 3.4.2) and therefore in the absence of the creation of a DNA strand containing MAX codons, suggested these products resulted either from amplification of the template strand, or else from unanticipated self-priming events between other, component oligonucleotides. When considering the constant oligonucleotides used to generate repeat 4, a product of 90bp could only be produced by a

cross-priming event between the upstream oligonucleotide and the NNN template strand, because of the directionality of DNA extension (Figure 3.4.3.1).

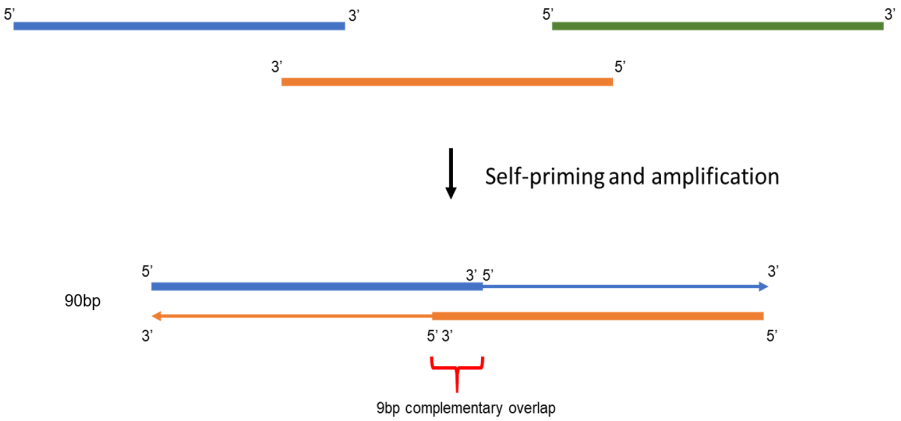

Figure 3.4.3.1: Diagrammatic representation of a self-priming event between repeat 4 MAX oligonucleotides.

A self-priming event between the upstream repeat 4 MAX oligonucleotide (blue) and the NNN template (orange), with a complementary region of 9bp (red bracket), (excluding the downstream oligonucleotide (green)) would generate a product of 90bp (blue/orange construct).

To investigate this experimentally, the MAX randomisation constituent oligonucleotides were mixed to form all possible combinations and PCR amplified (2.2.3.1). PCR products were visualised via agarose gel electrophoresis (2.2.1.3).

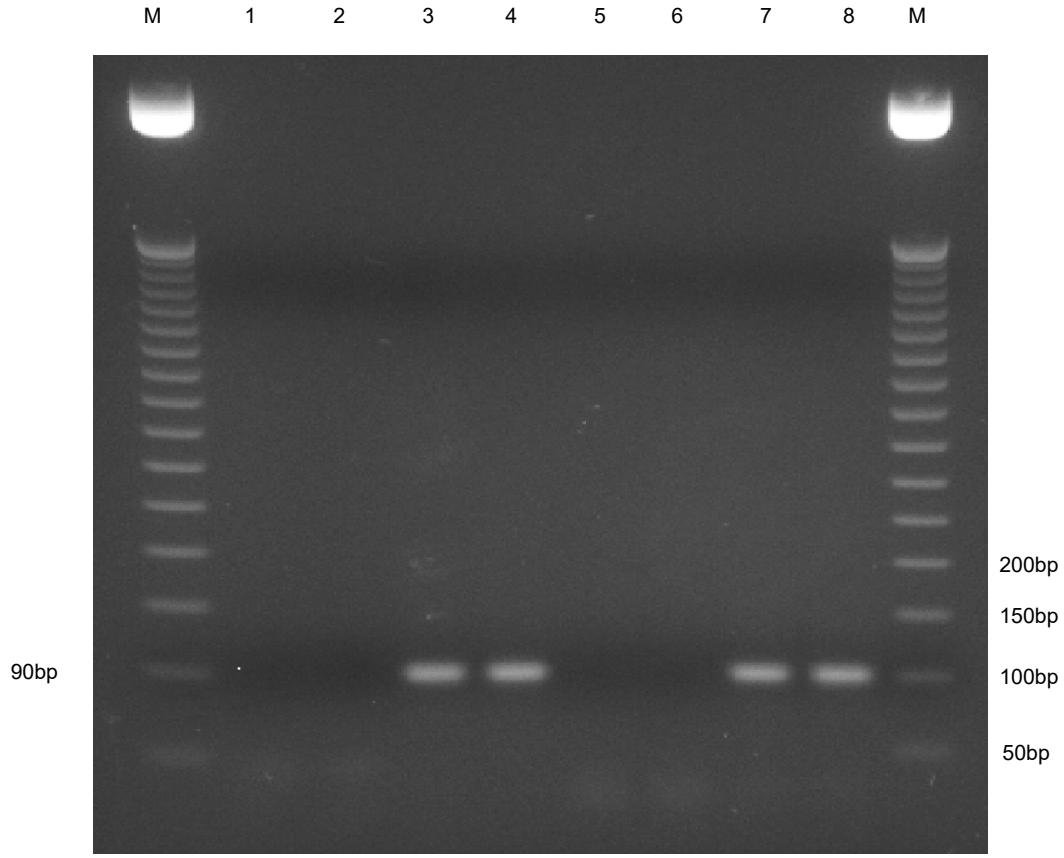

Figure 3.4.3.2: PCR amplification of different MAX oligonucleotide combinations to investigate self-priming events with 9bp complementary regions.

Four individual PCR reactions were made and sampled in duplicate, with each reaction investigating the possibility of self-priming events between a different combination of the constant selection oligonucleotides. Lanes: M= 50bp MW ladder, 1-8, PCRs with oligonucleotide combinations of: 1) upstream flanking oligonucleotide and downstream flanking oligonucleotide, 2) duplicate of 1), 3) upstream flanking oligonucleotide and NNN template, 4) duplicate of 3), 5) downstream flanking oligonucleotide and NNN template, 6) duplicate of 5), 7) upstream flanking oligonucleotide, downstream flanking oligonucleotide and NNN template, 8) duplicate of 7).

The 90bp product in Figure 3.4.3.2, was only generated in the presence of the upstream flanking oligonucleotide and the NNN template strand, validating the assumption of self-priming events causing the 90bp construct production (2.2.3.1).

The production of a 90bp product from self-priming did not account for the 132bp product seen in Figure 3.4.2, but within the non-ligase MAX randomisation control PCR amplification, the forward and reverse primers for the MAX codon containing DNA strand were also present. Figure 3.4.3.3, shows how these primers could be responsible for integrating the

downstream flanking oligonucleotide with the 90bp product generated by self-priming of the upstream flanking oligonucleotide and NNN template.

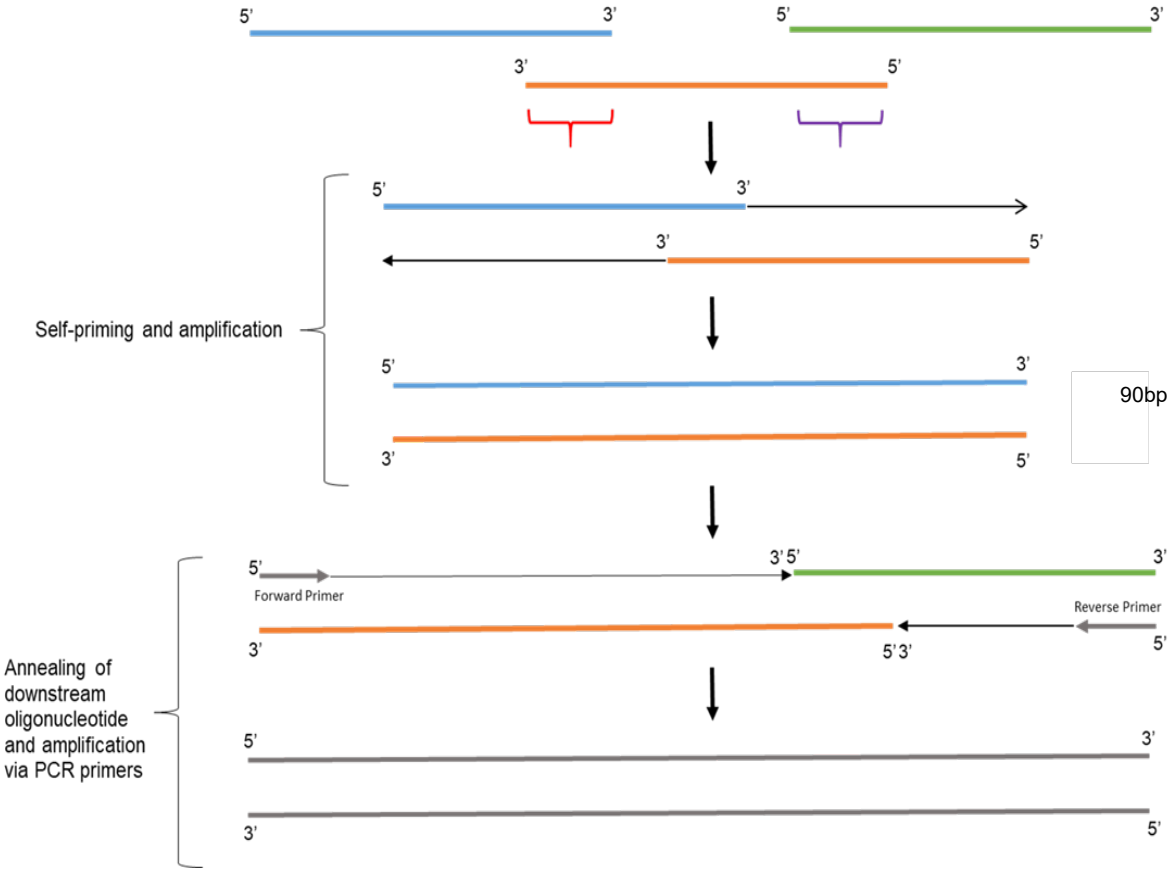

Figure 3.4.3.3: Diagrammatic representation of the series of events between repeat 4 MAX oligonucleotides, resulting in a 132bp product in the absence of ligase.

A self-priming event is possible between the upstream repeat 5 MAX oligonucleotide (blue) and the NNN template (orange), with a complementary region of 9bp (red bracket), generating a product of 90bp (blue/orange construct). This 90bp construct can then anneal to the downstream flanking oligonucleotide (green) via their complementary region of 9bp (purple bracket). PCR primers then facilitate the extension and downstream amplification to generate the 132bp product.

3.5.4 Repeat 4 design alteration to eliminate products in non-ligase control

To prevent the production of construct in the non-ligase control amplification, the 9bp complementary region between the NNN template and the flanking oligonucleotides was shortened to 6bp, by reducing the length of the NNN template strand from 45bp to 39bp. (Figure 3.5.4.1).

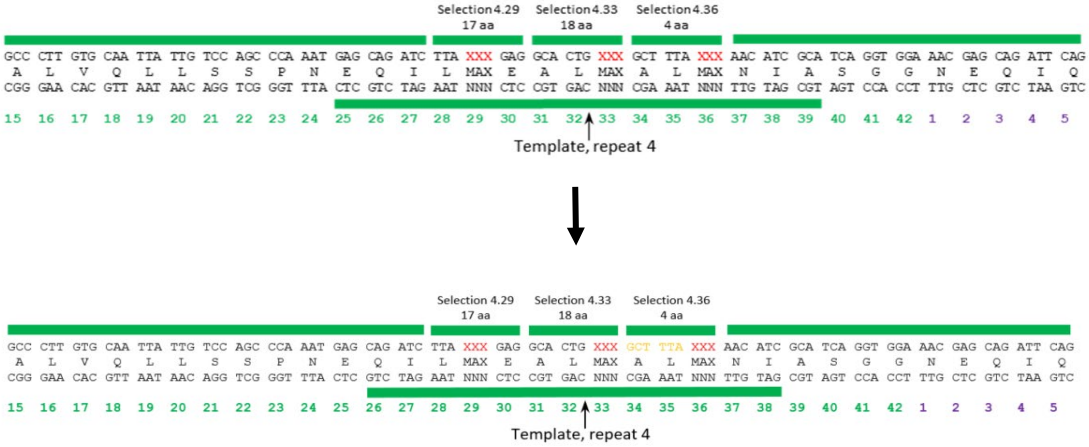

Figure 3.5.4.1: Diagrammatic representation of library design alternation for repeat 4 construct, by the shortening of the NNN template strand.

The complementary region between the NNN template strand and the flanking oligonucleotides was reduced from 9bp to 6bp, by shortening the NNN template strand by 3bp at both the 5' and 3' end, reducing the oligonucleotide from 45bp to 39bp.

This would facilitate a two-stage approach in reducing undesirable product formation by first reducing the possibility of self-priming between the upstream oligonucleotide and the NNN template producing the 90bp construct and secondly, reducing the chance of annealing between the downstream flanking oligonucleotide and the 90bp construct. The new shortened template oligonucleotide was combined with the upstream and downstream oligonucleotides and tested as before, to investigate if a 90bp product could still be formed. (2.2.3.1). The PCR products were visualised via agarose gel electrophoresis (2.2.1.3).

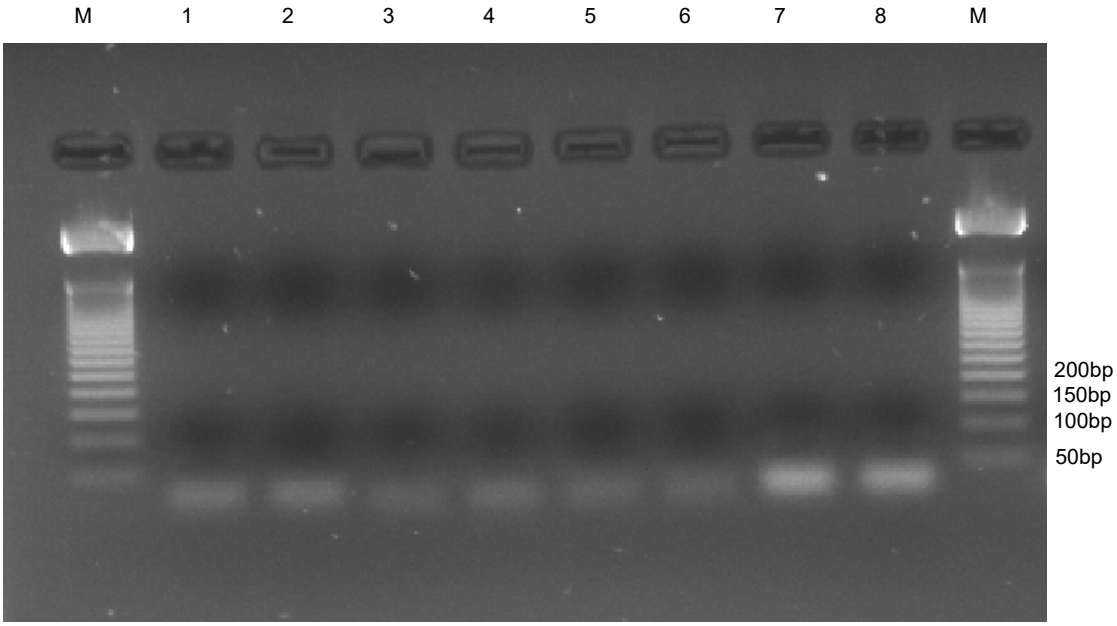

Figure 3.5.4.2: PCR amplification of different MAX oligonucleotide combinations to investigate self-priming events with 6bp complementary regions.

Four individual PCR reactions were made and sampled in duplicate, with each reaction investigating the possibility of self-priming events between a different combinations of the constant selection oligonucleotides. Lanes: M= 50bp MW ladder, 1-8, PCRs with oligonucleotide combinations of: 1) upstream flanking oligonucleotide and downstream flanking oligonucleotide, 2) duplicate of 1), 3) upstream flanking oligonucleotide and NNN template, 4) duplicate of 3), 5) downstream flanking oligonucleotide and NNN template, 6) duplicate of 5), 7) upstream flanking oligonucleotide, downstream flanking oligonucleotide and NNN template, 8) duplicate of 7).

With the oligonucleotide combinations no longer producing the 90bp product (Figure 3.5.4.2), the MAX randomisation process to generate the repeat 4 construct template was re-performed now using the shortened NNN template oligonucleotide. Both positive and ligase-negative control MAX randomisations were carried out (2.2.2.2). These products were then used as template in identical PCR amplifications (2.2.4.1) as those using the 9bp overhang containing MAX randomisation product as template. PCR products were visualised via agarose gel electrophoresis (2.2.1.3).

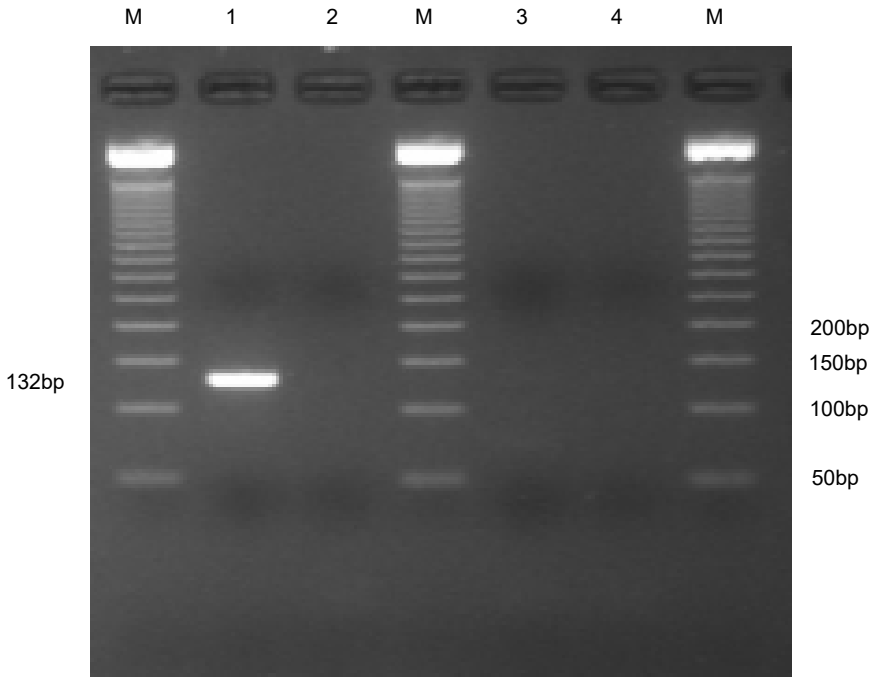

Figure 3.5.4.3: Comparison between the PCR amplification of positive and ligase-negative control repeat 4 MAX randomisation products.

Two individual PCR amplifications were performed using either 1/10 diluted repeat 4 MAX randomisation product or 1/10 ligase-negative control MAX randomisation repeat 4 products as template. Both PCR amplifications used an annealing temperature of 45°C. Lanes: M=50bp MW ladder, **1**. PCR amplification using repeat 4 MAX randomisation product as template, **2**. No template control, **3**. PCR amplification using ligase-negative repeat 4 MAX randomisation product as template, **4**. No template control.

A bright single repeat 4 construct band (132bp) was seen for the PCR amplification using the positive repeat 4 MAX randomisation product as template (Figure 3.5.4.3), with no band visible in the equivalent reaction using the non-ligase MAX randomisation product template. This meant that any product formed resulted from the specific, asymmetric amplification of the MAX codon-containing DNA strand.

3.6 Production of the repeat 5 construct for incorporation into a randomised arginine pocket DNA library

The production of each library construct was performed independently of each other as shown in Figure 3.2; the repeat 5 construct (purple, Figure 3.2) was next engineered.

3.6.1 Determining the optimal repeat 5 MAX randomisation product template dilution and annealing temperature for PCR amplification

In order to generate a high quality PCR product the optimal conditions for the repeat 5 PCR amplification had to be determined. As with repeat 4, this occurred in a two-pronged approach, ascertaining the optimal annealing temperature for the amplification while at the same time finding the optimal template dilution. The repeat 5 MAX randomisation product was diluted using milliQ water to form a template dilution range of 1/10, 1/100, 1/1000 – fold dilutions. Neat template and each of the template dilutions were used in an individual annealing temperature gradient PCR amplification, allowing for the determination of the optimal annealing temperature and template dilution in one set of experiments (2.2.4.1). PCR products were visualised via agarose gel electrophoresis (2.2.1.3).

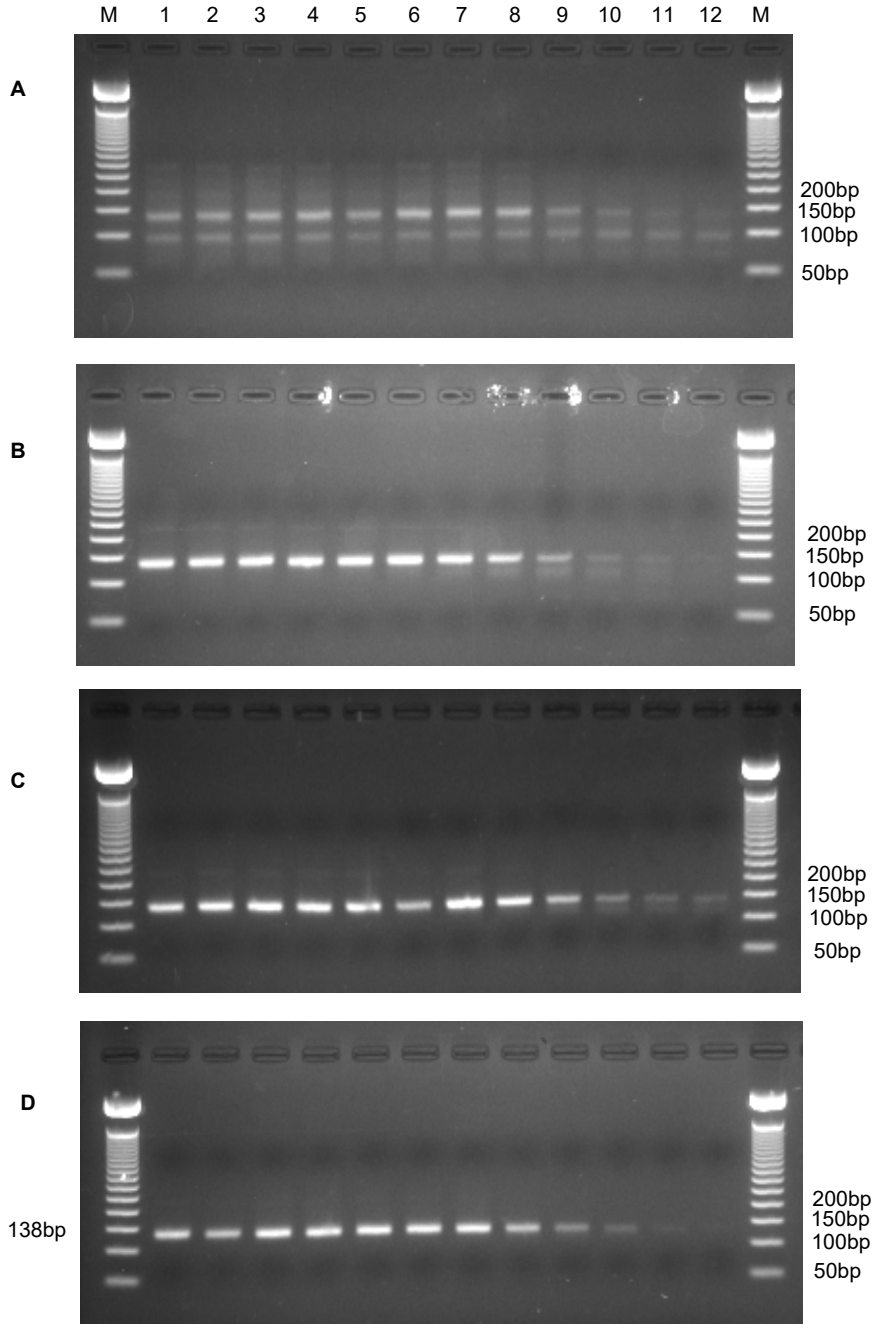

Figure 3.6.1: Four annealing temperature gradients using different MAX randomisation product as template at varying dilutions to determine optimal annealing temperature and template dilution for repeat 5 PCR amplification.

Four individual PCR reactions were made using (**A**) neat repeat 5 MAX randomisation product, (**B**) 1/10 repeat 5 MAX randomisation product, (**C**) 1/100 repeat 5 MAX randomisation product and (**D**) 1/1000 repeat 5 MAX randomisation as template and divided into 12 equal reactions. Annealing occurred at the temperatures described below. Lanes: M= 50bp MW ladder, annealing temperatures; **1**.45.0, **2**. 45.4, **3**.46.5, **4**.48.5, **5**.51.1, **6**.53.7, **7**.56.1, **8**.58.7, **9**.61.2, **10**.63.2, **11**.64.4,**12**. 65.0 (°C).

All the annealing temperatures tested in the amplification of the neat repeat 5 MAX randomisation product template, showed a prominent lower band of approximately 90bp (Figure 3.6.1A). Multiple bands larger than the repeat 5 construct (138bp) were present for temperatures 45.0- 58.7°C, becoming more defined as the annealing temperature increased. These upper bands were not present/very faint for annealing temperatures 61.2-65.0°C, but the 90bp product was consistent across the entire gradient. Accordingly, the use of neat template was not optimal and was therefore abandoned. The annealing temperature gradient using the 1/10 diluted repeat 5 MAX randomisation product as template had considerably less unwanted products than the neat using annealing temperature gradient, with only a faint band visible at ~225bp for annealing temperatures 45.0-56.1°C and a band at ~120bp seen for temperatures 56.1-65.0°C (Figure 3.6.1B). The upper band at ~225bp was also visible in Figure 3.6.1C, using 1/100 diluted template for temperatures 45-58.7°C, with the same lower band of ~120bp seen in Figure 3.6.1B also visible in Figure 3.6.1C for temperatures 61.2-65.0°C. These unwanted products meant both 1/10 and 1/100 dilutions were not optimal template dilutions for producing the repeat 5 construct. In comparison with the other template dilutions used, 1/1000 diluted template produced single bands of the correct size, 138bp (Figure 3.6.1D), identifying 1/1000 as the optimal template dilution. With 1/1000 dilution determined as the optimal template dilution for repeat 5, the annealing temperature 56.1°C was chosen based on Figure 3.6.1D as optimal, as it produced the brightest band.

3.6.2 Testing asymmetry of MAX randomisation of repeat 5 under optimised conditions

To determine if the 138bp product generated was a result of the specific asymmetric amplification of the MAX codon containing DNA strand, a ligase-negative control was constructed (2.2.2.2) and amplified using the optimal conditions of 56.1°C annealing and a 1/1000 dilution of MAX randomisation product as template (2.2.4.1). The PCR products were visualised via agarose gel electrophoresis (2.2.1.3).

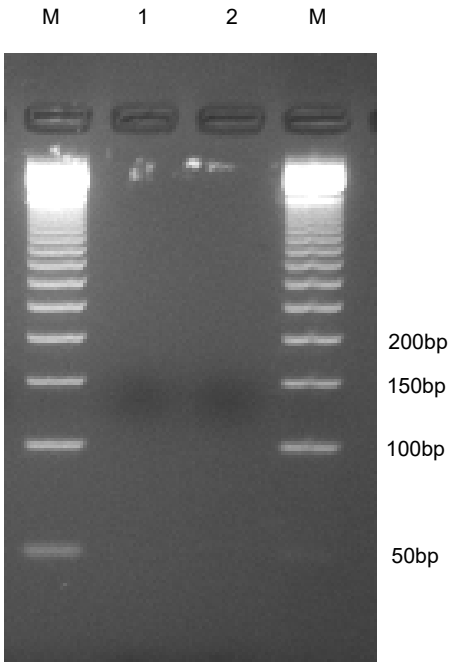

Figure 3.6.2: PCR amplification to examine asymmetric amplification of repeat 5.

A PCR amplification using 1/1000 diluted ligase-negative repeat 5 as template, with an annealing temperature of 56.1°C. Lanes: M=50bp MW ladder, **1**. Ligase-negative control, **2**. No template control.

Under these conditions no product was formed from the ligase-negative control (Figure 3.6.2), meaning repeat 5 production could be attributed to the specific, asymmetric amplification of the MAX codon-containing DNA strand of repeat 5.

3.6.3 Repeat 5 construct production for full length arginine library overlap PCR

The optimal PCR conditions for repeat 5 PCR amplification (2.2.4.1) were then used to generate sufficient PCR product of repeat 5 for the eventual overlap PCR between the constituent full length library constructs (Figure 3.2) to form the full library product (2.2.4.2). The PCR product was visualised via agarose gel electrophoresis (2.2.1.3).

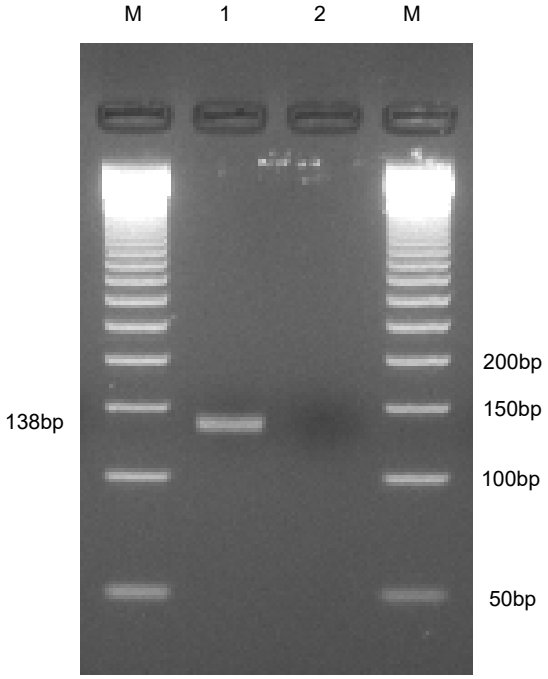

Figure 3.6.3: PCR amplification to generate repeat 5 construct.

A PCR amplification using 1/1000 diluted repeat 5 MAX randomisation product as template with an annealing temperature of 56.1°C. Lanes: M= 50bp MW ladder, **1**. Positive PCR reaction, **2**. No template control.

A bright single band of 138bp was generated, with a clean no-template control lane.

3.7 Production of the conserved region construct for incorporation into a randomised arginine pocket DNA library

As shown in Figure 3.2, with both repeat 4 and repeat 5 (green and purple respectively) successfully engineered, the final construct to be made was the conserved region (grey).

3.7.1 Determining the optimal conserved region annealing temperature for PCR amplification

To generate the conserved region, firstly an overlap PCR of its two constituent oligonucleotides was performed (2.2.4.2). This reaction would generate the template for the subsequent PCR amplification. The overlap PCR product was diluted to achieve a 1/1000 dilution template used for an annealing temperature gradient PCR to determine the optimal annealing temperature for generating the conserved region construct (2.2.4.1). A 1/1000 dilution was used as template to match that of the repeat 5 template dilution used. PCR products were visualised via agarose gel electrophoresis (2.2.1.3).

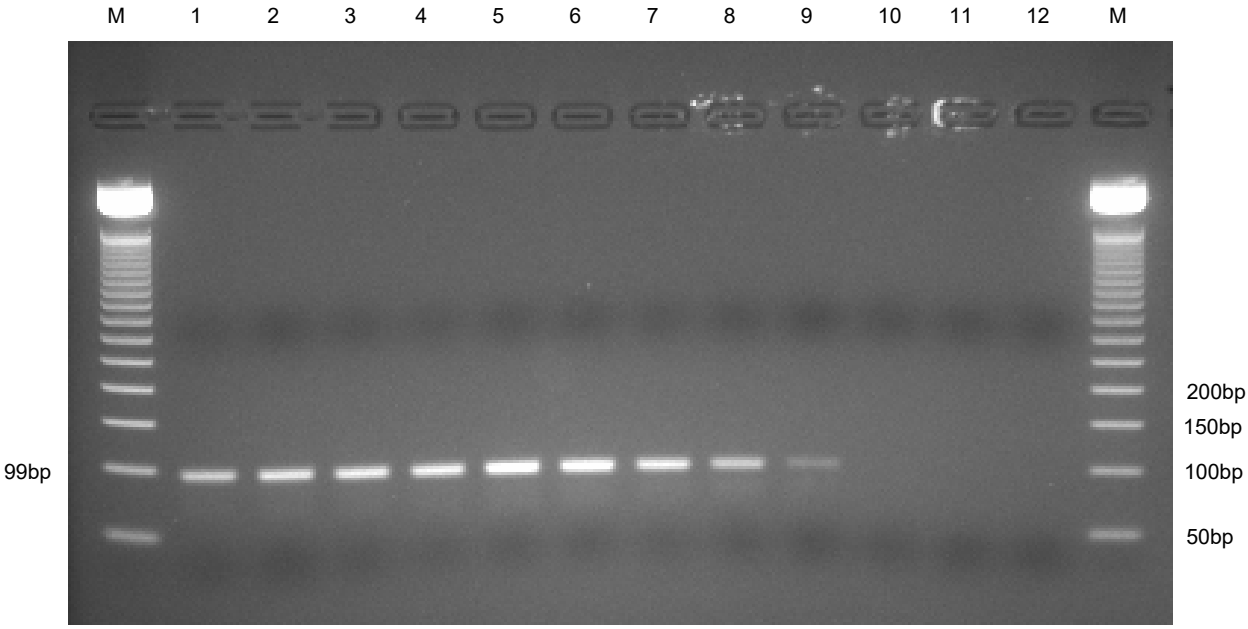

Figure 3.7.1: Annealing temperature gradient using conserved region overlap PCR product as template, to determine optimal conditions for PCR amplification.

A single PCR master mix was made, using 1/1000 diluted conserved region overlap PCR product as template across a range of annealing temperatures as described below. Lanes: M= 50bp MW ladder, annealing temperatures; **1**.45.0, **2**. 45.4, **3**.46.5, **4**.48.5, **5**.51.1, **6**.53.7, **7**.56.1, **8**.58.7, **9**.61.2, **10**.63.2, **11**.64.4, **12**. 65.0 (°C).

A bright 99bp product band was seen for annealing temperatures 45.0-56.1°C with the construct band becoming fainter for 58.7 and 61.2°C (Figure 3.7.1). No band was seen for temperatures 63.2-65.0°C. The brightest band was achieved using an annealing temperature of 51.1°C.

3.7.2 Conserved region construct for full length arginine library overlap PCR

With the optimal annealing temperature for conserved region PCR amplification determined, a final PCR amplification was performed (2.2.4.1) to generate sufficient conserved region construct product volume for the overlap PCR (2.2.4.2) to construct the full length arginine library cassette. The PCR product was visualised via agarose gel electrophoresis (2.2.1.3).

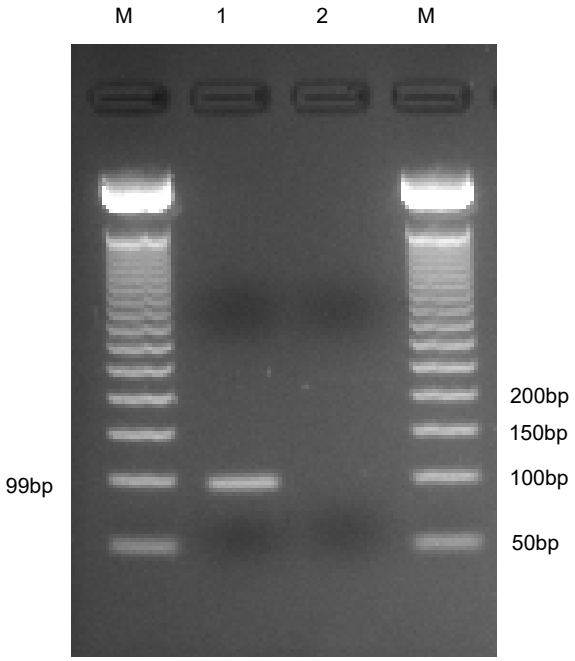

Figure 3.7.2: PCR amplification to generate conserved region construct.

A PCR amplification using 1/1000 diluted conserved region overlap PCR product as template with an annealing temperature of 51.1°C. Lanes: M= 50bp MW ladder, **1**. Positive PCR reaction, **2**. No template control.

A bright single band of 99bp was generated, with a clean no-template control lane.

3.8 Construction of complete randomised arginine DNA library

With the successful production of the conserved region, repeat 4 and repeat 5 (Figure 3.2, grey, green and purple respectively) the three individual constructs now needed to be joined together. This was achieved via an overlap PCR reaction (2.2.4.2) which exploited the complementary regions indicated by the orange brackets in Figure 3.2.

3.8.1 Determining the optimal template dilution and annealing temperature for PCR amplification of the complete cassette

The product formed from the overlap PCR between the three constructs (2.2.4.2) was diluted to form neat, 1/10, 1/100 and 1/1000 dilutions, to be used as template for PCR amplification (2.2.4.1).

To determine both optimal annealing temperature and template dilution, four individual annealing temperature gradients were conducted, each using a different template dilution (2.2.4.1). The PCR products were visualised via agarose gel electrophoresis (2.2.1.3).

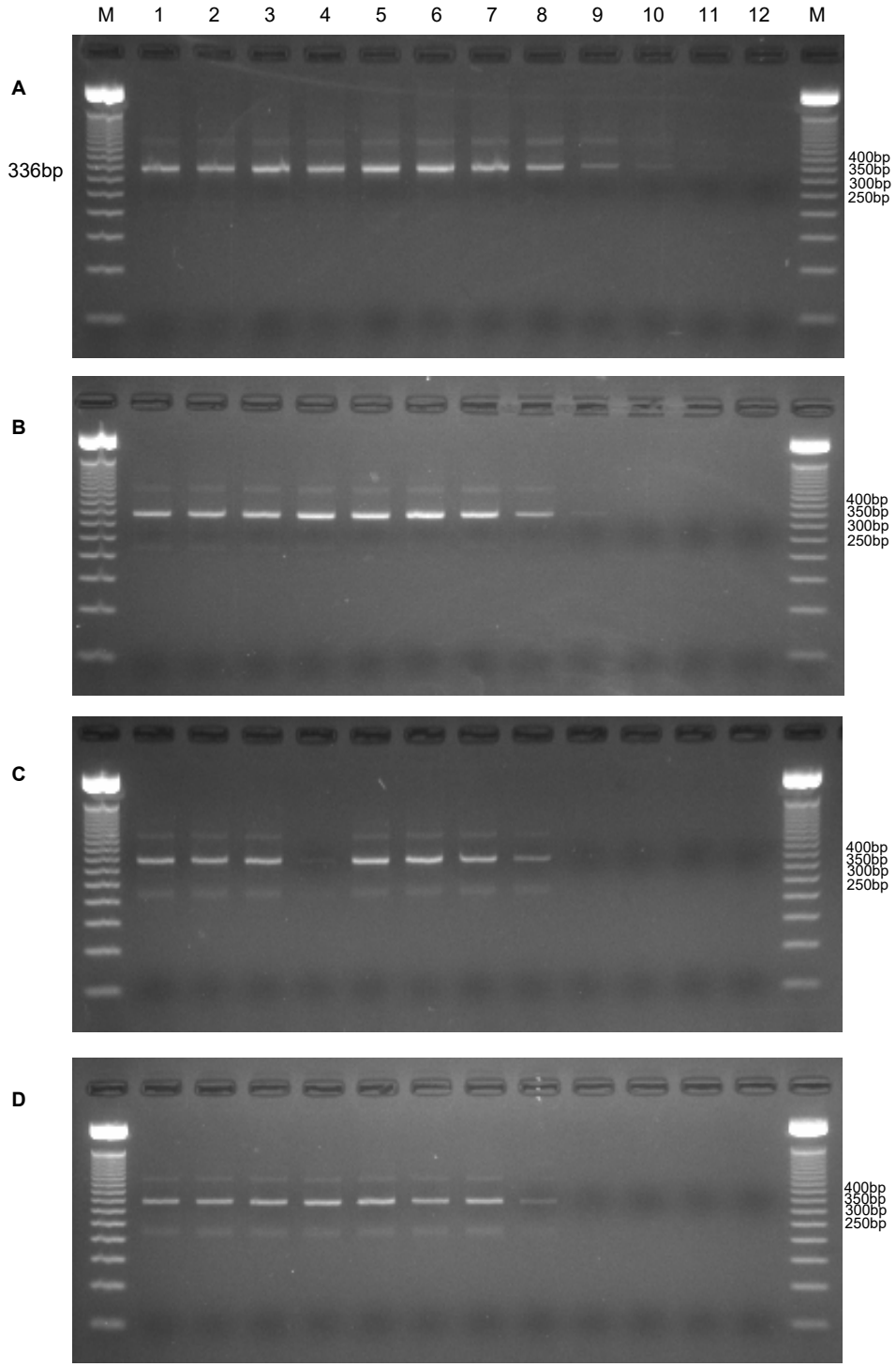

Figure 3.8.1: Four annealing temperature gradients using different full length library overlap product as template at varying dilutions to determine optimal annealing temperature and template dilution for complete library construct amplification. Four individual PCR reactions were made using neat overlap product (**A**), 1/10 diluted overlap product (**B**), 1/100 diluted overlap product (**C**) and 1/1000 diluted overlap product as template (**D**) and divided into 12 equal reactions. Annealing occurred at the temperatures described below. Lanes: M= 50bp MW ladder, annealing temperatures; **1**.45.0, **2**. 45.4, **3**.46.5, **4**.48.5, **5**.51.1, **6**.53.7, **7**.56.1, **8**.58.7, **9**.61.2, **10**.63.2, **11**.64.4,**12**. 65.0 (°C).

Upon comparison of each annealing gradient, all annealing temperatures producing a library product of 336bp also produced an upper band approximately 500bp. The annealing temperature gradient using neat template dilution (Figure 3.8.1A) also showed more smearing compared to the other annealing gradients, for annealing temperatures producing a library construct band. Because of this, the neat template was abandoned. A prominent lower band at approximately 225bp, could be seen in the annealing gradients using 1/10, 1/100 and 1/1000 diluted overlap product template (Figures 3.8.1B, 3.8.1C, 3.8.1D respectively). This lower band was faintest in the annealing temperature gradient using 1/10 diluted overlap product template, potentially caused by the lower band being hidden by the dye front (Figure 3.8.1B). The brightness of the library product band generated using a 1/10 template dilution was stronger than that of 1/100 and 1/1000, meaning the template dilutions of 1/100 and 1/1000 were abandoned, leaving the 1/10 dilution as the optimal template dilution. The annealing temperatures 51.1, 53.7 and 56.1°C in the 1/10 template dilution annealing gradient, produced a bright library construct band with a very faint 225bp band. The upper 500bp band had a consistent intensity, so determination of the optimal annealing temperature was based on the library construct band and the 225bp band intensity (Figure 3.8.1B). Based on this, the annealing temperature 56.1°C was selected as optimal.

3.8.3 Optimisation of PCR amplification of a randomised arginine DNA cassette

With none of the conditions tested thus far providing a complete library PCR product of a high enough quality, the selected conditions of 1/10 template dilution with an annealing temperature of 56.1°C were then used as the basis for downstream optimisation.

3.8.3.1 Reduction in PCR extension time to remove non-library products

In an attempt to remove the upper 500bp band seen in Figure 3.8.1B, the extension times within the PCR cycling were reduced (2.2.4.1) by 50%. The subsequent PCR product was visualised via agarose gel electrophoresis (2.2.1.3).

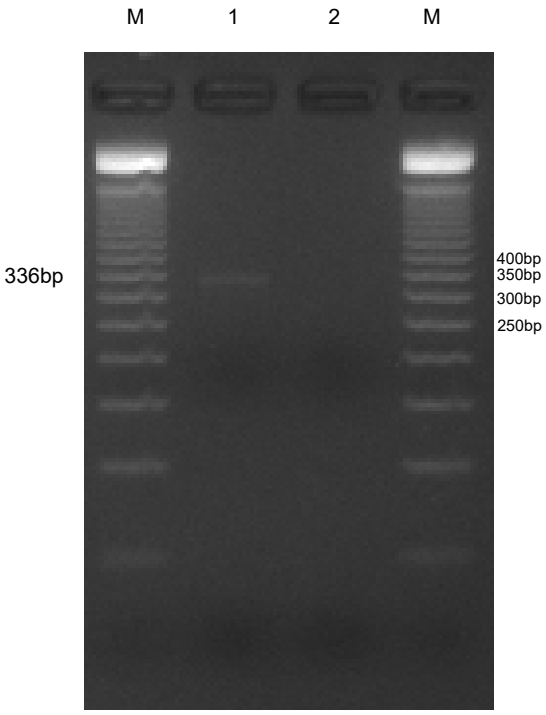

Figure 3.8.3.1: PCR amplification to generate full length library construct.

A PCR amplification using 1/10 full length library overlap PCR product as template using an annealing temperature of 56.1°C with reduced extension times of 10 seconds, with a final extension of 30 seconds. Lanes: M= 50bp MW ladder, **1**. Positive PCR reaction, **2**. No template control.

The reduction in extension time may have removed the upper 500bp band, but this could not be concluded with confidence (Figure 3.8.3.1) since a considerable decrease in library band intensity meant the upper band could still have been produced but was just too faint for agarose gel visualisation (2.2.1.3).

3.8.3.2 PCR cycle number and dNTP concentration investigations to improve production of randomised arginine library construct.

The poor library construct intensity had now become the greatest concern, so in attempt to improve product yield, two variables were investigated: an increase in cycle number and an increase in cycle number with a 50% reduction in dNTP concentration (2.2.4.1), with the decrease in dNTP concentration investigated alongside increased cycle numbers to pre-emptively attempt to reduce the production of non-product bands caused by extra PCR cycles, as the designed PCR amplification of the library would outcompete the amplification

of unwanted products. PCR products were visualised via agarose gel electrophoresis (2.2.1.3).

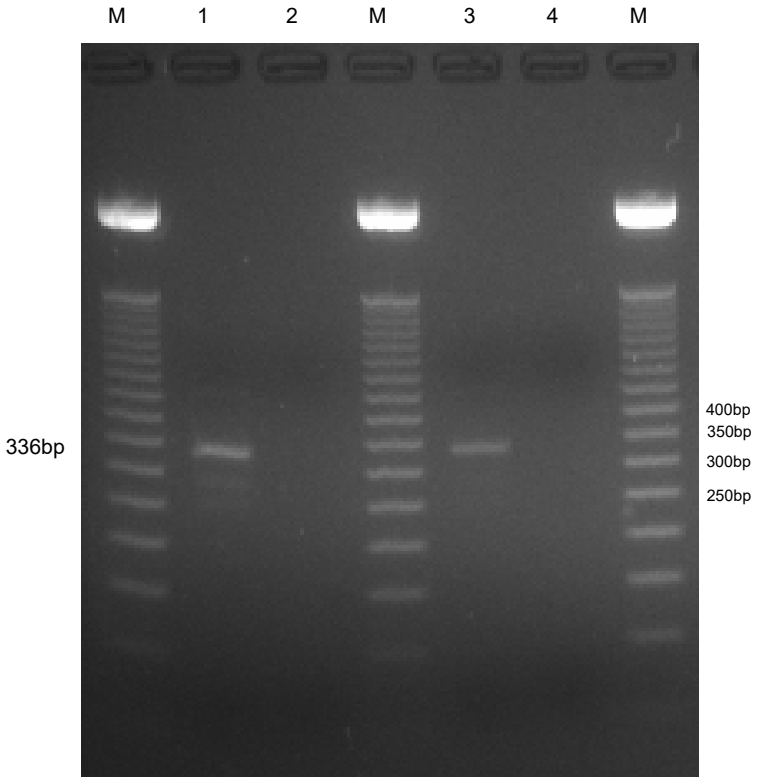

Figure 3.8.3.2: Comparison between PCR amplifications generating full length library product, using different end dNTP concentrations.

Two individual PCR amplifications were made both using; 1/10 full length library overlap PCR product as template, an annealing temperature of 56.1°C, reduced extension times of 10 seconds, 32 cycles and a final extension of 30 seconds. The only variable between the two PCRs being the final concentration of dNTPs. Lanes: M= 50bp MW ladder, **1**. Positive PCR reaction, **2**. No template control, **3**. Positive PCR reaction with 50% dNTPs concentration reduction, **4**. No template control.

A full length library product could be seen for both PCR amplifications (Figure 3.8.3.2). The 336bp construct produced using 32 cycles alone (lane 1 of Figure 3.8.3.2) was brighter than the band produced using 32 cycles and reduced dNTP concentration (lane 3 of Figure 3.8.3.2), but was accompanied by several other bands at approximately 260bp, 280bp and 425bp. The band produced using 32 cycles and the reduced dNTP concentration was a good

quality single band, thus the reduced dNTP concentration and 32 cycles conditions were used for downstream PCR amplification (2.2.4.1).

3.8.4 Complete randomised arginine DNA cassette production

A final, large-scale PCR amplification of the complete cassette was performed (2.2.4.1), with the PCR product visualised via agarose gel electrophoresis prior to purification (2.2.1.3). This was to produce an adequate volume of PCR product for purification (2.2.1.4), which was then quantified (2.2.1.5). The purified PCR product was then sent for Next Generation Sequencing (2.2.1.6).

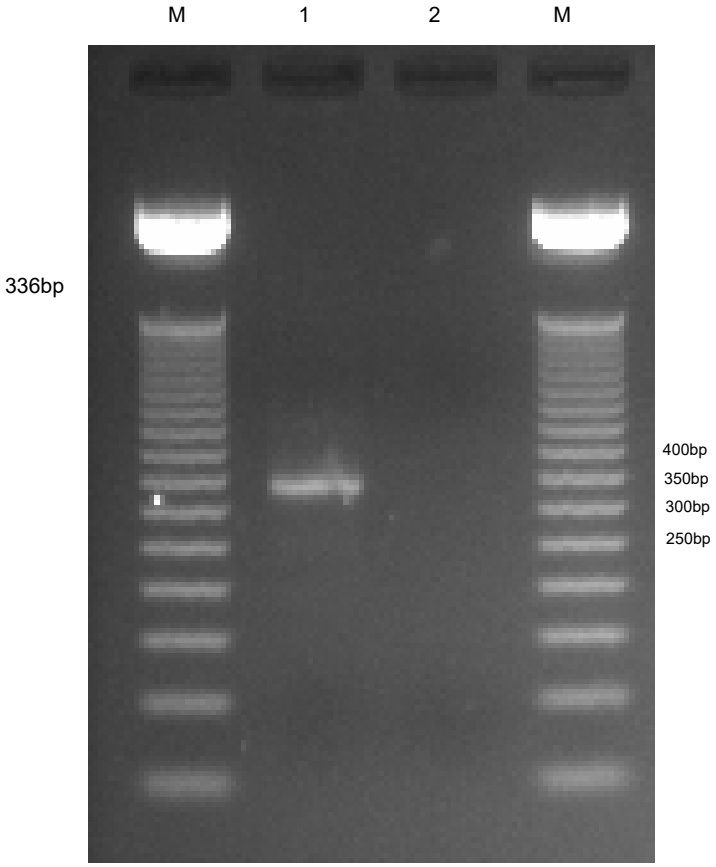

Figure 3.8.4: PCR amplification to generate full length library construct.

A PCR amplification using 1/10 full length library overlap PCR product as template and a 50% reduction in final dNTP concentration. The annealing temperature of 56.1°C was used, alongside reduced extension times of 10 seconds within the 32 cycles and a final extension of 30 seconds. Lanes: M= 50bp MW ladder, **1**. Positive PCR reaction, **2**. No template control.

3.9 Analysis of Next Generation Sequencing data to assess amino acid representation at saturated positions in an arginine DNA cassette

The arginine library cassette PCR product was sequenced at Genewiz (2.2.1.6) using the Amplicon EZ service, which utilises two 250bp reads running 5'-3 and 3'-5 to provide full library coverage. The data output of the Miseq Illumina sequencing was processed and analysed (2.2.6) then presented graphically (Figure 3.9).

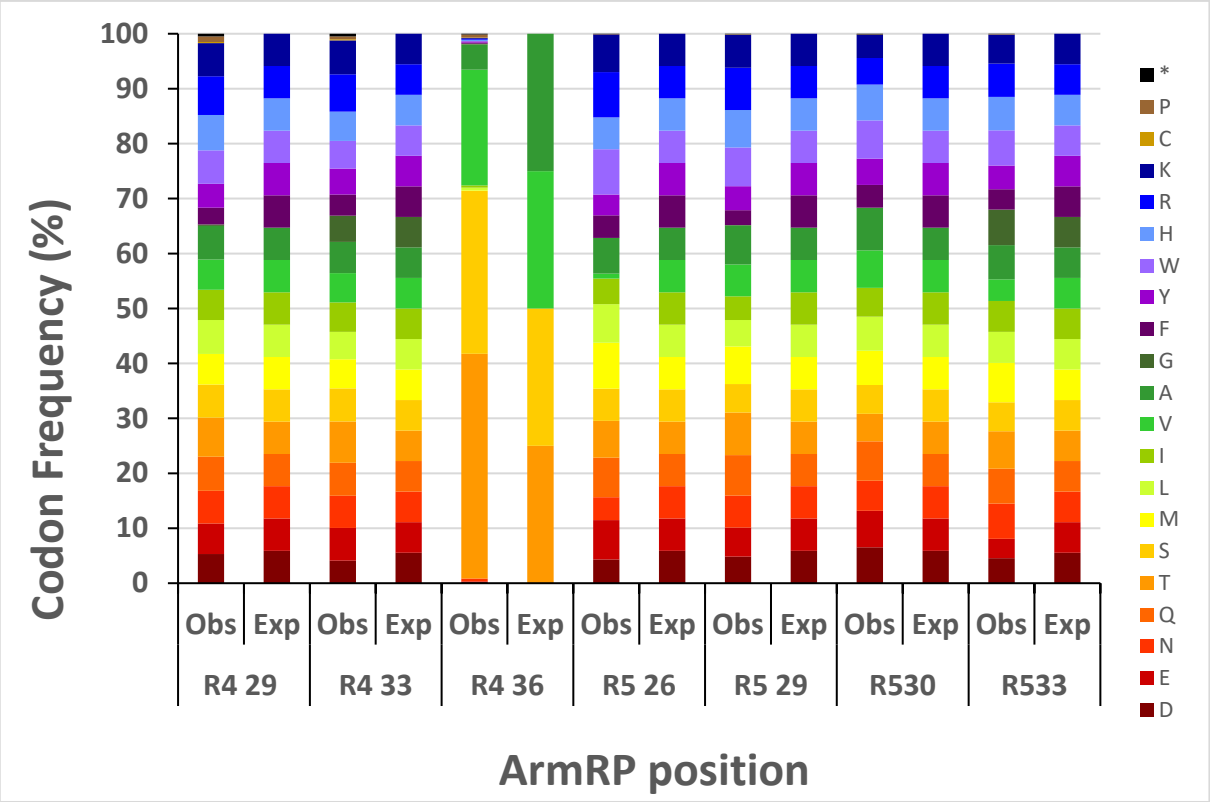

Figure 3.9: Observed vs expected amino acid distribution for positions randomised in an arginine designed armadillo repeat protein library.

Letters in the legend correspond to universal amino acids abbreviations in the genetic code: P: proline; C: cysteine; K: lysine; R: arginine; H: histidine; W: tryptophan; Y: tyrosine; F: phenylalanine; G: glycine; A: alanine; V: valine; I: isoleucine; L: leucine; M: methionine; S: serine; T: threonine; Q: glutamine; N: asparagine; E: glutamic acid; D: aspartic acid; *: stop codon. Raw data from Miseq Illumina sequencing. (See Appendix 2 for raw count data)

Upon visual inspection of the observed and expected frequencies, a primary evaluation that the observed and expected codon frequencies matched well for positions R4-29, R4-33, R5-26, R5-29, R5-30 and R5-33 was made (Figure 3.9). This was assessed statistically using the Chi-square Goodness-of-fit Test. In order to use the Chi-square Goodness-of-fit Test to determine if the observed frequencies of each amino acid at each of the randomised positions statistically fit the expected relevant distribution, the raw counts were used (Appendix 2). The amino acids observed at each position, but were not expected (mainly glycine, cysteine, proline and stop codons), had to be removed from the analysis, as part of the calculation involves dividing by the expected value, not possible if the value is 0. The Chi-square Goodness-of-fit Test was performed using Graph Pad on each position individually. Each of the 7 tests resulted in a two-tailed P value less than 0.0001 meaning the difference between the observed and expected proportions was significant and not due to chance.

This was suspected, as the MAX randomisation process is an incredibly sensitive process and to have a statistically perfect amino acid representation is practically impossible; this is why the library randomisation was assessed primarily through a visual representation (Figure 3.9). An example of this sensitivity is the over representation of threonine at position R4-36, most likely caused by a minute unavoidable increased threonine MAX oligonucleotide concentration in the MAX selection oligonucleotide pool, with the opposite the potential cause for the alanine under representation at the same position. As the randomised single arginine library was to be used to discover novel binder hits, the key concern was any obvious issue with the library for example, a large amount of stop codons which would have been detrimental in protein screening, while maximising library diversity.

As there were no concerning discrepancies in the library seen in the visual comparison of the observed and expected frequencies (Figure 3.9), the library quality was considered satisfactory for use in protein binder investigations.

3.10 Screening outputs from the randomised single arginine DNA library

With the successful completion of the randomised single arginine cassette, investigating the resulting protein library began. Using yeast surface display and *Kd* determination, the Plückthun group identified multiple binders of interest. The wild-type specificity of the pocket, originally arginine, had been successfully switched to tyrosine (*Kd*: 7 Nm, sequence: KEVLIRG) and tryptophan (*Kd*: 64 nm, sequence: TATAWRT). Other library hits included a binder that recognised the large hydrophobic residues: isoleucine, leucine and methionine (with *Kd* values: 28 nM, 43 nM and 61 nM respectively) and another that recognised the hydrogen bond donors histidine and threonine (*Kd* values: 142 nM and 162 nM respectively).

An important observation in the threonine binder hits was the common occurrence of leucine at position R4-33 and glutamine at position R4-36, matching the computational predictions of the Höcker group who provided the computational predictions for a new threonine binder.

Because of the feedback loop used by the PRe-ART consortium, the data acquired by the Plückthun group from the single arginine library was then used by the Höcker group as both source material for new computational predictions as well as experimental validation of already existing predictions.

3.11 Concluding the mutagenesis of the designed armadillo repeat protein arginine pocket

MAX randomisation was successfully used to engineer a saturated designed armadillo cassette, containing seven positions of saturation spanning two internal repeats of the protein, with a theoretical library size of $1.08x10^8$. Screening of the resulting protein library, via yeast display, conducted by the Plückthun group, identified multiple successful binders, demonstrating altered specificity from the wild-type pocket's arginine preference. *Kd* values in the nM region were determined for novel tyrosine, tryptophan, isoleucine, leucine, methionine, histidine and threonine binders, providing experimental data for the next rounds of binder engineering and validation material for computational designs from the Höcker group.

Chapter 4 Mutagenesis of a synthetic designed armadillo repeat protein arginine pocket to generate an Arg→Thr binding pocket: the Atligator-Threonine Library

4.1 Introduction

As described in Chapter 3, the $Y_{II}M_5A_{II}$ designed armadillo repeat protein contains five internal M repeats responsible for the recognition and binding of the target peptide sequence. The target peptide sequence of the original $Y_{II}M_5A_{II}$ designed armadillo repeat protein was $(KR)_5$, with each H3 groove recognising a KR dipeptide. This chapter focuses on creating a DNA library engineered to generate a 'threonine-specific pocket', i.e. switching the specificity from KR to KT in the context of a longer peptide (target sequence changed from KRKRKRKRKR to KRKRK**T**KRKR). The protein library design was based on computational predictions generated by ATLIGATOR. ATLIGATOR (Atlas-based LIGand binding site ediTOR) is a computational prediction programme specifically designed to investigate protein-peptide interactions with the overall aim to design new binding interfaces, by allowing the analysis of interaction modes of two or more amino acids. ATLIGATOR uses Protein Data Bank submissions, filtered by the user based on factors such as: protein family, binder/ligand length and presence of secondary structure along with the assumption that natural pairwise interactions have evolved to be most favourable, to create an 'atlas' from structures relevant to the protein-peptide interaction of interest. Distances associated with ionic, aromatic, hydrogen and other interaction types are used to determine if a genuine interaction is occurring, with the relevant amino acids then assigned co-ordinates allowing for pair wise interaction patterns to be mapped. ATLIGATOR is then capable of extracting common pairwise interactions for a single ligand residue, forming a 'pocket' which can then be grafted to binder scaffolds and investigated by the user.

Accordingly, this construction was named the ATLIGATOR-Threonine (At-Thr) library. The positional targets were identical to the original single arginine library (Figure 4.1), with the major divergences predicted by ATLIGATOR, being the focused amino acid selections for positions R4-33 (positionally fixed for leucine) and R4-36 (fixed for asparagine or glutamine).

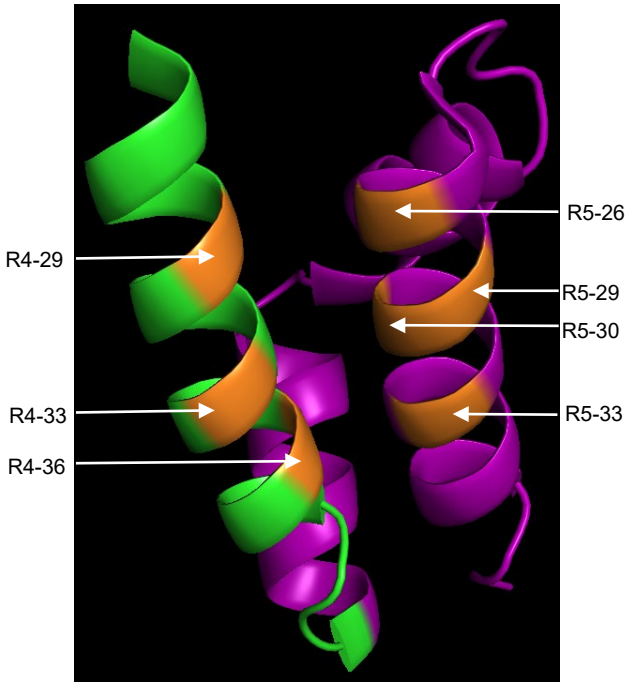

Figure 4.1: Visualisation of the binding groove formed from the H3 helices of repeat 4 and repeat 5 of the designed armadillo repeat protein. Key residues for peptide binding coloured orange in H3 helix of repeat 4 (green) (positions 29,33,36) and in the H3 helix of repeat 5 (purple) (positions 26, 29, 30, 36). Adapted from (Hansen *et al*., 2016).

4.2 At-Thr Library design to facilitate MAX randomisation

The At-Thr library design was a focused version of the single arginine library, with the only difference occurring in the MAX selection oligonucleotide pools for saturating target positions. Because of this, the At-Thr library contained an identical conserved region from repeat three of the designed armadillo repeat protein to allow for downstream homologous recombination (grey), a repeat 4 and repeat 5 region (green and purple respectively), with all three individual constructs to be joined via overlap PCR using complementary regions between constructs (orange brackets) to form the full length library product (Figure 4.2).

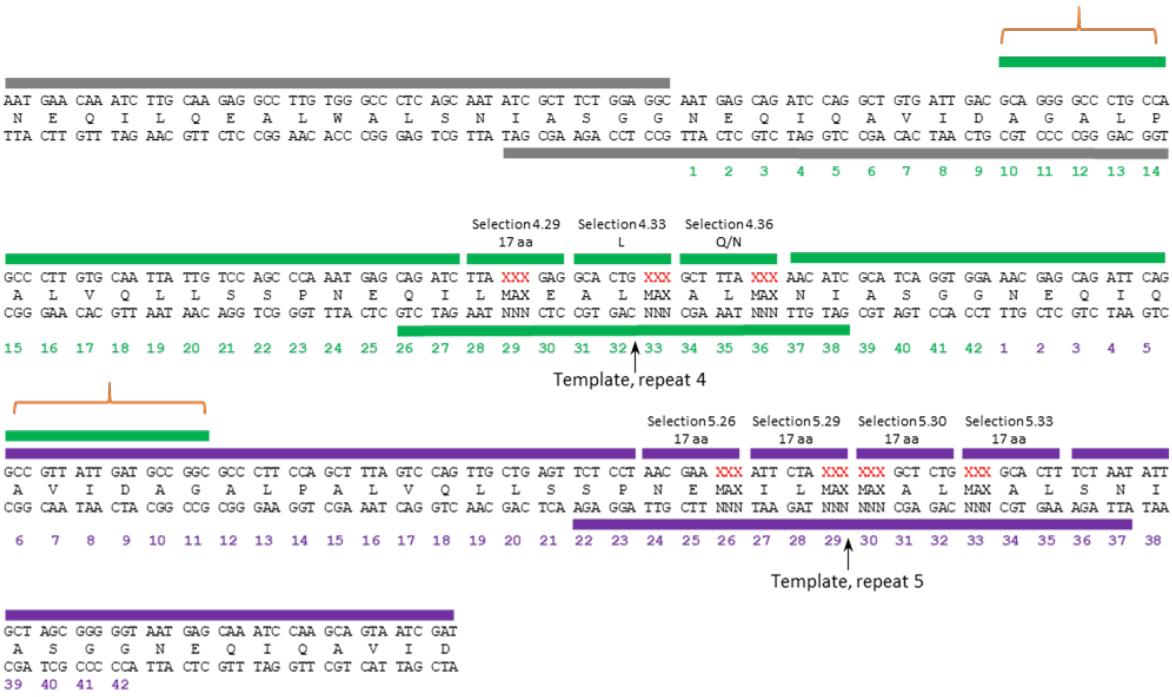

Figure 4.2: DNA sequence for internal repeat 3, 4 and 5 of the designed armadillo repeat protein, overlaid with the MAX randomisation library design. The DNA sequence is split into three separate constructs, a conserved region (grey), repeat 4 (green) and repeat 5 (purple), with positions of saturation highlighted red. Overlapping regions between constructs are indicated by orange brackets.

4.3 Generation of MAX selection oligonucleotide pools for positional saturation of key binding residues in the At-Thr library

Before commencing library construction, pools of MAX selection oligonucleotides were needed for each targeted position. As *S. cerevisiae* was the organism of choice for protein expression, codons selected to represent each amino acid were based on *S. cerevisiae* preference. Determination of codons was done using GenScript Codon Frequency Table(chart) Tool- *Saccharomyces cerevisiae* (gbpln) (2.2.2.1). The codons with the highest usage frequency were chosen except for lysine, where the preferred AAA codon was substituted for AAG, in an attempt to combat any ligation bias caused by excessive consecutive adenine bases during MAX randomisation. The required MAX selection oligonucleotides to represent the desired amino acids at each position (Table 4.3; noting that in this library, glycine, cysteine and Proline were excluded from all positions) were mixed to

form seven different MAX selection oligonucleotide pools, one for each target position (2.2.2.1). (See Appendix 1 for non-MAX selection oligonucleotide oligonucleotides used for engineering the randomised At-Thr library).

	Position	R4-29	R4-33	R4-36	R5-26	R5-29	R5-30	R5-33
	Invariant sequence showing MAX position	TTAMAXGAG	GCACTGMAX	GCTTTAMAX	AACGAAMAX	ATTCTAMAX	MAXGCTCTG	MAXGCACTT
Identity of MAX codon	Aspartic acid	GAT			GAT	GAT	GAT	GAT
	Glutamic acid	GAA			GAA	GAA	GAA	GAA
	Aspargine	AAT		AAT	AAT	AAT	AAT	AAT
	Glutamine	CAA		CAA	CAA	CAA	CAA	CAA
	Threonine	ACT			ACT	ACT	ACT	ACT
	Serine	TCT			TCT	TCT	TCT	TCT
	Methionine	ATG			ATG	ATG	ATG	ATG
	Leucine	TTG	TTG		TTG	TTG	TTG	TTG
	Isoleucine	ATT			ATT	ATT	ATT	ATT
	Valine	GTT			GTT	GTT	GTT	GTT
	Alanine	GCT			GCT	GCT	GCT	GCT
	Phenylalanine	TTT			TTT	TTT	TTT	TTT
	Tyrosine	TAT			TAT	TAT	TAT	TAT
	Tryptophan	TGG			TGG	TGG	TGG	TGG
	Histidine	CAT			CAT	CAT	CAT	CAT
	Arginine	AGA			AGA	AGA	AGA	AGA
	Lysine	AAG			AAG	AAG	AAG	AAG
	Glycine							
	Cysteine							
	Proline							

Table 4.3: Table of MAX selection oligonucleotides used to saturate target positions in repeat 4.

Each MAX selection oligonucleotide was received at 50 µM in water. To make the MAX selection oligonucleotide pool for each position, an equal volume and therefore concentration of each MAX selection oligonucleotide was mixed. The resulting end concentration of each selection pool was 50 µM, with individual MAX selection oligonucleotide concentration at 2.94 µM (2. dp) for R4-29, R5-26, R5-29, R5-30 and R5-33, 50.0 µM (2. dp) for R4-33 and 25.0 µM (2.dp) for R4-36.

4.4 Production of the repeat 4 construct for incorporation into a randomised ATLIGATOR-threonine DNA library

Each of the three library components were to be made separately as shown in Figure 4.2, with initial focus on the repeat 4 construct (green, Figure 4.2).

4.4.1 Determining the optimal At-Thr repeat 4 MAX randomisation product template dilution and annealing temperature for PCR amplification

To achieve successful production of a high quality PCR product, the optimal repeat 4 PCR amplification conditions had again to be determined. Ascertaining the optimal annealing temperature for the amplification and finding the optimal template dilution were once more the first stages of optimisation. The repeat 4 MAX randomisation product template was produced via MAX randomisation (2.2.2.2) and then diluted in milliQ water to form a template dilution range of 1/10, 1/100, 1/1000 – fold dilutions. Individual annealing temperature gradient PCR amplifications, each using one template dilution from the dilution range allowed for the determination of the optimal annealing temperature and template dilution in one set of experiments (2.2.4.1). PCR products were visualised via agarose gel electrophoresis (2.2.1.3).

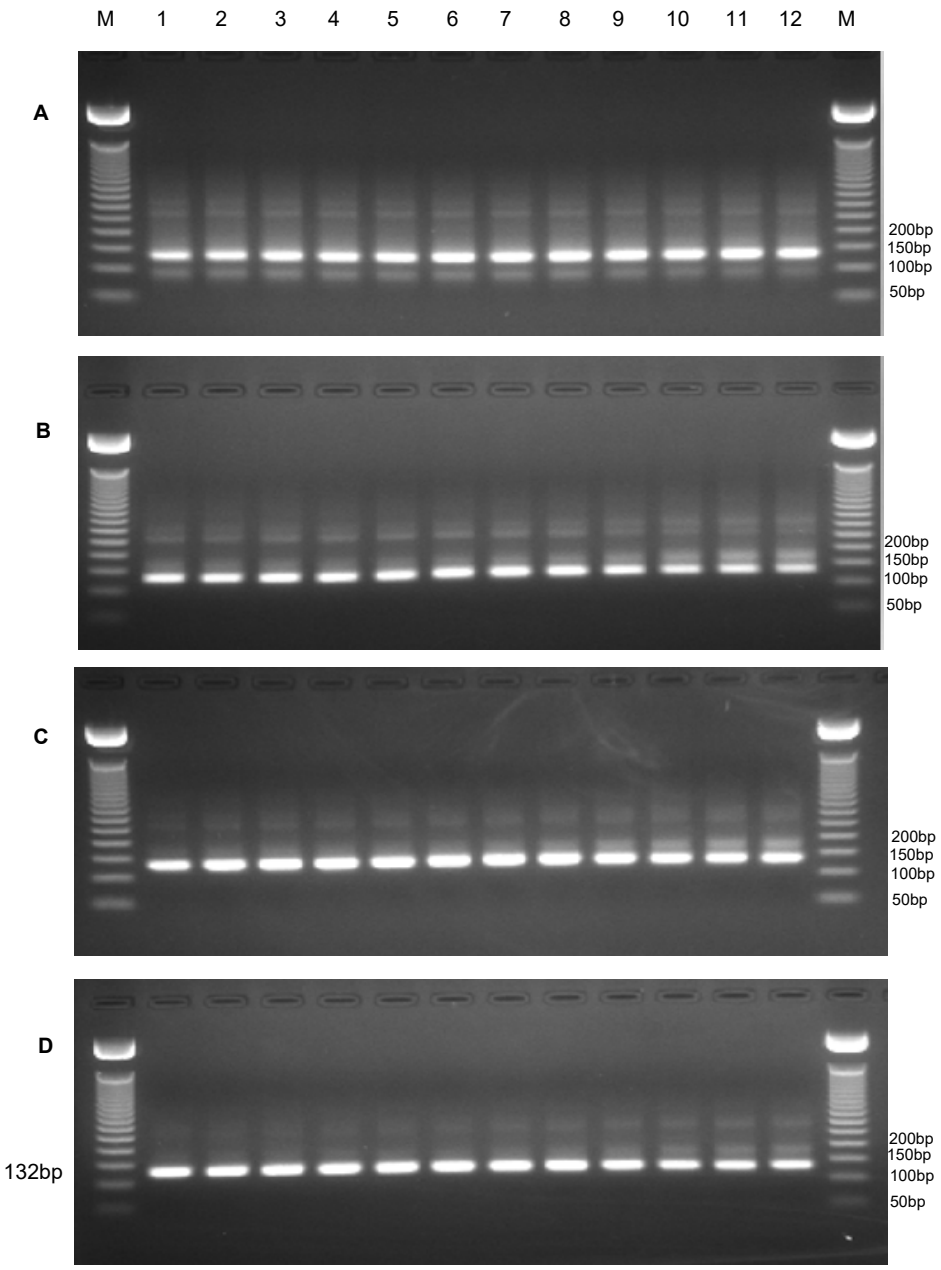

Figure 4.4.1.1: Four annealing temperature gradients using different MAX randomisation product as template at varying dilutions to determine optimal annealing temperature and template dilution for repeat 4 PCR amplification.

Four individual PCR reactions were made using neat MAX randomisation product (**A**), 1/10 diluted MAX randomisation product (**B**), 1/00 diluted MAX randomisation product (**C**) and 1/1000 diluted positive MAX randomisation as template (**D**) and divided into 12 equal reactions. Annealing occurred at the temperatures described below, with the in-cycle extension lasting 1 minute and a final extension of 5 minutes. Lanes: M= 50bp MW ladder, annealing temperatures; **1**.45.0, **2**. 45.4, **3**.46.5, **4**.48.5, **5**.51.1, **6**.53.7, **7**.56.1, **8**.58.7, **9**.61.2, **10**.63.2, **11**.64.4,**12**. 65.0 (°C).

A bright product band of 132bp was seen across the entire annealing temperature gradient using each of the four different MAX randomisation product template dilutions. Figure 4.4.1.1A shows two other prominent products at 90bp and 275bp for all annealing temperatures tested, with extensive smearing present. The smearing intensity seen in Figure 4.4.1.1A becomes fainter as the template dilution increases seen across Figures 4.4.1.1B, C and D, with individual bands becoming visible at 175bp, 260bp, and 310bp. The intensity of these bands are therefore faintest in Figure 4.4.1.1D using 1/1000 diluted template. This meant 1/1000 was selected as the optimal template dilution for downstream PCRs. The intensity of these larger non-repeat 4 constructs increased as the annealing temperature rose. This meant alongside the template dilution of 1/1000, the annealing temperature 45.0°C was determined to be optimal.

These optimal conditions were then used to produce repeat 4 construct (2.2.4.1) for downstream use in generating the full At-Thr library construct. PCR products were visualised via agarose gel electrophoresis (2.2.1.3).

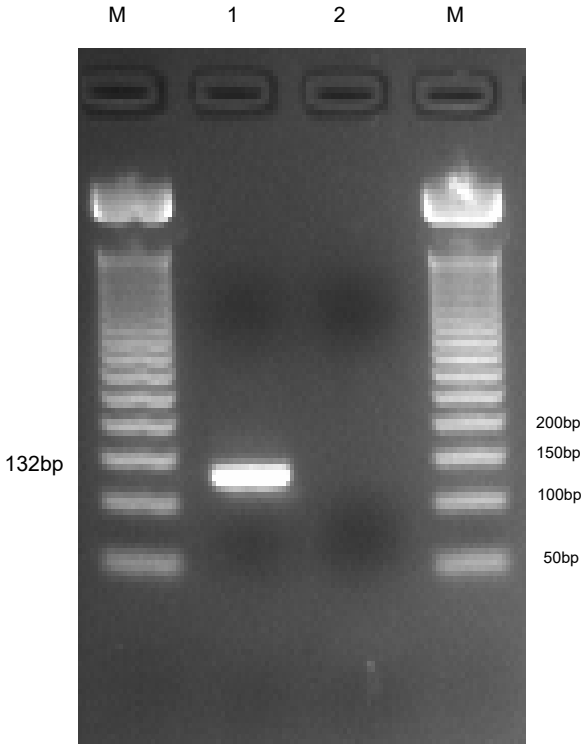

Figure 4.4.1.2: PCR amplification to generate repeat 4 construct.

A PCR amplification using 1/1000 diluted repeat 4 MAX randomisation product as template with an annealing temperature of 45.0°C, with the in-cycle extension lasting 1 minute and a final extension of 5 minutes. Lanes: M= 50bp MW ladder, **1**. Positive PCR reaction, **2**. No template control.

A bright single band of 132bp was seen, with a clean no template control.

4.5 Production of the repeat 5 construct for incorporation into a randomised At-Thr DNA library

With the completion of repeat 4 construction, the production of repeat 5 began (purple, Figure 4.2).

4.5.1 Determining the optimal At-Thr repeat 5 MAX randomisation product template dilution and annealing temperature for PCR amplification

As with the production of the repeat 4 construct, ascertaining the optimal annealing temperature for the amplification and determining the optimal template dilution were the first stages of PCR optimisation. The repeat 4 MAX randomisation product template was produced via MAX randomisation (2.2.2.2) and then diluted in milliQ water to form a template dilution range of 1/10, 1/100, 1/1000 – fold dilutions. Four individual annealing temperature gradient PCR amplifications were then performed with each using one template dilution from the dilution range. This allowed for the determination of the optimal annealing temperature and template dilution simultaneously (2.2.4.1). PCR products were visualised via agarose gel electrophoresis (2.2.1.3).

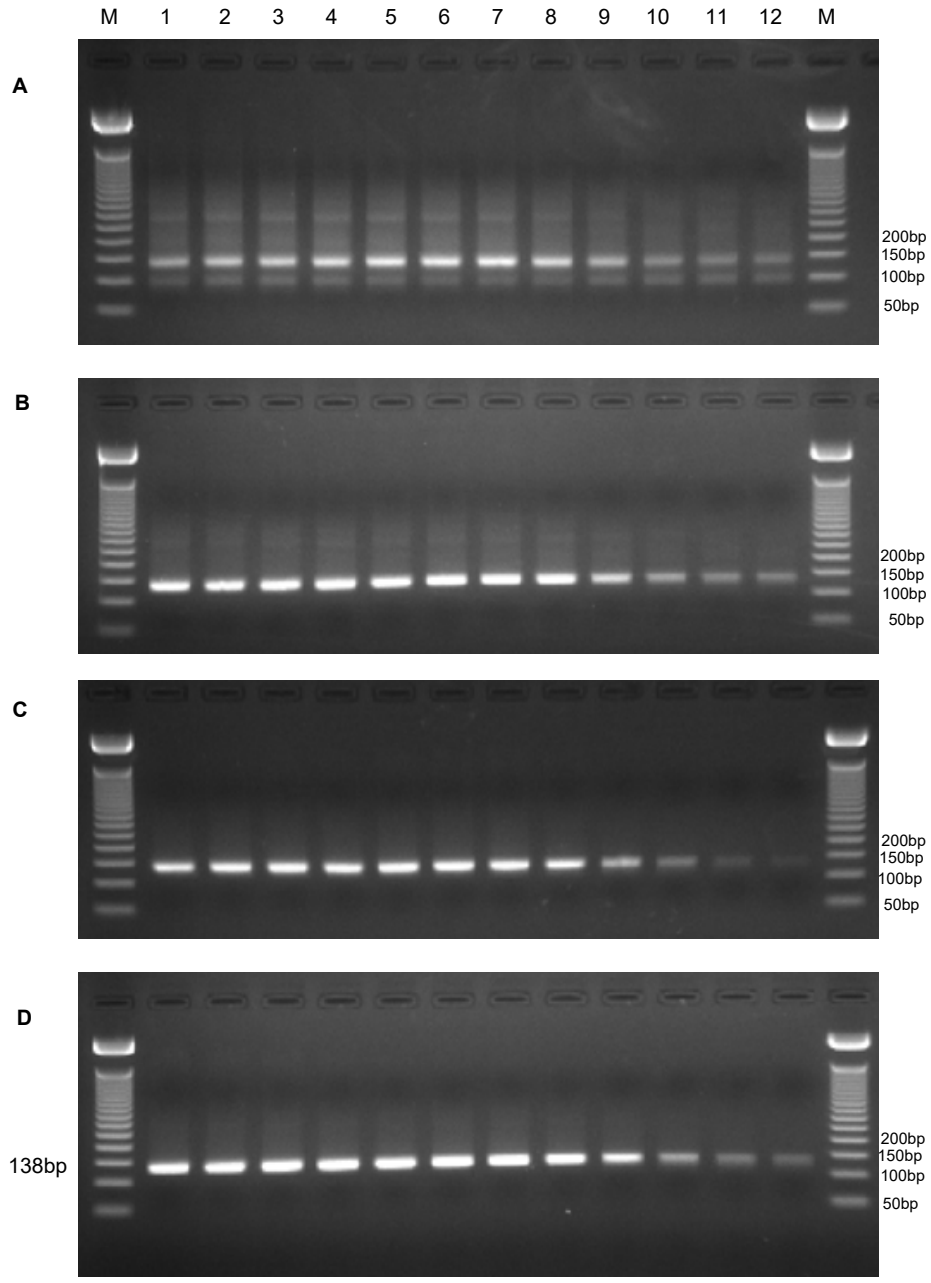

Figure 4.5.1: Four annealing temperature gradients using different MAX randomisation product as template at varying dilutions to determine optimal annealing temperature and template dilution for repeat 5 PCR amplification.

Four individual PCR reactions were made using neat MAX randomisation product (**A**), 1/10 diluted MAX randomisation product (**B**), 1/100 diluted MAX randomisation product (**C**) and 1/1000 diluted positive MAX randomisation as template (**D**) and divided into 12 equal reactions. Annealing occurred at the temperatures described below, with the in-cycle extension lasting 1 minute and a final extension of 5 minutes. Lanes: M= 50bp MW ladder, annealing temperatures; **1**.45.0, **2**. 45.4, **3**.46.5, **4**.48.5, **5**.51.1, **6**.53.7, **7**.56.1, **8**.58.7, **9**.61.2, **10**.63.2, **11**.64.4,**12**. 65.0 (°C).

A 138bp library construct band could be seen for all annealing temperatures tested, across all four different template dilutions (Figure 4.5.1). The brightness of this library construct followed the same pattern in each of the annealing gradients, in that the band intensity decreased between 63.3-65.0°C. All of the annealing temperatures producing a bright library construct band (temperatures 45.0-61.2°C) across all four annealing gradients, possessed some degree of smearing. The level of smearing was the distinguishing factor used to identify the optimal template dilution to be used in subsequent repeat 5 PCR amplifications. The prominent 100bp band across the entire annealing gradient in Figure 4.5.1A, along with the extensive smearing and visible 275bp band within the smear, meant using neat template for amplification was not viable. Even with the smearing being reduced in comparison to the neat template annealing gradient (Figure 4.5.1A), the annealing gradient using 1/10 template dilution (Figure 4.5.1B) was also abandoned, as the smearing present was still substantial. The library construct bands and the degree of smearing produced using 1/100 (Figure 4.5.1C) and 1/1000 (Figure 4.5.1D) template dilutions were comparable. The library construct bands produced using 1/1000 diluted repeat 5 MAX randomisation product as template, were slightly brighter than their 1/100 counterparts, thus the 1/1000 template dilution was identified as the optimal template dilution. With no substantial difference in band quality for annealing temperatures 45.0-58.7°C visible in the annealing temperature gradient using 1/1000 template (Figure 4.5.1D), the highest of these temperatures, 58.7°C was chosen as the annealing temperature for downstream PCR amplifications.

The optimal conditions of a 58.7°C annealing temperature using a 1/1000 dilution of repeat 5 MAX randomisation product as template, were then used to produce the repeat 5 construct (2.2.4.1) for the eventual full library overlap PCR (2.2.4.2), to generate the full library product. PCR products were visualised via agarose gel electrophoresis (2.2.1.3).

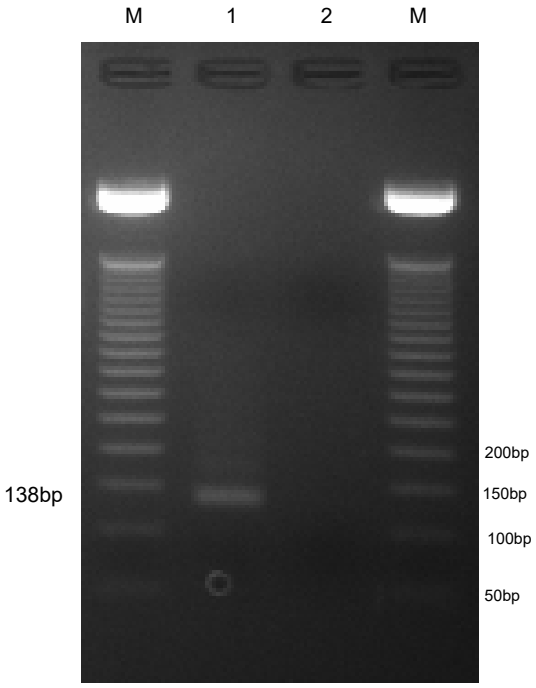

Figure 4.5.2: PCR amplification using repeat 5 MAX randomisation product as template.

PCR amplification using 1/1000 diluted MAX randomisation repeat 4 product as template, with an annealing temperature of 58.7°C with the in-cycle extension lasting 1 minute and a final extension of 5 minutes.. Lanes: M=50bp MW ladder, **1**. Positive PCR reaction, **2**. No template control.

The PCR amplification using 1/1000 dilution of repeat 5 MAX randomisation as template produced an expected 138bp band (Figure 4.5.2). A small amount of smearing was also seen, but it was considerably fainter than the repeat 5 band and therefore not a concern. The brightness of the repeat 5 construct was not as bright as repeat 4 (Figure 4.4.1.2), but was not a concern as enough PCR product had been visualised and therefore been produced to cover all of the randomised codon combinations (4 positions each saturated with 17 different MAX selection oligonucleotides: 17^4 = 83521 combinations).

4.6 Production of the conserved region construct for incorporation into a randomised At-Thr DNA library

To generate the conserved region, firstly an overlap PCR of its two constituent oligonucleotides was performed (2.2.4.2), (Figure 4.2). This reaction would generate the template for the subsequent PCR amplification (2.2.4.1).

4.6.1 Using optimal conditions to produce the conserved region for the At-Thr library

Since the conserved region was identical to that produced for the randomised single arginine pocket library, the conditions previously optimised (Figure 3.7.1) i.e. an annealing temperature of 51.1°C and a 1/1000 dilution of the template were repeated (2.2.4.1) and the product visualised via agarose gel electrophoresis (2.2.1.3).

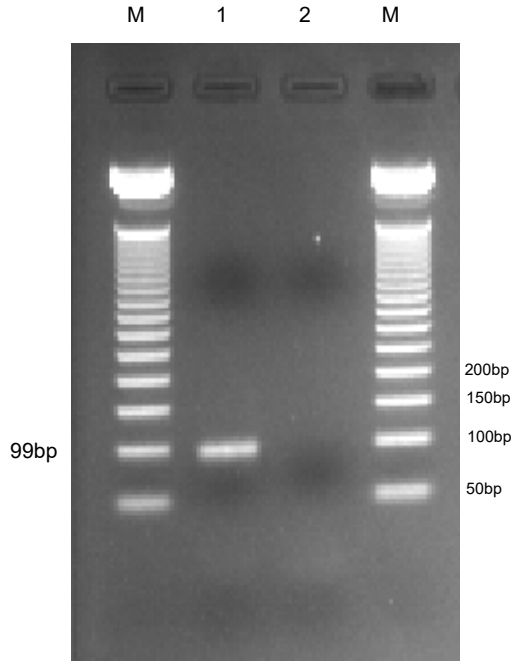

Figure 4.6.1: PCR amplification to generate conserved region construct.

A PCR amplification using 1/1000 diluted conserved region overlap PCR product as template with an annealing temperature of 51.1°C. Lanes: M= 50bp MW ladder, **1**. Positive PCR reaction, **2**. No template control.

A bright single 99bp was seen (Figure 4.6.1), with a clean no template control.

4.8 Construction of the complete randomised At-Thr DNA library cassette

With the successful production of the conserved region, repeat 4 and repeat 5 (Figure 4.2, grey, green and purple respectively) the three individual constructs were joined via an overlap PCR reaction using the incorporated complementary regions between the individual constructs indicated by the orange brackets in Figure 4.2 (2.2.4.2).

4.8.1 Determining the optimal template dilution and annealing temperature for PCR amplification of the complete cassette

The product formed from the overlap PCR between the three individual constructs (2.2.4.2) was diluted to form neat, 1/10, 1/100 and 1/1000 dilutions, to be used as template for PCR amplification (2.2.4.1). As with the optimisation of repeat 4 and repeat 5, four individual annealing temperature gradients were performed, each using a different template dilution, to determine both optimal annealing temperature and template dilution in one set of experiments (2.2.4.1). PCR products were visualised via agarose gel electrophoresis (2.2.1.3).

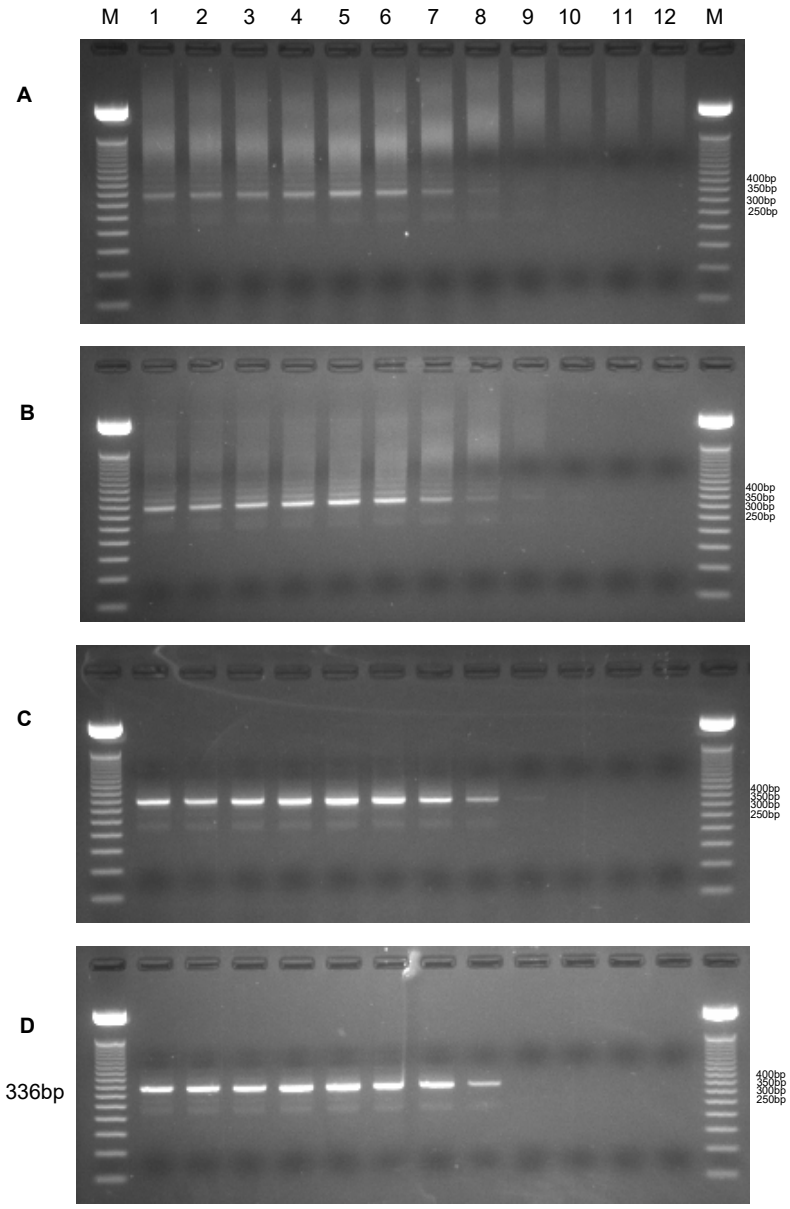

Figure 4.8.1: Four annealing temperature gradients using different full length construction overlap PCR product as template at varying dilutions to determine optimal annealing temperature and template dilution for full library construct amplification.

Four individual PCR reactions were made, each using a different dilution of full length library construct overlap PCR product as template: Neat (**A**), 1/10 diluted overlap PCR product (**B**), 1/100 overlap PCR product (**C**) and 1/1000 overlap PCR product (**D**) and divided into 12 equal reactions. Annealing occurred at the temperatures described below, with the in-cycle extension lasting 1 minute and a final extension of 5 minutes. Lanes: M= 50bp MW ladder, annealing temperatures; **1**.45.0, **2**. 45.4, **3**.46.5, **4**.48.5, **5**.51.1, **6**.53.7, **7**.56.1, **8**.58.7, **9**.61.2, **10**.63.2, **11**.64.4,**12**. 65.0 (°C).

The substantial smearing seen in the annealing gradients using neat and 1/10 diluted overlap PCR product as template (Figure 4.8.1A and 4.8.1B respectively) meant these template dilutions were immediately abandoned. The annealing gradients using 1/100 diluted overlap PCR product and 1/1000 diluted overlap PCR product as templates (Figures 4.8.1C and 4.8.1D respectively) were very similar. Both followed the same pattern, where the intensity of the library construct band (336bp) started to decrease at the annealing temperature 58.7°C onwards. The annealing gradient using 1/100 dilute template (Figure 4.8.1C) showed slightly more smearing than the 1/1000 template gradient (Figure 4.8.1D), thus identifying the 1/1000 diluted full length library overlap PCR product as the optimal template dilution. Using only the 1/1000 template annealing gradient to ascertain the optimal annealing temperature, meant comparing the balance of library construct intensity with the amount of non-library construct bands produced for each annealing temperature. 58.7°C produced the lowest intensity library construct band, but also showed the smallest amount of additional products, while 56.1°C produced a much brighter library construct band but also produced more intense additional bands.

To try and balance both the negative and desirable properties these two annealing temperatures were producing, a mean average of the two was calculated, 57.4°C, and used for downstream amplification to generate full length library product.

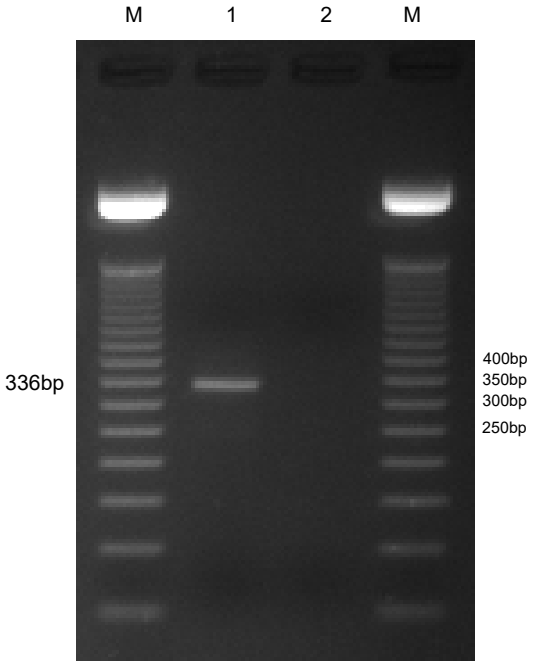

Figure 4.8.2: PCR amplification to generate full length library construct.

A PCR amplification using 1/1000 full length library overlap PCR product as template, using an annealing temperature of 57.4°C, with the in-cycle extension lasting 1 minute and a final extension of 5 minutes. Lanes: M= 50bp MW ladder, **1**. Positive PCR reaction, **2**. No template control.

A good quality single band of 336bp was produced, while the no template control was clean (Figure 4.8.2). These conditions were then used to produce a sufficient volume of PCR product for purification (2.2.1.4) and quantification (2.2.1.5) which was sent for Next Generation Sequencing (2.2.1.6).

4.9 Analysis of Next Generation sequencing data to assess amino acid representation at saturated positions in the At-Thr DNA cassette

The At-Thr library cassette PCR product was sequenced at Genewiz (2.2.1.6) using the Amplicon EZ service, a Miseq Illumina sequencing service, which utilises two 250bp reads running 5'-3 and 3'-5 to provide full library coverage. The sequencing data was processed and analysed (2.2.6), then presented graphically (Figure 4.9).

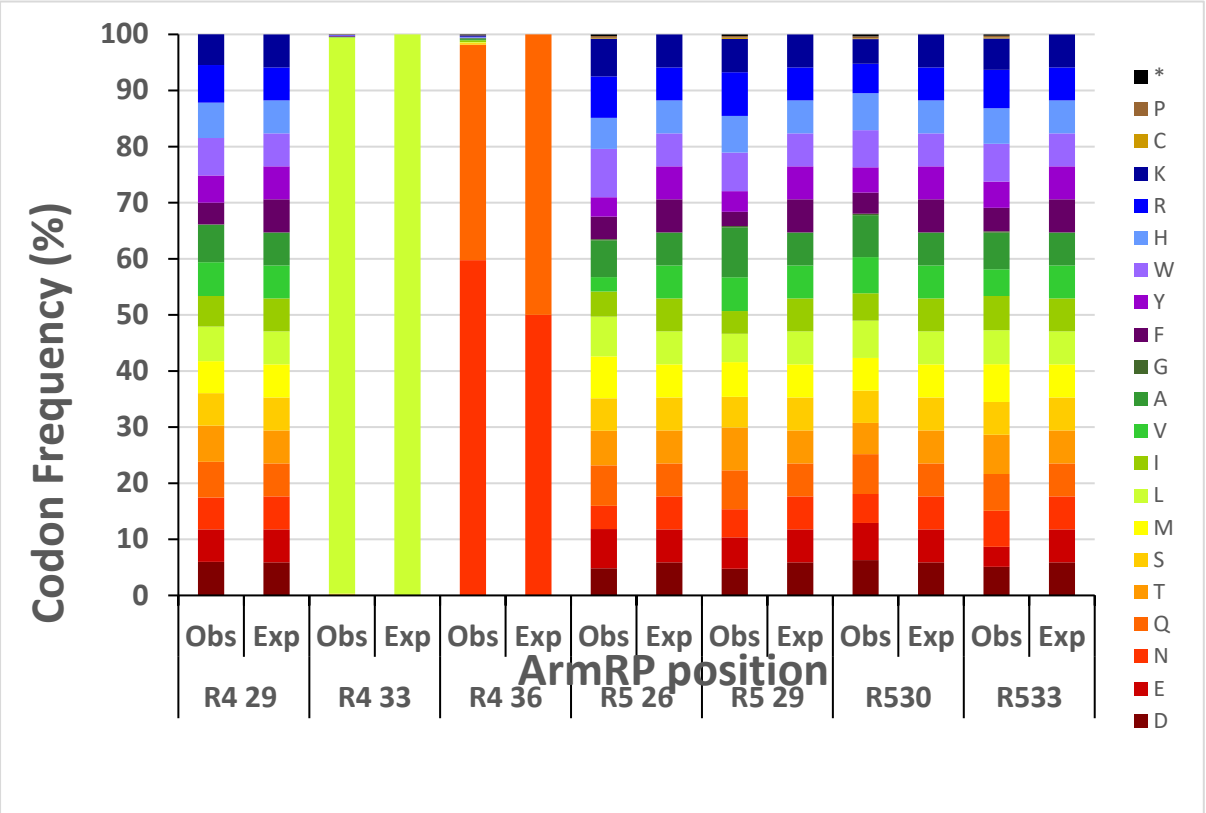

Figure 4.9: Observed versus expected amino acid distribution for positions randomised in the At-Thr library.

Letters in the legend correspond to universal amino acids abbreviations in the genetic code: P: proline; C: cysteine; K: lysine; R: arginine; H: histidine; W: tryptophan; Y: tyrosine; F: phenylalanine; G: glycine; A: alanine; V: valine; I: isoleucine; L: leucine; M: methionine; S: serine; T: threonine; Q: glutamine; N: asparagine; E: glutamic acid; D: aspartic acid; *: stop codon. Raw data from Miseq Illumina sequencing. (See Appendix 3 for raw count data).

Upon visual inspection of the observed and expected frequencies, a primary evaluation that the observed and expected codon frequencies matched well for all the positions was made (Figure 4.9). This was assessed statistically using the Chi-square Goodness-of-fit Test. In order to use the Chi-square Goodness-of-fit Test to determine if the observed frequencies of each amino acid at each of the randomised positions statistically fit the expected relevant distribution, the raw counts were used (Appendix 3). The amino acids observed at each position, but were not expected, had to be removed from the analysis, as part of the calculation involves dividing by the expected value, not possible if the value is 0. As position R4-33 was only saturated with leucine, a goodness of fit test could not be performed. The Chi-square Goodness-of-fit Test was performed using Graph Pad on each position individually. Each of the remaining 6 tests resulted in a two-tailed P value less than 0.0001

meaning the difference between the observed and expected proportions was significant and not due to chance.

This was suspected, as the MAX randomisation process is an incredibly sensitive process and to have a statistically perfect amino acid representation is practically impossible; this is why the library randomisation is assessed primarily through a visual representation (Figure 4.9). An example of this sensitivity being the slight over representation of asparagine at position R4-36, most likely caused by a minute unavoidable increased concentration of the asparagine MAX oligonucleotide in the selection oligonucleotide pool. As the At-Thr library was to be used to discover novel threonine binders, the key concern was any obvious issue with the library for example, a large amount of stop codons which would have been detrimental in protein screening, while maximising diversity.

As there were no concerning discrepancies in the library seen from the visual comparison of the observed and expected frequencies (Figure 4.9), the library quality was considered satisfactory for use in protein binder investigations.

4.10 Screening outputs from the randomised At-Thr DNA library

With the successful completion of the At-Thr library, investigating the resulting protein library began, attempting to find an improved threonine binder. As with the single arginine library, screening occurred via yeast surface display, crystallography and *Kd* determination conducted by the Plückthun group. From the library, a binder hit was found with the following amino acids at the saturated positions: R4-29= phenylalanine, R4-33= leucine), R4-36= glutamine R5-26= leucine, R5-29 = methionine, R5-30 = isoleucine and R5-33 = arginine. The crystal structure of the protein was determined, with Figure 4.10 focusing upon the randomised pocket. The threonine side chain of the target peptide is seen clearly interacting with the pocket, providing a promising outlook for *Kd* determination, which is ongoing.

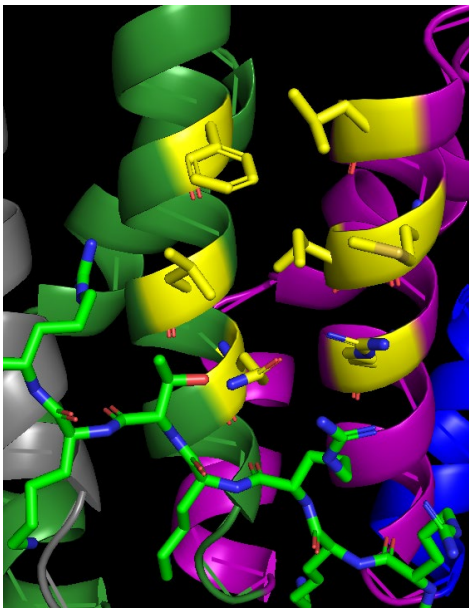

Figure 4.10 Visualisation of the repeat 4/5 binding groove of the binder hit from the At-Thr randomised library, interacting with the target peptide KRKRKTKRKR.

Depicted are helix 3 of repeat 4 (green) and repeat 5 (purple), with the key residues for peptide binding coloured yellow. The peptide target coloured green with side group oxygen coloured red and nitrogen blue. (P Mittl 2021, personal communication, 16 December).

4.11 Concluding the mutagenesis of the At-Thr library

MAX randomisation was successfully used to engineer a randomised designed armadillo cassette, containing seven positions of saturation spanning two internal repeats of the protein, with the amino acid saturation pattern determined by the computational predication programme ATLIGATOR. The positional saturation varied from the single arginine library, most prominently at position R4-33, fixed for leucine, and R4-36 which was fixed for asparagine and glutamine, based on ATLIGATOR predictions, resulting in a theoretical library size of $2.8x10^6$. Subsequent screening of the resulting protein library by the Plückthun group, identified a promising binder with the crystal structure bound determined, with *Kd* determination still in progression.

5.0 The design and optimisation of an excel based NGS analysis technique to calculate codon representation at saturated positions in an At-Thr DNA library

5.1 Introduction

To efficiently alter the original binding specificities of the designed armadillo repeat protein H3 groove binding pockets, high throughput saturation mutagenesis of key amino acid residues important for target peptide recognition and binding, was performed. These saturated DNA libraries were engineered using MAX (2.2.2.2) or ParaMAX randomisation (2.2.2.3 – see Chapter 6), to ensure near-even amino acid encoding within the resulting DNA libraries. To determine the success of this non-degenerate positional saturation, each completed, engineered DNA library was sequenced via Illumina sequencing, a sequencing by synthesis technique, with the resulting data processed and analysed.

The analysis of the Illumina data for each engineered library required an accurate count of each amino acid's encoding at each saturated position. With existing DNA analysis tools not catering for codon counting at specific positions, a new methodology was required.

The analysis of the Illumina data for each engineered library required an accurate count of each amino acid encoded at each saturated position. With existing DNA analysis tools not catering for codon counting at specific positions, a new methodology was required. Previously, the raw NGS data was retrieved without any quality processing and counts to determine amino acid representation were performed at each of the randomised positions, using an inefficient and complicated delimiting methodology within Excel. This approach was unstandarised, time consuming and required a considerable knowledge of Excel functions.

The ideal methodology required a user friendly platform which was cost effective and time efficient, capable of processing the raw Illumina sequencing data and outputting it into a workable format for non-bioinformaticians. To accommodate for these requirements the open source web based bioinformatics platform, Galaxy (https://usegalaxy.org/), was used to process the raw Illumina sequencing data, while Microsoft Excel was used for the codon counting and determining the amino acid representation at saturated positions.

This chapter focuses on the development of a NGS processing and codon counting methodology for libraries containing no contiguous randomised regions. The model library used was the At-Thr library (Chapter 4). The At-Thr library contained 7 non-contiguous positions of saturation, randomised using MAX randomisation (2.2.2.2).

5.2 Processing of raw Illumina sequencing data of the At-Thr library using Galaxy

The Galaxy processing was divided into four stages shown in Figure 5.2, with each subsequent step investigated using the function outputs of the previous optimised stages.

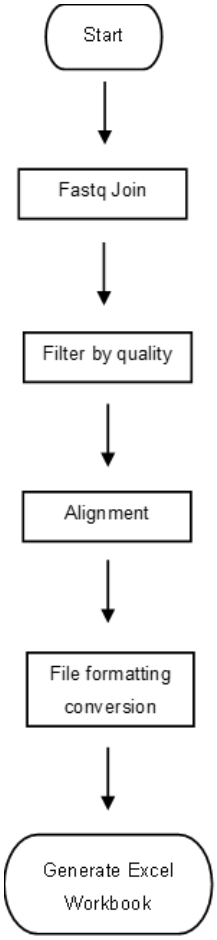

Figure 5.2: Flow diagram depicting pre-count processing stages of NGS data using the free online bioinformatics server, Galaxy.

Fastq join: joining of individual read mates, Filter by quality: using the associated quality score for each read to determine if the quality is satisfactory, Alignment: aligning the filtered reads with a reference sequence. Alignment output is then formatted to be used in an Excel Workbook.

5.2.1 Joining the directional mates of the At-Thr library using Galaxy Fastq Join function

The At-Thr library was sequenced using Miseq, with two 250bp reads 5'-3' and 3'-5', meaning the 336bp library was sequenced in both directions to a length of 250bp, coined the forward and reverse mates. These forward and reverse mates were provided in two Fastq files (flow cell output file type), the output file type from Illumina sequencing. In order to generate a full length read of the At-Thr library these two files had to be joined.

The Fastq Join function attempts to join mates on their overlapping end, scoring the join using: distance/len, where distance refers to the hamming distance (number of mismatches) and length being the attempted join length between the two mates. This scoring method is weighted towards longer length joins, with fewer mismatches (lower hamming distance) scoring the lowest, and therefore better joining scores. The successful joins (best scoring joins) are then scrutinised using the inbuilt quality check within the Fastq Join function. Figure 5.2.1 demonstrates a hypothetical join between two mates, the scoring of the join and the Fastq Join quality checking, based on Aronesty (2013).

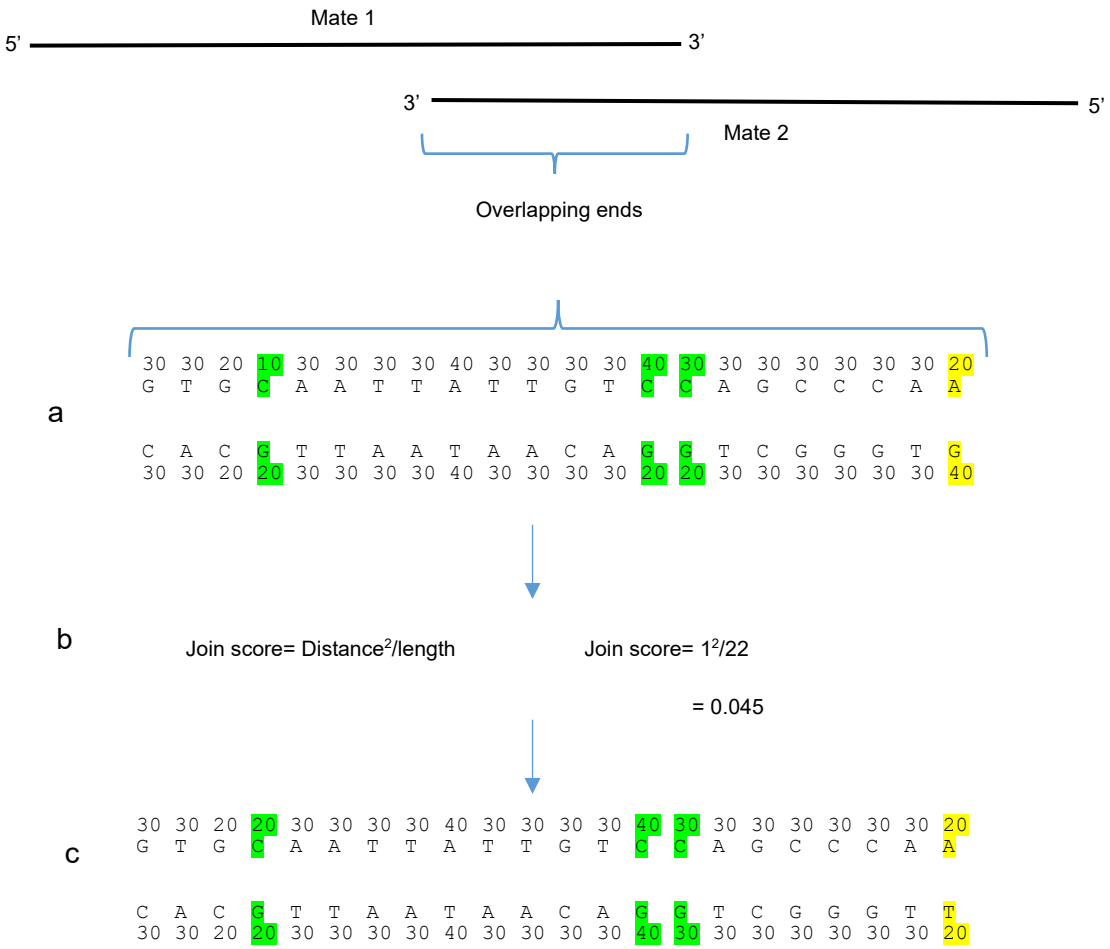

Figure 5.2.1: Example join overlap demonstrating Fastq Join stages.

Fastq files containing mate 1 and mate 2 are uploaded to the Galaxy server. A) Potential join in mate overlapping region is recognised by the Fastq Join function. Differences in quality scores between complementary bases (green) are identified. Any mismatching bases are also identified (yellow). B) Join score is calculated. Hamming distance equalled 1. The length of the mate join was 22bp, giving a join score of 0.045. C) Fastq Join quality control, corrects inconsistencies in quality scores for aligned bases, by adjusting lower quality score to match the higher (green). Base mismatches (yellow) are corrected by assigning the base with the highest quality score as correct. The opposite base of the other mate is changed so the base pairs are complementary and the new base alignment quality score assigned to the pair is equal to the difference of the original bases: 20-40 in a) (yellow) is changed to c) 20 (yellow)

Galaxy Fastq Join allows for user input in the following parameters: maximum percentage difference between matching segments and minimum length of matching segments. Using a

value of 5% for the maximum difference between matching segments was justifiable as this would convey a 95% confidence in segment complementarity; alongside a minimum length of 9bp for matching segments, as 9bp sequences would allow enough specificity between matching segments, considering 9bp oligonucleotides are used for the accurate saturation of target positions in MAX randomisation (2.2.2.2). The Fastq Join's quality control stage, where confidence scores for aligned bases are adjusted to reflect the higher confidence score, ensures the largest proportion of valid reads are accepted. The correction in base confidence is valid due to the absoluteness of base alignments, A-T and C-G. The alignment of C-G in Figure 5.2.1A has the confidence of C as 10 while G is 20. Having the bases correctly aligned allows for the transitive adjustment of C's confidence score to match G's. The Fastq Join function outputted 440928 joined reads from a total of 468540 (94.11%). These joined reads were then used as the input for Filter by Quality (Figure 5.2).

5.2.2 Quality control using Filter by Quality function on full length At-Thr joined reads

Once full length At-Thr library reads had been made via Fastq Join, the reads were subjected to the Filter by Quality function. The Filter by Quality function discards reads based on two parameters defined by the user; the 'quality cut-off value' and the 'percent of bases in sequence that must have quality equal to/higher than cut-off value' (2.2.5.2.2).

A base quality score of 30, correlated to a 1/1000 chance that a base identity call was an error (Illumina, 2014) so 30 was assigned to the 'quality cut-off value'. As quality scores are representative of the confidence in a base calling being correct, non-identifiable bases ('N's) will have an intrinsically lower base call confidence and will therefore impact on the total read quality. As the Filter by Quality function would be assessing each base quality in the library sequence (including randomised positions), adjustments to 'the percent of bases in sequence that must have quality equal to/higher than cut-off value' parameter were made. This accommodation for randomisation was done by calculating the percentage of randomised bases in the library (At-Thr library: 21bp/336bp*100= 6.25%) and lowering the value for 'percent of bases in sequence that must have quality equal to/higher than cut-off value' to the nearest whole number under the randomised base percentage (At-Thr: 93%). The Filter by Quality function performed on the joined At-Thr reads using a 'quality cut-off value' of 30 and 100% of bases in sequence must have quality equal to/higher than cut-off value resulted in a read output of 85203 with 355725 (80.68%) reads discarded. In comparison, using the adjusted percentage of 93% of bases needing at least a quality value of 30, an output of 398703 reads with 42225 reads being discarded (9.58%) was achieved, demonstrating the importance of adjusting for the randomised bases.

5.2.3 Using Bowtie2 to align full length At-Thr library reads to a reference sequence

In order to align the library reads an alignment tool and reference sequence were required. The alignment was performed using Bowtie2, a tool capable of matching raw read sequences to a user defined reference sequence, even if the reference sequence contains undetermined bases (Ns). This sequence was the At-Thr library sequence with MAX codons substituted for NNN at randomised positions made using UGENE, software that enables the visualization of DNA sequences that can be saved in a FASTA format, compatible with Galaxy processing (2.2.5.1). This reference sequence was inputted into the alignment function along with the 398703 Filter by Quality outputted reads (2.2.5.2.4). The Bowtie2 output file was then reformatted (2.2.5.2.6) so the read data could be used in Excel.

5.3 Invariant anchor count methodology to determine codon frequencies at saturated positions in the At-Thr library

As all saturated positions in the At-Thr had been targeted using MAX randomisation (2.2.2.2) (Figure 4.2), they had therefore been randomised using predefined MAX selection oligonucleotides. Accordingly, a count technique applying the specificity of these selection oligonucleotides was developed to assess codon (and therefore encoded amino acid) representation at each targeted position in the At-Thr library. The NGS data would be subjected to a specific amino acid count at each of the 7 individual saturated position, using the Excel countif function. Table 5.3 shows the countif functions used to count codon frequencies at position R4-29 in the At-Thr library as an example.

Amino acid	Excel function used to count codon frequencies
Aspartic acid	SUM(COUNTIFS(range,{"*TTAGATGAG*","*TTAGACGAG*"}))
Glutamic acid	SUM(COUNTIFS(range,{"*TTAGAAGAG*","*TTAGAGGAG*"}))
Aspargine	SUM(COUNTIFS(range,{"*TTAAATGAG*","*TTAAACGAG*"}))
Glutamine	SUM(COUNTIFS(range,{"*TTACAAGAG*","*TTACAGGAG*"}))
Threonine	SUM(COUNTIFS(range,{"*TTAACTGAG*","*TTAACCGAG*","*TTAACAGAG*","*TTAACGGAG*"}))
Serine	SUM(COUNTIFS(range,{"*TTATCTGAG*","*TTATCCGAG*","*TTATCAGAG*","*TTATCGGAG*","*TTAAGTGAG*","*TTAAGCGAG*"}))
Methionine	SUM(COUNTIFS(range,{"*TTAATGGAG*"}))
Leucine	SUM(COUNTIFS(range,{"*TTACTTGAG*","*TTACTCGAG*","*TTACTAGAG*","*TTACTGGAG*","*TTATTAGAG*","*TTATTGGAG*"))
Isoleucine	SUM(COUNTIFS(range,{"*TTAATTGAG*","*TTAATCGAG*","*TTAATAGAG*"))
Valine	SUM(COUNTIFS(range,{"*TTAGTTGAG*","*TTAGTCGAG*","*TTAGTAGAG*","*TTAGTGGAG*"))
Alanine	SUM(COUNTIFS(range,{"*TTAGCTGAG*","*TTAGCCGAG*","*TTAGCAGAG*","*TTAGCGGAG*"))
Glycine	SUM(COUNTIFS(range,{"*TTAGGTGAG*","*TTAGGCGAG*","*TTAGGAGAG*","*TTAGGGGAG*"))
Phenylalanine	SUM(COUNTIFS(range,{"*TTATTTGAG*","*TTATTCGAG*"))
Tyrosine	SUM(COUNTIFS(range,{"*TTATATGAG*","*TTATACGAG*"))
Tryptophan	SUM(COUNTIFS(range,{"*TTATGGGAG*"))
Histidine	SUM(COUNTIFS(range,{"*TTACATGAG*","*TTACACGAG*"))
Arginine	SUM(COUNTIFS(range,{"*TTACGTGAG*","*TTACGCGAG*","*TTACGAGAG*","*TTACGGGAG*","*TTAAGAGAG*","*TTAAGGGAG*"))
Lysine	SUM(COUNTIFS(range,{"*TTAAAAGAG*","*TTAAAGGAG*"))
Cysteine	SUM(COUNTIFS(range,{"*TTATGTGAG*","*TTATGCGAG*"))
Proline	SUM(COUNTIFS(range,{"*TTACCTGAG*","*TTACCCGAG*","*TTACCAGAG*","*TTACCGGAG*"))
Stop	SUM(COUNTIFS(range,{"*TTATAGGAG*","*TTATGAGAG*","*TTATAAGAG*"))

Table 5.3 Countif function to determine summed frequency of each codons for each amino acid at position R4-29 of the At-Thr library.

Counting of all codons for each amino acid, ensured the maximum number of reads were used in determining amino acid abundance and that any 'non-MAX' codons would be included in the calculations. Using the selection oligonucleotide as a wildcard (by using ** in the formula), meant positional changes caused by additions, deletions or sequencing errors outside of the invariant region were not important as the position of the anchor in relation to the codon being counted was fixed. One consideration to be made before counting with non-MAX selection oligonucleotides (oligonucleotides that were not used to represent an amino acid in the MAX selection oligonucleotide pools) was whether that oligonucleotide sequence was present anywhere else in the library sequence. This was the case for GCTTTAGTC, when counting valine abundance at position R4-36. The count for GCTTTAGTC was 358400. As there were only 369406 total reads, and valine was not expected at this position (saturation of R4-36 had only been performed using Q and N encoding MAX selection oligonucleotides) it was suspected this invariant-non MAX codon combination was present in the conserved region (Figure 5.3).

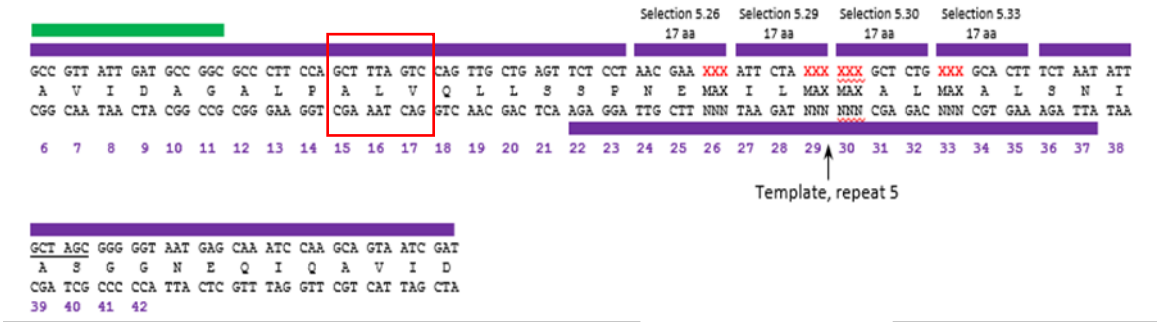

Figure 5.3 Determining if the countif wildcard GCTTTAGTC representing valine was present in the conserved region of the At-Thr library.

Find function searching for 'GCTTTAGTC' confirmed the oligonucleotide sequence was present in the conserved region of the repeat 5 construct in the At-Thr library (red box).

With the presence of GCTTAGTC confirmed in a conserved region of the library, the count using GTC for valine at position R4 36 was removed from the sum(countif) function for determining valine abundance. As valine was not used to saturate R4-36 and further still GTC was not the assigned valine MAX codon, the impact of removing GTC from the valine count was suspected to be similarly insignificant as the other valine encoding amino acids GTG, GTA and GTT (counts of 66, 69 and 129 respectively). If the count for a MAX codon was not possible due to the count criteria being unspecific in the sequence, an adjustment to the criteria could be made, for example the addition of bases to ensure criteria specificity.

5.4 Analysing the amino acid distribution at saturated positions in a Galaxy processed, Bowtie2 aligned At-Thr library

Once all counts had been performed at each of the 7 saturated positions in the At-Thr library, the raw codon counts were converted to codon frequency (%) and compared graphically (Figure 4.9) with a visual inspection showing no concerning mismatches between observed and expected amino acid ratios, while statistically they were different (further explanation provided in section 4.9).

5.5 Investigating the impact of changing alignment tool from Bowtie2 to BWA-MEM has on the observed amino acid distribution at saturated positions in the At-Thr library

As none of the existing alignment tools available on the Galaxy server were designed to handle DNA sequences containing multiple positions of randomisation, a second alignment using the Filter by Quality output and the UGENE generated reference sequence was performed, this time using BWA-MEM, an alignment tool capable of aligning sequences greater than 100bp via local alignments (2.2.5.2.5).

The observed codon frequencies at the 7 randomised positions where compared to the expected distribution (Figure 5.5).

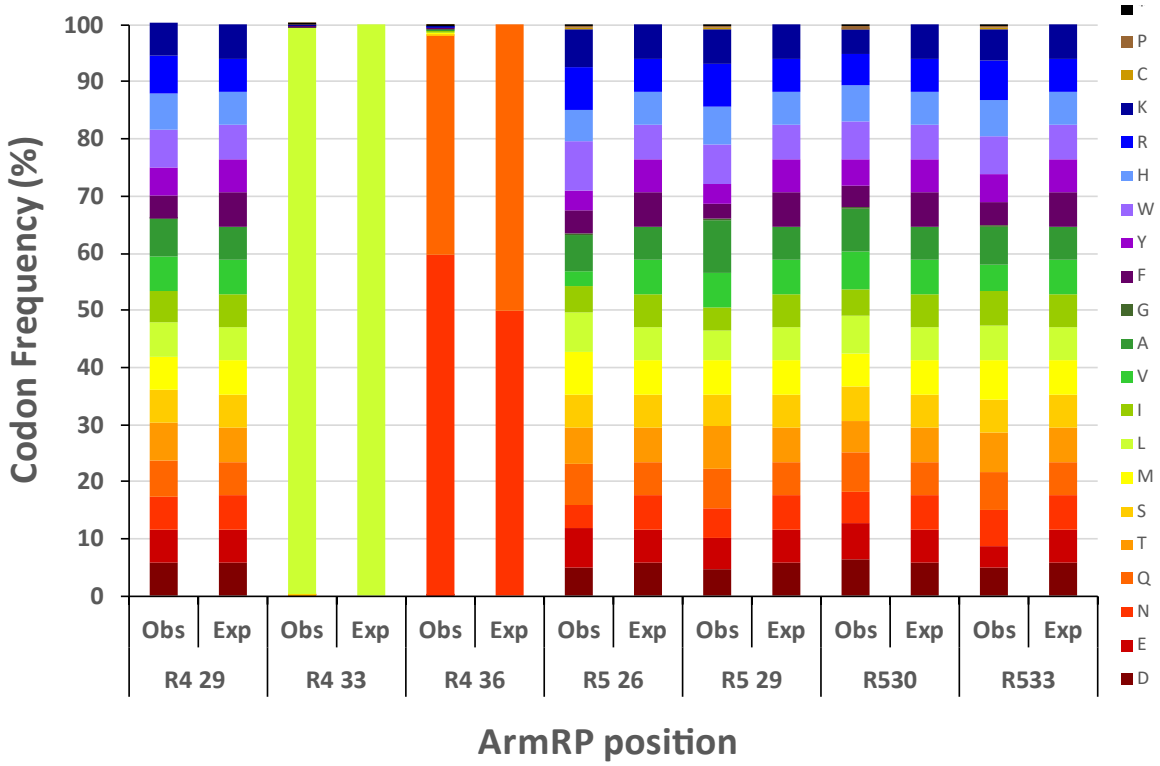

Figure 5.5: Observed vs expected amino acid distribution for positions randomised in a BWA-MEM aligned At-Thr designed armadillo repeat protein library.

Letters in the legend correspond to universal amino acids abbreviations in the genetic code: P: proline; C: cysteine; K: lysine; R: arginine; H: histidine; W: tryptophan; Y: tyrosine; F: phenylalanine; G: glycine; A: alanine; V: valine; I: isoleucine; L: leucine; M: methionine; S: serine; T: threonine; Q: glutamine; N: asparagine; E: glutamic acid; D: aspartic acid; *: stop codon. Raw data from Miseq Illumina sequencing. (See Appendix 4 for raw count data)

Upon visual inspection of the amino acid distribution after Bowtie2 alignment (Figure 4.9) compared with the distribution with a BWA-MEM alignment (Figure 5.5), no obvious differences were apparent. To determine any statically significant changes, a paired t-test using Graph-pad was conducted between the observed percentages of each amino acid at each randomised position (Table 5.5).

Randomised Position	Two-Tailed P-value (2.dp)
R4 29	1.00
R4 33	0.49
R4 36	0.90
R5 26	0.86
R5 29	0.94
R5 30	0.72
R5 33	0.90

Table 5.5: Paired t-test results comparing amino acid observed percentages between Bowtie2 aligned NGS reads and BWA-MEM aligned NGS reads, at each of the seven randomised position of the At-Thr library

Observed codon frequency percentages for each amino acid at each of the 7 randomised positions in the At-Thr library were inputted into a paired t-test using Graph-pad, and the two-tailed P-value (2.dp) recorded.

The assumption based on the visual inspection, that there was no significant difference in amino acid distribution caused by alignment tool (Figure 4.9 and Figure 5.5) was supported by the 7 paired t-tests performed, as none of the two-tailed P-values were considered significant (Table 5.5). Differences in the P-value away from 0 (the hypothesised value for an exact match between the two groups being tested) within the implemented 95% confidence range are associated with random chance and were therefore not significant results. As none of the P-values calculated were outside of this 95% interval, no significant differences between the amino acid percentages were observed. This meant the choice of alignment tool had no impact on the codon counts in the aligned reads and therefore the observed amino acid distributions at any position in the At-Thr library.

5.6 Investigating the impact of length filtering aligned reads in Excel on observed amino acid distribution at saturated positions in the At-Thr library

With no significant difference in amino acid distribution observed between alignments of the At-Thr reads using Bowtie2 or BWA-MEM, the Bowtie2 reads were used for investigating the impact of read length filtering. The output of the At-Thr Galaxy processing (using a Bowtie2 alignment) was 369407 reads. With only 189129 reads being the expected length of 336bp (51.2% (2.dp)) removing incorrect read lengths and investigating the impact on amino acid distribution was necessary. The identical read output data set as the Bowtie2 counts (Figure 4.9) was used, with all reads not 336bp in length deleted. The remaining 189129 reads were subjected to codon counting (2.2.6.3), with the codon frequency % compared to the expected distribution (Figure 5.6).

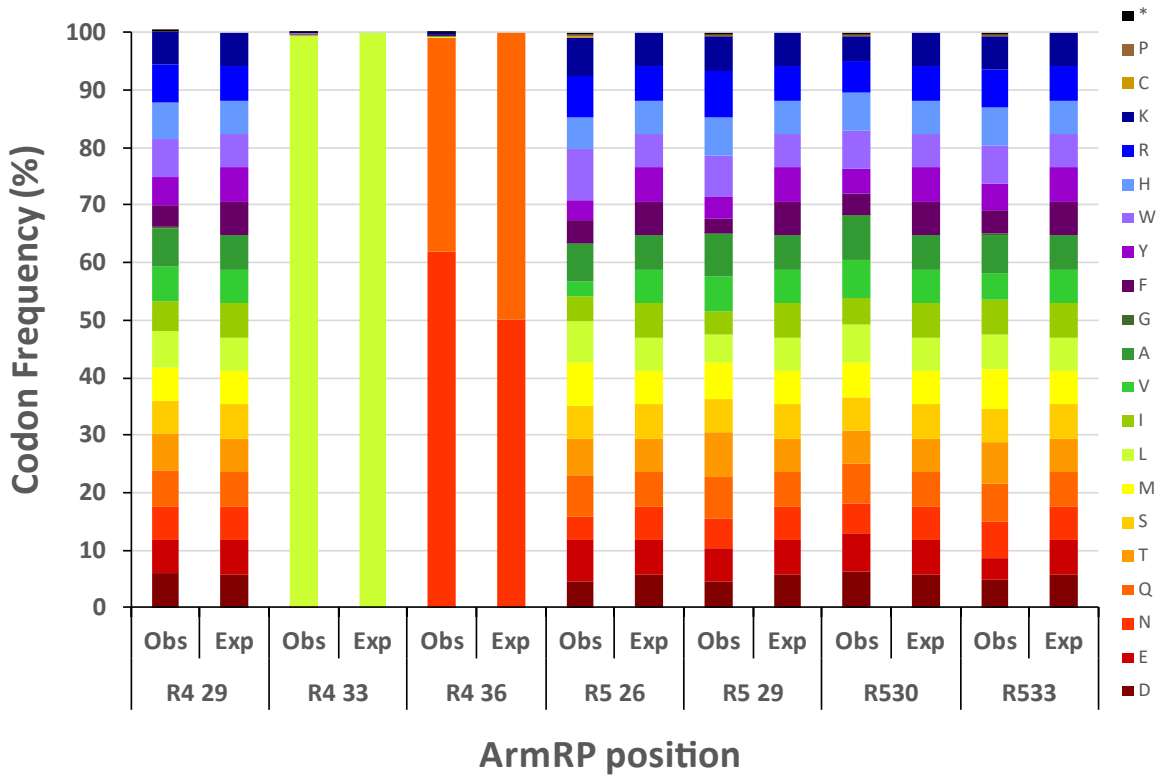

Figure 5.6: Observed vs expected amino acid distribution for positions randomised in a Bowtie2 aligned At-Thr designed armadillo repeat protein library, with non-library length reads deleted from data set.

Letters in the legend correspond to universal amino acids abbreviations in the genetic code: P: proline; C: cysteine; K: lysine; R: arginine; H: histidine; W: tryptophan; Y: tyrosine; F: phenylalanine; G: glycine; A: alanine; V: valine; I: isoleucine; L: leucine; M: methionine; S: serine; T: threonine; Q: glutamine; N: asparagine; E: glutamic acid; D: aspartic acid; *: stop codon. Raw data from Miseq Illumina sequencing. (See Appendix 5 for raw count data)

Upon visual inspection of the amino acid distribution with no read length filtering (Figure 4.9) compared to the distribution with read length filtering (Figure 5.6), the only noticeable difference was the marginally increased asparagine representation at position R4-36 in Figure 5.6. To determine if there were any statistically significant changes as a result of read length filtering, a paired t-test using Graph-pad was conducted between the observed percentages of each amino acid at each randomised position (Table 5.6).

Randomised Position	Two-Tailed P-value (2.dp)
R4 29	0.63
R4 33	0.69
R4 36	0.99
R5 26	0.98
R5 29	1.00
R5 30	0.96
R5 33	1.00

Table 5.6: Paired t-test results comparing amino acid observed percentages between performing read length filtering or not, at each of the seven randomised position of the At-Thr library

Observed percentages for amino acids at each position were inputted into a paired t-test using Graph-pad, and the two-tailed P-value (2.dp) recorded.

Differences in the P-value away from 0 (the hypothesised value for an exact match between the two groups being tested) within the implemented 95% confidence range are associated with random chance and were therefore not significant results. As none of the P-values calculated were outside of this 95% interval, no significant difference was seen between the length filtered and non-filtered codon counts and therefore amino acid distribution at any of the 7 randomised At-Thr library positions. This meant removing non-library length reads was not necessary before performing codon counts.

A closer inspection of the Fastq Join and Filter by Quality outputs was conducted to investigate the causation for only 51.2% (2.dp) of the analysed reads being the expected 336bp. Using a FilterFastq run (2.2.5.2.3) (no quality parameters implemented), the output of the Fastq Join function (2.2.5.2.1) was queried for the number of reads at a length of 336bp first. Only 46.85% of the Fastq Join outputted reads were 336bp (206559/440928). To investigate the impact of the Filter by Quality function on the number of correct length reads,

these 440928 ‘successfully’ joined reads were used in the Filter by Quality function (2.2.5.2.2) with the output queried using the Filter Fastq function (2.2.5.2.3) (no quality parameters implemented). The Filter by Quality function outputted 398703/440928 (90.42%) reads that had met the quality parameters, but the Filter Fastq query revealed only 47.44% (189156 of 398703) of the filtered joined reads were of the correct length. When comparing the number of 336bp reads outputted from the Fastq Join and Filter by Quality function (189156) to the number of 336bp reads in the Excel file (189129) revealed 99.99% of these 336bp reads were in the Excel file. This highlighted the cause for the low percentage of 336bp reads as the Fastq Join function. Mohsen *et al* (2019) demonstrated the importance of base quality scores in the overlapping region for successful joins, and how low quality scores can negatively impact on the Fastq Join success. To investigate any quality issues within the overlap region, the last 50bps of the Mate 1 reads for the At-Thr library were assessed using Filter Fastq (2.2.5.2.3) (kept bases 200-250bp, no length parameters implemented) with a minimum quality score of 30 per base required. Of the 492701 reads inputted, only 75321 (16.26%) met the Filter Fastq quality parameter. As there are no significant differences in codon representation at the saturated positions between reads filtered by length or not for the At-Thr library, the performance of the Fastq Join function did not impact on the library analysis, meaning the relative amino acid distributions observed were reliable.

5.7 Concluding the development of an excel based methodology for determining amino acid distribution at specific positions in an At-Thr library

The use of NGS for quality control is an essential step prior to protein screening. The observed amino acid ratios at the positions of saturation were calculated to assess the success of randomisation and to also assist with eventual protein library analyses, where a significant over/under representation of an amino acid could be compensated for. The uniqueness of the libraries engineered using MAX randomisation, possessing specific codons saturated with up to 20 different codons across the length of the DNA sequence, required the invention of a standardised raw data processing pathway and novel analysis approach. The MAX invariant sequence, used for the accurate localisation of the MAX oligonucleotide, provided the ideal criteria to enable specific MAX counts. Processing of the raw NGS output files into a format accessible in Excel, allowed the use of the ‘Countif function’. Processing the raw Illumina data using a quality control stage before amino acid distribution calculations, coupled with the utilisation of the novel count methodology, significantly decreased NGS data processing time, while increasing analysis accuracy and reproducibility.

6.0 ParaMAX: a novel contiguous codon randomisation technique

6.1 Introduction

The $Y_{II}M_5A_{II}$ designed armadillo repeat protein protein contains five internal M repeats responsible for the recognition and binding of the target peptide sequence, with the adjacent H3 helices of each M repeat forming a binding groove with the original peptide recognition and binding target being a KR_5 peptide. Each of these H3 generated binding grooves recognise a single dipeptide (originally KR). With PRe-ART's ultimate aim of generating 400 individual binders to bind any dipeptide combination, the shallow lysine pocket had to be engineered to accommodate more bulky amino acids. Work in the Plückthun lab determined that this would be best done by introducing a "loop", a stretch of 2-8 amino acids between the end of one M repeat and the beginning of the next (Figure 6.1.1).

Figure 6.1.1: Visualisation of M3, M4 and M5 of the designed armadillo repeat protein, $Y_{II}M_5A_{II}$, highlighting location of novel amino acid stretch addition. Each M repeat coloured differently M3 (blue), M4 (purple) and M5 (green), with location for potential addition of novel amino acid stretch highlighted using dashed red circle (end of M3 (blue) and beginning of M4 (purple)) . Adapted from (Hansen *et al.*, 2016)

The investigation of this novel stretch of amino acids would involve both target binding and specificity determination as well as exploring any potential structural effects on the protein. Accordingly, the most efficient approach was high throughput saturation mutagenesis to

generate a DNA library encoding all possible amino acid combinations in various loop lengths.

Previous libraries (Chapters 3 and 4) were constructed using MAX randomisation. However, MAX randomisation is not designed to saturate more than two consecutive codons, owing to the addressing sequence required for hybridisation of the MAX selection oligonucleotides to the NNN-containing template strand (Hughes *et al.*, 2003). Since a 6bp conserved region would not be available in the middle of the introduced codon stretch, an alternative saturation mutagenesis technique was required. ProxiMAX randomisation (Ashraf *et al.,* 2013; Poole *et al*., 2018) is capable of saturating many consecutive codons, but employs long, high quality oligonucleotides and is best automated (Frigotto *et al.*, 2015). Accordingly, a new saturation mutagenesis technique was designed to combine the advantages of MAX randomisation (a manual process that employs standard grade oligonucleotides) and ProxiMAX randomisation, to create a manual process that employs standard oligonucleotides to saturate multiple consecutive codons. Herein, construction of a loop encoding an insertion of four additional amino acids using the novel ParaMAX technique was investigated, using repeat 3 and repeat 4 of the $Y_{III}M_5A_{II}$ DNA as the backbone for the loop insertion (Figure 6.1.2).

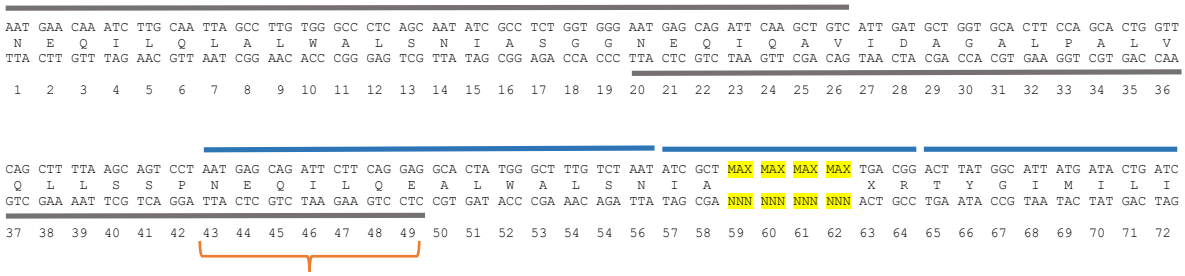

Figure 6.1.2 : DNA sequence of a ParaMAX model library based on repeat 3 and repeat 4 of a designed armadillo repeat protein, overlaid with the ParaMAX randomisation library design. The DNA sequence was split into two separate overall constructs, a conserved region (grey) and the contiguous region containing the inserted construct "the Quad "(blue), with positions of saturation highlighted (yellow). The overlapping region between constructs is indicated by an orange bracket.

This chapter will explain the full process of ParaMAX randomisation, via the construction of the model library shown in Figure 6.1.2, with the positional saturation assessed via Illumina sequencing.

6.2 Schematics of generating a stretch of contiguous randomised positions using ParaMAX

ParaMAX is designed to use MAX randomisation to saturate two postions at a time, to generate a randomsied couplet, which is digested with *MlyI* and then blunt-end ligated to a second, similarly-processed randomsied couplet. These ransomised couplets employ unique, expendable flanking oligonucleotide sequences to enable PCR amplification, but these flanking oligonucleotides can also contain a *MlyI* site (cryptic, if required) to allow for specific restriction and exposure of the randomised postions after PCR amplification. In the context of a designed armadillo repeat protein library, the couplet's position in the contiguous region, impacts upon elements of the couplet sequence. For example, as illustrated in Figure 6.2.1, couplet 59+60 would be at the 5' end of the contiguous codons meaning that for successful addition of couplet 61+62, the 3' flanking region of couplet 59+60 would need to be removed, to expose the MAX codons.

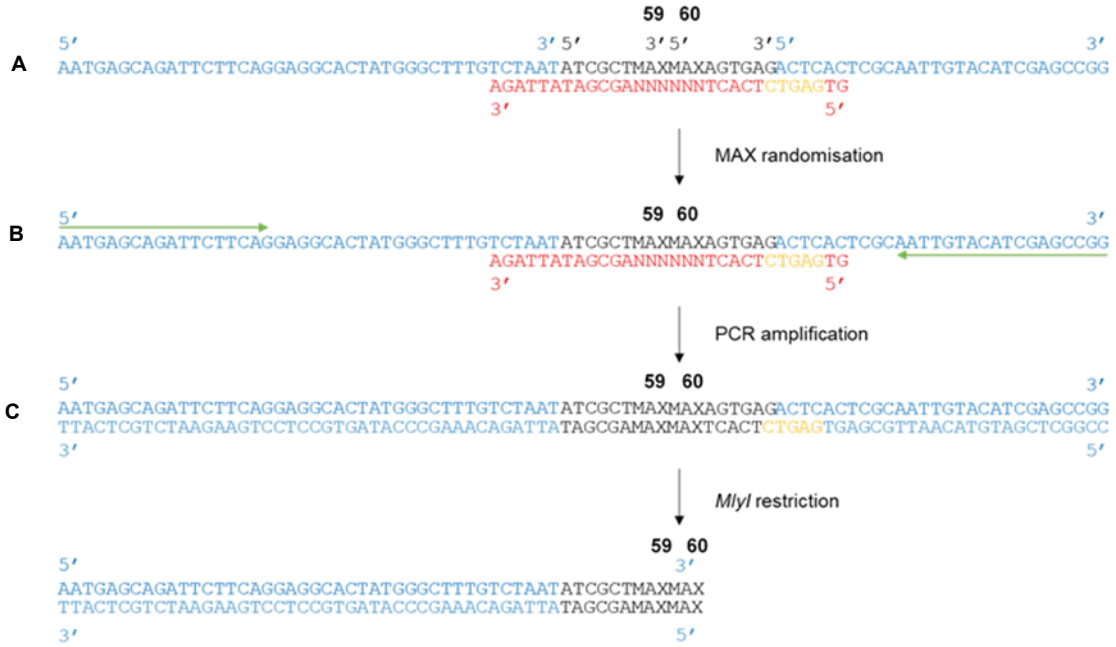

Figure 6.2.1 : ParaMAX schematic to form position 59 and 60 couplet, and prepare for couplet to couplet ligation.

ParaMAX process divided into three stages: A) MAX randomisation and blunt end ligation of constituent MAX selection oligonucleotides and flanking oligonucleotides. Flanking oligonucleotides (blue) and MAX selection oligonucleotides (black) anneal to complementary region of the NNN template strand (red), which contains a 3'-5' directed *MlyI* site. B) Asymmetric PCR amplification of MAX codon containing DNA strand, using 18mer primers (green arrows). C) *MlyI* restriction using 3'-5' directed *MlyI* site, to expose MAX codons.

In contrast to couplet 59+60, codons 61+62 are at the 3' end of the contiguous codons, meaning the 5' flanking oligonucleotide of the couplet needs to be removed. Positioning an *Mly*I site upstream of the MAX codons, would allow the subsequent restriction to expose couplet 61+62's MAX codons (Figure 6.2.2), prior to ligation of the individual couplets to generate a larger region of randomisation.

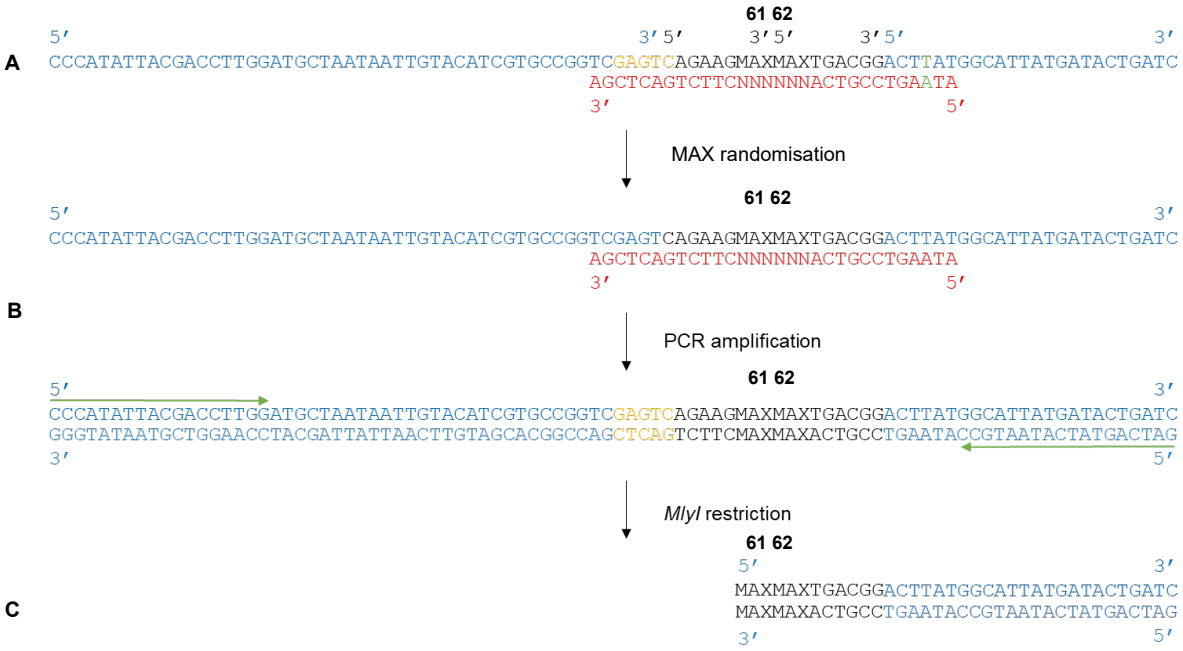

Figure 6.2.2 : ParaMAX schematic to form position 61 and 62 couplet, and prepare for couplet to couplet ligation.

ParaMAX process divided into three stages: A) MAX randomisation and blunt end ligation of individual oligonucleotides, flanking oligonucleotides (blue) and MAX selection oligonucleotides (black) anneal to complementary region of the NNN template strand (red). Sequence contains a 5'-3' directed *MlyI* site (orange). Cryptic *MlyI* site available via mutation of green labelled base pair. B) Asymmetric PCR amplification of MAX codon containing DNA strand, using 18mer primers (green arrows). C) *MlyI* restriction using 5'-3' directed *MlyI* site, to expose MAX codons.

With both couplets exposing their MAX codons on the necessary ends, blunt end ligation and PCR amplification of the product would be used to generate the contiguous region containing four MAX codons, coined the Quad (Figure 6.2.3).

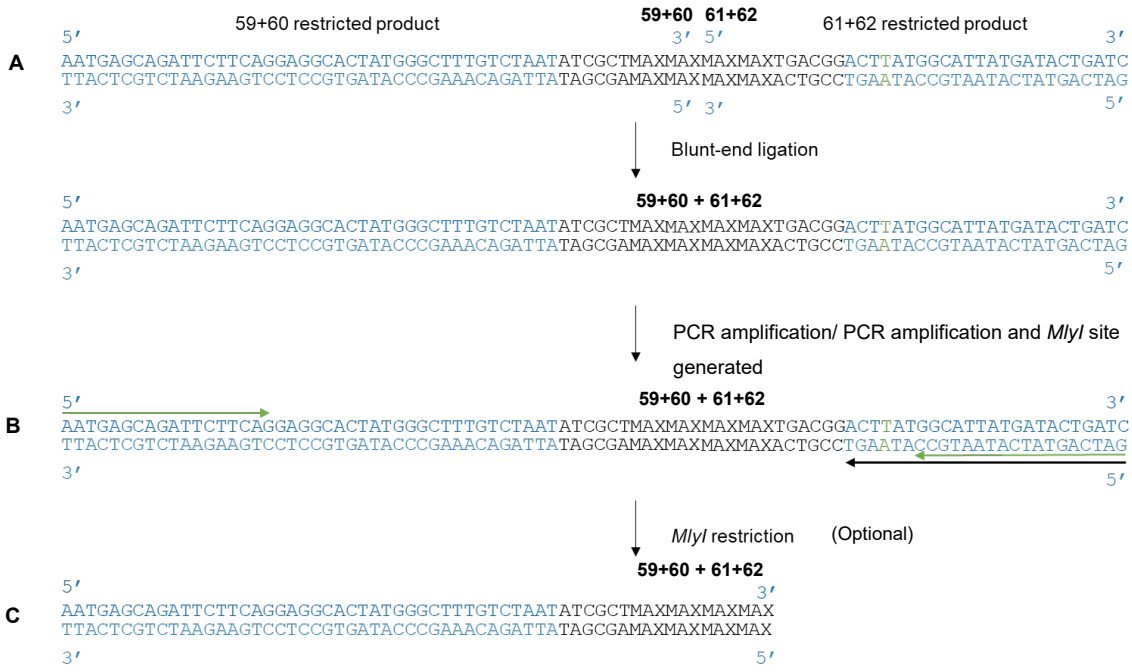

Figure 6.2.3 : ParaMAX schematic to form construct containing four contiguous randomised positions.

ParaMAX process divided into three stages: A) Blunt end ligation of *MlyI* restricted couplets, 59+60 and 61+62 at terminal MAX codons. MAX selection oligonucleotides (black) flanked by each couplet's remaining flanking oligonucleotide (blue). B) PCR amplification of ligation product from A, using 18mer primers (green arrows) or reverse 24mer primer to introduce a *MlyI* restriction site (fixing the point mutation highlighted by the green base pair). C) Possibility of *MlyI* restriction using 3'-5'' directed *MlyI* site generated using 24mer primer in B, to expose MAX codons.

With couplet 61+62 at the 3' end of the contiguous region, a method of removing its 3' flanking oligonucleotide and thus enabling the addition of another processed couplet (or quad) to form a longer contiguous region is a necessary option. The addition of a third couplet was not performed, but a point mutation (green, Figure 6.2.3), shows the position for a site directed mutation, which can be introduced via a primer mismatch in the amplification stage, capable of generating a new *MlyI* site. Further randomised codon additions would require new couplets to mimic couplet 61+62, by having a 5' *MlyI* site in respect to its randomised positions and a cryptic 3' *MlyI* site to allow for further couplet additions.

6.3 Generation of MAX selection oligonucleotide pools for positional saturation of positions in the ParaMAX model

Before commencing ParaMAX model construction, MAX selection oligonucleotides pools were needed for each contiguous position to be randomised. As the downstream expression organism would be *S. cerevisiae* as with previous libraries, the codons selected to represent each amino acid were based on *S. cerevisiae* preference determined using GenScript Codon Frequency Table(chart) Tool- *Saccharomyces cerevisiae* (gbpln) (2.2.2.1). As with the engineered single arginine and At-Thr designed armadillo repeat protein libraries, lysine encoding was switched from the more popular AAA to AAG in an attempt to alleviate any potential ligation bias during MAX randomisation that can be caused by consecutive adenine bases. The required MAX selection oligonucleotides to represent the desired amino acids at each position were mixed to form four different MAX selection oligonucleotide pools, one for each target position (2.2.2.1). (See Appendix 6 for non-MAX selection oligonucleotides used for model ParaMAX library engineering).

	Position	59	60	61	62
	Invariant sequence showing MAX position	ATCGCTMAX	MAXAGTGAG	CAGAAGMAX	MAXTGACGG
Identity of MAX codon	Aspartic acid	GAT	GAT	GAT	GAT
	Glutamic acid	GAA	GAA	GAA	GAA
	Aspargine	AAT	AAT	AAT	AAT
	Glutamine	CAA	CAA	CAA	CAA
	Threonine	ACT	ACT	ACT	ACT
	Serine	TCT	TCT	TCT	TCT
	Methionine	ATG	ATG	ATG	ATG
	Leucine	TTG	TTG	TTG	TTG
	Isoleucine	ATT	ATT	ATT	ATT
	Valine	GTT	GTT	GTT	GTT
	Alanine	GCT	GCT	GCT	GCT
	Phenylalanine	TTT	TTT	TTT	TTT
	Tyrosine	TAT	TAT	TAT	TAT
	Tryptophan	TGG	TGG	TGG	TGG
	Histidine	CAT	CAT	CAT	CAT
	Arginine	AGA	AGA	AGA	AGA
	Lysine	AAG	AAG	AAG	AAG
	Glycine	GGT	GGT	GGT	GGT
	Cysteine				
	Proline	CCA	CCA	CCA	CCA

Table 6.3: Table showing the encoded amino acids in the MAX selection oligonucleotide pools created for saturation of each target position in the randomised ParaMAX model.

Each cell highlighted indicates that amino acid's exclusion in the corresponding MAX selection oligonucleotide pool. Each MAX selection oligonucleotide was received at 100 µM in water. To make the MAX selection oligonucleotide pool for each position, an equal volume and therefore concentration of each MAX selection oligonucleotide was mixed. The resulting end concentration of each selection pool was 100 µM, with individual MAX selection oligonucleotide concentration at 5.26 µM in their respective selection pools.

6.4 Production of couplet 59+60 for generation of randomised contiguous region in the ParaMAX model library

As each construct of the model ParaMAX library, was to be made individually (Figure 6.1.2), couplet 59+60 was engineered first.

6.4.1 Determining the optimal couplet 59+60 MAX randomisation product template dilution and annealing temperature for PCR amplification

The first stage in PCR optimisation, to produce the couplet 59+60 construct, was to determine the optimal annealing temperature and template dilution. This would be achieved using four individual annealing temperature gradients (2.2.4.1), each using a different template dilution. The couplet 59+60 MAX randomisation construct produced via MAX randomisation (2.2.2.2) was diluted using milliq water to achieve the four different template dilutions to be used, neat, 1/10, 1/100 and 1/1000. PCR products were visualised via agarose gel electrophoresis (2.2.1.3).

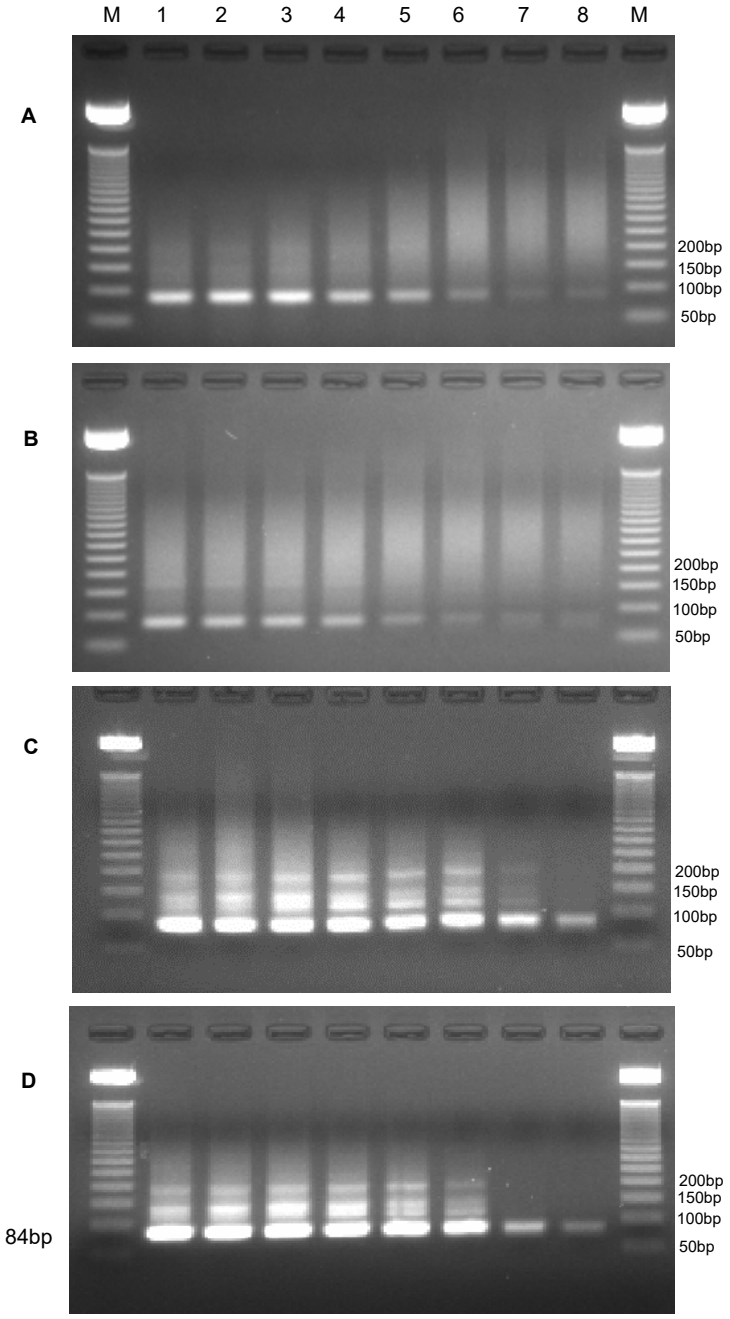

Figure 6.4.1: Four annealing temperature gradients using couplet 59+60 MAX randomisation product as template at varying dilutions to determine optimal annealing temperature and template dilution for PCR amplification.

Four individual PCR reactions were made using (**A**) neat MAX randomisation product, (**B**) 1/10 diluted MAX randomisation product, (**C**) 1/100 diluted MAX randomisation product and (**D**) 1/1000 diluted positive MAX randomisation as template and divided into 8 equal reactions. Annealing occurred at the temperatures described below with an in cycle extension time of 30 seconds. Lanes: M= 50bp MW ladder, annealing temperatures; **1**.50.0, **2**. 51.2, **3**.53.4, **4**.56.3, **5**. 59.9, **6**.62.8, **7**.64.9, **8**.66.0 (°C).

All annealing temperatures across all four template dilutions shown in Figure 6.4.1 produced the desired 84bp couplet 59+60 PCR product, with the brightest bands occurring using 1/100 and 1/1000 couplet 59+60 MAX randomisation product dilutions as template (Figure 6.4.1 C and D respectively). Because of this, Figure 6.4.1A and Figure 6.4.1B were not considered optimal and the corresponding template dilutions of neat and 1/10 were no longer considered. Figure 6.4.1 C and Figure 6.4.1D were very similar, both producing defined non-couplet 59+60 constructs approximately 125bp, 175bp and showing considerable smearing for annealing temperatures 50.0-62.8°C. Figure 6.4.1 C using the 1/100 dilution template, did produce slightly more smearing compared to Figure 6.4.1 D using the 1/1000 dilution template, with undesirable upper band products visualized in Figure 6.4.1 C using the annealing temperature 64.9°C, not present in Figure 6.4.1 D. Based on this, the template dilution of 1/1000 was deemed preferable and was the only temperature gradient used to determine optimal annealing temperature. The annealing temperatures 64.9°C and 66.0°C produced considerably fainter couplet product bands (Figure 6.4.1D), so were not considered optimal. The brightness of the undesirable upper bands seen using annealing gradients 50.0-56.3°C was greater than annealing temperatures 59.9°C and 62.8°C so were also deemed not optimal. A mean average between 59.9°C and 62.8°C was calculated (61.35°C) and rounded to 61.5°C, with this temperature considered to be the likely optimal annealing temperature.

6.4.2 Testing asymmetry of MAX randomisation of couplet 59+60 under optimised conditions

As described previously, MAX randomisation relies on the specific asymmetric amplification of the MAX codon-containing DNA strand. The optimal couplet 59+60 PCR conditions were used to amplify the 59+60 couplet in the absence of ligase (2.2.2.2). The PCR products were visualised via agarose gel electrophoresis (2.2.1.3).

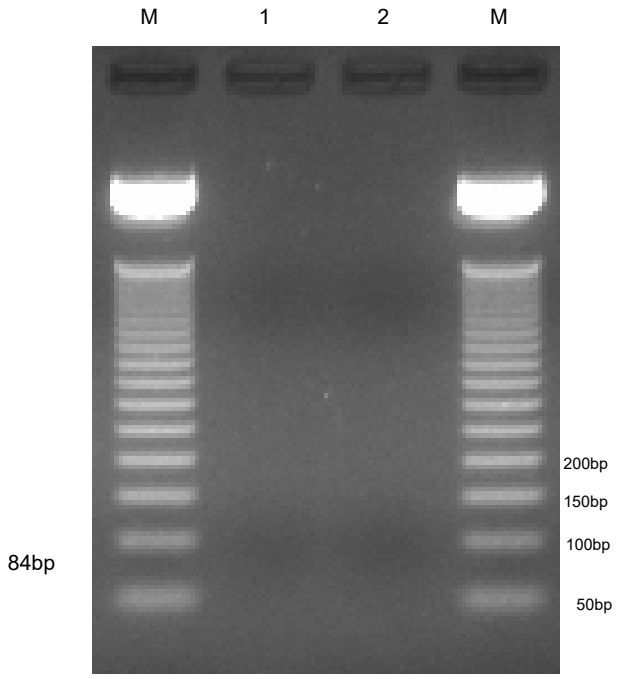

Figure 6.4.2: Couplet 59+60 negative control to assess asymmetric amplification. PCR amplification of 1/1000 diluted no-ligase Couplet 59+60 MAX randomisation product using optimised annealing temperature of 61.5°C, with an in cycle extension time of 10 seconds**.** Lanes: M=50bp MW ladder, **1**. Ligase-negative control, **2**. No template control.

No product was visualised from the amplification of the control MAX randomisation template (Figure 6.4.2), alongside a clean no template control. This meant using the annealing temperature 61.5°C alongside the 1/1000 dilution of couplet 59+60 MAX randomisation product as template were viable conditions for the specific asymmetric amplification of the MAX codon containing strand.

6.4.3 Couplet 59+60 construct for *MlyI* restriction for Quad cassette construction

With the optimal annealing temperature for couplet 59+60 amplification determined, a final PCR amplification was performed (2.2.4.1) to generate enough construct product volume for *MlyI* restriction (2.2.4.2) to expose the MAX randomised positions, ready for ligation with couplet 61+62 to form the Quad cassette. The PCR product was visualised via agarose gel electrophoresis (2.2.1.3).

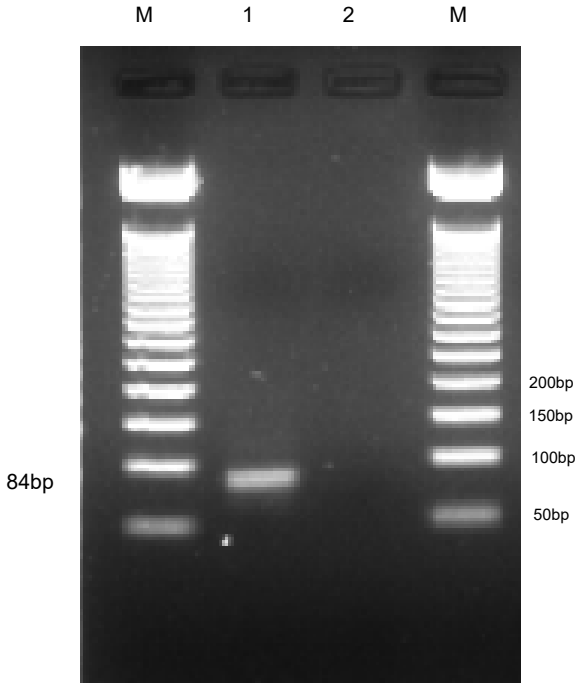

Figure 6.4.3: PCR amplification to generate 59+60 couplet construct.

A PCR amplification using 1/1000 diluted 59+60 MAX randomisation product as template with an annealing temperature of 61.5°C and in cycle extension time of 30 seconds. Lanes: M= 50bp MW ladder, **1**. Positive PCR reaction, **2**. No template control.

A bright single band of 84bp was produced (Figure 6.4.3), with a clean no template control. The PCR product was made in sufficient volume for purification (2.2.1.4) and quantified (2.2.1.5) for *MlyI* restriction.

6.4.4 *MlyI* restriction of couplet 59+60

As shown in Figure 6.2.1, for successful blunt-end ligation of the individual couplets, the 3' flanking oligonucleotide of couplet 59+60 had to be removed. The unique sequence flanking oligonucleotide of the couplet was designed to contain a 3'-5' directed *MlyI* restriction site (Figure 6.2.1) to facilitate the removal of the oligonucleotide resulting in the exposure of the couplet's MAX codons. The couplet 59+60 PCR product was restricted using *MlyI* (2.2.1.1) with the restriction products visualised via agarose gel electrophoresis (2.2.1.3)

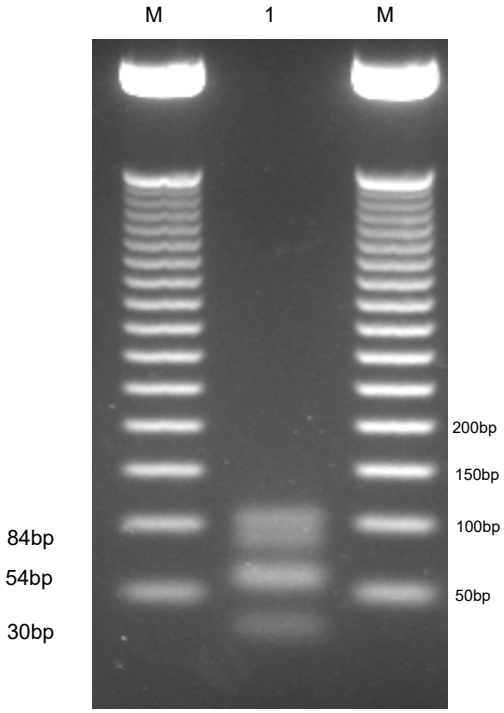

Figure 6.4.4: *MlyI* restriction of couplet 59+60.

Lanes: M= 50bp MW ladder, **1**. *MlyI* restriction of couplet 59+60. Restriction of couplet 59+60 (84bp) designed to result in two fragments of sizes, 54bp and 30bp.

Expected products of 30bp and 54bp were visualised, with some uncut couplet 59+60 construct remaining (84bp) (Figure 6.4.4), meaning the restriction was not 100% successful. A larger band approximately 105-110bp in size was also visualised, not present in the PCR amplification of the 59+60 construct (Figure: 6.4.3). As the 54bp product was the MAX containing product and therefore required downstream, the appearance of the 105-110bp band was not a concern as a mis-ligation using this band would have resulted in an visually larger band than was expected from the PCR amplification of the quad ligation product (90bp, Figure: 6.6.3).

6.5 Production of couplet 61+62 for generation of randomised contiguous region in the ParaMAX model library

With the construction and restriction of couplet 59+60 completed, the engineering of couplet 60+61 commenced. As before MAX randomisation (2.2.2.2) was used to generate the MAX codon containing DNA strand that would act as PCR template for couplet 61+62 PCR optimisation.

6.5.1 Determining the optimal couplet 61+62 MAX randomisation product template dilution and annealing temperature for PCR amplification

As with couplet 59+60, determining the optimal template dilution and annealing temperature was the first stage of PCR optimisation, achieved using four individual annealing temperature gradients (2.2.4.1), each using a different template dilution. The couplet 61+62 MAX randomisation construct, produced via MAX randomisation (2.2.2.2) was diluted using milliq water to achieve the four different template dilutions to be used: neat, 1/10, 1/100 and 1/1000. PCR products were visualised via agarose gel electrophoresis (2.2.1.3).

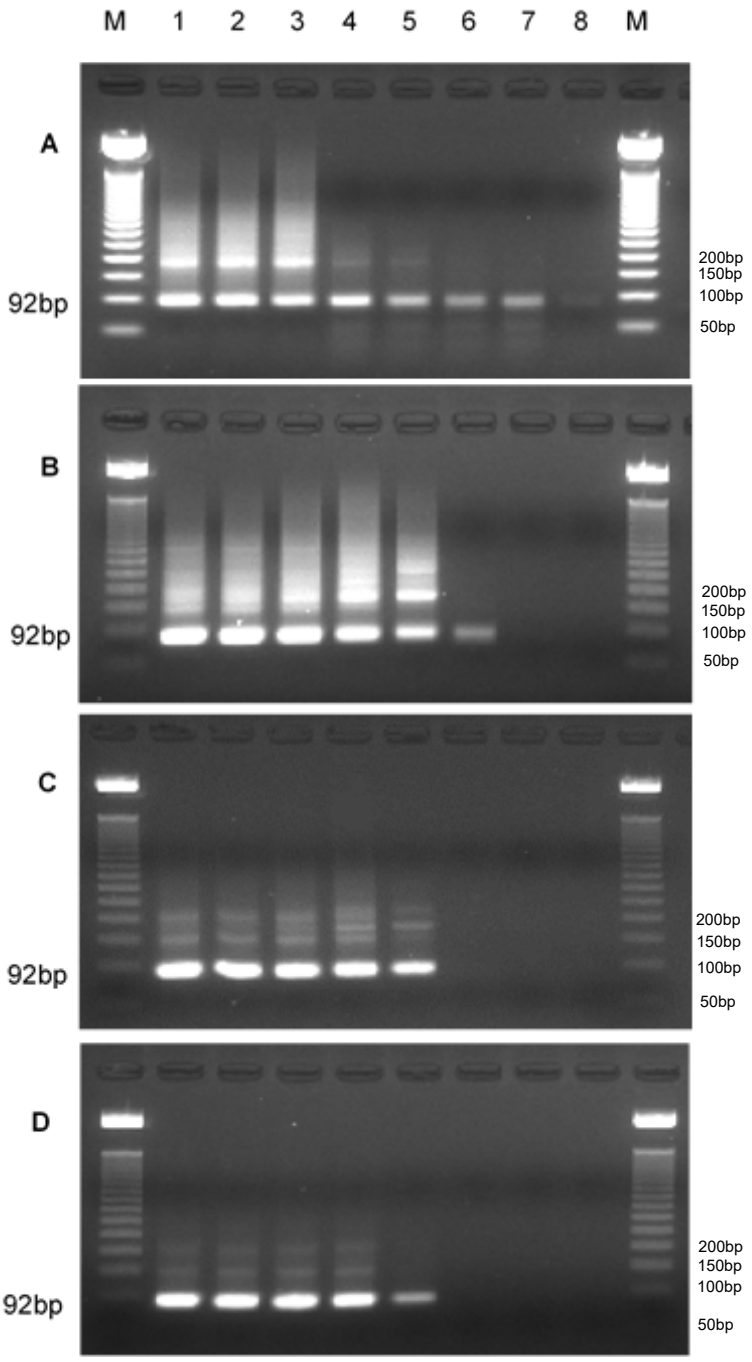

Figure 6.5.1: Four annealing temperature gradients using couplet 61+62 MAX randomisation product as template at varying dilutions to determine optimal annealing temperature and template dilution for PCR amplification.

Four individual PCR reactions were made using (**A**) neat MAX randomisation product, (**B**) 1/10 diluted MAX randomisation product, (**C**) 1/100 diluted MAX randomisation product and (**D**) 1/1000 diluted positive MAX randomisation as template and divided into 8 equal reactions. Annealing occurred at the temperatures described below with an in cycle extension time of 30 seconds. Lanes: M= 50bp MW ladder, annealing temperatures; **1**.50.0, **2**. 51.2, **3**.53.4, **4**.56.3, **5**. 59.9, **6**.62.8, **7**.64.9, **8**.66.0 (°C).

All template dilutions used across the four annealing temperature gradients produced the desired 92bp product (Figure 6.5.1). Across all of the annealing temperature gradients, the higher the annealing temperature used, the fainter the couplet 61+62 product became. This meant determining the optimal annealing temperature would be achieved by comparing the lower temperatures of each temperature gradient. Extensive smearing in Figure 6.5.1 A and B (template dilutions neat and 1/10 respectively) for the early annealing temperatures used, meant these template dilutions were not optimal and were no longer considered. A comparison between Figure 6.5.1 C and D, showed Figure 6.5.1C (1/100 template dilution) had noticeably more smearing and was therefore deemed not optimal. Figure 6.5.1D showed one annealing temperature that produced a single band (Lane 5, 59.9°C). This band was considerably fainter than those in lane 1-4 so was no longer considered. Lanes 1-4 of Figure 6.5.1 D, were incredibly similar so a mean average of the annealing temperature was taken between Lane 3 and 4 (53.4°C and 56.4°C) resulting in the annealing temperature of 54.9°C. This annealing temperature was to be used for downstream couplet 61+62 production alongside the 1/1000 diluted MAX randomisation product as template.

6.5.2 Testing asymmetry of MAX randomisation of couplet 61+62 under optimised conditions

As with couplet 59+60, the negative couplet 61+62 MAX randomisation product (generated in the absence of ligase) (2.2.2.2), was amplified using the optimal PCR conditions of 54.9°C using a 1/1000 dilution of the negative MAX randomisation product as template (2.2.4.1). The PCR products were visualised via agarose gel electrophoresis (2.2.1.3).

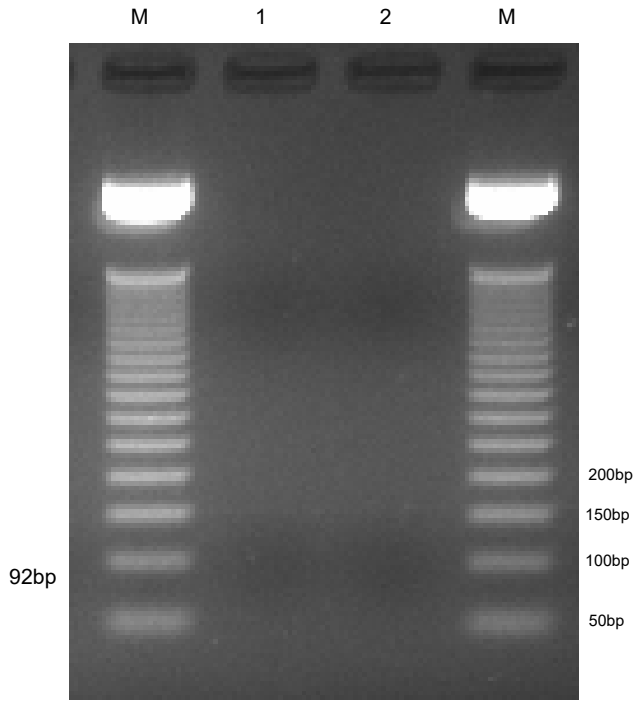

Figure 6.5.2 : Couplet 61+62 negative control to assess asymmetric amplification. PCR amplification of 1/1000 diluted no-ligase couplet 61+62 MAX randomisation product using optimised annealing temperature of 54.9°C, with an in cycle extension time of 10 seconds**.** Lanes: M=50bp MW ladder, **1**. Ligase-negative control, **2**. No template control. No 92bp product seen.

No product was visualised from the amplification of the control MAX randomisation template (Figure 6.5.2), alongside a clean no template control. This meant using the annealing temperature 54.9°C alongside the 1/1000 dilution of couplet 61+62 MAX randomisation product as template were viable conditions for the specific asymmetric amplification of the MAX codon containing strand.

6.5.3 Couplet 61+62 construct for *MlyI* restriction for Quad cassette construction

With the optimal annealing temperature for couplet 61+62 amplification determined a final PCR amplification was performed (2.2.4.1) to generate sufficient construct product volume for *MlyI* restriction (2.2.4.2) to expose the MAX randomised positions, ready for ligation with couplet 59+60 to form the Quad cassette. The PCR product was visualised via agarose gel electrophoresis (2.2.1.3).

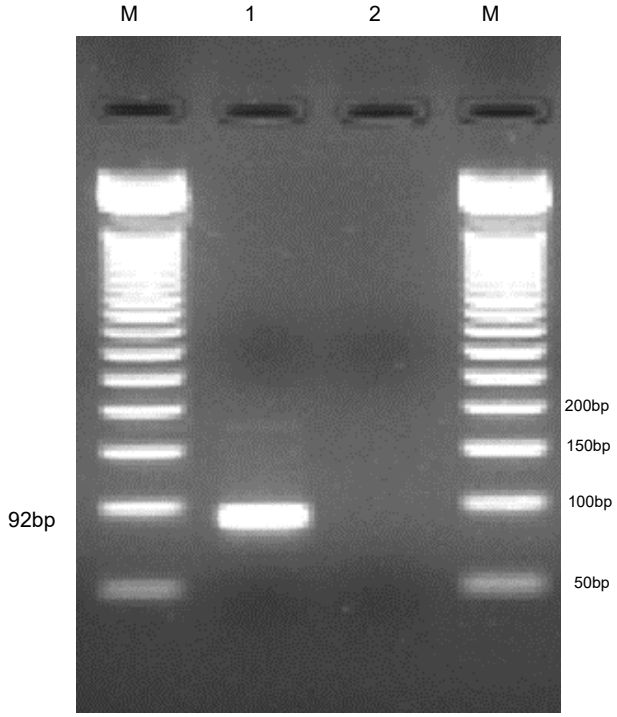

Figure 6.5.3: PCR amplification to generate 61+62 couplet construct.

A PCR amplification using 1/1000 diluted 61+62 MAX randomisation product as template with an annealing temperature of 54.9°C and an in cycle extension time of 30 seconds. Lanes: M= 50bp MW ladder, **1**. Positive PCR reaction, **2**. No template control. Expected 92bp product generated in positive reaction and a faint upper band of 175bp, while the no template control was clear.

A bright product band was seen at 92bp correlating to couplet 61+62 (Figure 6.5.3), albeit with a very faint smear of larger product. The no template lane was clean. The PCR product was made in sufficient volume for purification (2.2.1.4) and quantified (2.2.1.5) for *MlyI* restriction.

6.5.4 *MlyI* restriction of couplet 61+62

As shown in Figure 6.2.2, for successful blunt-end ligation of the individual couplets, the 5' flanking oligonucleotide of couplet 61+62 had to be removed. The sequencing of the couplet was designed to contain a 5'-3' directed *MlyI* restriction site (Figure 6.2.2) to facilitate the removal of the oligonucleotides and expose the couplet's MAX codons. The couplet 61+62 PCR product was restricted using *MlyI* (2.2.1.1) with the restriction products visualised via agarose gel electrophoresis (2.2.1.3).

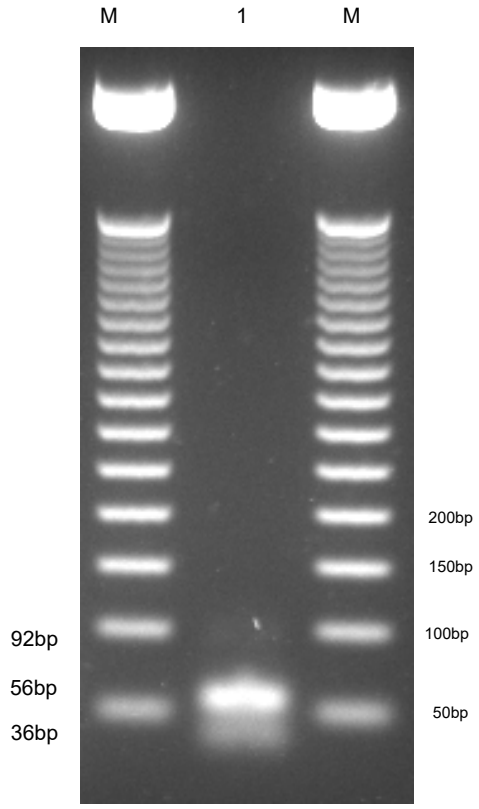

Figure 6.5.4 *: MlyI* restriction of couplet 61+62.

Lanes: M= 50bp MW ladder, **1**. *MlyI* Restriction of couplet 61+62 (92bp) resulted in two fragments of size 56bp and 36bp. Minimal full length couplet 60+61 product (92bp) seen.

The expected product bands of 56bp and 36bp were seen with some smearing between the bands present (Figure 6.5.4), suggesting a longer electrophoresis run time would have been beneficial. The *MlyI* restriction of couplet 61+62 was considered successful as two bands could be identified, with the MAX containing 36bp product used downstream.

6.6 Production of the Quad cassette construct for incorporation into the ParaMAX model library

With the successful construction and restriction of both couplets, the construction of the Quad could take place. As shown in Figure 6.2.3 the Quad was generated using blunt end ligation (2.2.1.2), between the restricted couplets, with the ligation product, acting as template for subsequent PCR optimisation and amplification (2.2.4.1).

6.6.1 Determining the optimal Quad ligation product template dilution and annealing temperature for PCR amplification

As with the individual couplet construction, the first stage in Quad PCR optimisation was to determine the optimal annealing temperature and template dilution. This was achieved using four individual annealing temperature gradients (2.2.4.1), each using a different template dilution. The Quad ligation product was diluted using milliq water to achieve the four different template dilutions to be used, neat, 1/10, 1/100 and 1/1000. PCR products were visualised via agarose gel electrophoresis (2.2.1.3).

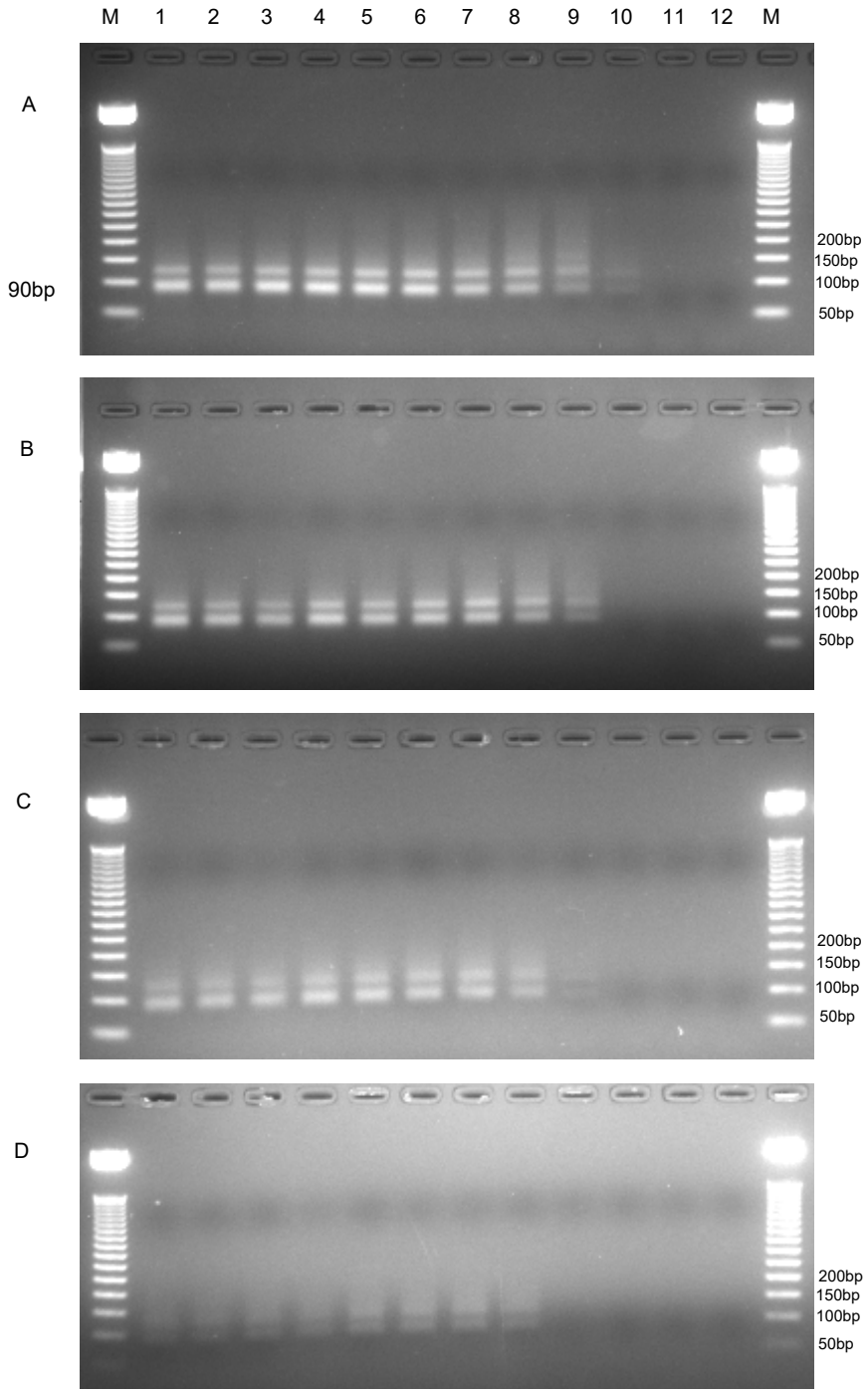

Figure 6.6.1: Four annealing temperature gradients using Quad ligation product as template at varying dilutions to determine optimal annealing temperature and template dilution for PCR amplification.

Four individual PCR reactions were made using (**A**) neat Quad ligation product, (**B**)1/10 diluted Quad ligation product, (**C**) 1/100 diluted Quad ligation product and (**D**) 1/1000 diluted Quad ligation as template and divided into 12 equal reactions. Annealing occurred at the temperatures described below with an in cycle extension time of 10 seconds. Lanes: M= 50bp MW ladder, annealing temperatures; **1**.45.0, **2**.45.4, **3**.46.5, **4**.48.5, **5**.51.1, **6**.53.7, **7**.56.1, **8**.58.7, **9**.61.2, **10**.63.2, **11**.64.4, **12**.65.0(°C).

All template dilutions generated a 90bp product with varying degrees of success across lanes 1-8 (45.0-58.7°C) (Figure 6.6.1) with template dilutions 1/10 and 1/100 (Figure 6.6.1 C and D respectively) also showing Quad product in lane 9 (61.2°C), yet noticeably fainter. The neat template dilution annealing temperature gradient was also able to produce the Quad construct using a temperature of 63.2°C (lane 10, Figure 6.6.1 A). All annealing temperatures capable of generating the 90bp Quad product across all template dilutions, also produced an upper band approximately 125bp in size, of a similar brightness to the Quad band. As this upper band was a constant, the brightness of the Quad construct band was the deciding factor in determining the optimal template dilution and annealing temperature for downstream PCR optimisation. The brightest Quad product bands were seen using the neat template dilution (Figure 6.6.1 A), so was the only one considered when identifying the optimal annealing temperature. Quad construct bands of acceptable brightness were seen in lanes 1-6 (Figure 6.6.1 A), with a marginal increase in smearing as the annealing temperature increased. Because of this, a balance between smearing and band brightness had to be made, so the midrange temperature 48.5°C was chosen as the downstream annealing temperature to be used alongside the neat template dilution.

6.6.2 Using specific denaturation temperatures to eliminate a higher molecular weight construct formed during Quad PCR amplification

The 125bp product seen across Figure 6.6.1, had a consistent brightness relative to the 90bp Quad product it accompanied, meaning as the Quad product band changed across the different templates used, or the annealing temperature range, so did the 125bp product in the same way. This meant the traditional PCR optimisation techniques of investigating the optimal template dilution and annealing temperature failed to give a single Quad construct band (Figure 6.6.1). A method of removing the higher molecular weight band was required. The assumption that the 125bp product would require a higher denaturation temperature when compared with the 90bp Quad was the foundation for using a denaturation temperature gradient. The denaturation temperature gradient would identify a temperature high enough to break open the shorter Quad DNA strands allowing for their amplification, while not being sufficiently high to separate the 125bp DNA strands. Using the neat Quad ligation product as template and an annealing temperature of 48.5°C, a denaturation gradient was performed (2.2.4.1), with the PCR products visualised via agarose gel electrophoresis (2.2.1.3).

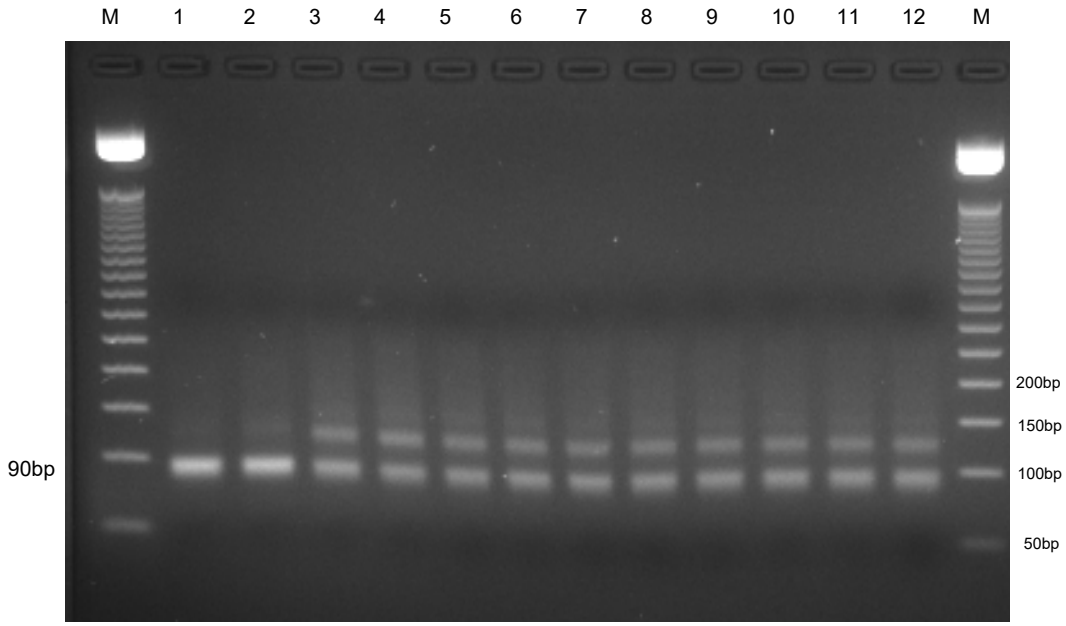

Figure 6.6.2.1: Denaturation temperature gradient using neat Quad ligation product as template to determine the optimal denaturation temperature for PCR amplification.

A single PCR reaction was divided into 12 equal reactions. Denaturation occurred at the temperatures listed below, with annealing at 48.5°C with an in cycle extension time of 10 seconds. Lanes: M= 50bp MW ladder, annealing temperatures; **1**.80.0, **2**.80.3, **3**.81.2, **4**.82.6, **5**.84.6, **6**.86.5, **7**.88.4, **8**.90.3, **9**.92.2, **10**.93.7, **11**.94.6, **12**.95.0 (°C).

The 125bp band brightness reduced as the denaturation temperature decreased, with the upper band almost completely removed in lane 1 of Figure 6.6.2.1, using a denaturation temperature of 80.0°C. This Quad band alongside a very similar lane 2 result, was also considerably brighter than the Quad products seen in lanes 3-12. This correlation between reduced upper band brightness and decreasing denaturation temperature, prompted investigation of a lower denaturation temperature range between 70°C and 80°C (2.2.4.1).

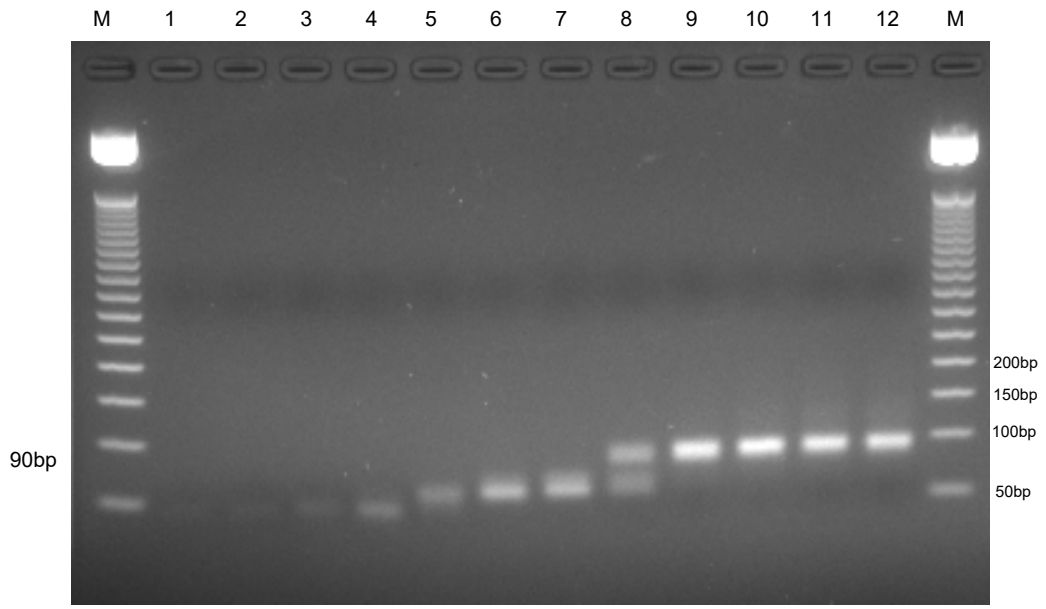

Figure 6.6.2.2: Lower denaturation temperature gradient using neat Quad ligation product as template to determine the optimal denaturation temperature for PCR amplification.

A single PCR reaction was divided into 12 equal reactions. Denaturation occurred at the temperatures listed below, with annealing at 48.5°C with an in cycle extension time of 10 seconds. Lanes: M= 50bp MW ladder, annealing temperatures; **1**.70.0, **2**.70.2, **3**.70.8, **4**.71.8, **5**.73.1, **6**.74.4, **7**.75.6, **8**.76.9, **9**.78.2, **10**.79.2, **11**.79.8, **12**.80.0 (°C).

The lower range denaturation gradient (Figure 6.6.2.2) was successful in removing the upper 125bp band seen in the upper range denaturation gradient (Figure 6.6.2.1). Lanes 9-12 (denaturation temperatures 78.2-80.0°C) produced a bright Quad construct band in the absence of the 125bp construct, so a midrange temperature (79.2°C) was chosen as the optimal denaturation temperature and used in the downstream production of the Quad construct. With the upper 125bp band removed, no investigations into the source of the band were conducted.

6.6.3 Production of the Quad cassette construct using optimised PCR conditions

Using the optimised denaturation temperature, annealing temperature and template dilution, a final PCR amplification to produce the Quad construct was performed (2.2.4.1).

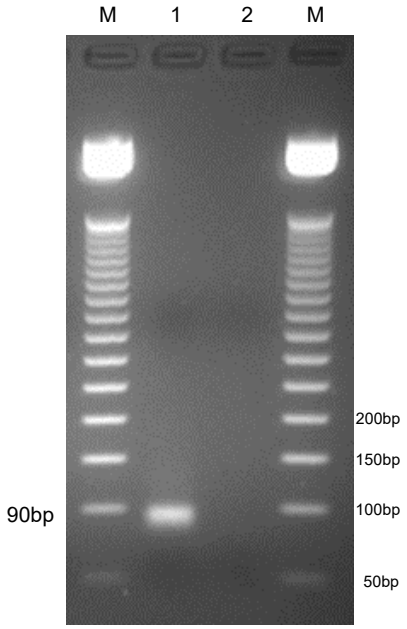

Figure 6.6.3: PCR amplification to generate Quad construct

A PCR amplification using 1/1000 diluted Quad ligation product as template with a denaturation temperature of 79.2°C, an annealing temperature of 48.5°C and an in cycle extension time of 10 seconds. Lanes: M= 50bp MW ladder, **1**. Positive PCR reaction, **2**. No template control.

A single bright Quad construct band was produced, with a clean no template control (Figure 6.6.3), meaning the construction of the Quad was completed successfully.

6.7 Production of the conserved region for incorporation in the ParaMAX model library

With the successful construction of the Quad cassette, the generation of the conserved region began. The conserved region was constructed via overlap PCR (2.2.4.2), between its two constituent oligonucleotides seen in Figure 6.1.2. The full length conserved construct was then diluted to form a 1/1000 dilution which was used as template in downstream PCR optimisation and amplification (2.2.4.1).

6.7.1 Determining the optimal annealing temperature for conserved region PCR amplification

To determine the optimal annealing temperature for the conserved region PCR amplification, an annealing gradient using a 1/1000 dilution of the conserved overlap PCR product as template was performed (2.2.4.1). PCR products were visualised via agarose gel electrophoresis (2.2.1.3).

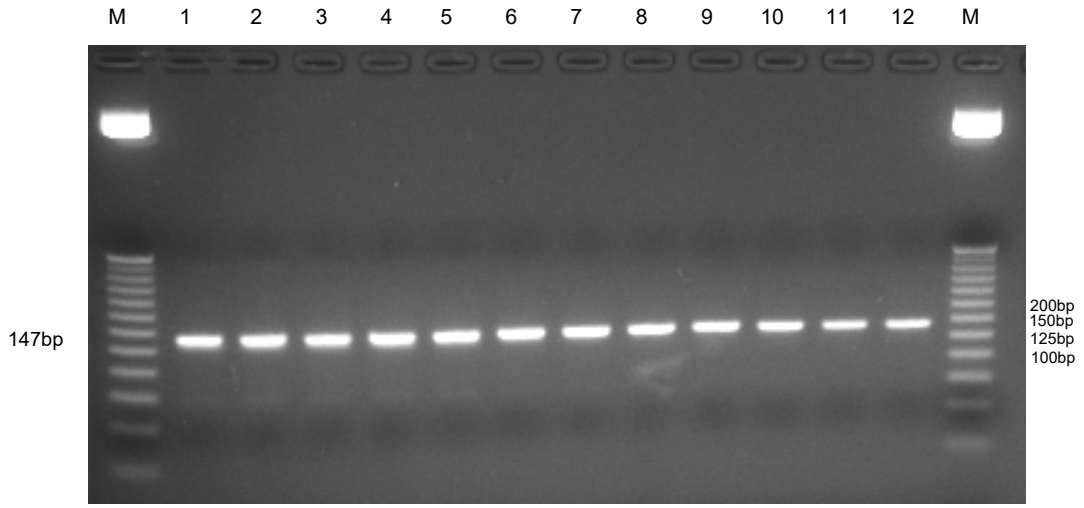

Figure 6.7.1: Annealing temperature gradient using 1/1000 diluted conserved region overlap product as template to determine optimal annealing temperature for PCR amplification.

Annealing occurred at the temperatures described below with an in cycle extension time of 10 seconds. Lanes: M= 25bp MW ladder, annealing temperatures; **1**.45.0, **2**.45.3, **3**.46.2, **4**.47.6, **5**.49.6, **6**.51.5, **7**.53.4, **8**.55.3, **9**.57.2, **10**.58.7, **11**.59.6, **12**.60.0(°C).

A bright conserved region construct band (147bp) was seen across the entire annealing temperature range (Figure 6.7.1), with band thickness decreasing as the annealing temperature increased. Lanes 1-7 produced a lower molecular weight product (approximately 70bp), which decreased in brightness as the annealing temperature increased. A balance between band quality and avoiding an annealing temperature that produced non-conserved region products, meant the annealing temperature 57.2°C (lane 9 Figure 6.7.1) was chosen as the optimal condition.

Using a 1/1000 dilution conserved region product as template and an annealing temperature of 57.2°C, the conserved region was then generated in sufficient volume for downstream applications (2.2.4.1).

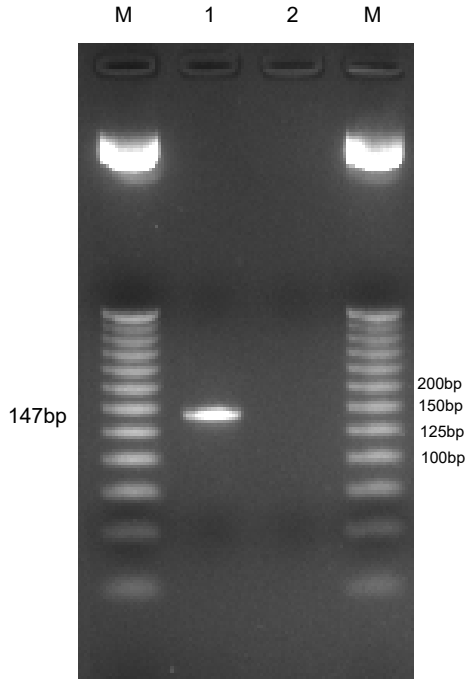

Figure 6.7.2: PCR amplification to generate the conserved region construct.

A PCR amplification using 1/1000 diluted conserved region overlap PCR product as template with an annealing temperature of 57.2°C with an in cycle extension time of 10 seconds. Lanes: M= 25bp MW ladder, **1**. Positive PCR reaction, **2**. No template control.

A single bright conserved region band was produced, with a clean no template control (Figure 6.7.2).

6.8 Construction of the complete ParaMAX model DNA library

With the successful construction of both the Quad and conserved region, the generation of the full ParaMAX model library began. The ParaMAX model full length library was constructed via overlap PCR (2.2.4.2), between the conserved region and Quad constructs (Figure 6.1.2).

6.8.2 Determining the optimal ParaMAX model library overlap product template dilution and annealing temperature for PCR amplification

Determining the optimal annealing temperature and template dilution was the first stage of PCR optimisation and was achieved using four individual annealing temperature gradients (2.2.4.1), each using a different template dilution. The ParaMAX model library overlap PCR product was diluted using milliq water to achieve the four different template dilutions to be used: neat, 1/10, 1/100 and 1/1000. PCR products were visualised via agarose gel electrophoresis (2.2.1.3).

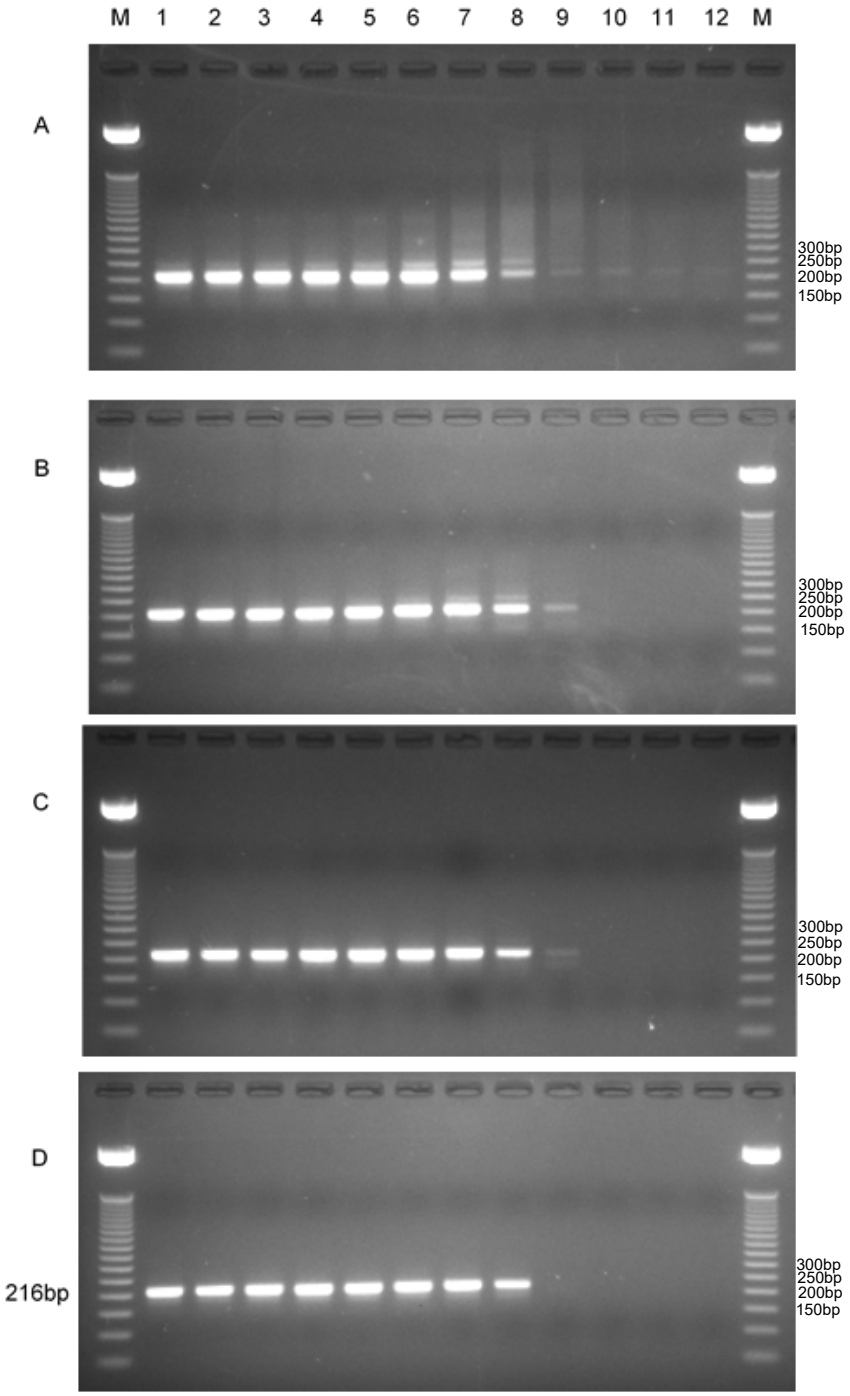

Figure 6.8.2: Four annealing temperature gradients using full length construct overlap product as template at varying dilutions to determine optimal annealing temperature and template dilution for PCR amplification.

Four individual PCR reactions were made using (**A**) neat overlap product, (**B**)1/10 diluted overlap product, (**C**) 1/100 diluted overlap product and (**D**) 1/1000 diluted overlap product as template and divided into 12 equal reactions. Annealing occurred at the temperatures described below with an in cycle extension time of 10 seconds. Lanes: M= 50bp MW ladder, annealing temperatures; **1**.45.0, **2**.45.4, **3**.46.5, **4**.48.5, **5**.51.1, **6**.53.7, **7**.56.1, **8**.58.7, **9**.61.2, **10**.63.2, **11**.64.4, **12**.65.0(°C).

A bright 216bp product band was seen across all template dilutions for lanes 1-7 (Figure 6.8.2), an annealing temperature range between 45.0-56.1°C inclusive, with the band quality decreasing in lanes 8-12. Because of this decrease in band quality as the annealing temperature increased, only lanes 1-5 across the template dilutions were from this point used to determine optimal template dilution and annealing temperature to avoid any potential band quality reduction. The smearing seen using the neat template dilution (Figure 6.8.2 A), meant that template dilution was not considered optimal. Lanes 1-5 of the 1/10, 1/100 and 1/1000 template dilutions (Figure 6.8.2 B,C and D respectively) were all very similar. Ultimately the annealing temperature of 45.0°C using a 1/1000 template dilution (Figure 6.8.2 D) were determined to be the optimal conditions for downstream PCR reactions.

6.8.3 Using optimal conditions to produce the complete ParaMAX model cassette

A final, large-scale PCR amplification of the complete ParaMAX model cassette was performed (2.2.4.1), with the PCR product visualised via agarose gel electrophoresis prior to purification (2.2.1.3). The PCR product was purified (2.2.1.4), quantified (2.2.1.5) and sent for Next Generation Sequencing (2.2.1.6).

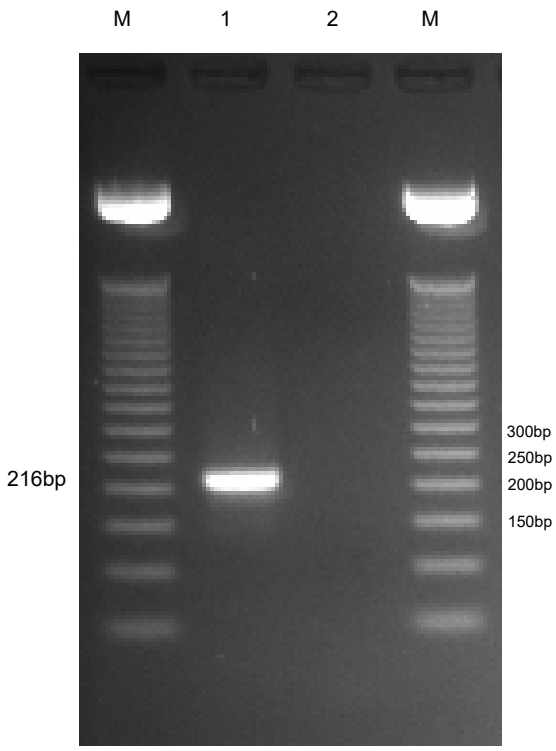

Figure 6.8.3: PCR amplification to generate full length ParaMAX model library construct.

A PCR amplification using 1/1000 diluted full length construct overlap product as template with an annealing temperature of 45.0°C with an in cycle extension time of 10 seconds. Lanes: M= 50bp MW ladder, **1**. Positive PCR reaction, **2**. No template control.

6.8.4 Concluding ParaMAX randomisation for the saturation of four contiguous positions in a modified designed armadillo repeat protein DNA backbone

The necessity for engineering randomised DNA libraries containing contiguous saturated positions is vital for the success of PRe-ART. The requirement to investigate the impact of saturating stretches of adjacent codons in an efficient manner, is perfectly demonstrated by the potential introduction of loops to maximise the depth of the wild-type lysine binding pocket, to accommodate larger amino acids. The invention of ParaMAX was the ideal solution.

The randomised cassette generated using ParaMAX randomisation was based upon the designed armadillo repeat protein DNA sequence, but unlike the single arginine library (Chapter Three) and the At-Thr Library (Chapter Four) did not possess the flanking regions required for homologous recombination to generate the full DNA sequence used to produce

and then screen the protein libraries. The focus of the ParaMAX library was to analyse the success of the positional randomisation of the contiguous codons, with no intentions of this randomised cassette to be used downstream in binder investigations. After determining the success of ParaMAX randomisation, a library design for extension of the single lysine pocket by introducing a novel stretch of randomised codons, would have been created and ParaMAX implemented in engineering the randomised pocket.

As seen in Figure 7.4, this preliminary attempt to engineer a contiguous randomised region of four codons, revealed areas requiring optimisation and further investigation, primarily the *MlyI* restrictions involved to form the Quad cassette. Further work within the PRe-ART group is underway and discussed extensively throughout section 7.5.

7.0 The optimisation of an excel based NGS analysis technique to calculate codon representation at saturated positions in the model ParaMAX DNA library

7.1 Introduction

With the construction of the ParaMAX model library (Figure 6.8.3) complete, determining the amino acid representation at each of the four newly introduced positions, would determine how successful the non-degenerate saturation via ParaMAX randomisation had been. An accurate count of amino acid representation at each of the four novel positions, would need to be performed. As the model ParaMAX library was structurally different to the At-Thr library, in that it contained a contiguous region of randomisation, modifications and subsequent optimisations of the methodology used to analyse the At-Thr library were required. This new methodology, still used the open source web based bioinformatics platform, Galaxy (https://usegalaxy.org/), to process the raw Illumina sequencing data and Microsoft Excel to perform codon counts to determine the amino acid distribution at saturated potions.

This chapter focuses on the development of a NGS processing and codon counting methodology for libraries containing contiguous randomised regions. The library used was the model ParaMAX library (Chapter 6), containing a contiguous region of four saturated positions, randomised using ParaMAX (2.2.2.3).

7.2 Processing the raw Illumina sequencing data of the model ParaMAX DNA library using Galaxy

Initial Galaxy processing was divided into three stages (Figure 7.2), with each processing stage optimised using the output of the previous function.

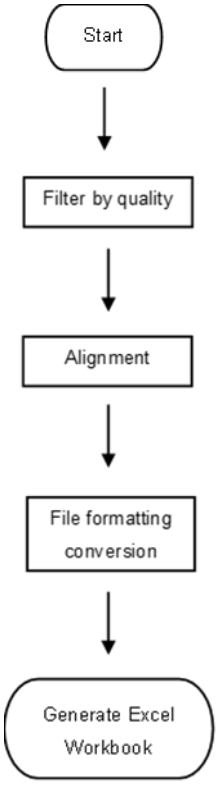

Figure 7.2: Flow diagram depicting pre-count processing stages of NGS data using the free online bioinformatics server, Galaxy, for the model ParaMAX library.

Filter by quality: using the associated quality score for each read to determine if the quality is satisfactory, Alignment: aligning the filtered reads with a reference sequence. Alignment output is then formatted to be used in an Excel Workbook.

The model ParaMAX library was sequenced using Miseq, with two 250bp reads 5'-3' and 3'-5', as this was the default sequencing option provided by Genewiz (2.2.1.6). The production of both a 250bp forward read (mate 1) and 250bp reverse read (mate 2) was not necessary for the analysis of the model ParaMAX library as the complete construct length was 216bp. Because of this, only one directional mate was required for full library sequence analysis (read 1 was chosen), making the Fastq Join function redundant.

7.2.1 Quality control using Filter by Quality function on mate 1 reads of the model ParaMAX library

The Filter by Quality function uses two user defined parameters to determine if a read is discarded; the ‘quality cut-off value’ and the ‘percent of bases in sequence that must have quality equal to/higher than cut-off value’ (2.2.5.2.2).

As in the analysis of MAX randomisation libraries, a base quality score of 30 correlated to a 1/1000 chance of a base identity call being incorrect (Illumina, 2014), so was used as the ‘quality cut off value’. Quality scores are representative of the confidence in a base calling, so non-identifiable bases (Ns) will have an intrinsically lower base call confidence and will therefore impact on the total read quality. With the Filter by Quality function assessing the quality score of every base in a read sequence (including randomised positions), adjustments to ‘the percent of bases in sequence that must have quality equal to/higher than cut-off value’ parameter were made. This accommodation for randomisation was done by calculating the percentage of randomised bases in the library and lowering the value for ‘percent of bases in sequence that must have quality equal to/higher than cut-off value’ to the nearest whole number under the randomised base percentage. The model ParaMAX library contained 12 randomised bases 12/216*100=5.56 (2.dp), so 94% was used. Using the Filter by Quality with a quality cut off value of 30 and a 94% requirement of bases in the sequence to have a quality score of 30 or more resulted in an output of 186127 reads with 44079 (19.15%) discarded. When considering the mate 1 reads had not been subjected to the quality control stages involved in the Fastq Join function (Figure 5.2.1), discarding 19.15% of the total number of reads was deemed acceptable.

7.2.2 Using Bowtie2 to align model ParaMAX library reads to a reference sequence

In order to align the library reads using Bowtie2, a reference sequence was required. This sequence was the model ParaMAX sequence with MAX codons substituted for NNN at the randomised positions, made using UGENE (2.2.5.1). This reference sequence was inputted into the alignment function along with the 186127 filter by quality outputted reads (2.2.5.2.4). The Bowtie2 output file was then reformatted (2.2.5.2.6) so the read data could be used in Excel.

7.3 Invariant anchor count methodology to determine codon frequencies' at saturated positions in the model ParaMAX library

The randomised region in the model ParaMAX library was made of four contiguous randomised positions. This meant using the MAX selection oligonucleotide invariant region as a count anchor (as was used in the At-Thr counts (Table 5.3)) was not possible for positions 60 and 61 (Figure 6.1.2). Instead, the 12bp sequence adjacent to position 59 on the 5' end of the randomised region (TCTAATATCGCT) was used as the invariant anchor. The choice of a 12bp anchor corresponded to the length of the randomised region, reducing the possibility of the anchor appearing in the randomised region to less than 0.0008% as each of the positions was saturated using 19 different MAX selection oligonucleotides during ParaMAX randomisation (Table 6.3) (2.2.2.3) ((((1/19)^4)*100= 0.0008 (2.dp)). To count codon frequencies at positions 60, 61 and 62, the same 12bp anchor was used, alongside '???', which in Excel represents any three characters (sequence bases). An example of this is shown in Table 7.3, where aspartic acid is counted for at each of the four randomised positions, with the use of '???' to represent the previously counted codon, moving the focus of the Countif function to the next randomised position.

Randomised position	Excel function used to count aspartic acid frequency
59	SUM(COUNTIFS(N:N,{"*TCTAATATCGCTGAT*","*TCTAATATCGCTGAC*"}))
60	SUM(COUNTIFS(N:N,{"*TCTAATATCGCT???GAT*","*TCTAATATCGCT???GAC*"}))
61	SUM(COUNTIFS(N:N,{"*TCTAATATCGCT??????GAT*","*TCTAATATCGCT??????GAC*"}))
62	SUM(COUNTIFS(N:N,{"*TCTAATATCGCT?????????GAT*","*TCTAATATCGCT?????????GAC*"}))

Table 7.3: Countif function to determine summed frequency of aspartic acid at each of the four saturated positions in the model ParaMAX library.

7.4 Analysing the amino acid distribution at saturated positions in a Galaxy processed Bowtie2 aligned model ParaMAX library

Once all counts had been performed at each of the four saturated positions in the model ParaMAX library, the raw codon counts were converted to codon frequency (%) and analysed graphically (Figure 7.4).

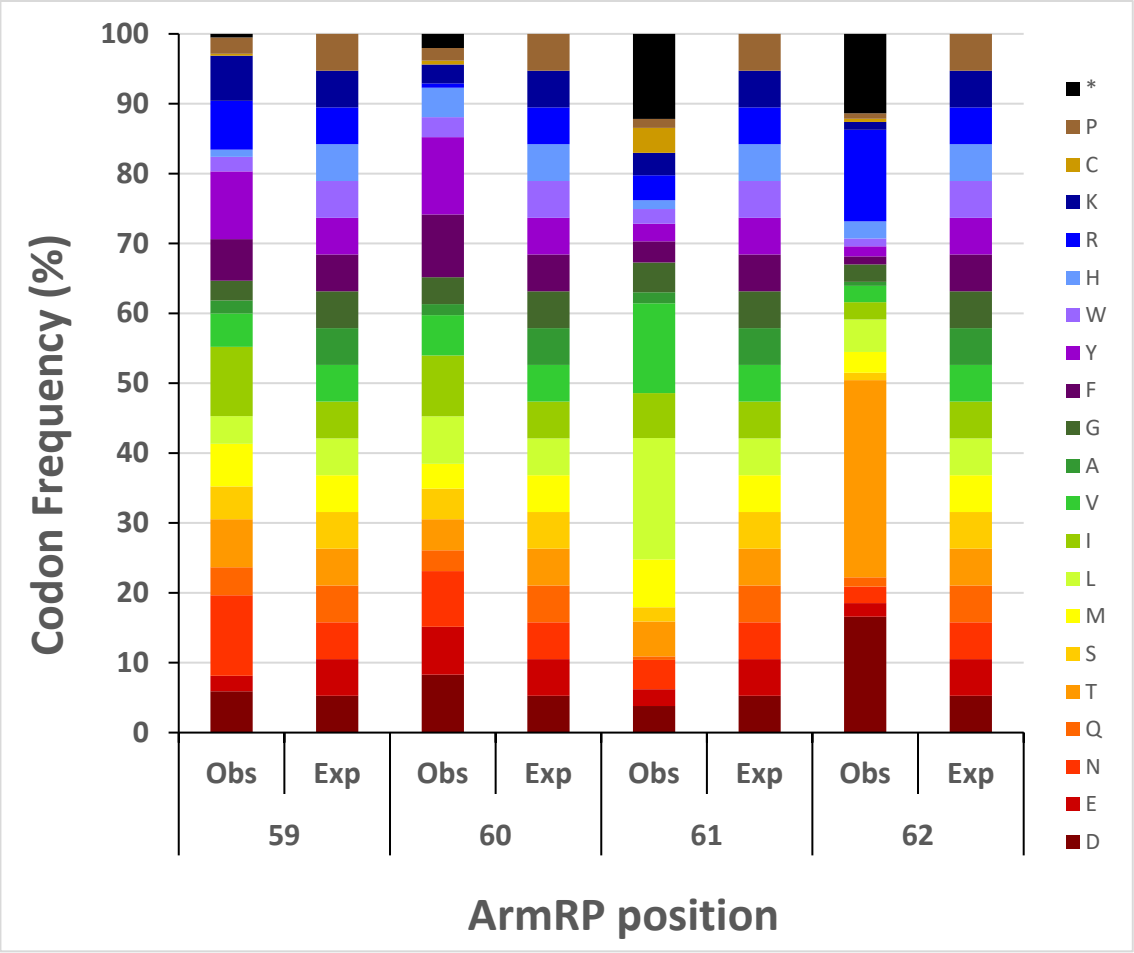

Figure 7.4: Observed vs expected amino acid distribution for positions randomised in the Bowtie2 aligned model ParaMAX library.

Letters in the legend correspond to universal amino acids abbreviations in the genetic code: P: proline; C: cysteine; K: lysine; R: arginine; H: histidine; W: tryptophan; Y: tyrosine; F: phenylalanine; G: glycine; A: alanine; V: valine; I: isoleucine; L: leucine; M: methionine; S: serine; T: threonine; Q: glutamine; N: asparagine; E: glutamic acid; D: aspartic acid; *: stop codon. Raw data from Miseq Illumina sequencing. (See Appendix 7 for raw count data).

A comparison between the observed and expected codon frequency and therefore the amino acid distribution, across all four randomised positions showed an over or under representation of multiple amino acids (Figure 7.4). The most concerning observation was the unexpectedly large percentage of termination codons present at positions 61 and 62 (black bars, Figure 7.4). Further investigation was required to determine the cause of the poor correlation between the observed and expected amino acid frequencies at these positions.

7.5 Investigating potential contributing factors to the observed amino acid distribution at randomised positions in the model ParaMAX library

Clearly, the application of ParaMAX to randomise the four novel positions in an unbiased manner, particularly at positions 61 and 62, was unsuccessful. The problem might lie either with the construction of the DNA cassette itself, or else with the in silico analysis of the sequencing data, using the parameters selected. Since the data was already available, the in silico analysis was first examined in detail. Several possible factors were identified as potentially contributing to the amino acid distribution observed in Figure 7.4, with each requiring their own separate investigations.

7.5.1 Determining the impact of count anchor on the observed amino acid distribution at randomised positions in the model ParaMAX library

The success of the codon counts relies on the count anchor selected. The uniqueness of the sequence is imperative to accurate counts as demonstrated during the At-Thr library analysis (Figure 5.3). A search of the model ParaMAX library sequence for the 12bp anchor sequence (TCTAATATCGCT) was performed and the uniqueness of the sequence was confirmed, eliminating the anchor as a contributing factor to the amino acid distribution.

7.5.2 Assessing the quality of the Illumina sequence data for the model ParaMAX library.

Before any investigations into the Galaxy processing stages occurred, the raw Illumina data was scrutinised using the Galaxy function, Filter FASTQ. This function had two parameters, each with a subset of user defined values that could be used to investigate the raw FASTq sequencing file: read length and base quality score. When scrutinising the raw data using length, user defined values for minimum and maximum read length could be input, while a minimum and maximum per base quality score was also user determined. Each parameter and the impact on the raw sequencing data, was investigated individually.

7.5.2.1 Investigating the impact of quality score filtering on the raw Illumina sequencing data of the model ParaMAX library

The quality scoring aspect of the Filter FASTQ function allowed a user defined minimum and maximum quality score, as well as the number of bases allowed outside of the quality range (2.2.5.2.3). The quality score range was set to 30 or more, (1/1000 chance of the base call being incorrect, ((Illumina 2014) with a 12 base allowance outside of the range to accommodate for the 12 randomised bases in the model ParaMAX library (Figure 6.1.2). Of the 230206 inputted reads in the mate 1 FASTq file, 186371 reads (80.96% (2.dp)) were within the Filter FASTq quality scoring filter parameters. The stringent filter parameter of only having 12 bases outside of the quality score range, meant a single base change in the conserved region of the model ParaMAX sequence would result in a read being discarded. In this light, 80.96% of reads having no such base substitutions was deemed acceptable. As the original Filter by Quality (2.2.5.2.2) performed on the model ParaMAX library using a quality score of 30 and a 94% sequence requirement to have that score or more, resulted in a read output of 186127 from the same input file (slight variation as Filter by Quality uses percent of bases instead of a defined number), it was concluded the quality of the raw Illumina reads was not a contributing factor to the amino acid distribution seen in Figure 7.4.

7.5.2.2 Investigating the impact of length filtering on the raw Illumina sequencing data of the model ParaMAX library

With the quality of the raw Illumina reads within an acceptable level, the impact of filtering the raw read 1 sequencing data by length was investigated. The full length model ParaMAX library construct was 216bp. The raw non-quality checked mate 1 FASTq file was inputted into the Filter FASTQ function, with the minimum and maximum length read parameters both set to 216bp, with the quality range set to accept all base scores (2.2.5.2.3). Of the 230206 input reads, 46618 were 216bp (20.25%). Using the same length filter parameters, but instead inputting the Filter Fastq quality score filtered reads resulted in 37724 of 186127 being kept (20.24%). This considerable reduction in the number of reads kept, highlighted read length of the raw Illumina reads as a potential factor for the amino acid distribution observed in Figure 7.4 and warranted further investigation.

7.5.3 Analysing the amino acid distribution at saturated positions in a Filter FASTq processed Bowtie 2 aligned model ParaMAX library

The key variation between the Galaxy processes outlined in Figure 7.2, which generated the amino acid distribution observed in Figure 7.4, and the new model ParaMAX library processing to investigate the impact of read length on amino acid distribution was the additional filtering by length step. Unlike the previous Galaxy processing (Figure 7.2), the filter by quality and the new filter by length was performed simultaneously by the Filter FASTq function (2.2.5.2.3) (Figure 7.5.3.1).

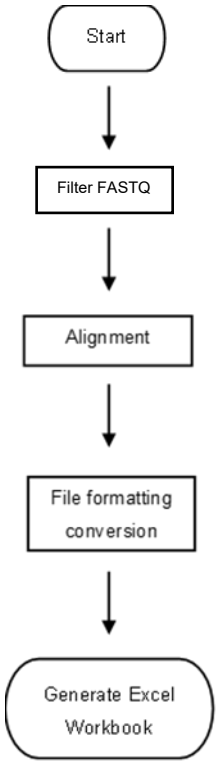

Figure 7.5.3.1: Flow diagram depicting the refined pre-count, Galaxy processing stages for the model ParaMAX library NGS data.

Filter FASTq: using the associated quality score for each read to determine if the quality is satisfactory and filtering by read length, Alignment: aligning the filtered reads with a reference sequence. Alignment output is then formatted to be used in an Excel Workbook.

The 37724 outputted reads of the Filter FASTq function (2.2.5.2.3) (read length of 216bp only, Q=30 or more with a 12 base exception) (16.39% (2.dp) of the inputted 230206 reads), were then inputted alongside the model ParaMAX library reference sequence generated using UGENE (2.2.5.1) into the Bowtie2 aligner (2.2.5.2.4). The Bowtie2 output file was reformatted (2.2.5.2.6) and the codon counts for each of the four randomised positions performed. The raw counts were changed to codon frequency per position and displayed graphically (Figure 7.5.3.2).

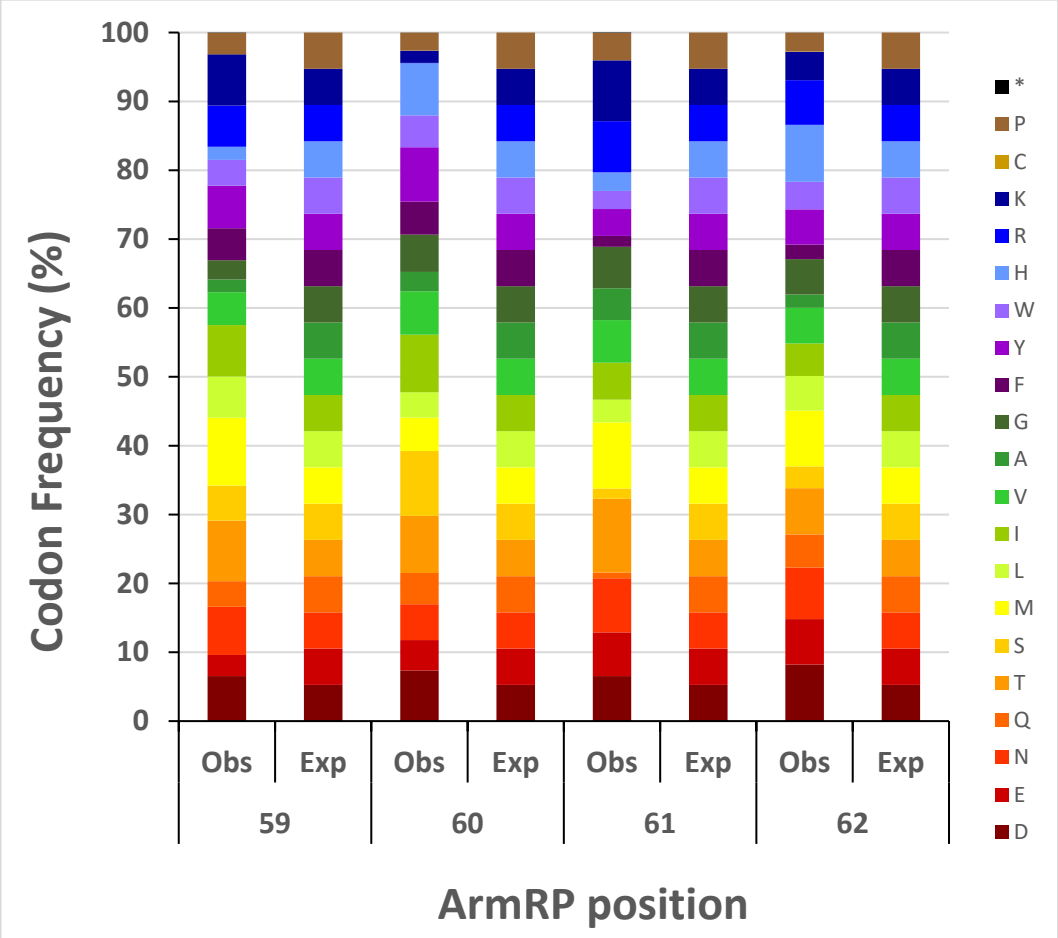

Figure 7.5.3.2: Observed vs expected amino acid distribution for positions randomised in the Filter FASTq processed and Bowtie2 aligned model ParaMAX library.

Letters in the legend correspond to universal amino acids abbreviations in the genetic code: P: proline; C: cysteine; K: lysine; R: arginine; H: histidine; W: tryptophan; Y: tyrosine; F: phenylalanine; G: glycine; A: alanine; V: valine; I: isoleucine; L: leucine; M: methionine; S: serine; T: threonine; Q: glutamine; N: asparagine; E: glutamic acid; D: aspartic acid; *: stop codon. Raw data from Miseq Illumina sequencing. (See Appendix 8 for raw count data)

The observed amino acid distribution in Figure 7.5.3.2, particularly in comparison with Figure 7.4, demonstrated a good agreement between the observed and expected codons, with the most prominent feature being the substantial reduction in the presence of termination codons. By achieving a good match between the observed and expected amino acid distribution using only 216bp length reads demonstrated clearly, that read length (and thus the construction of the library), was the factor responsible for the unsatisfactory amino acid distribution observed in Figure 7.4. Manual inspection of the conserved region downstream of the randomised contiguous region showed poor alignment between the reference sequence and the Illumina reads. As the alignment upstream of the randomised region had been successful and no issues in alignment had been observed in any conserved region of the single arginine or At-Thr library downstream of randomised positions, the conclusion that the library construction was the cause for the amino acid distribution observed in Figure 7.4 was supported.

7.5.4 Investigating possible deletions caused by *MlyI* star activity in couplet 59+60

Any potential star activity of *MlyI*, resulting in non-specific restrictions would cause unexpected base deletions, with any deletions less than 10bp not necessarily visible via agarose gel electrophoresis.

To investigate the possibility of *MlyI* star activity during couplet 59+60 restriction, a range of codon counts (2.2.6.3) were performed pertaining to specific bp deletions, using all of the Bowtie2 outputted reads as used to generate Figure 7.4. The deletion of a MAX codon at position 62 would result in a 3bp deletion, meaning TGA would now be present at position 62 (Figure 6.1.2). A count for TGA at position 62 (2.2.6.3) using the countif criteria, “*TCTAATATCGCT?????????TGA*”, outputted 17921 reads. A deletion of both MAX codons from couplet 59+60 (or 61+62) would result in codons TGA (a termination codon) and CGG (arginine) occupying positions 61 and 62 respectively (Figure 6.1.2), with the countif criteria for this scenario being, “*TCTAATATCGCT??????TGACGG*” (2.2.6.3). The number of reads meeting this criteria was 17708. A 5bp deletion counted using the criteria “*TCTAATATCGCT???????TGACG*” (2.2.6.3), outputted 41181 reads while a 4bp deletion counted using “*TCTAATATCGCT????????TGAC*” (2.2.6.3) outputted 21532 reads. The number of reads meeting the count criteria for a 6bp, 5bp, 4bp and 3bp deletion collectively was 98324 reads, 53.71% of reads (183054 reads total). This strongly supports the suspicion of *MlyI* induced deletions being the causative factor for the varied read lengths, causing the unsatisfactory distribution observed in Figure 7.4.

7.5.5 Conclusion regarding the success of an Excel based analysis tool for determining amino acid distribution in the contiguous saturated positions of the ParaMAX library

The amino acid distributions at the four saturated contiguous positions in the ParaMAX library were successfully determined using the novel Excel based count methodology. These counts highlighted an unexpectedly high amount of stop codons in the randomised region. Eliminating in silico causes for the stop codon ratios observed, coupled with further library counts, strongly supported the suggestion of deletions within the randomised region occurring during the library generation as the cause; most likely associated with suboptimal *MlyI* restrictions. Work in the PRe-ART group is ongoing to further examine this phenomenon, with the first stage using a fixed library at the four randomised positions to determine the location of the deletions. Using the fixed codons; AAA, TTT, CCC, GGG, to eliminate the potential of upstream bases of adjacent codons compensating in the event of deletions, alongside sequencing of the couplet products and then the quad product, would enable the precise location of all deletions to be determined, providing evidence of the suspected *MlyI* causation.

8.0 Discussion and Conclusion

8.1 Summary of results

The overall aim of this project was to engineer several randomised designed armadillo repeat protein DNA libraries which would be incorporated into a designed armadillo repeat protein backbone construct, for downstream protein production and screening to discover novel peptide binders. Each of these randomised DNA libraries had their own specific amino acid representation requirements at their targeted positions, which needed analysing to determine library success.

The single arginine library was successfully engineered first, saturating seven positions within the DNA sequence. The DNA library was divided into three individual constructs, each designed to possess complementary regions with its adjacent constructs. The conserved region was successfully generated by separating the region into two shorter sequences that were purchased as synthetic oligonucleotides and joined via overlap PCR. The two remaining regions contained saturation targets and were therefore engineered using MAX randomisation. The full-length DNA library was produced by PCR amplification of the overlap PCR cassette formed from the three individual constructs. Illumina sequencing of the

randomised DNA library was performed using the amplicon EZ service of Genewiz. The uniqueness of the sequencing data, in the sense that no main stream protocols exist for counting codon frequencies at specific positions, meant a novel sequencing analysis technique was required. Formatting of the sequencing data was performed using Galaxy, the online bioinformatics server, with the processed data accessible in Excel. The specificity of MAX selection oligonucleotides to each saturated position made them ideal count criteria for the Excel function countif, allowing the determination of codon frequencies at each targeted position. A comparison between the expected codon frequencies and those counted, showed a close resemblance meaning the single arginine randomised library was successful.

Based on computational predictions provided by the Höcker Lab of the University of Bayreuth generated by ATLIGATOR, a randomised DNA library schematic for a specific threonine binding designed armadillo repeat protein was provided. This library utilised the same binding pocket of the single arginine library and therefore DNA sequence, but required saturation of the target positions with more specific amino acid subsets. Library construction mimicked the process used in the single arginine engineered library, with the full length cassette eventually submitted to Genewiz for Illumina sequencing. As the targeted positions were identical to that of the single arginine, the same protocol for determining codon frequencies at each saturated position was used with the resulting observed frequencies compared to the expected. A close resemblance between observed and expected codon frequencies was observed, indicating the successful production of the library.

The H3 groove of each armadillo repeat in the designed armadillo repeat protein possesses two binding pockets originally targeting arginine and lysine. The shallowness of the lysine pocket was a potential issue downstream as this pocket would need to bind specifically and with a high affinity to more bulky amino acids, than could be accommodated for at present. The idea to extend the pocket by introducing a loop between the H3 helix of repeat three and H1 of repeat four was proposed by Dr Erich Michel at the University of Zurich. Since this region would be completely novel, the traditional approach of using consensus sequences to direct amino acid identities for the loop was not possible and therefore investigating all amino acid combinations via saturation mutagenesis would be the most practical approach. This would mean a contiguous region of codons would need to be engineered, making MAX randomisation redundant. A new technique for randomising contiguous codons was required. ParaMAX randomisation would use the principles of MAX randomisation to saturate two adjacent codons, with production of multiple cassettes occurring simultaneously. These cassettes would contain *MlyI* restriction sites, to allow the exposure of the saturated positions which could be joined via blunt end ligation, generating a construct containing a contiguous region of randomised codons. A proof of concept library was designed containing four randomised contiguous codons. The library sequence was separated into a conserved region and two randomised cassettes. The upstream conserved region was generated by overlap

PCR between two synthetic oligonucleotides encoding the whole region. The contiguously randomised codon-containing cassettes were engineered as described above and joined via overlap PCR to the conserved region. Genewiz provided Illumina sequencing data of the complete model ParaMAX library cassette. Modifications to the Galaxy processing of the sequence data were made to accommodate for the contiguous region, primarily the avoidance of joining directional mates and the processed data was then subjected to codon counting in Excel. As the second and third positions in the contiguous region no longer possessed their MAX invariant region, a new codon count technique was required. To accommodate for the 12bp-randomised region, the 12bp upstream of the first randomised codon was used as a count anchor to minimise the chance of the anchor being present in the randomised region. To move the count focus along the contiguous region, Excel's ignorance to sets of '???' was used, meaning the countif function ignored those base characters.

The codon counts showed an unexpectedly high frequency of termination codons, especially that of TGA. As TGA was the adjacent codon downstream of position 62, it was suspected deletions within the randomised region had caused a shift resulting in TGA occupying position 62 more frequently. Codon counts relating to deletions in the randomised contiguous region revealed a range of deletions had occurred. Upon reflection, an issue in the library construction was the most likely cause for these deletions, most probably suboptimal *MlyI* restrictions as opposed to sequencing errors, which would not have been confirmed until codon counting in the Illumina sequencing data. Certainly, such small deletions would not be detectable via agarose gel electrophoresis.

8.2 Methodological considerations and future directions

8.2.1 ParaMAX

The suspected issue of *MlyI* restriction specificity and the range of deletions that were introduced into the model ParaMAX library, were only identified when analysing the Illumina sequencing data. In future, each new construct generated and amplified using PCR, instead of being visualised using agarose gel electrophoresis, should be visualised using polyacrylamide gel electrophoresis. This would allow the visualisation of different length products with higher resolution (fewer bp differences) determining if each stage of ParaMAX had been successful. A library design change, introducing a longer flanking oligo one side of the randomised couplets would also allow easier differentiation by size after restriction.

If *MlyI* restriction proved a non-viable method of exposing the randomised codons, a different type IIs endonuclease could be used. As *MlyI* is unique in that it produces no overhang, the ParaMAX protocol would need to accommodate for a change from blunt end ligation to sticky end ligation when joining randomised cassettes. These cassettes would still be randomised using MAX randomisation with the type IIs recognition site incorporated into the expendable flanking oligonucleotides. To maintain non-degeneracy of MAX randomisation, a type IIs restriction enzyme that produces overhangs in multiples of three would be required. Two endonucleases fall into the category, *BspQI* (GCTCTTC(1/4)) and *EarI* CTCTTC(1/4), with *EarI* used in the Slonomics process (Van den Brulle *et al*., 2008). As both enzymes only produce an overhang of 3bp, an acceptor cassette containing two randomised positions engineered by MAX randomisation could be extended by one MAX codon per cycle (n), with the MAX codon donor sequence containing the NNN overhang upon restriction for the complementary annealing of the next MAX codon in the n+1 cycle.

Alternatively, as *MlyI* restriction has been successfully used in ProxiMAX randomisation (Ashraf *et al*., 2013), instead of a methodological change to a different restriction enzyme, a thorough investigation to determine the basis of *MlyI* star activity in the model ParaMAX library could be performed.

The potential for ParaMAX in future protein engineering projects, both within PRe-ART and other applications makes the optimisation of the methodology a worthwhile endeavour. The ability to introduce novel regions into DNA sequences, with a user defined degree of randomisation and length in regards to PRe-ART will be vital in generating binding pockets capable of accommodating larger amino acids, necessary to achieve the final project aim of generating a library of binding modules specific to all 400 di-amino acid combinations.

8.2.2 Processing of randomised DNA library NGS data

Closer inspection of the aligned data for the model ParaMAX library filtered by quality and length using the FilterFastq tool (Figure 7.5.3.2), (the data set containing the correct length reads only), showed the conserved region downstream of the contiguous randomised region did not successfully align. The Bowtie2 tool used to align these reads to the reference sequence can manage alignments containing N (ambiguous bases) but no clear limit of Ns and the impact of their placement (point mutations vs contiguously randomised bases) is discussed (bowtie-bio.sourceforge.net, n.d; Langmead & Salzberg, 2012). Considering the alignment is incredibly accurate upstream of the contiguous region of the Filter Fastq processed model ParaMAX library, the poor alignment of the conserved region downstream of the randomised positions, suggests an incompatibility of Bowtie2 and therefore other existing aligners, as they are not designed to align such niche randomised DNA libraries.

Alignment programmes available are more commonly used instead to find regions of similarity to identify homologues and functional similarities, investigating phylogenetic trees. A well-known example being BLAST (Basic Local Alignment Search Tool (Altschul,1990).

A specific alignment tool script would ideally be produced to handle randomised DNA libraries with regions of randomisation longer than 6bp. 6bp randomised regions were observed in both the single arginine and At-Thr libraries (positions R5-29 and R5-30) with no consequence to the alignment accuracy of the conserved region as codon counts for R5-30 were performed using its invariant anchor (downstream of the MAX codon).

A potential design for an alignment tool, would include the criteria to ignore the randomised region completely, instead of negatively scoring matches against the Ns of the reference sequence. The alignment process would begin simultaneously in both the upstream and downstream direction from the 5' and 3' ends of the randomised region (red arrows Figure 8.2.2) instead. Alignment matches in the adjacent upstream and downstream conserved regions equal to the length of the randomised region (green and yellow brackets Figure 8.2.2) would be scored more favourably compared to the rest of the sequence to make alignments where these conserved regions and therefore the contiguous region by association would be correctly aligned (Figure 8.2.2).

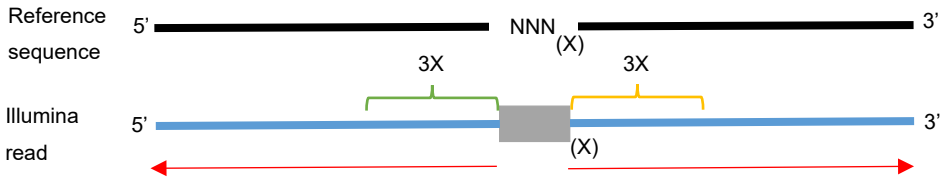

Figure 8.2.2 Diagrammatic representation of optimal read alignment programme for DNA libraries containing contiguous randomised positions.

Illumina read of the randomised DNA library would be aligned to the reference sequence (black) simultaneously in both the upstream and downstream direction from the randomised region (grey box) (red arrows), ignoring the NNN region of the reference sequence corresponding to the randomised region. (X) depicts the number of codons randomised in the contiguous region. The base match scores for the conserved region measuring 3X bp up and downstream of the contiguous region (green and yellow brackets) would be scored higher relative to matches outside of these regions to make alignments with these conserved regions and therefore the contiguous region in the correct orientation more favourable.

Careful consideration would be required to ensure the biased scoring of the adjacent conserved regions to the randomised region (green and yellow brackets, Figure 8.2.2), did not force alignments that would otherwise be incorrect. An example being if the conserved region sequences were not unique in the library, an important factor for repeat protein DNA sequences.

The invention of an invariant anchor based count methodology will significantly increase NGS analysis efficiency and the confidence in library quality before progressing to the next stage of protein expression and screening. To fully meet the requirements for processing such unique libraries, the development of a new alignment tool with the basic criteria outlined above would enable the accurate analysis of libraries with truncated regions of continuous randomisation, for example the introduction of two novel loops via ParaMAX.

8.3 Conclusion

Two high quality DNA libraries were generated containing non-contiguous randomised positions, with non-degeneracy successfully analysed using a novel codon count technique. Methodological developments for randomising and then analysing contiguous codons was initiated, but posed a greater challenge. These challenges however highlighted other methodological approaches and the lack of analysis tools available for amino acid ratio determination at specific DNA library positions.

8.4 References

Acevedo-Rocha, C.G., Reetz, M.T. and Nov, Y. (2015). Economical analysis of saturation mutagenesis experiments. *Scientific Reports*, 5(1). https://doi.org/10.1038/srep10654.

Aihajj, M & Farhana, A. (2022). Enzyme Linked Immunosorbent Assay. [Online]. Treasure Island: StatPearls Publishing. Accessed 2nd April 2022.

Altschul, S. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215(3), pp.403–410. doi: https://doi.org/10.1016/S0022-2836(05)80360-2

Ashraf, M., Frigotto, L., Smith, Matthew E., Patel, S., Hughes, Marcus D., Poole, Andrew J., Hebaishi, Husam R.M., Ullman, Christopher G. and Hine, Anna V. (2013). ProxiMAX randomization: a new technology for non-degenerate saturation mutagenesis of contiguous codons. *Biochemical Society Transactions*, 41(5), pp.1189–1194. doi: https://doi.org/10.1042/BST20130123.

Bahmanyar, S., Kaplan, D.D., DeLuca, J.G., Giddings, T.H., O'Toole, E.T., Winey, M., Salmon, E.D., Casey, P.J., Nelson, W.J. and Barth, A.I.M. (2008). β-Catenin is a Nek2 substrate involved in centrosome separation. *Genes & Development*, 22(1), pp.91–105. doi: https://doi:10.1101/gad.1596308.

Boder, E. T. and Wittrup, K. D. (1997). Yeast surface display for screening combinatorial polypeptide libraries. *Nature Biotechnology*, 15(6), pp.553–557. https://doi.org/10.1038/nbt0697-553

Crook, Z.R., Nairn, N.W. and Olson, J.M. (2020). Miniproteins as a Powerful Modality in Drug Development. *Trends in Biochemical Sciences*, 45(4), pp.332–346. https://doi.org/10.1016/j.tibs.2019.12.008

Isogenica: manufacturing better proteins for industry (2015). Available at: https://webarchive.nationalarchives.gov.uk/ukgwa/20210901102727/https://bbsrc.ukri.org/research/impact/isogenica-manufacturing-better-proteins-for-industry/ (Accessed: March 2022).

Berglund, L., Björling, E., Oksvold, P., Fagerberg, L., Asplund, A., Al-Khalili Szigyarto, C., Persson, A., Ottosson, J., Wernérus, H., Nilsson, P., Lundberg, E., Sivertsson, Å., Navani, S., Wester, K., Kampf, C., Hober, S., Pontén, F. and Uhlén, M. (2008). A Genecentric Human Protein Atlas for Expression Profiles Based on Antibodies. *Molecular & Cellular Proteomics*, 7(10), pp.2019–2027. doi:https://doi.org/10.1074/mcp.R800013-MCP200

Boder, E. T. and Wittrup, K. D. (1997) "Yeast surface display for screening combinatorial polypeptide libraries," Nature Biotechnology, 15(6), pp. 553–557. doi: 10. 1038/nbt0697-553

Bowtie 2: Manual (n.d). Available at: http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml. (Accessed: September 2021).

Borowicz, P., Chan, H., Hauge, A. and Spurkland, A. (2020). Adaptor proteins: Flexible and dynamic modulators of immune cell signalling. *Scandinavian Journal of Immunology*, 92(5). doi: https://doi.org/10.1111/sji.12951.

Bradbury, A. and Plückthun, A. (2015). Reproducibility: Standardize antibodies used in research. *Nature*, 518(7537), pp.27–29. doi: https://doi.org/10.1038/518027a.

Briney, B., Inderbitzin, A., Joyce, C. and Burton, D.R. (2019). Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature, 566(7744), pp.393–397. doi: https://doi.org/10.1038/s41586-019-0879-y

Cadwell,R.C. and Joyce,G.F. (1994) Mutagenic PCR. *PCR Methods Appl*, 3, pp:136–40. doi:10.1101/gr.3.6.s136

Chen, C.K.-M., Chan, N.-L. and Wang, A.H.-J. (2011). The many blades of the β-propeller proteins: conserved but versatile. *Trends in Biochemical Sciences*, 36(10), pp.553–561. doi: https://doi.org/10.1016/j.tibs.2011.07.004

Coates, J. (2003). Armadillo repeat proteins: beyond the animal kingdom. *Trends in Cell Biology*, 13(9), pp.463–471. doi: https://doi.org/10.1016/S0962-8924(03)00167-3

Coco,W.M. (2003) RACHITT: Gene family shuffling by Random Chimeragenesis on Transient Templates. *Methods Mol. Biol*., 231, pp: 111–127. Doi: https://doi.org/10.1385/1-59259-395-X:111

Reprinted from Structure, 8 (3), Conti, E. and Kuriyan, J. (2000). Crystallographic analysis of the specific yet versatile recognition of distinct nuclear localization signals by karyopherin α. pp.329–338. Doi: https://doi.org/10.1016/S0969-2126(00)00107-6. Copyright (2022) with permission from Elsevier.

D'Andrea, L. (2003). TPR proteins: the versatile helix. *Trends in Biochemical Sciences*, 28(12), pp.655–662. doi: https://doi.org/10.1016/j.tibs.2003.10.007

Dimitrov D.S. (2012) Therapeutic Proteins. In: Voynov V., Caravella J. (eds) Therapeutic Proteins. *Methods in Molecular Biology (Methods and Protocols)*, 899. https://doi.org/10.1007/978-1-61779-921-1_1

Ebersbach, H., Fiedler, E., Scheuermann, T., Fiedler, M., Stubbs, M.T., Reimann, C., Proetzel, G., Rudolph, R. and Fiedler, U. (2007). Affilin–Novel Binding Molecules Based on Human γ-B-Crystallin, an All β-Sheet Protein. *Journal of Molecular Biology*, 372(1), pp.172–185. doi: https://doi.org/10.1016/j.jmb.2007.06.045.

Ems-McClung, S.C., Zheng, Y. and Walczak, C.E. (2004). Importin α/β and Ran-GTP Regulate XCTK2 Microtubule Binding through a Bipartite Nuclear Localization Signal. *Molecular Biology of the Cell*, 15(1), pp.46–57. doi: https://doi.org/10.1091/mbc.e03-07-0454

Reprinted (adapted) with permission from {Ernst, P., Zosel, F., Reichen, C., Nettels, D., Schuler, B. and Plückthun, A. (2020). Structure-Guided Design of a Peptide Lock for Modular Peptide Binders. *ACS Chemical Biology*, 15(2), pp.457–468. doi: https://doi.org/10.1021/acschembio.9b00928}. Copyright {2022}.

Forsström, B., Bisławska Axnäs, B., Rockberg, J., Danielsson, H., Bohlin, A. and Uhlen, M. (2015). Dissecting Antibodies with Regards to Linear and Conformational Epitopes. *PLOS ONE*, 10(3). doi: https://doi.org/10.1371/journal.pone.0121673.

Fournier, D., Palidwor, G.A., Shcherbinin, S., Szengel, A., Schaefer, M.H., Perez-Iratxeta, C. and Andrade-Navarro, M.A. (2013). Functional and Genomic Analyses of Alpha-Solenoid Proteins. *PLoS ONE*, 8(11). doi: https://doi.org/10.1371/journal.pone.0079894

Frigotto, L., Smith, M., Brankin, C., Sedani, A., Cooper, S., Kanwar, N., Evans, D., Svobodova, S., Baar, C., Glanville, J., Ullman, C. and Hine, A. (2015). Codon-Precise, Synthetic, Antibody Fragment Libraries Built Using Automated Hexamer Codon Additions and Validated through Next Generation Sequencing. *Antibodies*, 4(2), pp.88–102. doi: https://doi.org/10.3390/antib4020088

Harmsen, M.M. and De Haard, H.J. (2007). Properties, production, and applications of camelid single-domain antibody fragments. *Applied Microbiology and Biotechnology*, 77(1), pp.13–22. doi: https://doi.org/10.1007/s00253-007-1142-2

Reprinted (adapted) with permission from {Hansen, S., Tremmel, D., Madhurantakam, C., Reichen, C., Mittl, P.R.E. and Plückthun, A. (2016). Structure and Energetic Contributions of a Designed Modular Peptide-Binding Protein with Picomolar Affinity. *Journal of the American Chemical Society*, 138(10), pp.3526–3532. doi: https://doi.org/10.1021/jacs.6b00099}. Copyright {2022}.

Holm, L. (1986). Codon usage and gene expression. *Nucleic Acids Research* 14(7), pp.3075–3087. Doi: https://doi.org/10.1093/nar/14.7.3075

Huber, A.H., Nelson, W.James. and Weis, W.I. (1997). Three-Dimensional Structure of the Armadillo Repeat Region of β-Catenin. *Cell*, 90(5), pp.871–882. doi: https://doi.org/10.1016/S0092-8674(00)80352-9

Reprinted from Journal of Molecular Biology, 331 (5), Hughes, M.D., Nagel, D.A., Santos, A.F., Sutherland, A.J. and Hine, A.V. (2003). Removing the Redundancy From Randomised Gene Libraries. pp.973–979. doi: https://doi.org/10.1016/S0022-2836(03)00833-7. Copyright (2022) with permission from Elsevier.

Hyman, E.D. (1988). A new method of sequencing DNA. *Analytical Biochemistry*, 174(2), pp.423–436. doi: https://doi.org/10.1016/0003-2697(88)90041-3

Illumina (2011) Quality scores for Next Generation Sequencing. Avaliable at https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf (Accessed 23rd August 2021).

Kaplan, D.D., Meigs, T.E., Kelly, P. and Casey, P.J. (2004). Identification of a Role for β-Catenin in the Establishment of a Bipolar Mitotic Spindle. *Journal of Biological Chemistry*, 279(12), pp.10829–10832. doi: https://doi.org/10.1074/jbc.C400035200

Kille, S., Acevedo-Rocha, C.G., Parra, L.P., Zhang, Z.-G., Opperman, D.J., Reetz, M.T. and Acevedo, J.P. (2012). Reducing Codon Redundancy and Screening Effort of Combinatorial Protein Libraries Created by Saturation Mutagenesis. *ACS Synthetic Biology*, 2(2), pp.83–92. doi: https://doi.org/10.1021/sb300037w

Kim, W.K., Kwon, Y., Jang, M., Park, M., Kim, J., Cho, S., Jang, D.G., Lee, W.-B., Jung, S.H., Choi, H.J., Min, B.S., Il Kim, T., Hong, S.P., Paik, Y.-K. and Kim, H. (2019). β-catenin activation down-regulates cell-cell junction-related genes and induces epithelial-to-mesenchymal transition in colorectal cancers. *Scientific Reports*, 9(1). doi: https://doi.org/10.1038/s41598-019-54890-9

Köhler, G. and Milstein, C. (1975). Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature*, 256(5517), pp.495–497. doi: https://doi.org/10.1038/256495a0

Kutay, U., Bischoff, F.Ralf., Kostka, S., Kraft, R. and Görlich, D. (1997). Export of Importin α from the Nucleus Is Mediated by a Specific Nuclear Transport Factor. *Cell*, 90(6), pp.1061–1071. doi: https://doi.org/10.1016/S0092-8674(00)80372-4

Langmead, B. and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), pp.357–359. doi: https://doi.org/10.1038/nmeth.1923

Li, J., Mahajan, A. and Tsai, M.-D. (2006). Ankyrin Repeat: A Unique Motif Mediating Protein−Protein Interactions. *Biochemistry*, 45(51), pp.15168–15178. doi: https://doi.org/10.1021/bi062188q

Liu, J.K.H. (2014). The history of monoclonal antibody development – Progress, remaining challenges and future innovations. *Annals of Medicine and Surgery*, 3(4), pp.113–116. doi: https://doi.org/10.1016/j.amsu.2014.09.001

Lu, J., Wu, T., Zhang, B., Liu, S., Song, W., Qiao, J. and Ruan, H. (2021). Types of nuclear localization signals and mechanisms of protein import into the nucleus. *Cell Communication and Signaling*, 19(1). doi: https://doi.org/10.1186/s12964-021-00741-y

Lutz,S., Ostermeier,M. and Benkovic,S.J. (2001) Rapid generation of incremental truncation libraries for protein engineering using alpha-phosphothioate nucleotides. *Nucleic Acids Res*, 29, pp:16. doi: https://doi.org/10.1093/nar/29.4.e16

MacDonald, B.T., Tamai, K. and He, X. (2009). Wnt/β-Catenin Signaling: Components, Mechanisms, and Diseases. *Developmental Cell,* 17(1), pp.9–26. doi: https://doi.org/10.1016/j.devcel.2009.06.016

Madhurantakam, C., Varadamsetty, G., Grütter, M. G., Plückthun, A. and Mittl, P. R. (2012) Structure-based optimization of designed armadillo-repeat proteins. *Protein Sci,* 21, pp.1015-1028. doi: https://doi.org/10.1002/pro.2085

Makowski, L. and Soares, A. (2003). Estimating the diversity of peptide populations from limited sequence data. *Bioinformatics*, 19(4), pp.483–489. doi: https://doi.org/10.1093/bioinformatics/btg013

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P. and Jando, S.C. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057), pp.376–80. doi: https://doi.org/10.1038/nature03959

Miyamoto, Y., Yamada, K. and Yoneda, Y. (2016). Importin α: a key molecule in nuclear transport and non-transport functions. *Journal of Biochemistry*, 160(2), pp.69–75. doi: https://doi.org/10.1093/jb/mvw036

Mohsen, A., Park, J., Chen, Y.-A., Kawashima, H. and Mizuguchi, K. (2019). Impact of quality trimming on the efficiency of reads joining and diversity analysis of Illumina paired-end reads in the context of QIIME1 and QIIME2 microbiome analysis frameworks. *BMC Bioinformatics,* 20(1), pp.581. doi: https://doi.org/10.1186/s12859-019-3187-5

Murakami, H., Hohsaka, T. and Sisido, M. (2002) Random insertion and deletion of arbitrary number of bases for codon-based random mutation of DNAs. *Nat. Biotechnol*, 20, 76–81. doi https://doi.org/10.1038/nbt0102-76

Muyldermans, S. (2020). A guide to: generation and design of nanobodies. *The FEBS Journal*, 288 (7), pp. 2084-2102. doi: https://doi.org/10.1111/febs.15515.

Neylon, C. (2004). Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: library construction methods for directed evolution. *Nucleic Acids Research*, 32(4), pp.1448–1459. doi: https://doi.org/10.1093/nar/gkh315

Nord, K., Gunneriusson, E., Ringdahl, J., Ståhl, S., Uhlén, M. and Nygren, P.-Å. (1997). Binding proteins selected from combinatorial libraries of an α-helical bacterial receptor domain. *Nature Biotechnology*, 15(8), pp.772–777. doi: https://doi.org/10.1038/nbt0897-772.

Nyrén P.I., Lundin A. (1985). Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Anal Biochem*, 151(2), pp.504–509. doi: https://doi.org/10.1016/0003-2697(85)90211-8

Oka, M. and Yoneda, Y. (2018). Importin α: functions as a nuclear transport factor and beyond. *Proceedings of the Japan Academy, Series B*, 94(7), pp.259–274. doi: https://doi.org/10.2183/pjab.94.018

Ostermeier,M., Shim,J.H. and Benkovic,S.J. (1999) A combinatorial approach to hybrid enzymes independent of DNA homology. *Nat. Biotechnol*., 17, pp; 1205–1209. doi: https://doi.org/10.1038/70754

Pardon, E., Laeremans, T., Triest, S., Rasmussen, S.G.F., Wohlkönig, A., Ruf, A., Muyldermans, S., Hol, W.G.J., Kobilka, B.K. and Steyaert, J. (2014). A general protocol for the generation of Nanobodies for structural biology. *Nature protocols*, 9(3), pp.674–693. doi: https://doi.org/10.1038/nprot.2014.039.

Reprinted from *J.Mol.Biol, 376,* Parmeggiani, F., Pellarin, R., Larsen, A. P., Varadamsetty, G., Stumpp, M. T., Zerbe, O., Caflisch, A. and Plückthun, A. (2008). Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational optimization of the hydrophobic core. pp: 1282-1304. doi: https://doi.org/10.1016/j.jmb.2007.12.014. Copyright (2022), with permission from Elsevier.

Perrimon, N. and Mahowald, A.P. (1987). Multiple functions of segment polarity genes in Drosophila. *Developmental Biology*, 119(2), pp.587–600. doi: https://doi.org/10.1016/0012-1606(87)90061-3

Pikkemaat, M.G. and Janssen, D.B. (2002) Generating segmental mutations in haloalkane dehalogenase: a novel part in the directed evolution toolbox. *Nucleic Acids Res*, 30, p35. doi: https://doi.org/10.1093/nar/30.8.e35

Poole, A.J., Frigotto, L., Smith, M.E., Baar, C., Ivanova-Berndt, G., Jaulent, A., Stace, C., Ullman, C.G. and Hine, A.V. (2019). A C-terminal cysteine residue is required for peptide-based inhibition of the NGF/TrkA interaction at nM concentrations: implications for peptide-based analgesics. *Scientific Reports*, 9(1). doi: https://doi.org/10.1038/s41598-018-37585-5

Reichen, C., Madhurantakam, C., Plückthun, A. and Mittl, P. R. (2014) Crystal structures of designed armadillo repeat proteins: Implications of construct design and crystallization conditions on overall structure. *Protein Sci,* 23, pp.1572-1583. doi: https://doi.org/10.1002/pro.2535

Reichen, C., Hansen, S. and Plückthun, A. (2014). Modular peptide binding: From a comparison of natural binders to designed armadillo repeat proteins. *Journal of Structural Biology*, 185(2), pp.147–162. doi: https://doi.org/10.1016/j.jsb.2013.07.012.

Rivera-Calzada, A., Fronzes, R., Savva, C.G., Chandran, V., Lian, P.W., Laeremans, T., Pardon, E., Steyaert, J., Remaut, H., Waksman, G. and Orlova, E.V. (2013). Structure of a bacterial type IV secretion core complex at subnanometre resolution. *The EMBO Journal*, 32(8), pp.1195–1204. doi: https://doi.org/10.1038/emboj.2013.58.

Rousseau, F., Schymkowitz, J. and Itzhaki, L.S. (2013). Implications of 3D Domain Swapping for Protein Folding, Misfolding and Function. *Adv Exp Med Biol*, 747, pp.137-52. doi: https://doi.org/10.1007/978-1-4614-3229-6_9

Sanger, F., Nicklen, S. and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12), pp.5463–5467. doi: https://doi.org/10.1073/pnas.74.12.5463

Service, F. (2022). Software-designed miniproteins could create new class of drugs. Available at: https://www.science.org/content/article/software-designed-miniproteins-could-create-new-class-drugs. (Accessed 2nd April 2022)

Shang, S., Hua, F. and Hu, Z.-W. (2017). The regulation of β-catenin activity and function in cancer: therapeutic opportunities. *Oncotarget*, 8(20). doi: https://doi.org/10.18632/oncotarget.15687

Slatko, B.E., Gardner, A.F. and Ausubel, F.M. (2018). Overview of Next-Generation Sequencing Technologies. *Current Protocols in Molecular Biology*, 122(1). doi: https://doi.org/10.1002/cpmb.59

Stemmer,W.P.C. (1994) DNA shuffling by random fragmentation and reassembly: In vitro recombination for molecular evolution. *Proc. Natl Acad. Sci. USA*, 91, pp: 10747–10751. doi: https://doi.org/10.1073/pnas.91.22.10747

Tang, L., Gao, H., Zhu, X., Wang, X., Zhou, M. and Jiang, R. (2012). Construction of “small-intelligent” focused mutagenesis libraries using well-designed combinatorial degenerate primers. *BioTechniques*, 52(3), pp.149–158. doi: https://doi.org/10.2144/000113820

Tang, L., Wang, X., Ru, B., Sun, H., Huang, J. and Gao, H. (2014). MDC-Analyzer: A novel degenerate primer design tool for the construction of intelligent mutagenesis libraries with contiguous sites. *BioTechniques*, 56(6). doi: https://doi.org/10.2144/000114177

Tolwinski, N.S. and Wieschaus, E. (2004). A Nuclear Function for Armadillo/β-Catenin. *PLoS Biology*, 2(4). doi: https://doi.org/10.1371/journal.pbio.0020095

Tsuchiya, Y. and Mizuguchi, K. (2016). The diversity of H3 loops determines the antigen-binding tendencies of antibody CDR loops. *Protein Science*, 25(4), pp.815–825. https://doi.org/10.1002/pro.2874

Turcatti, G., Romieu, A., Fedurco, M. and Tairi, A.-P. (2008). A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Research*, 36(4). doi: https://doi.org/10.1093/nar/gkn021

Valenta, T., Hausmann, G. and Basler, K. (2012). The many faces and functions of β-catenin. *The EMBO Journal*, 31(12), pp.2714–2736. doi: https://doi.org/10.1038/emboj.2012.150

Van den Brulle, J., Fischer, M., Langmann, T., Horn, G., Waldmann, T., Arnold, S., Fuhrmann, M., Schatz, O., O'Connell, T., O'Connell, D., Auckenthaler, A. and Schwer, H. (2008). A novel solid phase technology for high-throughput gene synthesis. *BioTechniques*, 45(3), pp.340–343. doi: https://doi.org/10.2144/000112953

Varadamsetty, G., Tremmel, D., Hansen, S., Parmeggiani, F. and Plückthun, A. (2012) Designed armadillo repeat proteins: library generation, characterization and selection of peptide binders with high specificity. *J. Mol. Biol*. 424, pp.68-87. doi: https://doi.org/10.1016/j.jmb.2012.08.029

Vasilevsky, N.A., Brush, M.H., Paddock, H., Ponting, L., Tripathy, S.J., LaRocca, G.M. and Haendel, M.A. (2013). On the reproducibility of science: unique identification of research resources in the biomedical literature. *PeerJ*, 1, pp:148. doi: 10.7717/peerj.148.

Vazquez-Lombardi, R., Phan, T.G., Zimmermann, C., Lowe, D., Jermutus, L. and Christ, D. (2015). Challenges and opportunities for non-antibody scaffold drugs. *Drug Discovery Today*, 20(10), pp.1271–1283. doi: https://doi.org/10.1016/j.drudis.2015.09.004.

Vercruysse, T., Pawar, S., De Borggraeve, W., Pardon, E., Pavlakis, G.N., Pannecouque, C., Steyaert, J., Balzarini, J. and Daelemans, D. (2011). Measuring cooperative Rev protein-protein interactions on Rev responsive RNA by fluorescence resonance energy transfer. *RNA Biology*, 8(2), pp.316–324. doi: https://doi.org/10.4161/rna.8.2.13782.

Verhaar, E.R., Woodham, A.W. and Ploegh, H.L. (2020). Nanobodies in cancer. *Seminars in Immunology*. pp.101425. doi: https://doi.org/10.1016/j.smim.2020.101425.

Virnekäs, B., Ge, L., Plückthun, A., Schneider, K.C., Wellnhofer, G. and Moroney, S.E. (1994). Trinucleotide phosphoramidites: ideal reagents for the synthesis of mixed oligonucleotides for random mutagenesis. *Nucleic Acids Research*, 22(25), pp.5600–5607. doi: https://doi.org/10.1093/nar/22.25.5600

Voelkerding, K.V., Dames, S.A. and Durtschi, J.D. (2009). Next-generation sequencing: from basic research to diagnostics. *Clinical chemistry*, 55(4), pp.641–58. doi: https://doi.org/10.1373/clinchem.2008.112789

Voskuil, J. (2014). Commercial antibodies and their validation [version2; peer review: 3 approved]. *F1000Research*, 3, p.232. doi: https://doi.org/10.12688/f1000research.4966.2

Wieschaus, E. and Riggleman, R. (1987). Autonomous requirements for the segment polarity gene armadillo during Drosophila embryogenesis. *Cell*, 49(2), pp.177–184. doi: https://doi.org/10.1016/0092-8674(87)90558-7

Xing, Y., Takemaru, K.-I., Liu, J., Berndt, J.D., Zheng, J.J., Moon, R.T. and Xu, W. (2008). Crystal Structure of a Full-Length β-Catenin. *Structure*, 16(3), pp.478–487. doi: https://doi.org/10.1016/j.str.2007.12.021

Yang, E.Y. and Shah, K. (2020). Nanobodies: Next Generation of Cancer Diagnostics and Therapeutics. *Frontiers in Oncology*, 10. doi: https://doi.org/10.3389/fonc.2020.01182.

Young, A.D. and Gillung, J.P. (2019). Phylogenomics — principles, opportunities and pitfalls of big-data phylogenetics. *Systematic Entomology*, 45(2), pp.225–247. doi: https://doi.org/10.1111/syen.12406

Zhao,H., Giver,L., Shao,Z., Affholter,J. and Arnold,F. (1998) Molecular evolution by staggered extension process (StEP) in vitro recombination. *Nat. Biotechnol*., 16, pp: 258–261. doi: https://doi.org/10.1038/nbt0398-258

Appendices

Appendix 1

Oligonucleotides used for the engineering of the single arginine and At-Thr randomised DNA libraries

Oligonucleotide name	Sequence (5'-3')
Repeat 4 MAX flanking oligo 1	GCAGGGGCCCTGCCAGCCCTTGTGCAATTATTGTCCAGCCC AAATGAGCAGATC
Repeat 4 MAX flanking oligo 2	[PHO]AACATCGCATCAGGTGGAAACGAGCAGATTCAGGCCG TTATTGATGCCGGC
Repeat 5 MAX flanking oligo 1	GCCGTTATTGATGCCGGCGCCCTTCCAGCTTTAGTCCAGTTG CTGAGTTCTCCT
Repeat 5 MAX flanking oligo 2	[PHO]TCTAATATTGCTAGCGGGGGTAATGAGCAAATCCAAGC AGTAATCGAT
Repeat 3 oligo 1	AATGAACAAATCTGGCAAGAGGCCTTGTGGGCCCTCAGCAAT ATCGCTTCTGGAGGC
Repeat 3 oligo 2 (RC)	TGGCAGGGCCCCTGCGTCAATCACAGCCTGGATCTGCTCATT GCCTCCAGAAGCGAT
Repeat 4 forward primer	GCA GGG GCC CTG CCA
Repeat 4 reverse primer	GCC GGC ATC AAT AAC GGC
Repeat 5 forward primer	GCC GTT ATT GAT GCC GGC
Repeat 5 reverse primer	ATC GAT TAC TGC TTG GAT
Repeat 3 forward primer	AAT GAA CAA ATC TTG CAA
Repeat 3 reverse primer	TGG CAG GGC CCC TGC GTC

Appendix 2

Raw amino acid counts at each saturated position in the single arginine library

	Saturated position						
Amino acid	**R4 29**	**R4 33**	**R4 36**	**R5 26**	**R5 29**	**R5 30**	**R5 33**
D	11886	9240	494	9832	10959	14693	10259
E	12416	13396	238	16227	11950	14969	7954
N	13273	13144	997	9465	13165	12296	14389
Q	13846	13531	219	16336	16668	16131	14339
T	15869	16881	91384	15162	17535	11235	15323
S	13463	13591	66286	13313	11750	11798	11866
M	12472	11895	204	19003	15469	14052	16112
L	13751	11245	1115	15919	10877	13969	12733
I	12325	12064	944	10590	9763	11756	12689
V	12279	11997	47112	2005	13064	15384	8819
A	13706	12782	10181	14745	16135	17314	14059
G	577	10719	310	50	95	136	14578
F	6834	8656	504	9290	6132	9290	8314
Y	9692	10635	522	8568	9903	10770	9676
W	13585	11306	28	18721	15909	15658	14445
H	14418	11985	608	13172	15425	14712	13743
R	15702	15292	399	18702	17431	10834	13645
K	13498	13882	290	15588	13618	9575	11890
C	472	468	177	43	111	41	132
P	2259	1162	1295	89	75	102	99
*	1070	1108	277	124	177	233	79
Total	223393	224979	223584	226944	226211	224948	225143

Appendix 3

Raw amino acid counts at each saturated position in the Bowtie2 non-length filtered At-Thr library

	Saturated position						
Amino acid	**R4 29**	**R4 33**	**R4 36**	**R5 26**	**R5 29**	**R5 30**	**R5 33**
D	21759	69	101	16975	16881	21656	17627
E	21038	88	131	24380	19826	22669	12326
N	20624	121	219980	14535	17884	17797	22126
Q	23094	126	141207	25157	24618	24327	22688
T	23389	222	230	21576	26999	19096	24158
S	21086	473	1055	20279	19520	19868	20221
M	20747	86	411	26098	21982	19937	23260
L	22332	355549	170	24806	17926	22826	20915
I	19823	252	1287	15588	14314	16751	21152
V	21856	138	264	8977	21379	22163	16380
A	22881	134	985	22999	31607	25675	22611
G	135	84	102	658	818	832	826
F	14052	489	179	14065	8996	13026	14490
Y	17581	89	110	12065	13131	15401	16130
W	24294	207	46	30165	24395	22767	23269
H	22925	105	325	19375	23195	22606	21854
R	24360	219	348	25821	27693	18038	23651
K	20616	146	908	23599	21018	15091	19410
C	176	81	42	418	819	412	548
P	287	168	186	1187	1011	1245	1117
*	217	185	346	993	1065	1232	796
Total	363272	358846	368413	349716	355077	343415	345555

Appendix 4

Raw amino acid counts at each saturated position in the BWA-MEM aligned At-Thr library

	Saturated position						
Amino acid	**R4 29**	**R4 33**	**R4 36**	**R5 26**	**R5 29**	**R5 30**	**R5 33**
D	23132	75	115	18253	18065	23234	18927
E	22321	89	138	26089	21231	24315	13164
N	21985	134	233017	15556	19133	19185	23672
Q	24580	133	149834	26872	26398	26077	24306
T	24913	234	257	22975	28970	20356	25829
S	22407	512	1164	21641	20890	21287	21642
M	22068	95	466	27870	23581	21249	25031
L	23789	376481	180	26421	19392	24411	22411
I	21041	266	1381	16705	15290	17868	22678
V	23247	187	283	9590	22750	23789	17578
A	24245	148	1052	24594	35842	27549	24282
G	137	89	114	699	879	898	917
F	14882	511	205	15036	9633	13873	15441
Y	18683	93	115	12914	13996	16532	17336
W	25731	218	54	32229	26003	24410	24874
H	24354	120	341	20781	24755	24120	23388
R	25927	233	362	27575	29638	19329	25514
K	21895	154	998	25159	22538	16246	20735
C	184	91	45	451	898	440	606
P	304	181	198	1259	1085	1321	1187
*	231	194	370	1072	1133	1308	849
Total	386056	380044	390689	373741	382100	367797	370367

Appendix 5

Raw amino acid counts at each saturated position in the Bowtie2 length filtered At-Thr library

	Saturated position						
Amino acid	**R4 29**	**R4 33**	**R4 36**	**R5 26**	**R5 29**	**R5 30**	**R5 33**
D	11299	40	67	8741	8853	11848	9445
E	10854	58	58	13384	10812	12520	6773
N	10774	74	117192	7690	9656	9544	11908
Q	12062	63	70010	13606	13619	13469	12583
T	12215	125	93	11653	14866	10761	13550
S	10918	256	125	10844	10521	10920	11036
M	10641	45	74	14173	12029	11050	13066
L	11692	186473	82	13544	9443	12591	11379
I	10264	147	349	8144	7393	9006	11514
V	11397	57	57	4785	11495	12185	8704
A	11894	75	88	12321	13962	14434	12541
G	88	40	26	378	379	411	360
F	7340	249	99	7428	4734	6934	7871
Y	9125	48	55	6525	6971	8295	8754
W	12715	54	26	16518	13653	12524	12714
H	11977	51	194	10240	12552	12494	12315
R	12522	147	82	13881	15024	9856	12774
K	10454	70	333	12571	11388	8088	10501
C	110	42	26	264	283	245	283
P	167	77	68	695	593	711	648
*	120	101	88	563	493	516	463
Total	188628	188191	189192	187948	188719	188402	189182

Appendix 6

Oligonucleotides used for the engineering of the model ParaMAX library

Couplet 59/60 forward primer	AATGAGCAGATTCTTCAG
Couplet 59/60 reverse primer	CCGGCTCGATGTACAATT
Couplet 59/60 MAX flanking oligo 1	AATGAGCAGATTCTTCAGGAGGCACTATGGGCTTTGTCTAAT
Couplet 59/60 MAX flanking oligo 2	[PHO]ACTCGCAATTGTACATCGAGCCGG
Couplet 59/60 MAX template	GCGAGTCTCACTNNNNNNAGCGATATTAGA[23ddC]
Conserved oligo 1	AAT GAA CAA ATC TTG CAA TTA GCC TTG TGG GCC CTC AGC AAT ATC GCC TCT GGT GGG AAT GAG CAG ATT CAA GCT GTC
Conserved oligo 2 (RC)	CTGAAGAATCTGCTCATTAGGACTGCTTAAAAGCTGAACCAGTGCTGG AAGTGCACCAGCATCAATGACAGCTTGAATCTGCTCATT
Couplet 61/62 forward primer	CCCATATTACGACCTTGG
Couplet 61/62 reverse primer	GATCAGTATCATAATGCC
Couplet 61/62 MAX flanking oligo 1	CCCATATTACGACCTTGGATGCTAATAATTGTACATCGTGCCGGTCGA GT
Couplet 61/62 MAX flanking oligo 2	[PHO]ACTTATGGCATTATGATACTGATC
Couplet 61/62 MAX template	ATAAGTCCGTCANNNNNNCTTCTGACTCGA[23ddC]

Appendix 7

Raw amino acid counts at each saturated position in the Bowtie2 aligned model ParaMAX library

	Saturated position			
Amino acid	**59**	**60**	**61**	**62**
D	9330	13067	5990	26128
E	3497	10825	3791	3079
N	18174	12542	6606	3714
Q	6304	4690	809	2118
T	10841	7016	7869	44480
S	7406	6907	3205	1677
M	9599	5607	10823	4663
L	6268	10729	27365	7341
I	15643	13708	10146	3922
V	7529	9122	20267	3659
A	2951	2482	2467	907
G	4430	6057	6741	3956
F	9407	14169	4763	1778
Y	15264	17444	3984	2292
W	3297	4530	3296	1780
H	1615	6638	1996	3855
R	11024	983	5630	20636
K	10178	4240	5059	1797
C	440	882	5651	719
P	3683	2845	1994	1199
*	805	3188	19201	17938
Total	157685	157671	157653	157638

Appendix 8

Raw amino acid counts at each saturated position in the FilterFastq processed model ParaMAX library

	Saturated position			
Amino acid	**59**	**60**	**61**	**62**
D	2456	2761	2457	3086
E	1151	1658	2389	2482
N	2636	1959	2952	2809
Q	1398	1705	308	1812
T	3301	3114	4032	2523
S	1910	3539	553	1189
M	3701	1825	3612	3039
L	2257	1380	1238	1890
I	2802	3150	2018	1780
V	1795	2365	2322	1947
A	694	1056	1744	720
G	1051	2039	2264	1929
F	1749	1794	593	800
Y	2325	2972	1464	1919
W	1401	1743	990	1499
H	715	2847	1007	3116
R	2259	13	2799	2436
K	2790	664	3323	1543
C	2	9	3	15
P	1170	982	1488	1037
*	15	3	22	3
Total	37578	37578	37578	37574