The impact of *Alu* elements on the human proteome

Sarah Elizabeth Atkinson

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

The University of Leeds
School of Chemistry

May 2019

Acknowledgements

Firstly, I would like to thank Prof. Paul Taylor, Dr Michael Webb and Dr Suzanne Dilly for their guidance and support throughout my PhD, as well as the University of Leeds and Valirx Plc for providing me with this wonderful opportunity.

Secondly, I would like to extend my deepest thanks to the members of Lab 1.49, past and present, for all the support they have provided throughout my PhD. In particular, I would like to thank Jack Caudwell for providing me with constant moral support and always keeping a smile on my face, as well as helping to provide the warm and social atmosphere that makes working in 1.49 so special. You made my entire PhD experience infinitely more enjoyable! To Ryan, I hope your next desk buddy is more of a pushover and your future desk take overs are more successful than they were with me. Additionally, I would like to thank Pablo, Devón and JoDo for providing me with friendship and support both inside and outside of the university.

It is, of course, customary to thank my parents, who have allowed me to remain in debt with the 'Bank of Mum and Dad' whilst I have postponed getting a 'real job' for as long as physically possible. On a more serious note, they have always provided me with the upmost support and it has never gone unappreciated.

Importantly, I would like to thank Dr Jennifer Miles who helped me leave my chemical roots and learn the 'nitty gritty' of biology. I certainly would have struggled to get my head round some of the trickier concepts without your help. I would also like to extend my gratitude to all the members of the Aspden group; in particular, Katerina who provided me with constant guidance during tissue culture and 'polysoming' and dedicated a large amount of time (often extending into the evening) to helping me and making me feel welcome in FBS.

Abstract

Approximately 45% of the human genome is comprised of mobile, or transposable, DNA elements (TEs). Of this, 11% is attributed to *Alu* elements. *Alu* elements are approximately 300 base pairs in length and are primarily located in the introns of non-coding DNA. However, in some cases, the introduction of an alternative splice site, as a result of an *Alu* insertion in a protein-coding region, leads to the exonisation of a partial *Alu* sequence. This exonisation can lead to the expression of an alternative protein isoform which may have disrupted or altered function and therefore, could have the potential to be cause disease.

Through the use of bioinformatics, this project firstly aimed to predict the extent of *Alu* exonisation and subsequent translation in the human proteome. Additionally, through the use of local sequence alignments, the nature of observed insertions could also be studied. Once prior aims were established, a series of techniques were used to study the possible effects of translated *Alu* insertions on the structure and function of proteins. A number of protein variants were expressed and purified from *E. coli.* Using biophysical techniques, such as ITC and CD, *Alu* structure and any effects of *Alu* insertions on the ligand binding and stability of MBP were studied. Additional binding experiments were performed as a means to explore a potential binding interaction between an *Alu*-like sequence with geldanamycin, an interaction which was initially observed using phage display.

A secondary avenue of research was performed in collaboration with the Aspden and Wurdak groups at the University of Leeds to investigate the difference in translation levels of '*Alu*' and 'non-*Alu*' mRNAs in human cells. Analysis was performed using a combination of polysome profiling, reverse transcription and quantitative PCR.

Abbreviations

AC	*Alu*-containing
AI	Auto-induction
AIP	Acute intermittent porphyria
ARMD	*Alu* recombination-mediated deletion
ASC-1	Activating signal cointegrator 1
ASCC1	Activating signal cointegrator 1 complex subunit 1
ATP	Adenosine triphosphate
BCAS3	Breast cancer amplified sequence 3
BCAS4	Breast cancer amplified sequence 4
BE	Barrett's esophagus
BLAST	Basic local alignment search tool
Bp	Base pair
BSA	Bovine serum albumin
CASC5	Kinetochore scaffold 1
CBPC2	Cytosolic carboxypeptidase 2
CBPC3	Cytosolic carboxypeptidase 3
CCNJL	Cyclin-J-like protein
CD	Circular dichroism
CDKL5	Cyclin-dependent kinase like 5
cDNA	Complementary DNA
CDNN	Circular dichroism analysis using Neural Networks
CK5P1	CDK5 regulatory subunit-associated protein 1

CNTLN	Centlein
Cp	Heat capacity
CP089	UBF0764 protein C16orf89
Da	Daltons
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DSB	Double strand break
DSC	Differential scanning calorimetry
dsDNA	Double-stranded DNA
DTT	Dithiothreitol
E. coli	*Escherichia coli*
EAC	Esophageal adenocarcinoma
EN	Endonuclease
ER	Endoplasmic reticulum
F193A	Protein FAM193A
FA	Fluorescence anisotropy
FAM	Fossil *Alu* monomer
FITC	Fluoroscein isothiocyanate
FLAM	Free left *Alu* monomer
FP	Fluorescence polarisation
FPLC	Fast protein liquid chromatography
FRAM	Free right *Alu* monomer
FT	Flow-through
FXL18	F-box/LRR-repeat protein 18

GBM1	Glioblastoma multiforme
GLYG2	Glycogenin-2
GM	Geldanamycin
GST	Glutathione-S-transferase
HSP90	Heat-shock protein 90
HTP	High-throughput
IPTG	Isopropyl-β-D-1-thiogalactopyranoside
ITC	Isothermal titration calorimetry
ITCH	E3 ubiquitin-protein ligase Itchy homolog
kDa	Kilodaltons
LB	Lysogeny broth
LINE	Long interspersed nuclear element
LTR	Long-terminal repeat
M4K1	Mitogen-activated protein kinase kinase kinase kinase 1
MAGI3	Membrane-associated guanylate kinase, WW and PDZ domain-containing protein 3
MBP	Maltose binding protein
MKNK1	MAP kinase-interacting serine/threonine-protein kinase 1
mRNA	Messenger RNA
MS	Mass spectrometry
MTO1	Mitochondrial protein homolog MTO1
MY15B	Unconventional myosin-15B
MYL10	Myosin regulatory light chain 10
nAC	Non-*Alu*-containing
NADPH	Nicotinamide adenine dinucleotide phosphate hydrogen

NCBI	National Centre for Biotechnology Information
NEK4	Never in mitosis gene A (NIMA)-related kinase 4
NHEJ	Non-homologous end-joining
Ni-NTA	Nickel-nitrilotriacetic acid
NPCL1	NPC1-like intracellular cholesterol transporter 1
Nrk	NIMA-related kinase
ORF	Open reading frame
PAGE	Polyacrylamide gel electrophoresis
PBGD	Porphobilinogen deaminase
PBS	Phosphate buffer saline
PCR	Polymerase chain reaction
PNK	Polynucleotide kinase
PPP5D1	PPP5 TPR repeat domain-containing protein 1
PRR34	Proline-rich protein 34
PTM	Post-translational modification
qPCR	Quantitative PCR
RABX5	Rab5 GDP/GTP exchange factor
REL	Proto-oncogene c-Rel
RF	Restriction-free
RI	Refractive index
RNA	Ribonucleic acid
RNAPIII	RNA polymerase III
rRNA	Ribosomal RNA
RT	Reverse transcriptase

RU	Resonance unit
SDM	Site-directed mutagenesis
SDS	Sodium dodecyl sulfate
SEC	Size exclusion chromatography
SINE	Short interspersed nuclear element
SPR	Surface plasmon resonance
SRP	Signal recognition particle
ssDNA	Single stranded DNA
SUMO	Small ubiquitin-like modifier
TB	Terrific broth
TE	Transposable element
TEV	Tobacco etch virus
TEX11	Testis-expressed protein 11
T_m	Melting temperature
TMM78	Transmembrane protein 78
TPRT	Target-primed reverse transcription
TRIP4	Thyroid hormone receptor interactor 4
TTF1	Transcription termination factor 1
TV	Transcript variant
TV23C	Golgi apparatus membrane protein TVP23 homolog C
Ub	Ubiquitin
UBP19	Ubiquitin carboxyl-terminal hydrolase 19
Ulp1	Ubiquitin-like-specific protease 1
UPP	Ubiquitin proteasome pathway

UV	Ultraviolet
WT	Wild-type
YS049	Zinc finger protein ENSP00000375192
ZMAT1	Zinc finger matrin-type protein 1
ZN195	Zinc finger protein 195
ZN415	Zinc finger protein 415
ZNF91	Zinc finger protein 91

Contents

Table of Figures

List of Tables

Chapter 1
An introduction to *Alu* elements

Overview

Retrotransposable DNA elements have been reported to contribute to a variety of different diseases, most interestingly, cancer. Among them, *Alu* elements, a subclass of retrotransposable elements previously considered to be 'junk DNA', have been observed to be exonised, resulting in the expression of alternative protein isoforms which have been implicated in disease. It has only recently been acknowledged that *Alu* elements may contribute more to the genome than mere expansion, and as a result they may also have a larger effect on the human proteome than previously thought.

1.1 Transposable elements

The human genome is made up of a large assortment of DNA, some of which is able to actively move to and replicate in different genomic locations. These mobile elements, also known as transposable elements (TEs), constitute 45% of the human genome.[1-2] They are a principal contributor to the genetic variation of an ever-changing genome and helped to mould the behaviour and formation of human genes.[3-4] As it stands, TEs are present in every eukaryotic genome that has been sequenced so far.

In the human genome, TEs are categorised into two main classes; Class I and Class II. Class II elements are referred to as DNA transposons and constitute 3% of the human genome.[5] In early evolution, DNA transposons replicated *via* a mechanism most easily described as being similar to the 'cut and paste' function on a modern computer (figure 1.1A).[6-7] That is to say that a DNA sequence was 'cut' from one part of the genome and 'pasted' into an entirely new genomic site, thus, moving the element. As far as it is known, there are currently no active DNA transposons left in the human genome.

Class I elements are referred to as retrotranspositional elements, or RNA retrotransposons.[8, 9, 10] RNA retrotransposons make up a much larger percentage of

the genome totalling approximately 42%.[11] Unlike DNA transposons, RNA retrotransposons replicate *via* a mechanism similar to the 'copy and paste' function of a computer and occurs through the generation of an RNA intermediate (figure 1.1B).[12] In other words, the DNA element is copied into RNA *via* transcription, relocated to a second genomic site where the RNA is converted back into cDNA (complementary DNA) *via* reverse transcription and inserted into the new location.[13]

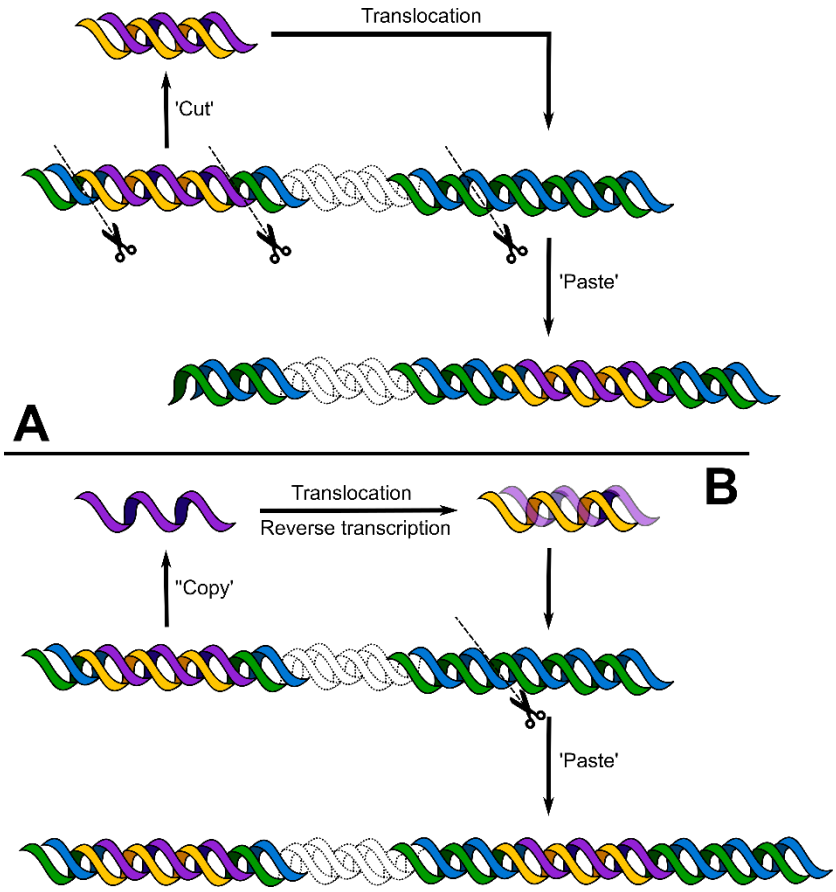


Figure 1.1 Mechanisms of DNA transposition and RNA retrotransposition

(A) DNA transposition. DNA is 'cut' from one genomic location, translocated and is 'pasted' in another region. (B) RNA retrotransposition. DNA is 'copied' into an RNA intermediate which moves to a new genomic region where the RNA is reverse transcribed into cDNA and inserted into the new location.

1.1.1 Retrotransposons in the human genome

Once DNA has been categorised as Class I or Class II TEs, Class I RNA retrotransposons can be further subcategorised (figure 1.2). The first level of sub-categorisation is the splitting of long terminal repeats (LTRs) and non-LTRs. LTR

simply refers to the presence of identical repeating units that flank the retrotransposon sequence.[14, 15, 16] Long terminal repeats are much less abundant (9%) than their non-LTR counterparts, which constitute 33% of the genome. It has been reported than non-LTR elements evolved as early as the first multicellular organisms.[17]

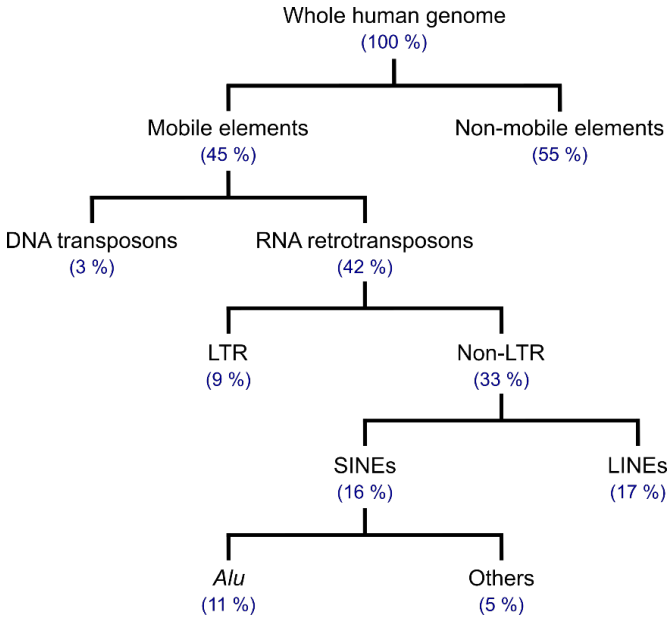


Figure 1.2 Categorisation of mobile DNA elements

Mobile DNA elements can be sorted into several subcategories dependent on their origin, features and replication mechanisms.

A further sub-categorisation of non-LTR elements splits DNA into Short- (SINEs) and Long Intersperse Nuclear Elements (LINEs), contributing 16% and 17% to the human genome, respectively.[18] LINE-1, or L-1, is the only autonomous mobile DNA that remains active in the human genome; however, there are fragments of extinct L-2 and L-3 which can be traced.[19] Approximately 12 million years ago, the expansion of L-1 elements in the human genome slowed, so most L-1 insertions are truncated, rearranged or mutated.[20] SINEs can be further subcategorised into *Alu*, SVA and other small elements, with *Alu* elements being the most abundant comprising 11% of the genome, with over 1×10^6 copies.[21] On average, 5% of new born babies will be born with a brand new retrotransposon insertion.

1.2 The origin of *Alu* elements

Alu elements, as we know them today, are the most abundant TEs in our genome, the majority of which were produced over 40 million years ago.[22] The first *Alu*-like sequence evolved from 7SL RNA and, through a series of evolutionary mutations, gave rise to the generation of many different *Alu* subfamilies.

1.2.1 Evolution from 7SL RNA

7SL RNA is a key component of the signal recognition particle (SRP), involved in the secretion and translocation of proteins in the endoplasmic reticulum (ER).[23] A study comparing the homology of insect 7SL RNA with mammalian 7SL RNA and *Alu* sequences revealed that the presence of the *Alu* sequence in the RNA arose prior to the divergence of mammals on the evolutionary tree of life.[24] The sequence of 7SL RNA is composed of an *Alu* sequence split by a 155-base pair (bp) sequence which is exclusive to 7SL RNA. This leaves approximately 100 bp and 45 bp at the 5' and 3' ends of 7SL RNA, respectively, which are homologous to the *Alu* right arm monomer.[25] This homology also accounts for the conservation of an RNA polymerase III promoter region.[26]

The first evolutionary step from 7SL RNA towards today's recognised *Alu* structure, involved the central deletion of the aforementioned 155 bp sequence of 7SL RNA giving the fossil *Alu* monomer (FAM).[27] This monomer is the oldest common ancestor of all SINEs derived from 7SL RNA.[28] Further mutation of FAM gave rise to the free right *Alu* monomer (FRAM) and the free left *Alu* monomer (FLAM), also known as the free *Alu* right and left arms, respectively. Combination of these monomers resulted in the consensus *Alu* sequence we know today (figure 1.3).[29]

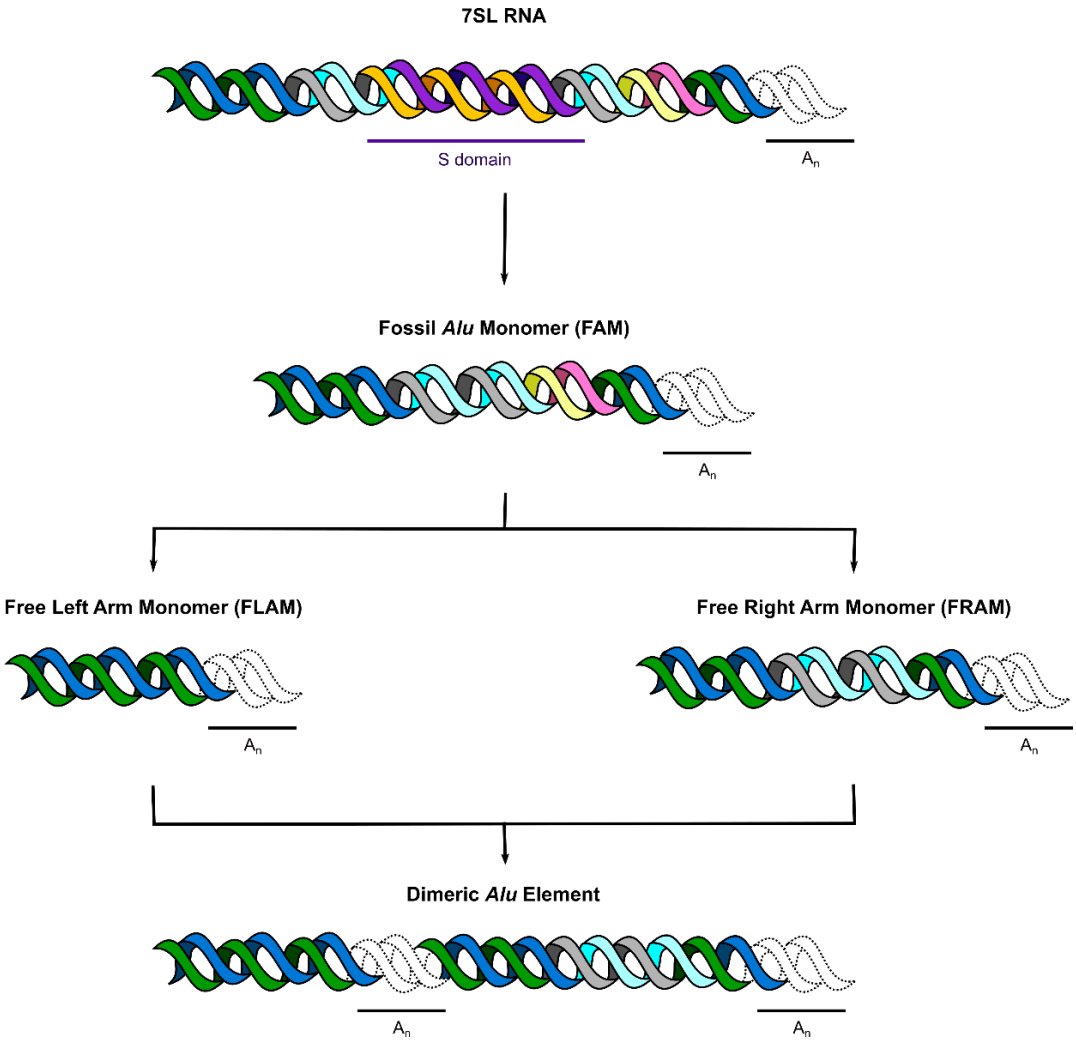


Figure 1.3 Evolution of *Alu* elements from 7SL RNA

The fossil *Alu* monomer (FAM) was formed through the deletion of the S-domain of 7SL RNA (purple/orange). Further sequence deletions resulted in the formation of the free left arm monomer (FLAM; deletion = grey/aqua, pink/yellow) and the free right arm monomer (FRAM; deletion = pink/yellow). The FLAM and FRAM combined to form a dimeric *Alu* element split by a poly-A sequence.

1.2.2 Alu subfamilies

Modern *Alu* elements are only observed in the genomes of primates, though they give rise to many different subfamilies within those genomes.[30] The broadest categorisation of *Alu* elements is into three main classes; *Alu*J, *Alu*S and *Alu*Y, which can then be subcategorised further, dependent on the location of base mutations (figure 1.4).[31] The *Alu*J class is the oldest dimeric *Alu* approximated to have arisen around 80 million years ago. *Alu*S subfamilies are of an intermediate age having

evolved approximately 30 – 50 million years ago and *Alu*Y subfamilies are the youngest subfamilies and are less than 15 million years old. Of these subfamilies, *Alu*S subfamilies are the most common, with *Alu*Sx being the most abundant *Alu* in primate genomes.[32] *Alu*Y subfamilies, due to their youth, are the only *Alu* class that are still retrotranspositionally active. That is to say, they are the only subfamilies that are still retrotransposed to create new insertions. For older subfamilies, which no longer undergo active retrotransposition, new *Alu*-like insertions are likely to have instead occurred through DNA recombination events.

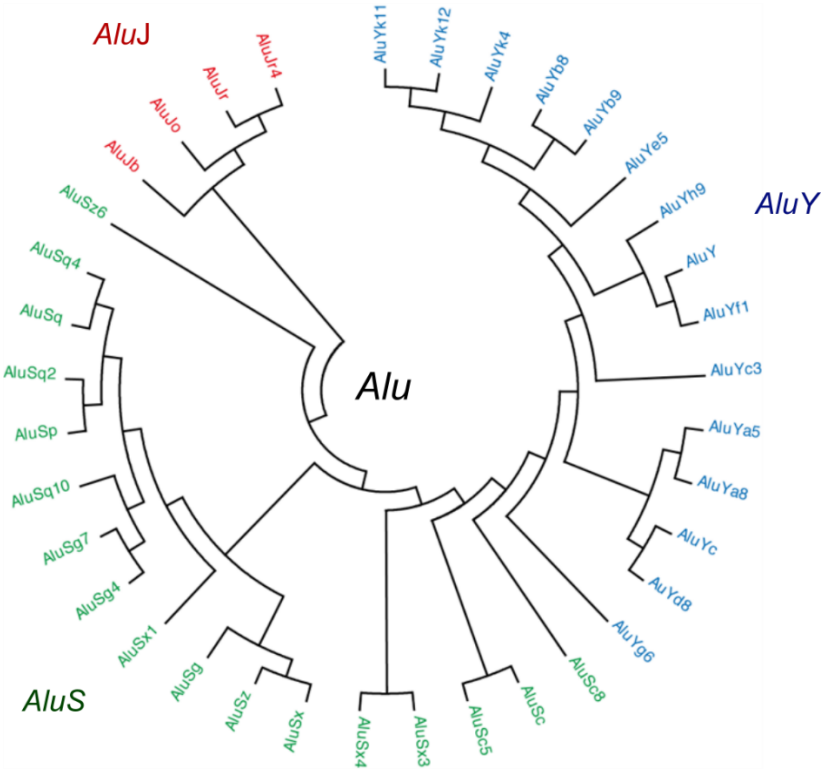


Figure 1.4 Divergence of *Alu* subfamilies

The *Alu*J class of *Alu* elements is the oldest, arising approximately 80 million years ago. From this, the *Alu*S class of subfamilies diverged 30 – 50 million years ago. *Alu*S subfamilies remain the most abundant *Alu* class in the human genome, with *Alu*Sx being the most common. *Alu*Y subfamilies are less than 15 million years old and remain the only active *Alu* class.

1.3 *Alu* structure

Alu elements (figure 1.5) are dimeric structures that are approximately 300 bp in length.[33] They are comprised of two monomers, left and right, which are identical to one another other than an approximately 34* base pair insert which is present only in the right monomer.[34, 35] The monomers are connected *via* a poly-A linker region; a second poly-A region acts as a tail at the 3' end of the element, which is responsible for active retrotransposition.[36] Younger *Alu*Y subfamilies have longer poly-A tails which account for their conserved activity. However, older classes, *Alu*J and *Alu*S, have significantly shortened tails which is assumed to contribute to their loss of activity.[37]

The left *Alu* monomer hosts an RNA polymerase III (RNAP III) promotor region composed of an A- (TGGCTCACGCC) and B- box (GTTCGAGAC).[38] The presence of the approximately 34 bp insert in the right monomer splits the B-box thus, the right monomer does not have RNAP III activity.

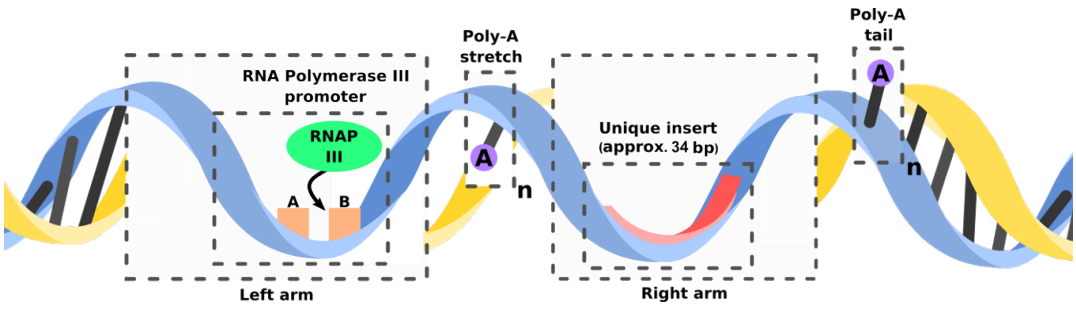


Figure 1.5 *Alu* element structure

Alu elements are approximately 300 base pairs in length, comprising of a left and right monomer. Monomers are split by a poly-A tract and similarly, have a poly-A tail at the 3' end. Monomers are identical to one another aside from an approximately 34 bp insert in the right arm monomer. The left arm hosts an RNA polymerase III promoter region required for retrotranspositional activity.

* Literature quotes insert length at approximately 31 bp; observations made upon alignments of Dfam *Alu* sequences calculated an average of 34 bp in this work.

1.4 *Alu* movement and replication in the genome

Alu elements are non-autonomous SINEs.[39] As a result, they do not host the 'machinery' required for active retrotransposition. For active *Alu*Y subfamilies, retrotransposition occurs through the 'hijacking' of LINE-1 (L-1) machinery.[40] L-1 elements are responsible for the insertion of over 1×10^6 non-autonomous SINEs.[41] The generation of new insertions arising from older, inactive classes, *Alu*J and *Alu*S, are usually a result of DNA recombination events which have used *Alu* sequences as homologous templates for DNA repair.[42]

1.4.1 Retrotransposition

The retrotransposition of active *Alu* elements occurs through utilisation of L-1 machinery. LINE-1 elements host ORF1 and ORF2 genes, which encode the ORF1p and ORF2p proteins, respectively, required for their own autonomous retrotransposition.[43, 44] ORF1 encodes the approximately 40 kDa (kilodalton) protein, ORF1p, about which relatively little is known. Its origins appear to be associated with retrotransposition, however, it shares very low sequence homology with any sequences in databases of known proteins.[45] ORF2 encodes a much larger, approximately 150 kDa, protein that exhibits both endonuclease[46] and reverse transcriptase activity.[47] ORF2p has two conserved domains, the Z-domain and a cysteine-rich (C-domain). The C-domain has unknown function, however, it has been observed that mutations in this domain eradicate L-1 retrotranspositional activity.[48] *Alu* elements require only ORF2p for retrotransposition, which they 'borrow' from LINE-1.[49] They achieve this through binding with the SRP9/14 subunit of the signal recognition particle (SRP) which localises them to the ribosome.[50, 51] This mode of action is supported by the fact that mutations in the *Alu* SRP 9/14 binding sequence eliminates its binding interaction.[52] It is this localisation which brings the *Alu* element within close enough proximity of L-1 elements to interact with ORF2p using both its L-1 endonuclease and L-1 reverse transcriptase activity.[53]

Since *Alu* elements use the same protein as L-1 elements for active retrotransposition, it is assumed that the process proceeds *via* the same mechanism (figure 1.6).[54]

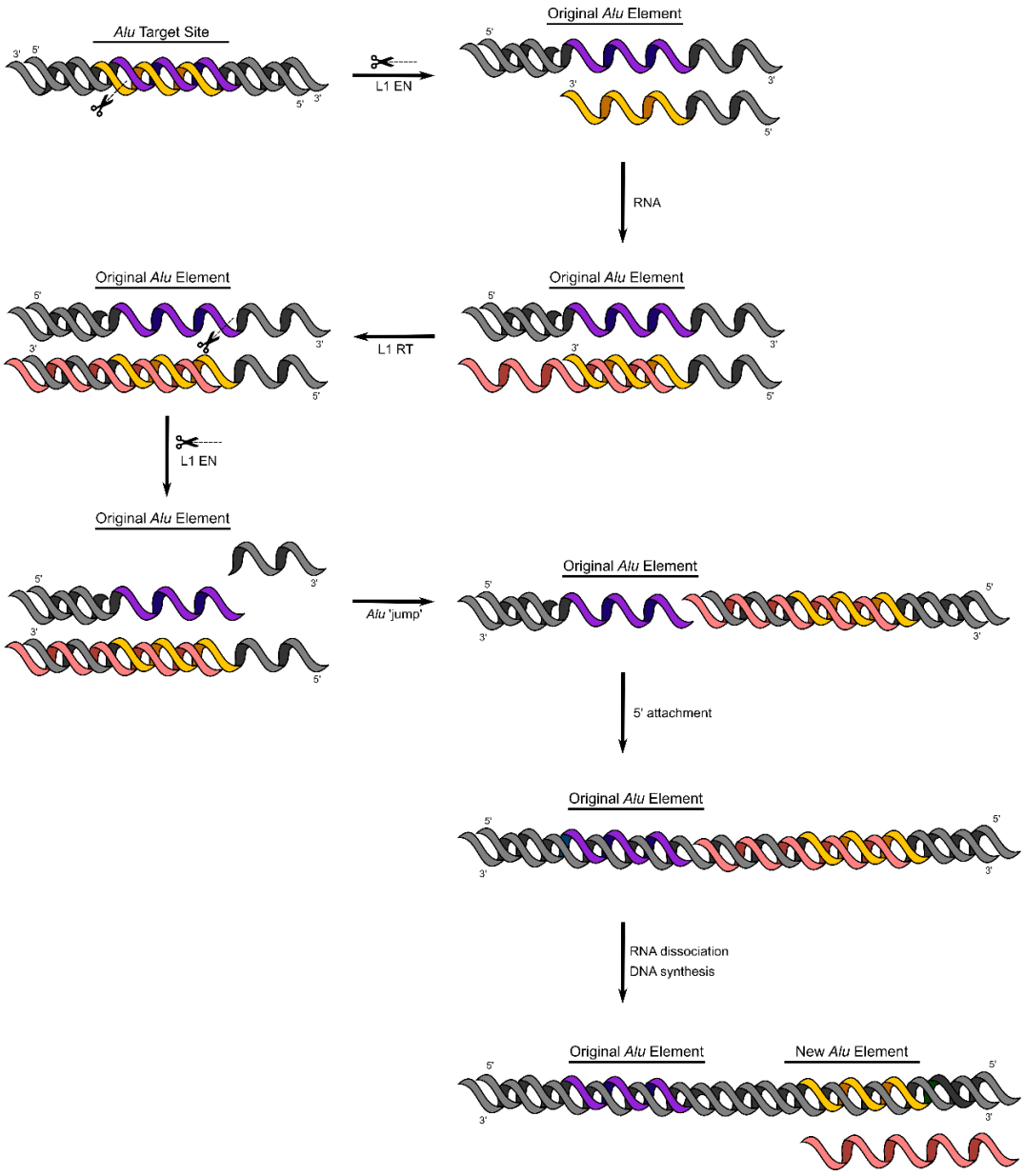


Figure 1.6 L-1 mediated retrotransposition of *Alu* elements

L-1 endonuclease (L1 EN; from N-terminus of ORF2p) cuts the reverse (-) strand of double stranded DNA (dsDNA) at the *Alu* target site. RNA anneals to the reverse stand and using L-1 reverse transcriptase (L1 RT) extends cDNA in the 5' – 3' direction, copying the second *Alu* strand. L1 EN cuts the forward (+) strand of the original DNA at a second target site located at the opposite end of the *Alu*. The original *Alu* 'jumps' along the DNA where the extended cDNA from the reverse strand anneals to the dsDNA of the insert. The RNA template dissociates, and cDNA is synthesised to fill in the gap.

Retrotransposition occurs through a series of steps. Firstly, L-1 endonuclease (L1 EN), which arises from the N-terminus of the ORF2p and shares sequence homology with AP endonuclease, determines the site of *Alu* insertion through the generation of a single strand break in the target DNA. This is usually in the form of a 5' TT|AAAA motif.[55] RNA anneals to the free single stranded (ss) *Alu* element and L-1 reverse transcriptase (L1 RT), the characteristic motifs for which lie at the C-terminus of ORF2p, initiates target-primed reverse transcription (TPRT) at the insertion site. Elongation uses the 3' end of the target site, which was released by L1-EN, as a primer. A second cut is made on the *Alu* forward (+) single strand at a second cut site located at the opposite end of the *Alu* element. The *Alu* 'jumps' along the gene and inserts itself into a new location. The 5' end of the reverse strand anneals at the first cut site. The RNA template dissociates, and the gaps are filled in by complementary DNA (cDNA).[56, 57]

As previously stated, only younger *Alu* subfamilies (*Alu*Y) still have the ability to actively undergo retrotransposition in this way. The rate of *Alu* retrotransposition is approximately one in every twenty births,[58] which is similar to that of LINE-1 elements.[59, 60, 61]

1.4.2 DNA recombination events

Despite no longer undergoing active retrotransposition with L1 machinery, new insertions from older, inactive *Alu* elements, such as *Alu*J and *Alu*S, still arise in the genome. When this occurs, it is usually the result of a DNA recombination event at the point of a DNA double strand break.[62]

In cases were DNA damage results in a double strand break (DSB), one of two mechanisms can occur to repair the damage; non-homologous end-joining (NHEJ) or homologous recombination.[63] NHEJ involves the direct ligation of broken DNA ends and therefore often results in deletions.[64] Homologous recombination events involve the use of homologous DNA template. Due to the high abundance of *Alu* elements in the human genome, it is likely that *Alu* elements share high sequence homology with certain genomic locations, also known as 'hot spots', and as a result, they are used as templates in DNA repair (figure 1.7).[65] Additionally, the similarity between the two *Alu* arms also provides a way in which mistakes in DNA repair could arise. The most common example of this is *Alu* mismatch, which is the mis-pairing of *Alu* elements for recombination and often results in either duplication or

deletion.[66] Homologous recombination starts with 5' resection to produce single stranded DNA (ssDNA) overhangs at the site of the DSB. The homologous *Alu* DNA then anneals and is used as a template for DNA repair, resulting in a new *Alu* insertion at the repair site.[67] In cases where an *Alu* mismatch has been made, this results in a sequence different to the original being introduced, or in some cases, parts of the sequence being missed out during repair, resulting in an *Alu* recombination-mediated deletion (AMRD).[68]

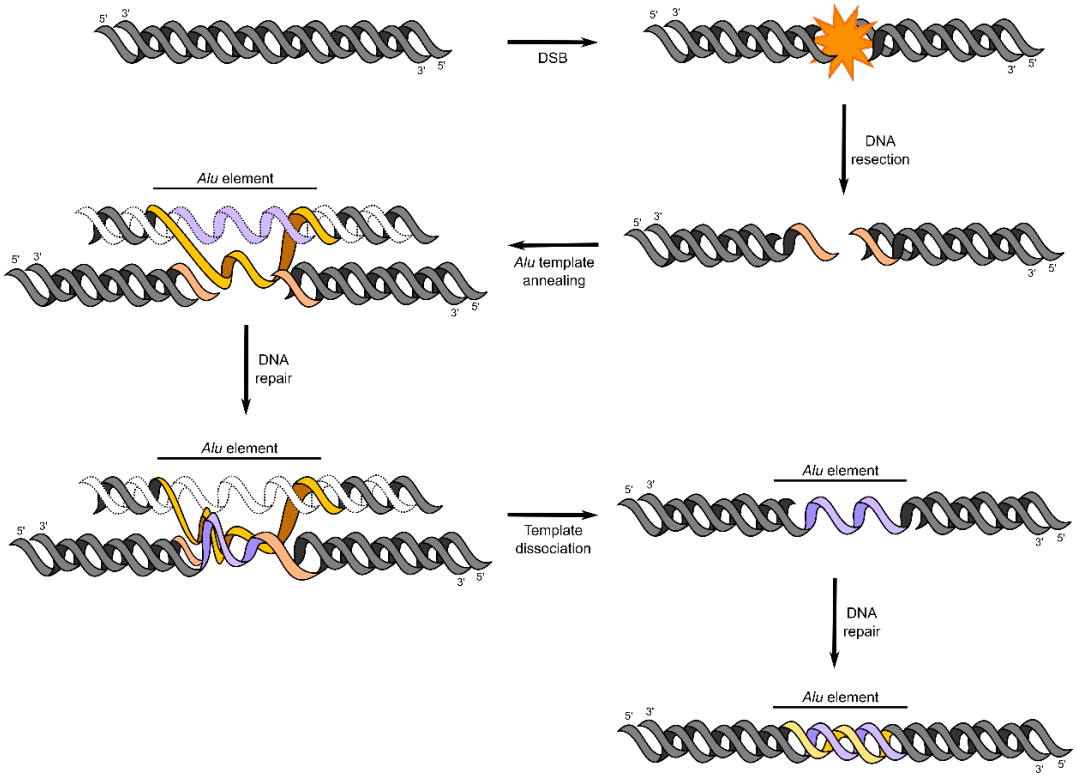


Figure 1.7 *Alu* insertion through DNA recombination events

DNA damage results in a double strand break (DSB). DNA resection occurs at the 5' ends of the break to give single stranded DNA (ssDNA) overhangs. An *Alu* element with high homology with the break site acts as a temple for DNA repair then dissociates leaving a copy of itself at the repair site.

1.5 Transcription and translation of *Alu* Elements

Proteins are produced *via* the central dogma of molecular biology, which states that there are two steps to protein production; transcription and translation. Transcription refers to the copying of DNA into an mRNA intermediate, and translation refers to the 'reading' of mRNA and production of the amino acid chain which makes up the protein. However, this description is some-what simplified, and the overall mechanism is much more complicated.[69]

1.5.1 Transcription of mRNA

Transcription refers to the process which copies DNA into pre-messenger RNA (mRNA). The process occurs through a reaction catalysed by RNA polymerase II (RNAP II). Transcription by RNAP II begins with the binding of regulatory transcription factors to the DNA strand at the site of transcription initiation.[70] Unlike DNA polymerases, RNA polymerases do not need primers to initiate synthesis.[71] RNAP II is positioned at the site *via* a promotor at which point the DNA helix is unwound to reveal 11 – 15 bases on a single stranded DNA template. Sequential addition of nucleoside triphosphates (ATP, GTP, CTP and UTP) and phosphodiester bond formation (catalysed by RNAP II) results in the extension of an mRNA coding strand in the 5' – 3' direction.[72] This strand is identical to the complementary strand of the DNA template with the minor difference that thymine (T) nucleotides are substituted for Uracil (U) nucleotides. Each RNA polymerase II is capable of adding 20 – 50 bases per second,[73] and with over 100 RNA polymerases working at any one time, it is possible to generate over 100 transcripts per hour.[74] Unlike with DNA synthesis, the synthesis mRNA does not remain hydrogen bonded to the transcript, instead it dissociates, allowing for the DNA helix to reform. Therefore, RNAP II moves along the DNA template unwinding the helix just ahead of its active site and allow it to reform behind it (figure 1.8).

RNA polymerase has an error rate of 1 in every 10^4 nucleotides, which is much higher than for DNA polymerases (1 in every 10^7 nucleotides). Errors made by RNA polymerases can lead the changes in exon splicing.[75]

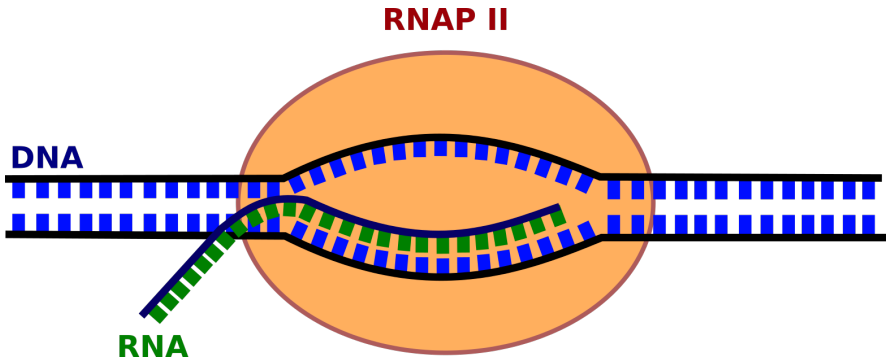


Figure 1.8 mRNA transcription by RNAP II

RNA polymerase II (RNAP II), moves along the DNA strand sequentially adding bases (A, U, G, C) to the extending mRNA chain. 11 – 15 bases on the DNA helix are unwound at any one time to reveal the DNA template. mRNA dissociates and the DNA helix reforms once the RNA polymerase has moved on.

1.5.2 The spliceosome

Simply put, splicing is the mechanism which removes introns from pre-mRNA transcripts resulting in a mature mRNA transcript which includes only exonised RNA. This mechanism occurs through cleavage at points called slice sites, which are conserved RNA sequences found at the 5' and 3' ends of introns.[76] The most common splice site is GU a the 5' end of an intron and AG at the 3' end. Splicing of the major (U2) class of introns generally occurs at CAG|G at the 3' end. Conversely, splicing tends to occur at X_1AG|GUX_2AGU at the 5' end, where X_1 and X_2 are A/C and A/G, respectively.[77] However, other classes of introns give rise to alternative splice sites.[78, 79, 80] Changing any one of the bases contributing to a splice site can result in complete inhibition of splicing.

Aside from the splice sites themselves, a secondary sequence, called the branch point, is also very important in the splicing mechanism. The branch point, which is located 18 – 40 nucleotides upstream of the 3' splice site, is responsible for the initiation of nucleophilic attack on splice sites.[81] The branch point has very loose conservation in comparison to the splice sites themselves; however, it always contains an adenine (A).

Splicing is a multi-step mechanism (figure 1.9) which is catalysed by small nuclear ribonucleoproteins (snRNPs), a major class of uridine (U)-rich non-coding RNAs which are bound by specific proteins to give an RNA-protein complex.[82] U1 snRNPs

attach to the 5' end of the complementary strand of the intron, at the splice site, and cleave it. The free end then attaches to the branch point through the pairing of guanine (G) to adenine (A) *via* transesterification to form a looped structure known as a lariat.[83] Additional snRNPs (U2 and U4/U6) aid in placing the 5' end and the branch point in close proximity of one another. Following initial transesterification, U5 snRNPs bring the 3' splice site within proximity of the 5' end and a second transesterification reaction cleaves the 3' end and joins it with the 5'.[84] The mechanism results in the joining of the two exons either side of the spliced intron and a lariat-snRNP complex which dissociates.

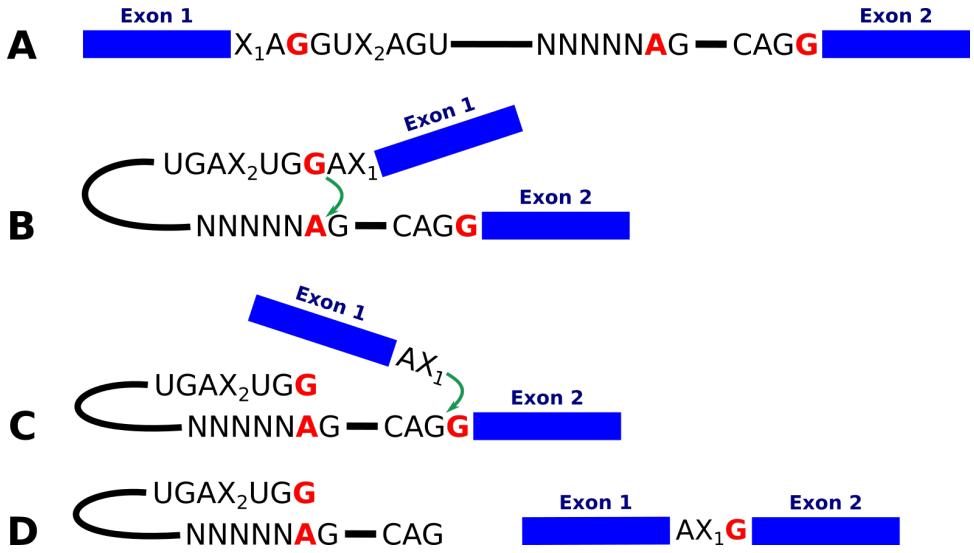


Figure 1.9 mRNA splicing

(A) An intron, flanked by two exons contains a 5' splice site, a branch point and a 3' splice site (left to right). (B) Catalysed by snRNPs, the 5' splice site is cut and joins to the branch point *via* transesterification. (C) By a second transesterification reaction, the 3' splice site is cut and joined to the free 5' end (D) to form a lariat and the two connected exons.

In addition to snRNPs, splicing is also regulated by a number of different splicing factors. These include trans- and cis-acting proteins, which can be activators or repressors, or silencers and enhancers, respectively.[85] Together, these factors determine how splicing occurs under different cellular conditions such as in different tissue types or under stress.

1.5.3 *Alu* inclusion in introns and exons

In most cases, *Alu* insertions occur in the introns of genes and therefore have no effect on the protein-coding DNA of a gene.[86] However, sometimes *Alu* insertions can introduce an alternative splice site into a protein coding region, leading to an alternative splicing event and hence, the partial exonisation of the *Alu* sequence.[87] This exonisation leads to the translation of the sequence as part of a protein, leading to the formation of an alternative protein isoform (figure 1.10).[88, 89, 90] In some cases, this insertion can lead to the formation of a disease-causing isoform.

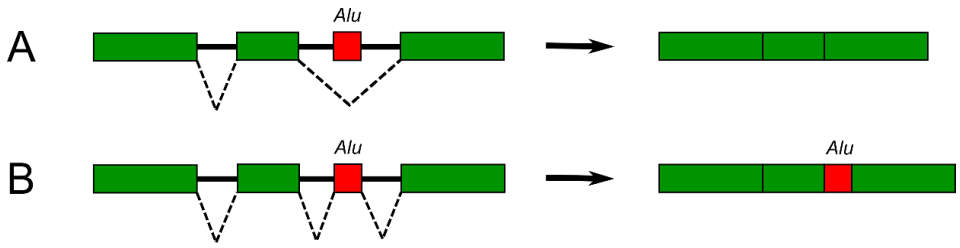


Figure 1.10 Alternative splicing of exonised *Alu* elements

Alu insertions can introduce a new splice site into protein coding regions. (A) The new splice site is ignored forming the non-*Alu* protein isoform. (B) The new splice site is incorporated resulting in translation of the *Alu* element into an alternative protein isoform.

Aside from the introduction of alternative splice sites, there are also other ways by which *Alu* elements can affect splicing and transcript formation and regulation. In some cases, the formation of inverted *Alu* pairs can lead to the circularisation of mRNA transcripts.[91] The inclusion of *Alu* elements in introns can sometimes interrupt silencers and enhancers and therefore affects the inclusion of downstream exons.[92] In the same way, they can affect activators and repressors altering the recognition of splice sites.[93]

1.5.4 Other effects of Alu mRNAs

Though many *Alu* elements are incorporated into introns and exons, free *Alu* RNAs are also generated.[94] Free *Alu* RNA has been observed to do a number of multiple things. Synthesis of synthetic *Alu* ribonucleoproteins (RNPs comprising of *Alu*-SRP/14 complexes) observed enhanced translation of reporter *Alu* mRNAs, but saw a general decrease in protein translation.[95, 96] *Alu* mRNAs have also been observed to have selective stimulation of translational expression.[97]

1.6 *Alu* elements in humans and their relation to disease

1.6.1 Current studies of *Alu* elements in humans

A large amount of interest with respect to *Alu* elements is the study of their polymorphisms as a means to greater understand human population genetics and diversity.[98] This is most easily observed in the introduction of younger, *Alu*Y, subfamilies which can only be traced in the human genome and not in that of other primates.[99] The incorporation of a new *Alu* mutation into a population is highly dependent on genetic drift. For example, if an insertion is introduced into a smaller population, it is less likely that the mutation will be lost. Additionally, the more an insertion is amplified over time, the more set it becomes in the genome. Polymorphisms of *Alu* elements can be traced through primate genomes. As previously mentioned, some polymorphisms are so new that they are only present in the human genome.[100] Though there are only a few, there are some cases where polymorphisms are present in the genome of one human but absent in another. These are known as *Alu*-insertion polymorphisms.[101] In rare cases, an individual *Alu* may be found in a single population, family or even individual (*de-novo* insertion) dependent on genetic drift.

A study into the frequency of *Alu* polymorphisms among populations revealed that the frequency varies between populations.[102] It was observed that the highest number of polymorphisms was observed in Africans, and the lowest in Europeans. As human evolution is known to be of African origin, this is result was consistent with the current model of human evolution. From a timescale perspective, it is difficult to say how often new *Alu* polymorphisms arise as the rate is constantly changing not across the entire human population but also in individual populations.

In most cases, there does not appear to be any negative impact of *Alu* elements with respect to genomic diversity; however, some have been implicated in disease.

1.6.2 *Alu* elements in disease

There are an increasing number of discoveries that link *Alu* elements to disease, with over sixty reported disease-causing *Alu* insertions.[103] In addition, the rate of retrotransposition of disease-based *Alu* elements is much higher than that of their evolutionary retrotransposition rate, indicating that they may be in a phase of higher activity than usual. *Alu* contribution to disease has be reported to occur at both the nucleotide level (mRNAs/DNA)[104], as a well as at the proteome level. However, there appears to have been less research into the level of effect *Alu* elements have on the proteome.

There are a number of reports that link *Alu* elements to disease whether it be through disease regulation[105] or the direct impact of exonised *Alu* elements in protein isoforms.[106] One example of the latter lies in *Alu*-specific deletions in the cyclin-dependent kinase like 5 (CDKL5) gene, which leads to a frame shift in the translated protein and leads to early-onset seizure disorder in females.[107] In another report, the insertion of an *Alu* element into the porphobilinogen deaminase (PBGD) gene disrupts its fifth exon leading to the expression of an alternative protein variant with abolished enzyme activity. This is a detrimental insertion that leads to acute intermittent porphyria (AIP).[108] There have been many other reports of *Alu* insertions leading to disease *via* the formation of alternative protein isoforms including links to Alzheimer's disease[109, 110], Apert Syndrome[111] and cancer.[112, 113]

1.7 Project aims

Currently, the majority of research into *Alu* elements is centred around the evolution of *Alu* elements and their impact at the nucleotide level. There are still may questions about the true impact of *Alu* elements on the human genome, and even more about their effects on the human proteome. Though a lot of research has been done on the impact of intronic *Alu* elements, very little has been done on exonised *Alu* elements. Moreover, extending the study of *Alu* exonisation to their incorporation into proteins and how they impact protein structure and function leaves a fairly open gap in the scientific field.

As a result, this project first aimed to analyse the extent of *Alu* exonisation in the human genome which leads to the formation of alternative *Alu*-containing protein isoforms. Secondary aims involved the study of how the insertion of such *Alu* elements in proteins affected their structure and function. Initial analysis into the abundance of *Alu* elements in protein coding exons was performed through the use of bioinformatic techniques based around the sequence alignment of both *Alu*-related nucleotide and protein sequences. Research into the effect of *Alu* elements on protein structure was performed through peptide binding assays, protein expression in *Escherichia coli* (*E. coli*) and a series of biophysical analytical techniques. Additional work to decipher the differing translation levels of *Alu* and non-*Alu* mRNAs arising from the same gene in ‘cancerous’ and ‘non-cancerous’ human cells was performed using a combination of polysome profiling, reverse transcription and qPCR.

Chapter 2
Bioinformatic analysis of *Alu* elements

Overview

Alu elements have been reported to be present in high abundance within the human genome, contributing approximately 11% to total genomic DNA. Until fairly recently, *Alu* elements were described as 'junk DNA';[114] however, recent studies have revealed this to be untrue. Although research by groups such as Mitchell *et al*[115] and Sorek *et al*[116] revealed that *Alu* elements in protein-coding regions can be exonised so as to be incorporated into proteins, the extent to which this happens remains unstudied.

Through use of a series of bioinformatic techniques, 57 protein hits were initially identified to contain translated *Alu*-like regions. This was refined with further bioinformatics to 46 different proteins, giving rise to 65 individual isoforms. Of the total 46 identified proteins, 32 could be translated as multiple isoforms which could be either *Alu*-containing or non-*Alu*-containing. Analysis of 65 *Alu*-containing protein isoforms concluded that sequences arising from *Alu* insertions were more common at protein termini than within internal protein regions. Bioinformatic studies of hits at the nucleotide level revealed that insertions primarily arose from the *Alu* left arm and corresponded to sequences copied from the antisense (-) strand of the parental *Alu*. Alignment of primary reading frames revealed a conserved amino acid sequence present in the majority of protein hits. Building upon the work of Lev-Maor *et al,*[117] six possible 3' AG splice sites were identified in hit mRNAs as an insight into the origin of the *Alu* insertions observed.

2.1 Identification of *Alu*-containing protein sequences

Most of the human genome is made up of non-coding DNA. As this research aimed to identify *Alu* insertions within protein coding-regions, consensus sequences were translated into open reading frames (ORFs) for analysis. At the time of this bioinformatic analysis, only the eight *Alu* consensus sequences obtained from the National Centre for Biotechnology Information (NCBI)[118] were readily available for

sequence comparison. Alignment of these consensus sequences showed 80% conservation between subfamilies with differences arising only from minor base mutations that had accumulated over evolutionary time (figure 2.1). Sequences were translated and screened against a database of human proteins to give a total of 57 protein hits. Due to the aforementioned sequence conservation between *Alu* subfamilies, in many cases the same protein matches were observed with multiple subfamilies. Upon translation of *Alu* consensus sequences, similar ORFs were yielded from each subfamily.

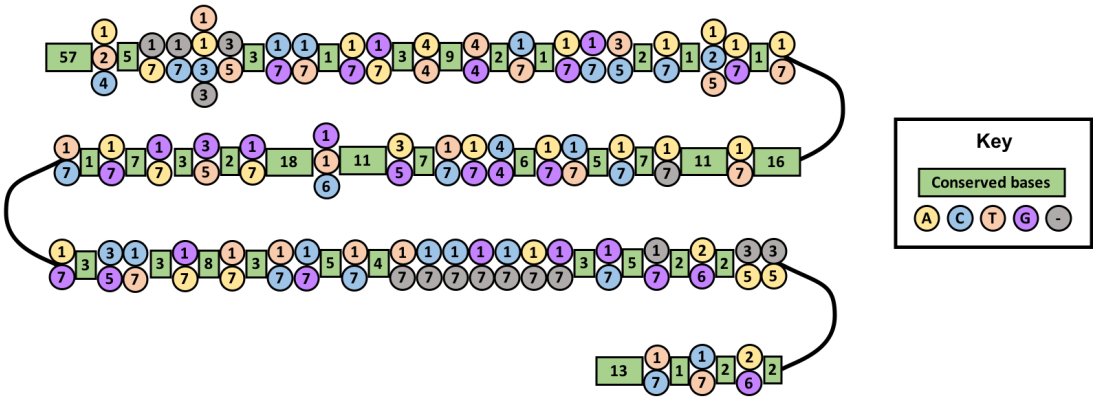


Figure 2.1 Summary of differences between *Alu* subfamilies upon alignment
Over eight *Alu* subfamilies, we see a conservation of a total of 235 bases. As the full length of each *Alu* subfamily ranges from 287 – 296, approximately 80% of the sequence is conserved between *Alu* subfamilies.

2.1.1 Creating a database of *Alu*-containing proteins

All eight of the *Alu* consensus sequences (J, Sx, Sp, Sq, Sc, Sb, Sb1, Yb) were translated into their respective six possible ORFs using ExPASy translate,[119] to give a total of 48 translated sequences. *Alu* consensus sequences can be found in Appendix 1. ORFs were individually screened against a library of known human proteins (UniProtKB/Swiss-Prot) using the NCBI basic local alignment search tool (BLAST).[120] A total of 57 protein hits with a sequence identity match[†] above 68% were identified. Search results with a sequence identity below 68% tended to be shorter sequences or include significant gaps between matched regions therefore,

[†] The amount of characters exactly matched between two sequences in relation to the shorter sequence, excluding gaps.

sequences matching less than 68% of the ORF sequence were disregarded as a way to reduce partial matches. Approximately 17 uncharacterised and putative proteins were omitted.

Protein	Matched *Alu*	Percentage Identity	Isoforms
NANGN	J Sx Sp Sq Sc Sb Sb1 Yb	97	1
PKP2	J Sx Sp Sq Sc Sb Sb1 Yb	95	2
OR1FC	Sx Sp Sq Sc Sb Sb1 Yb	95	1
ZN429	Sx Sp Sq Sc Sb Sb1 Yb	95	1
MOST1	J Sx Sp Sq Sc Sb Sb1 Yb	94	1
TV23C	J Sx Sp Sq Sb Sb1 Yb	94	3
HS905	J Sx Sp Sq Sc	94	1
RABX5	J Sx Sp Sq Sc Sb Sb1 Yb	93	4
MYL10	J Sx Sp Sq Sc Sb Sb1 Yb	93	1
ZN701	J Sx Sp Sq Sc Sb Sb1 Yb	92	2
GLOD4	J Sx Sq Sc Sb Sb1	92	3
CK5P1	J Yb	92	6
ZN283	J Sx Sp Sq Sc Sb Sb1 Yb	92	1
ZN415	J Sx Sp Sq Sc Sb Sb1 Yb	91	6

Table 2.1 The 15 hits with the highest percentage identity match to *Alu* ORFs
Many of the proteins listed have hits with multiple *Alu* subfamilies due to 80% sequence similarity between *Alu* consensus sequences. The recorded percentage identity is that of the best matched subfamily.

The 15 hits with the highest percentage identity match are recorded in table 2.1. A full list of hits is contained in Appendix 1. Of the 57 proteins identified, only five were found to match solely with one subfamily; TEX11, GLYG2, PRR34, YA021 and PPP5D1. These matches are likely to be shorter sequences and/or arise from less conserved regions of *Alu* elements. 15 proteins were found to match with all of the eight *Alu* subfamilies, though at different percentage identities. As previously discussed, this is unsurprising due to high sequence similarity between *Alu* subfamilies. The percentage identity values listed in table 2.1 correspond to those of the highest matching subfamily.

2.1.2 Database refinement

Protein hits with each *Alu* consensus sequence were individually analysed. All six ORFs were aligned with each isoform of their respective hits. In addition to a percentage identity of > 68%, hits were refined so as to only include those with an expect value (E-value)‡ of 1 x 10^{-8} or lower. These cut-offs were chosen so as to limit the overall number of hits to between 50 and 100. Though some hits may have been unidentified due to these cut-offs, in addition to those lost due to the use of a single database (UniProt KB/Swiss-Prot), for the purpose of this work the number of hits obtained was sufficient.

Database refinement resulted in a reduction of hits from 57 to 46 (figure 2.2A). Those lost included; CK5P1, CASC5, ZNF91, TEX11, MTO1, CBPC2, NPCL1, MY15B, PRR34, TTF1 and YA021. As noted previously, TEX11, PRR34 and YA021 were likely to have matched with shorter sequences due to only matching with one *Alu* subfamily. As a result, it is unsurprising that a stricter cut-off resulted in their loss. The 46 *Alu*-containing (AC) proteins gave rise to 65 isoforms which each contained an *Alu*-like insertion. In addition to this, there were also 68 non-*Alu*-containing isoforms (nAC) arising from the same list of proteins. Of the total 46 hits, 32 hits could be translated as both AC and nAC protein isoforms (figure 2.2B).

After database refinement, only two hits matched with a single *Alu* subfamily; PPP5D1 and GLYG2. Over half (54%) of all refined hits matched to six or more of the eight *Alu* subfamilies. As a result, the resulting translated ORFs observed to arise in proteins tended to arise from very similar sequences. This is discussed further in section 2.4.

‡ The number of hits expected to arise from chance in a database of a certain size. The closer the E-value lies to zero, the more significant the hit.

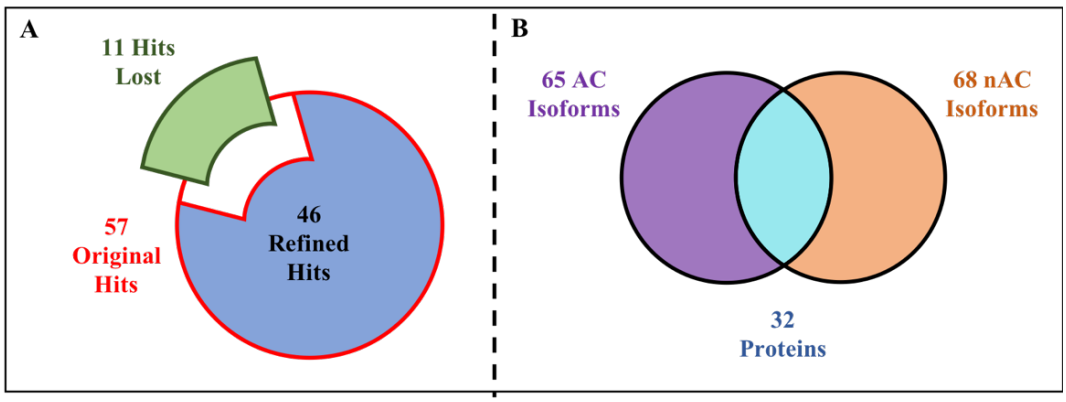


Figure 2.2 Summary of hits after database refinement

(A) Of a total of 57 original hits with above 68% identity match were refined to 46 after the loss of 11 hits that had an E-value of above 1 x 10^{-8}. (B) Although 46 proteins were identified to contain *Alu* insertions, this gave rise to a total of 65 *Alu*-containing (AC) isoforms. There was also a total of 68 non *Alu*-containing (nAC) isoforms for the same list of proteins. 32 of the 46 proteins hits could be translated as both AC and nAC isoforms.

2.2 Studying the locations of *Alu* insertions within proteins

In order to determine whether there was a trend in the locations of *Alu* insertions within proteins, 75 insertions from the refined database were studied.

Proteins were defined as having three distinct regions: N-terminal, internal and C-terminal. Insertions located in the first 20% of residues in the protein were defined as N-terminal, whereas those located in the last 20% were defined as C-terminal. Those arising in the middle 60% of residues (20 – 80%) were defined as internal.

2.2.1 Alignment of AC proteins with *Alu* ORFs

AC protein sequences were directly aligned with their corresponding *Alu* ORF using NCBI BLAST. As previously stated, the use of only one database, (UniProtKB/Swiss-Prot) may have resulted in a number of unidentified hits. Due to the nature of analysis performed in this project, this is unlikely to affect the overall results obtained. However, the list of hits obtained is likely only a partial representation of the full extent of *Alu* presence in the human proteome. The resulting alignments showed regions of similarity between the protein sequence and the *Alu* ORF alongside numbers defining the start and end residues of the sequence match (figure 2.3). Further alignments of protein hits and their translated *Alu* ORFs are contained in Appendix 1.

AluSx ORF 1	**13**	LECSGAISAHCNLRLPGSSDSPASASRVAGIT	**44**
		+ECSG I A CNLRLPGSSDSPASAS VAGIT	
BCAS4 ISO 1	**164**	VECSGTIPARCNLRLPGSSDSPASASQVAGIT	**195**

Figure 2.3 Example alignment of *Alu* ORF with protein isoform

Alignment of BCAS4 isoform 1 (ISO 1) and *Alu*Sx ORF 1 revealed a matched region as shown. Numbers on either end of each sequence indicate the residues at the start at end of the matched region. Conserved residues are shown in red. Blue (+) indicate where the two residues maintain similar characteristics and mutation between the two is unlikely to have a 'knock-on' effect.

As the number of residues of each protein hit were readily available, the region of the protein in which the insertion was located could be easily derived. This analysis was carried out for all AC hits for each ORF of all *Alu* subfamilies and results were combined. Multiple matches with the same protein hit were classed as a single hit unless there was a distinct shift in the location of the matched region (*i.e.* the matches arose from different *Alu* monomers) between subfamily, ORF or protein isoform. Taking this into account, a total of 75 different insertions were observed due to duplicated hits, despite only 68 different AC isoforms being identified in earlier stages of analysis.

2.2.2 Analysis of protein alignments to determine insertion site preference

For all 75 AC isoforms, the mid-point of the *Alu* insertion was calculated by taking the average of the start and end point of the insertion. The total size (number of residues) was obtained from the NCBI database and a ratio was calculated (table 2.2). Values lying below 0.199 (blue) were classed as N-terminal, those lying between 0.200 and 0.799 (green) were classed as internal and those above 0.800 (orange) were classed as C-terminal.

Protein	Isoform	Size AA	*Alu*J	
			Mid-point	Ratio
NEK4	1	841	479	0.57
	3	752	390	0.52
YS049	1	238	132	0.55
PPP5D1	1	171	145	0.85
TMM78	1	136	119	0.88
ZMAT1	1	638	25	0.04
ZN195	1	629	94.5	0.15
	5	606	94.5	0.16
	6	610	98.5	016
ASCC1	1	400	367	0.92
M4K1	1	833	814.5	0.98
SGT1	1	365	126	0.35
ZN701	1	531	19.5	0.04
ZN415	1	603	77	0.13
	2	567	40.5	0.07

Table 2.2 Example of ten proteins matching a single ORF of *Alu*J

All *Alu*-containing isoforms were analysed and ratios were calculated using the total sequence length and insertion mid-point. N-terminal, internal and C-terminal insertions are been colour-coded as blue, green and orange, respectively. In all the cases shown here, different isoforms of the same protein have the same insertion region.

In all cases, we see that different isoforms of the same protein contain *Alu* insertions that lie in the same region. For example, for the three AC isoforms of ZN195, all insertions are N-terminal. The calculated ratios differ in value due to a difference in total length of the isoform. It is expected that this would be the case as multiple AC isoforms of the same protein are translated from mRNA containing the same *Alu* insertion and differ due to alternate splicing elsewhere in the protein.

Combined results and deletion of duplicate hits resulted in a total of 75 different insertions. Calculated ratios were used to make a histogram (figure 2.4A). If no bias for insertion between the three defined regions; N-terminal, internal and C-terminal, was present, an even distribution of insertions would be observed throughout proteins. For the 75 studied insertions, 7 – 8 insertions would be predicted to arise in each 10% of the protein. Scaled up, 45 internal insertions would be expected, and 15 insertions would be predicted to arise at each of the N- and C-termini.

However, results reveal a distinct bias towards *Alu* insertions at protein termini, with a total of 29 and 23 insertions the N- and C-termini, respectively. This is equal to approximately 70% of the studied insertions. In the case of N-terminal insertions,

this is double the amount of insertions predicted for non-biased insertion. There also appears to be a small bias for insertions at the N-termini over the C-termini, but this is less prominent when studying individual subfamilies (figure 2.4B).

When studying insertion locations within individual subfamilies, again, an obvious bias towards terminal insertions is observed. There appears to be a preference for N-terminal insertions in older subfamilies (J, Sx, Sp, Sq and Sc); however, this seems to shift towards the C-terminus in younger subfamilies (Sb, Sb1 and Y). It should be noted that younger subfamilies had fewer protein hits than evolutionary older subfamilies, so it is unclear whether this observation is due to a true preference for C- over N- terminal insertions or whether it is due to analysis of limited data in these subfamilies.

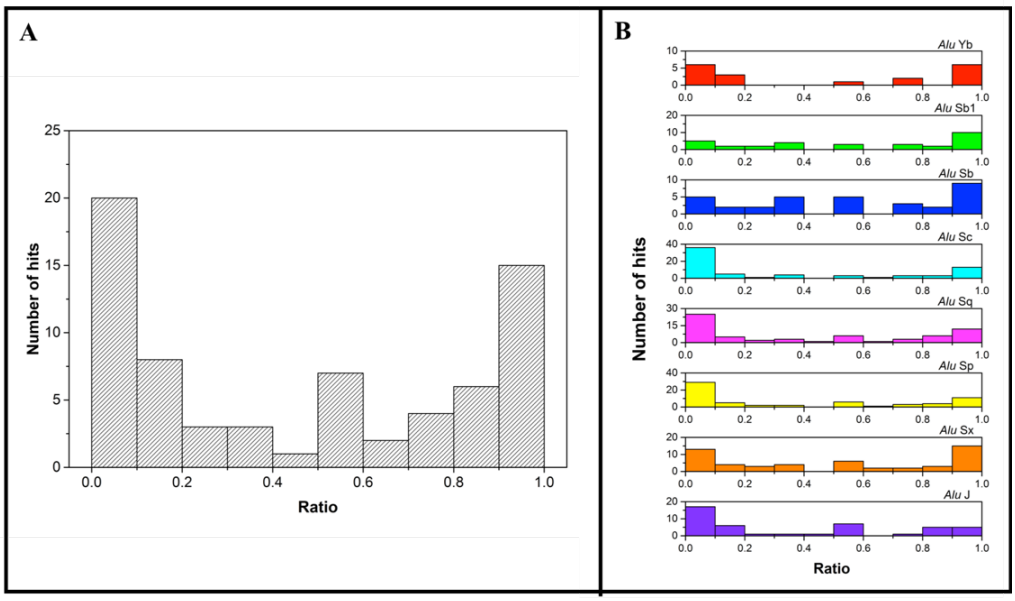


Figure 2.4 Histograms outlining the number of insertions found in different protein locations

(A) Combined insertions from all *Alu* subfamilies. (B) Histograms for insertions in individual subfamilies. If no insertion bias occurred, an even distribution of insertions across protein regions was expected. However, a lower number of insertions are observed at internal regions (0.2 – 0.8) and higher insertion numbers arise at each of the protein termini; N (0.0 – 0.2) and C (0.8 – 1.0).

2.2.3 Determining the locations of *Alu* insertions in protein secondary structure

Using data available from the RSCB Protein Data Bank (PDB)[121], the secondary structures of protein hits were examined and the site of *Alu* insertions within these structures was determined. For proteins which did not have structures available in the PDB, the Phyre^2 web portal[122] was used to predict the secondary structure of protein hits. Exceptions to this are the structures of *Alu* and non-*Alu* isoforms of ASCC1, BCAS4, NEK4 and ZMAT1 which were predicted using the I-TASSER[123] server (structures shown in Chapter 4). Due to the large number of hits and the high demand for this server, not all structures could be predicted this way. By identifying the locations of insertions in well predicted *Alu*-containing protein isoforms, or if a suitable prediction could not be obtained for the AC isoform, identifying where the insertion would lie by looking at a well-predicted (high confidence) non-*Alu*-containing isoform, the tolerance of *Alu* insertions in secondary structure elements such as α-helices, transmembrane helices or β-strands could be predicted. A summary of the predicted secondary structure of *Alu* insertions and *Alu* insertion sites within protein hits in shown in table 2.3.

Protein	Matched Region	Length of Insertion (AA)	Secondary Character of *Alu* Insertion	Secondary Character of Insertion Site	Source	AC/nAC
MOST1	68 – 98	31	α-helix (15)	α-helix	Phyre[2]	AC
NANGN	6 – 20	15	α-helix (11)	N/A	Phyre[2]	AC
ZN714	536 – 554	19	N/A	Coil	Phyre[2]	AC
PKP2	472 – 492	21	α-helix (19)	α-helix	PDB (ID: 3TT9)	AC
NEK4	456 – 502	47	α-helix (6), β-strand (2)	α-helix	I-TASSER	AC/nAC
YS049	94 – 179	86	β-strand (11)	Coil	Phyre[2]	AC
PPP5D1	119 – 171	53	α-helix (2)	Coil	Phyre[2]	AC
TMM78	102 – 136	35	β-strand (7)	α-helix	Phyre[2]	AC
ZMAT1	8 – 42	35	α-helix (9)	Coil	I-TASSER	AC/nAC
ZN195	76 – 113	37	N/A	Coil	Phyre[2]	nAC
ASCC1	347 - 387	41	α-helix (7), β-strand (2)	α-helix	I-TASSER	AC/nAC
M4K1	797 – 832	36	α-helix (3)	N/A	Phyre[2]	AC nAC
SGT1	110 – 142	33	α-helix (18)	Coil	Phyre[2]	AC
ZN701	2 – 37	36	N/A	Coil	Phyre[2]	AC
ZN415	61 – 93	33	N/A	N/A	Phyre[2]	nAC
PACRG	210 – 240	31	α-helix (5)	Coil	Phyre[2]	AC
UBP19	42 – 79	38	α-helix (4), β-strand (2)	N/A	Phyre[2]	AC
BCAS4	164 – 195	32	α-helix (4)	α-helix	I-TASSER	AC/nAC
REL	308 – 339	32	N/A	N/A	Phyre[2]	nAC
GLYG2	3 – 33	31	N/A	Coil	Phyre[2]	AC
CCNJL	99 – 122	24	α-helix (11)	α-helix	Phyre[2]	AC

Protein	Matched Region	Length of Insertion (AA)	Secondary Character of *Alu* Insertion	Secondary Character of Insertion Site	Source	AC/nAC
MKNK1	189 – 210	22	α-helix (1)	α-helix	PDB (ID: 2HW6)	AC
RABX5	377 – 409	33	N/A	N/A	PDB (ID: 4N3Z)	AC
MYL10	2 – 27	26	α-helix (6)	Coil	Phyre[2]	AC
AKD1A	495 – 522	28	α-helix (15)	Coil	Phyre[2]	AC
ZN283	3 – 40	38	N/A	Coil	Phyre[2]	AC
BEND2	80 – 126	47	N/A	N/A	Phyre[2]	AC
CP089	320 – 373	54	α-helix (6)	Coil	Phyre[2]	AC
OR1FC	308 – 328	21	N/A	Coil	Phyre[2]	AC
ZN429	646 – 667	22	N/A	Coil	Phyre[2]	AC
FTM	456 – 502	47	α-helix (45)	α-helix	Phyre[2]	AC
CBPC3	704 – 744	45	β-strand (6)	α-helix	Phyre[2]	AC/nAC
ITCH	162 – 204	43	N/A	Coil	Phyre[2]	AC
F193A	1172 – 1208	37	α-helix (3)	N/A	Phyre[2]	AC/nAC
CNTLN	1373 – 1396	24	α-helix (11)	α-helix	Phyre[2]	AC
TV23C	159 – 187	29	β-strand (4), TM helix (1)	α-helix	Phyre[2]	AC
DSCR8	28-49	22	N/A	α-helix	Phyre[2]	AC
HS905	2 – 19	18	β-strand (3)	Coil	Phyre[2]	AC
FXL18	755 – 784	30	β-strand (3)	β-strand	Phyre[2]	AC
MAGI3	363 – 387	25	N/A	Coil	Phyre[2]	AC
KANK3	794 – 840	47	α-helix (7), β-strand (3)	Coil	Phyre[2]	AC
GVQW1	100 – 151	52	β-strand (3)	N/A	Phyre[2]	AC
NEK5	530 – 549	20	α-helix (13)	N/A	Phyre[2]	AC
RGS3	2 – 27	26	α-helix (5)	Coil	Phyre[2]	AC

Protein	Matched Region	Length of Insertion (AA)	Secondary Character of *Alu* Insertion	Secondary Character of Insertion Site	Source	AC/nAC
GLOD4	38 – 49	12	N/A	Coil	PDB (ID: 3ZI1)	AC
LMO7D	68 - 105	38	β-strand (8)	N/A	Phyre[2]	AC

Table 2.3 Determination of *Alu* insertion sites within protein secondary structure

Secondary structures of proteins were obtained from the RSCB Protein Data Bank, or if no structure was available, secondary structures were predicted using Phyre[2] or I-TASSER. If a suitable prediction of the *Alu*-containing (AC) isoform could not be obtained, then the non-*Alu*-containing (nAC) isoform was used and assessed according to where the *Alu* insertion would arise. N/A is written for cases where secondary structure could not be obtained of was of low confidence. Results for the secondary character of *Alu* insertions are followed by the number of amino acids (AA) attributed to the secondary structure listed with high confidence. Unlisted residues are attributed to coiled/unstructured regions. The secondary character of the insertion site refers to the secondary structure present in the protein hit directly before the first residue of the *Alu* insertion. Note: TM helix refers to a transmembrane helix.

Alu insertions, on a whole, did not appear to have any obvious secondary structure specific to the insertion as a mix of α-helices and β-strands were observed in predicted structures. In many cases, any predicted secondary structure was part of a much larger coiled region and hence, had no distinct secondary structural elements. Additionally, in cases where a high confidence secondary structure could be obtained, *Alu* insertions seemed to occur in either coiled regions or α-helices. No *Alu* insertions were observed to be inserted in known functional/structural motifs (e.g. zinc fingers, active sites). This indicated that *Alu* insertions are generally well tolerated and may potentially mould to any secondary structure present in the protein, as long as the insertion site lies away from any well-established functional or structural motifs. This tolerance may account for the number of hits observed during these analyses.

Note: This data analysis was performed at the end of the PhD project and therefore, occurred after expression of the MBP-*Alu* mutants. As a result, it was not used to decide upon the insertion sites chosen in MBP. A retrospective rationale and discussion is given in Chapter 5.

2.3 Determining the origin of *Alu* insertions

In all of the observed matches, only partial insertions of *Alu* elements are observed. This is likely due to the introduction of an alternative splice site (discussed further in section 2.5). The aim of this analysis was to determine whether a conserved region of *Alu* elements was being inserted into protein-coding regions, through the study of insertions at the nucleotide level.

For the purpose of this analysis, the term *Alu* domain is used to refer to either the *Alu* left arm or the *Alu* right arm. The *Alu* left arm contains an RNA polymerase III promoter region, whereas the *Alu* right arm contains an approximately 34 bp sequence[§] unique to each *Alu* subfamily. At the time of this analysis, a new database of *Alu* sequences had been made available by Dfam.[124] This database allowed access to a total of 37 different *Alu* consensus sequences as opposed to the original eight available in our earlier analysis.

[§]This value differs from that of the 31 bp that is quoted in most literature. Alignment of the Dfam database of *Alu* consensus sequences (containing more sequences from younger *Alu* subfamilies than previously analysed) revealed longer 'inserts' averaging approximately 34 bp in length.

2.3.1 Re-alignment of hits with Dfam consensus sequences

As discussed in section 2.2.1, AC protein isoforms were directly aligned with their corresponding *Alu* ORF. Previously, focus was on the location of the protein in which the *Alu* insertion was present. Here, focus lay on the *Alu* insertion itself, and as such, *Alu* insertions were studied at the nucleotide level as well as the protein level.

Due to expansion of data from eight *Alu* consensus sequences to the 37 available from Dfam, *Alu* insertions were re-categorised according to the subfamily from which they arose (*i.e.* according to parental *Alu*). The Dfam search tool allows for direct alignment of *Alu* insertion sequences with the Dfam database of transposable elements, and as a result the most likely parental *Alu* sequence for each hit could be determined. This allowed to further analysis of the sequences which may have led to *Alu* exonisation as well as better analysis of *Alu* insertions at the nucleotide level. A ratio of the insertion mid-point and the full length *Alu* from which the insertion originated could then be calculated. Through knowledge of the parental strand orientation and mid-point, it could be determined whether each insertion arose from the *Alu* left or right arm (figure 2.4).

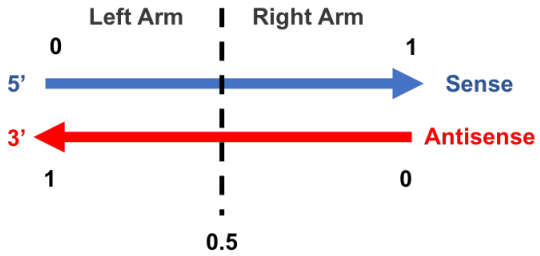


Figure 2.5 Determination of *Alu* left and right arm insertions

Ratios were calculated using the insertion mid-point and the length of the full *Alu* consensus. Sense and antisense insertions with a calculated value of < 0.5 and > 0.5, respectively, arise from the *Alu* left arm. In contrast, sense and antisense insertions of > 0.5 and < 0.5, respectively, arise from the *Alu* right arm.

Searches with the Dfam database resulted in a loss of five hits. RABX5, CBPC3, TV23C and MAGI3 yielded no match with the Dfam database of *Alu* sequences. This is likely due to the different parameters used in Dfam alignments (E-value < 1 x 10^5 and sequence identity > 75%). CNTLN was also discarded from the list of hits as a nucleotide sequence for the gene was not available from the NCBI database and therefore, the gene could not be analysed. As a result, a total of 41 genes were studied.

2.3.2 Insertion bias towards the *Alu* left arm

In previous analysis of insertions, additional hits were observed due to changes in ORF and protein isoform. However, in this case, there are no additional hits as insertions are being studied at the nucleotide level. As changes in ORF and isoforms rely on translation of the nucleotide sequence, they are irrelevant to this stage of analysis.

UBP19, MKNK1 and MYL10 did not directly match with any *Alu* subfamily. Instead they matched with the FLAM (free left *Alu* monomer) sequence; an evolutionary precursor to the *Alu* subfamilies known today. As a result, no ratios were calculated, and they were categorised as left arm insertions.

Database results showed a strong bias towards left arm *Alu* insertions in the protein-coding regions of hits (figure 2.6). Of a total of 41 *Alu* insertions, 36 arose from the *Alu* left arm, or FLAM. It was also observed that insertions tended to arise from the antisense strand of the parental *Alu* template, amounting to 34 of the total 41 insertions. Although 17% of insertions did arise from transcription of the *Alu* sense strand, they all corresponded to the *Alu* left arm.

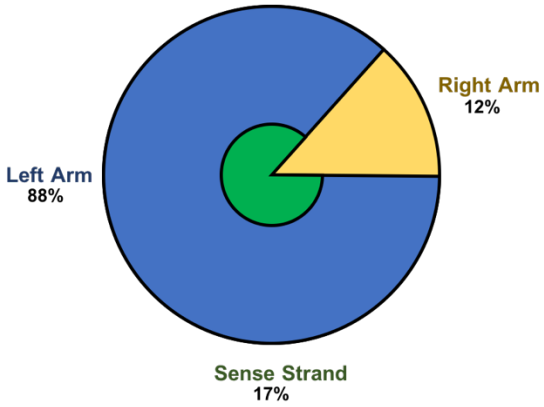


Figure 2.6 Analysis of *Alu* left arm and right arm insertions
88% of insertions arose from the *Alu* left arm revealing a distinct bias over insertions from the *Alu* right arm. 83% of insertions were observed to match the antisense strand of the parental *Alu*. Though 17% of insertions did arise from the sense strand of their parental *Alu*, all corresponded to the *Alu* left arm.

This observation, coupled with the known mechanism for gene splicing, could suggest that a new 5' splice site is introduced in the sense strand resulting in exonisation of the left arm but not the right arm. It could also indicate that though some splicing does occur to include right arm insertions, the efficiency of splicing out right arm-derived pre-mRNA is greater than that of splicing out left arm-derived pre-mRNA. Therefore, it is possible that any splice sites introduced by the *Alu* are well-recognised within the spliceosome. As *Alu* elements have been recognised to be promotors/repressors and have also been known to form *Alu* 'hotspots', it could be theorised that free *Alu* mRNAs within proximity of the splicing event may promote the exonisation of the *Alu* element either through promotion of the *Alu*-introduced splice site, or suppression of those downstream.

2.4 Identifying a conserved *Alu* insertion sequence in proteins

As already discussed, 80% sequence similarity between *Alu* consensus sequences results in the translation of very similar ORFs. In addition to this, it was observed that the majority (88%) of *Alu* insertions correspond to translation of the *Alu* left arm and 83% arise from sequences copied from the antisense strand of the parental *Alu*. As a result, a relatively well conserved *Alu*-encoded sequence in the majority of protein hits would be expected.

It should be noted that during the original search for *Alu*-containing proteins, a large majority of hits corresponded to a single ORF, read in the 3'5' direction. In all cases, this ORF yielded no stop codons, or had fewer stop codons when compared with alternate reading frames. The introduction of stop codons *via* frame shifts, produced shorter sequences which were unlikely to meet the match parameters used in database searches (figure 2.7). As a result, it is unsurprising that the majority of hits arose from a singular ORF which retained its similarity between subfamilies. This ORF was referred to as the primary reading frame.

Figure 2.7 Effect of frame shift on length of interrupted *Alu*J translation

Translation of the antisense strand of double stranded DNA element, *Alu*J, in the 3'5' direction gives three possible open reading frames (ORFs). ORF 1 is the primary reading frame and is uninterrupted by STOP codons. Frame shifts in ORF 2 and 3 result in the introduction of STOP codons, leading to shorter sequences which are less likely to be present in protein hits.

2.4.1 Alignment of *Alu* insertions in protein hits

Translated *Alu* insertion sequences were extracted from protein hits through application of matched region start and end points to full isoform sequences using ExPASy ProtParam. As *Alu* insertions in different isoforms of the same protein resulted in the same insertion sequence, only one *Alu* sequence was used per protein. This avoided biased alignment through identical repeats of the same sequence. A total of 46 insertion sequences were aligned.

Aligned sequences were subjected to JackHMMER[125] analysis and a model position diagram was generated highlighting the conservation of residues between sequences.

Model position diagrams show colour-coded residues in an alignment in which their size is dependent upon their conservation between sequences. Larger single-letter amino acid codes represent more conserved residues. From the model position generated using all *Alu* insertions sequences (figure 2.8), a conservation of residues in four regions (boxed) begins to become clear.

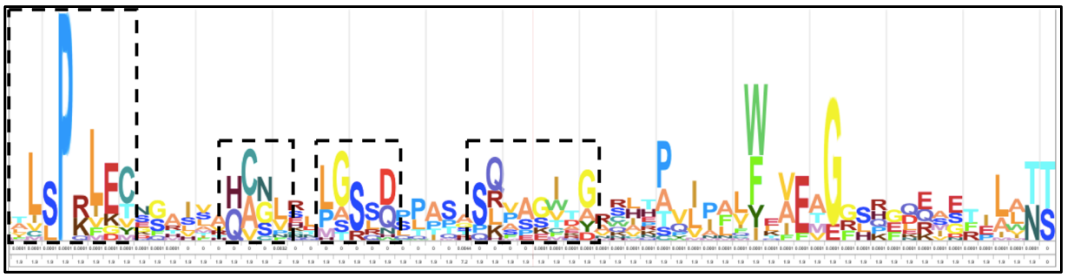


Figure 2.8 Model position for all *Alu* insertion sequences

Amino acids are represented by colour-coded single-letter amino acid codes. The larger the single-letter code, the more conserved the residue. At this stage of analysis, some conserved residues were clearly observed (boxed). The longer sequence shown in this model, which extends beyond the fourth boxed region, is due to the presence of longer insertion sequences in a low number of hits.

In order to try to identify a more conserved sequence, analysis was refined to study only the *Alu* insertion sequences from proteins found to match primary reading frames. Refinement gave rise to 26 insertion sequences. The generated model position (figure 2.9) shows a much more clear-cut conservation of residues between sequences.

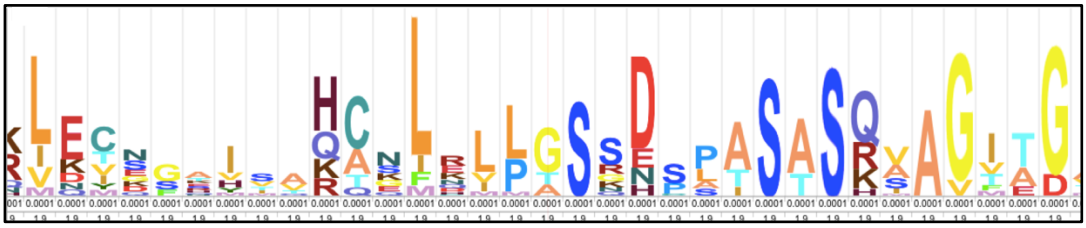


Figure 2.9 Model position for insertion sequences arising from proteins matching the primary reading frames of *Alu* subfamilies

Alignment of 26 insertion sequences results in a model position in which clear residue conservation can be observed. From this data, a conserved sequence of LEC-X_1-GAISAHCNLRLLGSSD-X_2-PASASQ-X_3-AGITG can be observed, where X_1, X_2 and X_3 could be N/S, S/P and V/A, respectively.

From these analyses the following conserved sequence was defined:

LEC-X_1–GAISAHCNLRLLGSSD-X_2–PASASQ-X_3-AGITG

where X_1 = N/S, X_2 = S/P and X_3 = V/A. Conserved residues are observed in the model position shown in figure 2.8 which lay beyond the 33 amino acid sequence identified in figure 2.8. These arise from the presence of longer sequence matches in a low number of protein hits which have been included in the model position generated through JackHMMER, but these additional residues are not present in the majority. The average insertion length, calculated from 46 insertion sequences, was calculated to be 34 amino acids. As a result, the conserved sequence was limited to 33 amino acids which was representative of the majority of protein hits.

From this, it was concluded that *Alu* insertions in our identified genes lead to the translation of a relatively well conserved insertion sequence in their expressed proteins. This is unsurprising due to the similarity between *Alu* subfamilies and agrees with previous conclusions of insertion origin. Direct alignment of this sequence with hits (figure 2.10) confirmed sequence conservation.

Conserved	**L**	**E**	**C**	**X**	**G**	**A**	**I**	**S**	**A**	**H**	**C**	**N**	**L**	**R**	**L**	**L**	**G**	**S**	**S**	**D**	**X**	**P**	**A**	**S**	**A**	**S**	**Q**	**X**	**A**	**G**	**I**	**T**	**G**
ASCC1	L	E	Y	N	D	A	I	S	A	H	C	N	L	C	L	P	G	S	S	D	S	P	A	S	A	S	Q	V	A	G	I	T	G
ZMAT1	L	E	C	S	G	A	I	S	A	H	C	S	L	H	L	P	G	S	S	D	S	P	A	S	A	S	Q	I	A	G	T	T	D
M4K1	L	E	C	S	G	T	I	S	P	H	C	N	L	L	L	P	G	S	S	N	S	P	A	S	A	S	R	V	A	G	I	T	G
NEK4	L	E	C	S	G	T	I	L	A	H	S	N	L	R	L	L	G	S	S	D	S	P	A	S	A	S	R	V	A	G	I	T	G
BCAS4	V	E	C	S	G	T	I	P	A	R	C	N	L	R	L	P	G	S	S	D	S	P	A	S	A	S	Q	V	A	G	I	T	G

Figure 2.10 Alignment of proposed conserved sequence with hits

Alignment of the proposed conserved sequence with five different hits confirmed conservation. For the proposed 33 residue sequence, 52% of residues are fully conserved (yellow) upon alignment with the above 5 proteins. Another 42% of residues are relatively well conserved (green) in that they can be one of two residues with respect to each site.

2.5 Identification of 3' splice sites leading to *Alu* exonisation

In 2003, Lev-Maor *et al* published that 'proximal' and 'distal' AG base doublets in *Alu* elements led to 3' splicing. Their study examined a dataset of exonised *Alu* elements in the human genome which led to the identification of 3' splice sites at positions 279 (proximal) and 275 (distal)[**]. It was proposed that *Alu* insertions into protein-coding genes may lead to the introduction of the same splice sites allowing for alternative splicing and as a result, providing an explanation for the formation of the alternate protein isoforms identified in early bioinformatic analyses. Observation of 3' splice sites would be concurrent with evidence for the conservation of sequences copied from the antisense *Alu* strand resulting in translation of the *Alu* left arm.

2.5.1 Identification of 3' splice sites in Dfam *Alu* subfamilies

The reverse complement of each Dfam *Alu* consensus sequence was obtained, giving the sequences for the antisense DNA strand of each *Alu* subfamily. Antisense sequences were aligned and compared to those of Lev-Maor *et al* to confirm the reported 3' AG splice sites (figure 2.10).

[**]Positions were numbered according to the alignments performed by Lev-Maor *et al* and differ slightly from those used in this research. Note that the same proximal and distal splice sites are being referenced despite slightly different numbering.

Splice site positions differ slightly from those reported; however, this is due to a difference in numbering due to different sequence lengths used in alignments and does not represent a different series of splice sites. Proximal (279) and distal (275) splice sites will from here on be numbered as 290 and 286, respectively. The proximal and distal splice sites observed match those reported by Lev-Maor. A shift of distal splice sites from position 286 to position 284 between *Alu*J and *Alu*S/*Alu*Y subfamilies is also observed. This is likely due to an A/G mutation at position 288 and is consistent with the observations reported by Lev-Maor.

					Proximal						Distal		
Alu	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1			
subfamily	296	295	294	293	292	291	290	289	288	287	286	285	284
Jo	-	-	T	G	A	G	A	C	A	G	G	G	T
Jb	-	-	T	G	A	G	A	C	A	G	G	G	T
Jr	T	T	T	G	A	G	A	C	A	G	G	G	T
Jr4	T	T	T	G	A	G	A	C	A	G	G	G	T
Sc	-	-	T	G	A	G	A	C	G	G	A	G	T
Sg	-	-	T	G	A	G	A	C	G	G	A	G	T
Sp	-	-	T	G	A	G	A	C	G	G	A	G	T
Sq	-	-	T	G	A	G	A	C	G	G	A	G	T
Sq10	T	T	T	G	A	G	A	C	G	A	A	G	T
Sx	-	-	T	G	A	G	A	C	G	G	A	G	T
Sz	-	-	T	G	A	G	A	C	G	G	A	G	T
Sz6	-	-	T	G	A	G	A	C	A	G	A	G	T
Y	-	-	T	G	A	G	A	C	G	G	A	G	T
Yc	-	-	T	G	A	G	A	C	G	G	A	G	T
Yk12	-	-	T	G	A	G	A	C	G	G	A	G	T

Figure 2.11 Proximal and distal 3' AG splice sites in Dfam *Alu* consensus sequences. The proximal splice site at 290 (red) and distal splice site at 286 (blue) match those reported by Lev-Maor (279 and 275). The distinct shift in distal splice site between *Alu*J and *Alu*S subfamilies, also reported, due to an A/G mutation at position 288 (yellow) and G/A mutation at 286, can also be observed. This shift continues throughout *Alu*Y subfamilies due to the same mutation. Note that not all aligned sequences are shown but were aligned. Shown subfamilies act as a representative sample. Numbers (-1 to -10) refer to distance from the distal splice site.

2.5.2 Analysis of *Alu*-derived 3' AG splice sites in protein-coding genes

Previous research into insertion bias (section 2.3) identified 35 insertions copied from the antisense *Alu* strand in protein-coding genes. Nucleotide sequences for genes containing these insertions were aligned with the consensus sequence of their parent *Alu*. 12 antisense insertions were found to contain the same 3' AG splice sites reported by Lev-Maor (figure 2.11).

ITCH, showed a range of mutations which resulted in none of the predicted proximal nor distal splice sites being present, instead a possible splice site at position 293 was observed. As seen previously, A/G mutations at position 288 lead to a shift in distal AG to position 284. This shift can be observed in 7 of the 13 sequences.

					Proximal					Distal			
	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1			
	296	295	294	293	292	291	290	289	288	287	286	285	284
AluJ	T	T	T	G	A	G	A	C	A	G	G	G	T
AluS	-	-	T	G	A	G	A	C	G	G	A	G	T
AluY	-	-	T	G	A	G	A	C	G	G	A	G	T
YS049	T	T	T	G	A	A	A	T	G	G	A	G	T
PPP5D1	C	C	C	C	T	A	C	C	A	G	G	G	T
TMM78	C	C	T	G	A	G	A	C	A	A	G	G	T
ZMAT1	G	A	G	A	A	G	A	T	G	G	A	G	T
ASCC1	T	A	T	T	T	T	A	A	A	G	A	G	C
CCNJL	T	T	G	C	A	A	A	C	G	G	A	G	T
CP089	C	A	A	T	T	T	C	C	A	G	A	T	T
ITCH	C	A	G	A	A	A	A	T	G	G	G	G	T
F193A	C	A	A	T	T	G	A	T	G	G	A	G	T
FXL18	T	T	T	G	A	G	A	C	A	G	A	G	T
GVQW1	T	T	T	C	A	G	A	T	G	G	A	G	T
NEK5	G	T	G	C	A	G	A	T	G	G	A	G	T
GLOD4	A	T	G	A	A	G	G	T	G	G	A	G	T

Figure 2.12 Proximal and distal 3' AG splice sites in antisense *Alu* insertions.

Sequences for *Alu*J, *Alu*S and *Alu*Y are representative of multiple subfamilies. 13 of 35 identified antisense insertions were found to have proximal (red) and/or distal (blue) 3' AG splice sites are observed to match those previously reported. As before, a mutation at position 288 (yellow) results a shift in distal AG. Several mutations in ITCH led to the observation of a potential 3' splice site at position 293 (grey).

Lev-Maor stated that if the distal AG is less than 4 nucleotides away from a proximal AG then preference is towards splicing at the distal AG, as is the case for FXL18. In cases such as those of YS049, PPP5D1, ASCC1, CCNJL, CP089 and F193A, a mutation that removes the proximal AG leads to splicing at the remaining distal site. Lev-Maor also stated that if the gap between proximal and distal AG is 6 or more nucleotides, then the proximal AG splice site is preferred, as is the case where A/G

mutations and position 288 have occurred. Exceptions to this arise when mutation has removed the proximal AG splice site. In the case of TMM78, where the distal AG is observed at 285, it is difficult to predict whether splicing would be proximal or distal. It is reported that splicing has a preference for the following triplets CAG > TAG > AAG > GAG.[126] This would lead to the prediction that the distal splice site would be preferred, especially as a preference for a distance of five nucleotides between proximal and distal AG splice sites is not reported by Lev-Maor.

2.5.3 Identification of new potential 3' AG splice sites in exonised *Alu* sequences

Although we see 12 exonised *Alu* sequences in genes which match the splice sites reported by Lev-Maor, and one additional sequence likely resulting from mutations around these sites, this accounts for less than half of the observed 35 antisense insertions. During analysis, three other potential 3' AG splices sites were identified downstream of those already identified. Potential splice sites were observed at positions 265, 258 and 116 (figure 2.12).

	270	269	268	267	266	265	264	263	262	261	260	259	258	257	256
AluJ	C	C	C	A	G	G	C	T	G	G	A	G	T	G	C
AluS	C	C	C	A	G	G	C	T	G	G	A	G	T	G	C
AluY	C	C	C	A	G	G	C	T	G	G	A	G	T	G	C
M4K1	C	C	C	A	G	G	C	T	G	G	A	G	T	G	C
ZN415	A	A	C	A	G	G	C	T	G	G	A	G	T	G	C
BCAS4	A	G	G	A	A	C	G	T	G	G	A	G	T	G	C
AKD1A	G	G	C	T	G	G	C	T	G	G	A	G	T	A	C
DSCR8	C	T	A	C	T	G	C	T	G	G	A	G	T	G	C
CASC5	G	G	A	A	G	G	C	T	G	C	A	G	T	G	C

	123	122	121	120	119	118	117	116	115	114	113	112	111	110	109
AluJ	T	T	T	G	T	A	G	A	G	A	C	G	G	G	G
AluS	T	T	A	G	T	A	G	A	G	A	C	G	G	G	G
AluY	T	T	A	G	T	A	G	A	G	A	C	G	G	G	G
NEK4	T	T	A	G	T	A	G	A	G	A	T	G	G	A	G
ZN195	C	A	T	C	C	A	G	A	A	T	G	G	G	A	T
REL	C	A	C	G	T	A	G	A	A	A	C	A	G	G	G

Figure 2.13 Potential 3' AG splice sites in exonisation *Alu* insertions.

Alignment of *Alu* consensus sequences led to the identification of potential 3' AG splice sites at positions 265 (purple), 258 (green) and 116 (orange). It is also possible that splicing could occur at positions 119 (gold) and 110 (pink), though the sites appear to be less conserved and may be due to single base mutations.

A total of six genes were identified to have a potential 3' AG splice site at position 258. Of these six, three appeared to have a potential alternative splice site at position 265. As CAG splice triplets are preferential to GAG triplets, it is likely that the AG at position 265 is preferentially spliced in the case of M4K1 and ZN415. By the same reasoning, that CAG is the preferred triplet for splicing, CASC5 is likely to be spliced at position 258.

Alignments of NEK4, ZN195 and REL seem to point towards splicing much further downstream at around position 116. There appear to be many possible AG splice sites around this region though position 116 appears to be more conserved.

2.5.4 Evidence of the expression of protein hits in human cells

Using data available from the Human Protein Atlas,[127, 128, 129] a brief search of evidence of the expression of identified hits within human cells was conducted. Of the 46 identified hits, 36 were identified to have evidence of expression at the protein level.[130, 131] Of these 36, 33 were identified to have evidence of both *Alu* and non-*Alu* isoform expression at the protein level. MOST1, YS049, UBP19, OR1FC and HS905 had no evidence of expression at either the transcript or protein level. NANGN, TMM78, AKD1A, GVQW1 and LMO7D has evidence of transcript expression, but not protein expression. Overall, most hits identified *via* bioinformatic analyses appear to be viable hits which have both AC and nAC isoforms expressed in human tissues.

2.6 Conclusions from bioinformatic analysis

Sequence alignment of eight *Alu* subfamilies (J, Sx, Sp, Sq, Sc, Sb, Sb1 and Yb), available from the NCBI database, showed 80% sequence conservation between subfamilies. Translation of the eight *Alu* consensus sequences into all six of their possible open reading frames, and subsequent BLAST analysis with a cut-off of > 68% identity match, identified 57 human proteins containing an *Alu*-like insertion. Further refinement through introduction of an additional parameter of E-value $< 1 \times 10^{-8,}$ reduced hits to a total of 46. A number of hits gave rise to multiple *Alu*-containing isoforms, totalling 65 isoforms over the 46 proteins. Of the identified proteins, 32 could be expressed as both AC and nAC isoforms.

Work by Lin *et al* (Genome Biology, 2016) used RNA-seq as a way to identify highly spliced putative coding *Alu* exons. The results obtained from their work

overlapped with the following proteins identified through the bioinformatic analysis used in this project: SGT1, NEK4, ZN415 and ZN195. The overlap between the lists, though minimal, shows that *Alu* exons do appear to be highly spliced in some cases. For both techniques, it is possible that many hits were not identified. However, this is likely due to the limitations of the databases used in each case. Lin *et al* used the PRIDE database, whereas this work used the Swiss-Prot/UniProtKB database. Dependent on the databases used, and the limitations of the chosen database it is possible that many exonised *Alu* sequences have not been identified.

Analysis of *Alu* insertions within protein hits revealed a clear bias towards insertions located at protein termini over those located internally. Non-biased insertion would be expected to result in even distribution of insertion sites throughout protein hits. However, we observed that only 30% of insertions were internal to protein hits and 70% occurred at protein termini. A bias towards N- over C-terminal insertions appeared to be present in older subfamilies (J, Sx, Sp, Sq and Sc) which seemed to slowly move towards a C-terminal preference in younger subfamilies (Sb, Sb1 and Yb). However, a lower number of hits for younger subfamilies may have contributed to this observed shift.

Analysis of the locations of *Alu* insertions within the secondary structures of protein hits using protein modelling software and the Protein Data Bank revealed that the majority of *Alu* insertions arose in coiled regions. Although some were observed to arise in α-helices, few were observed to arise in β-strands, and none of the insertions lay in or nearby defined structural motifs. Modelling of AC isoforms of protein hits revealed that the Alu insertions themselves did not appear to give rise to a definitive secondary structure.

The release of the Dfam database allowed for re-categorisation of insertions through the recognition of their parental *Alu*. Alignment of insertions with this new database, containing 37 *Alu* consensus sequences, revealed that 88% arose from the *Alu* left arm. It was also observed that 83% of insertions matched the antisense (-) strand of the parental *Alu*. Though 17% of hits matched the sense (+) strand of parental *Alu* sequence, it should be noted that all of these lead to left arm insertions. It is therefore likely that insertions copied from the sense strand lead to the introduction of a splice site which prevents the exonisation of the *Alu* right arm. Frame shifts between ORFs result in the introduction of premature STOP codons. As a result, a clear primary reading frame could be observed for each *Alu* subfamily

which gave rise to either no STOP codons, or fewer STOP codons when compared with alternative reading frames. Alignment of insertions arising from primary reading frames identified a conserved sequence between the majority of protein hits:

LECS-X_1-GAISAHCNLRLLGSSD-X_2–PASASQ-X_3-AGITG

The observation of this conserved sequence was in agreement with earlier conclusions that the majority of *Alu* insertions arose from the *Alu* left arm and were copied from the antisense strand of parental *Alu* DNA. Alignment of the proposed conserved sequence with insertion sequences from hits confirmed sequence conservation.

Building upon the work of Lev-Maor, 35 insertion sequences matching the antisense strand of *Alu* DNA were studied for identification of 3' AG splice sites. 12 genes were found to observe the same splices sites as those identified by Lev-Maor, at positions 290 and 286. In addition to these, an additional three potential 3' splice sites were identified at positions 265, 258 and 116.

Analysis of *Alu* insertions in the human genome at both the nucleotide and proteomic level provided insight into the nature of such insertions and identified a conserved sequence that can be observed in the majority of protein hits. This sequence provided a basis for *in-vitro* studies of *Alu* elements.

A brief search of the Human Protein Atlas database revealed 33 of the 46 refined hits had evidence of expression at the protein level in human tissues, for both AC and nAC isoforms.

During the last 20 years, a surge in *Alu* research has been observed as the elements have lost their "junk DNA" status. However, the majority of work which focuses of the study of *Alu* elements remains at the nucleotide level. For example, there is a large amount of work focussed around *Alu* methylation[132, 133, 134] and gene regulation. However, there is relatively little work which studies *Alu* elements at that protein level. In most cases, when *Alu* elements have been studied at the protein level it has been in the context of an insertion in a specific protein and as a result, research has been limited to this context.[135, 136]

The main exception this was the AluGene database (Dagan, *et al*. 2004), which looked at *Alu* elements incorporated within protein-coding genes. It would have been interesting to compare the database discussed in this chapter with the one produced

by AluGene, unfortunately, the AluGene database is no longer accessible. At this time, another database of *Alu*-containing proteins (other than the one produced in this work) is not available.

The database produced from this work provides a platform to study the extent of *Alu* presence at the protein level, an area which appears to be relatively understudied within the field. The further work contained in this chapter then begins to probe the nature of the insertions which, again, contributes a new angle to the research of *Alu* elements. Through identification of a sequence which is well conserved between many *Alu* insertions, it may be easier for others to identify protein regions which have originated from *Alu* elements. This may be beneficial in ruling out functional protein regions as we have observed that *Alu* insertions do not tend to occur in functional domains. By looking at the regions of the parental *Alu* elements from which insertions occur, this work begins to understand why the sequence observed in protein hit is observed. Additionally, by expanding on the work of Lev-Maor and studying the splice sites, further evidence of sequence conservation is observed.

Chapter 3
A potential binding interaction between a translated *Alu* and geldanamycin

Overview

Previous work by Taylor and Dilly used Magic Tag® immobilisation[137], an 'in-house' phage display technique, to identify a potential binding interaction between geldanamycin and a translated *Alu* sequence. As geldanamycin is a bioactive molecule with potent anti-cancer activity,[138] it was hypothesised that the translated *Alu* sequence could show potential for use as a therapeutic target or biomarker. Using the translated *Alu* sequence discussed in Chapter 2, several biophysical techniques were used to study the proposed interaction.

In initial binding experiments, the free *Alu* peptide, SEA-001, was used. Fluorescence intensity was measured using immobilised biotin-geldanamycin and FITC-labelled SEA-001. Fluorescence anisotropy measurements were also obtained using FITC-labelled derivatives of both geldanamycin and SEA-001. Due to limited amounts of each compound, it was determined that more sensitive techniques would be used to further study the potential binding interaction. In SPR studies, biotin-labelled derivatives of SEA-001 and geldanamycin were used. SPR relies on the immobilisation of one of the binding partners and, in this case, the high binding affinity of biotin to a streptavidin-functionalised chip was utilised. In ITC studies, labelled compounds were not required.

Further investigation into the binding interaction was performed using MBP (maltose binding protein) *Alu*-mutants, which allowed for the *Alu* peptide to be constrained in a way that might mimic the peptide displayed on the phage surface in Magic Tag®. Geldanamycin pull-down assays were performed using biotin-labelled geldanamycin immobilised on NeutrAvidin resin.

3.1 Geldanamycin and Magic Tag® immobilisation

3.1.1 Geldanamycin and its prominence in drug discovery

Geldanamycin (GM; figure 3.1) was the first known benzoquinone ansamycin antibiotic.[139, 140] Prior to this, all characterised ansamysin antibiotics tended to be macrocyclic compounds containing an aliphatic ansa bridge with a napthalenic linkage.[141] Geldanamycin is a natural product isolated from the gram-positive bacterium, *Streptomyces hygroscopicus* var. *-geldanus* and *–nova*.[142, 143, 144] GM is well-known for its potent anti-cancer activity.

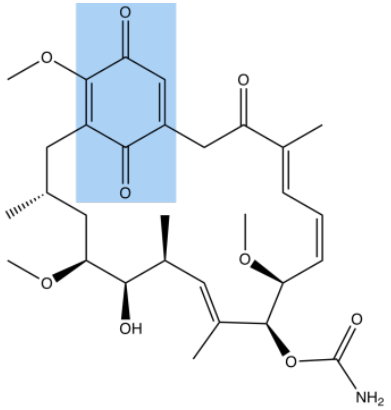


Figure 3.1 Chemical structure of geldanamycin (GM)

Geldanamycin is a benzoquinone (blue) ansomycin antibiotic isolated from *Streptomyces hygroscopius* var. *– geldanus* and *–nova.*

GM was labelled as a 'wonder drug' due to its capacity to target multiple hits[145] and it remains prevalent in current drug discovery programs as a scaffold for GM-derivatives used in cancer treatment.[146] The most noted derivative is 17-AAG, which maintains the potent anti-cancer activity of GM but gives a better toxicity profile.[147, 148] Though GM was originally thought to be a tyrosine kinase inhibitor,[149] it was later discovered that it had no direct effect on the Src kinase family and instead targeted the molecular chaperones in its extracellular environment.[150] It is now proposed that the anti-cancer properties of geldanamycin arise from its ability to inhibit the molecular chaperone HSP90 (heat-shock protein 90), through binding to its ATP-binding site[151] and by extension, restricting its conformational flexibility.[152]

HSP90 is responsible for the stabilisation of 'client' proteins during their folding process through the formation of an HSP90-client complex.[153] This complex protects the client from the ubiquitin proteasome pathway (UPP), which would lead to ubiquitination of the clients and subsequent protein degradation.[154] The HSP90 chaperone recognises hundreds of different types of client proteins[155] including signalling proteins, kinases, viral enzymes and telomerase components.[156] More importantly, it also recognises many oncogenic proteins and as a result, inhibition of oncogenic HSP90- client complexes can result in the depletion of oncogenic proteins to give clinical benefit.[157] HSP90 is highly expressed in several human cancers, including colon, prostate and breast cancer.[158, 159, 160] As geldanamycin is a strong inhibitor of HSP90, it has great importance in cancer drug discovery programs.

Geldanamycin binds within the binding pocket of HSP90 with the benzoquinone group binding at towards the entrance and the ansamycin ring pointing into the binding pocket (figure 3.2).[161]

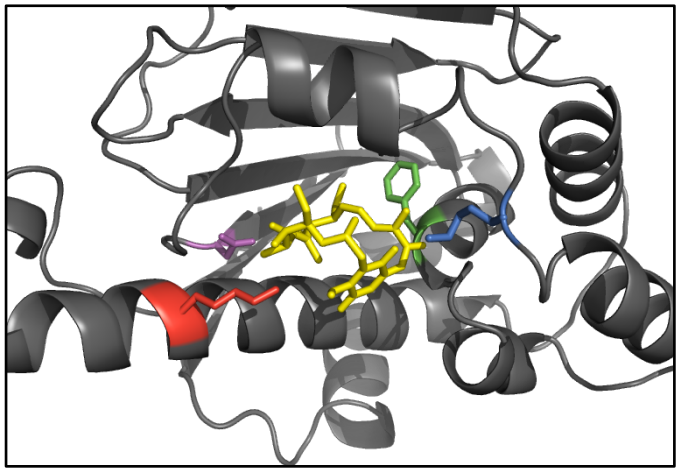


Figure 3.2 Crystal structure of HSP90 bound to geldanamycin

Geldanamycin bound to HSP90. Residues which interact *via* hydrogen bonding are highlighted. D93 (pink) at the base of the binding pocket interacts with the GM carbamate. K58 (red) interacts with both the hydroxy group and the methoxy group of the benzoquinone. K112 (blue) interacts with a carbonyl group of the benzoquinone and F138 interacts with the carbonyl adjacent to the benzoquinone. [PDB entry: 1YET.]

The binding pocket of HSP90 is largely hydrophobic with its hydrophobicity increasing with the depth of the pocket. However, the pocket does contain some polar and charged residues. Complementarity between the pocket residues and the GM allows for the presence of Van der Waal's interactions which aid in binding. There are five hydrogen bonds which contribute to the binding of GM to HSP90. The most important of these is between the carbamate group of geldanamycin and the Asp93 in the bottom of the binding pocket. Towards the centre of the pocket, an additional hydrogen bond is made between the O5 hydroxyl group of GM and the Lys58 residue of HSP90. Three of the five hydrogen bonds are formed at the pocket entrance. The first is formed between the carbonyl of the ansa ring with the backbone amide of Phe138. Additionally, one of the oxygens of the benzoquinone binds to Lys112 and the C29 methoxy group binds with Lys58 (figure 3.3).

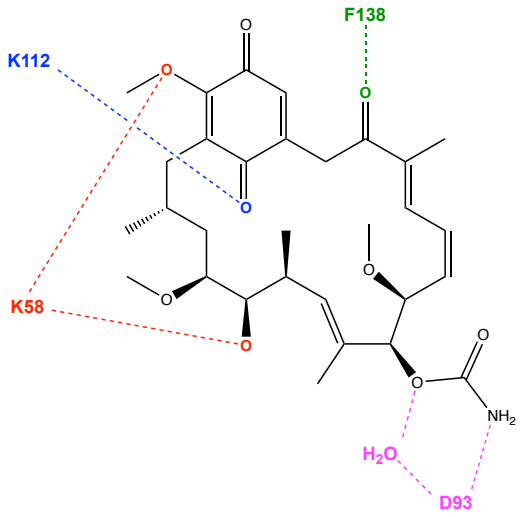


Figure 3.3 HSP90 residues which hydrogen bond to geldanamycin

The most important hydrogen bond is in the base of the HSP90 binding pocket and forms between the Asp93 residue and the carbamate of GM. The Lys58 residue of HSP90 contributes to hydrogen bonds to the methoxy group of the benzoquinone and the hydroxy group. An additional bond is made between the K112 residue and one of the benzoquinone carbonyls. The final bond is formed between Phe138 residue and the carbonyl group adjacent to the benzoquinone.

Though geldanamycin shows potent anti-cancer activity, it is inadequate for use at a clinical level due to problems with hepatotoxicity and solubility.[162] The source of its toxicity lies in the reduction of the GM benzoquinone moiety by NADPH-cytochrome P450 reductase to form a GM semiquinone and superoxide radicals, resulting in oxidative stress.[163]

In addition to its high toxicity, GM also has poor solubility in water, a poor quality in a drug-like molecule. As a result, many drug discovery programs focus on the synthesis of GM analogues with better drug-like properties.

3.1.2 Magic Tag® immobilisation of geldanamycin

Prior to the start of this project, the Magic Tag® chemical genomics tool was developed at the University of Warwick. The results obtained from this tool were not part of this project but did provide a basis for the work carried out. Although it was firstly used to work towards better understanding biochemical pathways in plants,[164] its use has since been extended to the field of biomedical research, more specifically, in the repurposing of medicinal drugs natural products.

Magic Tag® (figure 3.4) used light at a wavelength of 254 nm to photochemically immobilise natural products, in this case geldanamycin, to a pre-derivatised Corning® Stripwell 96-well plate. In previous work, prior to this project, plates were derivatised with one of five tags, of which the scaffolds cannot be discussed. Immobilised natural products were then screened against a T7® Select phage-display library of human lung cDNA,[165] representative of the proteome of the human lung. Phage display allows for the expression of short peptides, in this case, derived from human lung cDNA, to be expressed as a fusion product with phage surface proteins.[166] Hence, small peptides of 1 – 20 residues, encoded by cDNA libraries of choice, can be expressed on the surface on bacteriophages and screened for possible binding interactions.[167] The screen involved three rounds of bio-panning, with amplification in *E. coli* strain BLT5615 after each round. This was followed by PCR and subsequent sequencing to identify hits. The Magic Tag® method minimises non-specific binding through the use of oligo(ethylene glycol) groups which are described as 'protein resistant'.[168] Immobilisation of the natural product aims to maximise the specific interactions observed. The combination of photo-immobilisation and bio-panning has often been known to yield false positives.[169] These were minimised using a quick bioinformatic screen of all six possible open

reading frames of the observed 'hits' against the known protein database (NCBI) *via* protein-BLAST (blast-p). The translation of all reading frames of input cDNA accounts for possible frame slips upon expression upon the phage surface. Sequences that had reasonable similarity with the human proteome and were of suitable length (> 100 amino acids) were deemed as acceptable hits. The hits identified by the bioinformatic screen were re-exposed to the Magic Tag® conditions in which the hit was originally observed. Those which could be selectively eluted were considered to be true positive results. By using a cDNA library only partial sequences from genes are expressed and can be expressed as one of six open reading frames. Therefore, 'hits' identified with this technique could correspond to shortened regions of any reading frame translated from the cDNA.[170] The hit identified by this method was shorter than the 100 amino acid limit of the bioinformatic screen as the sequence retrieved from phages post-screen corresponded to only the first half of the *Alu*S element before sequencing of output DNA became convoluted.

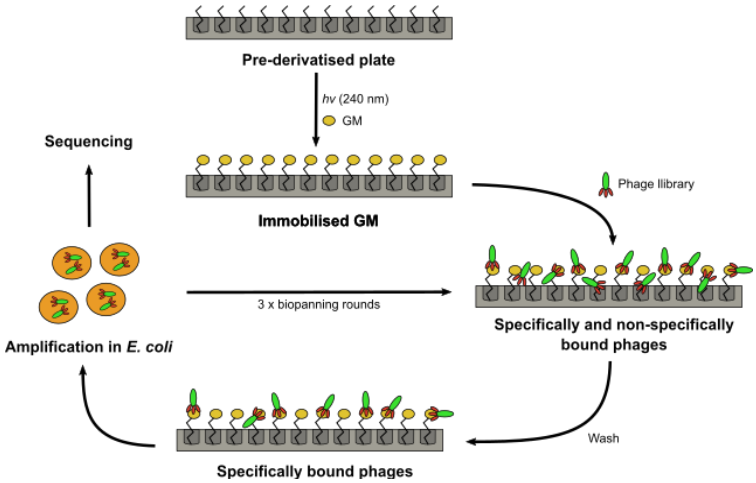


Figure 3.4 Outline of the Magic Tag® chemical genomics tool

Geldanamycin (GM) was photo-immobilised on a pre-derivatised 96-well plate using light at a wavelength of 254 nm. A T7® Select phage library was screened *via* three rounds of biopanning with intermediate wash and amplification steps. Successful binders were identified *via* PCR and subsequent sequencing.

A Magic Tag® screen of immobilised geldanamycin against a human lung library yielded a nucleotide hit corresponding to an *Alu* sequence (unpublished result). Through methods discussed in Chapter 2, a peptide sequence, SEA-001, representative of the translated *Alu* observed to bind geldanamycin using Magic

Tag® was obtained. Comparison between the original Magic Tag® hit and the peptide sequence ascertained from the bioinformatic analysis can be observed below. Residues where the sequences differ are shown in red.

Magic Tag® hit L E C S G A I S A H C K L R L P G S C H S P A S A S R V A G T T G
SEA-001 L E C S G A I S A H C N L R L L G S S D S P A S A S R V A G I T G

For the purpose of this work, SEA-001 was used in binding experiments as a way to represent a conserved sequence which could represent *Alu*-containing protein hits as a whole. As a result, the Magic Tag hit itself was not used in the binding assay. The Magic Tag hit is the translation of a single *Alu* subfamily (S) and therefore is not representative of hits which arise from J and Y subfamilies.

3.2 Using the free *Alu* peptide, SEA-001, to study the observed binding interaction with geldanamycin

The free *Alu* peptide, SEA-001 (shown below), was used in various binding studies in an attempt to re-create and investigate the observed binding interaction with geldanamycin discovered using Magic Tag®. The peptide and its fluorescein isothiocyanate (FITC)- and biotin-labelled derivatives were purchased from Proteogenix. Geldanamycin and its FITC- and biotin-labelled derivatives were purchased form Sigma Aldrich. For both labelled-GM derivatives, the biotin or FITC- label was attached at the C-terminus *via* a PEG linker.

Based on the interactions observed between HSP90 and geldanamycin, it is likely that any strong interactions between the translated *Alu* and GM would be due to hydrogen bonding with the carbonyl groups of, and adjacent to, the benzoquinone or the hydroxyl, methoxy groups of GM. These could be formed by the Asp residue of the peptide. Weaker interaction could be made between hydrophobic residues of the peptide; however, due to the undefined secondary structure of the free peptide, these may be weaker than those observed in a more conformationally locked structure such as HSP90.

SEA-001 H_2N – LECSGTISAHCNLRLPGSSDSPASASRVAGITG – COOH

The interaction between labelled and non-labelled derivatives of both compounds were studied using fluorescence assays, surface plasmon resonance and isothermal titration calorimetry.

3.2.1 Fluorescence studies of the binding interaction between SEA-001 and geldanamycin

Binding interactions were studied using a combination of fluorescence anisotropy and fluorescence intensity experiments. Assays were performed in standard black 96-well polypropylene plates (Greiner Bio-One), and streptavidin-coated black 96-well polypropylene plates (Sigma Aldrich), respectively, on an EnVision Multiplate reader (Perkin Elmer) accessible from the School of Chemistry (University of Leeds). Fluoroscein isothiocyanate (FITC) was used a fluorescent marker giving and excitation wavelength of 490 nm and an emission wavelength of 525 nm.[171]

3.2.1.1 Fluorescence anisotropy

Fluorescence anisotropy (FA), or fluorescence polarisation (FP), is a solution-based technique commonly used in drug discovery programs to study a broad range of molecular interactions.[172] Since its first theoretical description in 1926,[173] fluorescence polarisation has evolved from simple binding isotherms to high-throughput (HTP) screening assays to investigate complex enzymatic activity.[174]

FA involves the excitation of a fluorophore, in this case attached to a ligand, with polarised light, which leads to the subsequent emission of perpendicular polarised light. The term 'anisotropy' refers to the extent of polarisation of the emitted light.[175] FA can be observed due to the presence of emission and absorption transition moments which lie in certain directions within the fluorophore structure.[176] Prior to excitation with polarised light, fluorophores in the ground state are randomly orientated in solution. Upon excitation, fluorophores with transition moments which lie in the same direction as the incident light become selectively excited leading to an excited population which lies in one overall direction. As a result, the emitted light also lies predominantly in one direction, at 90° to the direction of the incident light.[177] The electrical vector of excitation light lies parallel to the z-axis.[178] Emission is measured through a polariser in two direction directions, parallel and perpendicular to the excitation light.[179] Combination of these measurements can be

used to calculate anisotropy (r),[180] using equation 3.1, where $I_{\parallel}$ and $I_{\perp}$ are the intensities of emitted parallel and perpendicular light, respectively.

Equation 3.1 $$r = \frac{I_{\parallel} - I_{\perp}}{I_{\parallel} + 2I_{\perp}}$$

Depolarisation of the fluorophore occurs due to the rotational diffusion of the species in solution. This in turn contributes to the average angular displacement of the fluorophore between absorption and emission, which corresponds to anisotropy.[181] Rotational diffusion is dependent on a number of factors, most notably the size and shape of the rotating species.[182] Generally, a protein bound to a ligand will be larger and will therefore rotate slower in solution than either of the species in the unbound state, leading to a higher fluorescence signal.[183]

3.2.1.2 Fluorescence anisotropy measurements of FITC-labelled geldanamycin and SEA-001

Buffer conditions for fluorescent anisotropy measurements were optimised so as to obtain the best conditions for monitoring the binding of FITC-geldanamycin (FITC-GM) and SEA-001, showing minimal non-specific binding to plates. Optimum conditions were determined by observation of a constant reading of fluorescence intensity of FITC-GM across all wells, whilst maintaining a constant concentration across the plate. The optimised buffer conditions used were phosphate buffered saline (PBS tablets; Sigma Aldrich) with either 0.1 mg/mL bovine serum albumin (BSA) or 0.05% Tween and 1 mM dithiothreitol (DTT). Three repeats of each dilution series were performed in order to minimise the impact of human error on the results.

The Envision Multiplate reader reads the relative amount of light refracted into the detector from the well, in comparison to the amount of light originally directed into the well. Light is measured in two directions to give an overall emission reading. As a result, the reading produced by the equipment must be manually converted into intensity (I) and anisotropy values using equations 3.2 and 3.3, respectively.

Equation 3.2 $$I = 2PG + S$$

Equation 3.3 $$r = \frac{S - PG}{I}$$

Readout data was processed in Microsoft Excel as above; where P is the perpendicular intensity, S is the parallel intensity and G is the instrumental factor (in this case, G=0.8).

In earlier experiments, a 33 nM concentration of FITC-GM was added to all wells. However, this concentration yielded weak intensity signals leading to an increase in concentration to 100 nM in order to obtain more observable readings. A serial dilution series of SEA-001 ranging from 0.95 nM to 500 μM resulted in small incremental increases in anisotropy as a result of increasing peptide concentration (figure 3.5A). However, due to the limited concentration range of peptide used, a binding curve representing only a partial S-shaped curve was obtained, as opposed to the full S-shaped curve usually obtained from successful FA experiments.

In an attempt to obtain a full S-shaped binding curve, a broader concentration range, 1.47 nM to 775 μM, of SEA-001 was used. However, a similar partial S-shaped curve was observed under these conditions (figure 3.5B).

It was determined that in order to obtain a full S-shaped curve for binding between the two species much higher concentrations of peptide would need to be used, indicating weak binding. It may be that binding was hindered by the presence of the FITC label on the geldanamycin. Weak binding may also be attributed to that fact that the peptide is free in solution and therefore, lacks any conformational locks that would have been present when expressed on the surface of the phage, as in the Magic Tag® experiments. It is also possible that there is simply a low binding affinity between the two species.

Due to the cost associated with the purchase of large amounts of peptide, no further fluorescence anisotropy experiments were carried out.

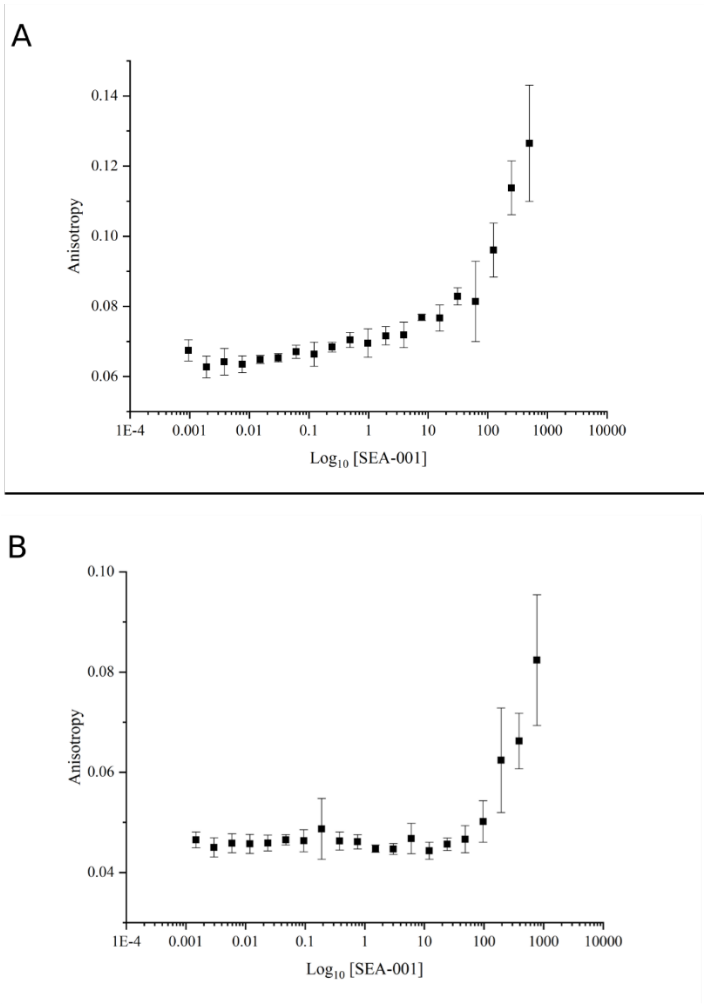


Figure 3.5 Fluorescence anisotropy plots of FITC-GM binding to SEA-001

(A) Addition of a serial dilution of 0.95 – 500 µM of FITC-GM to 100 nM SEA-001 yielded a partial S-shaped curve. (B) A broader concentration range of 1.47 nM – 775 µM of FITC-GM in 100 nM SEA-001 was used in an attempt to obtain more of the curve, however, a similar result was obtained.

3.2.1.3 Fluorescence intensity

Fluorescence intensity experiments rely on the binding of a fluorescently labelled species to another species that is immobilised on the surface of a 96-well plate, and the unbound species being washed away. Therefore, the higher the fluorescence intensity, the better the binding.

The experiments discussed here utilise the strong binding affinity between streptavidin and biotin.[184] Immobilisation was achieved through the binding of biotin-labelled SEA-001 (ProteoGenix) to streptavidin-coated 96-well black

polypropylene plates (Sigma Aldrich). As for the fluorescence anisotropy experiments, FITC-labelled geldanamycin was used.

3.2.1.4 Fluorescence intensity measurements of FITC-labelled geldanamycin and biotinylated SEA-001

Fluorescence intensity experiments were performed to determine whether an increase in fluorescence could be observed in correlation with increasing concentration of FITC-GM. This was achieved by immobilising biotinylated SEA-001 in experimental wells, washing any unbound peptide from the plate, loading FITC-labelled GM into the wells and washing any unbound FITC-GM from the plate. Fluorescence of wells was then measured.

Biotinylated peptide was immobilised on the plate at a concentration of 100 nM. A serial dilution of FITC-GM ranging from 0.24 μM to 1 mM was added. Results from non-specific binding was minimised though the use of relevant controls; a serial dilution of FITC-GM with no bound SEA-001, immobilised SEA-001 with buffer only, and a well containing only buffer. The buffer used was phosphate buffer saline (PBS tablets; Sigma Aldrich) containing 0.05% Tween-20 and 0.05% BSA.

The expected result, based on FA experiments, was that there would be an observable increase in fluorescence with increasing concentration due to higher binding to the immobilised species. However, the results obtained showed no sign of binding affinity between the two species. It is possible that either the FITC-label on the geldanamycin or the biotin-label on SEA-001 hindered binding or, as discussed earlier, the lack of conformational lock in the free peptide cannot recreate that binding interaction observed *via* Magic Tag®.

Fluorescence anisotropy experiments, discussed earlier, predicted only a weak binding interaction between the two species, and as such, the nature of this experiment may not be suitable for its observation.

3.2.1.5 Conclusions from fluorescence experiments with geldanamycin and SEA-001

From the above experiments, it was predicted that the binding interaction between geldanamycin and the free peptide, SEA-001 either did not exist or was too weak to be observed by these means. Increasing concentrations to study it further using these methods would not have been time- or cost-effective. As a result, more sensitive techniques were used to probe the binding interaction.

3.2.2 Studying the binding between SEA-001 and geldanamycin using SPR

Surface plasmon resonance (SPR) was performed on sensor chip SAs (GE Healthcare) which are pre-labelled with streptavidin for the immobilisation of biotinylated species. As discussed below, biotin-labelled variants of both geldanamycin (Sigma Aldrich) and SEA-011 (ProteoGenix) were used as the immobilised species. SPR experiments were performed on a Biacore 3000 SPR system (GE Healthcare) accessible through the Astbury Centre for Structural Molecular Biology (University of Leeds).

3.2.2.1 Surface Plasmon Resonance

Surface plasmon resonance (SPR) is an optical technique commonly used to detect biospecific interactions between proteins and ligands.[185] It relies on the oscillation of mobile electrons, or surface plasmons, on a reflective surface.[186] Polarised light is projected through a prism onto the thin metal surface of a sensor chip, which acts as a mirror and reflects the light and an angle termed the angle of incidence (θ).[187] By changing the angle of incidence and monitoring the intensity of reflected light, the intensity can be observed to pass through a minimum. It is at this minimum that the surface plasmons are excited which causes a dip in reflected intensity, resulting in surface plasmon resonance. The angle which leads to this minimum, or maximum loss of intensity, is termed the resonance angle (θ_{spr}).[188] The resonance angle can be affected by the surrounding system such as medium and temperature, as a result, both must be kept consistent throughout experiments.[189] More importantly, it is affected by changes in refractive index (RI) on the metal surface of the chip (i.e. the accumulation of molecules to the chip surface).[190] Although the RI of the light on the prism side of the system remains unchanged, the RI on the metal surface changes upon the accumulation of adsorbed molecules to the surface, or through the binding of a second species to that which is immobilised. This change in RI directly affects θ_{spr} (figure 3.6).

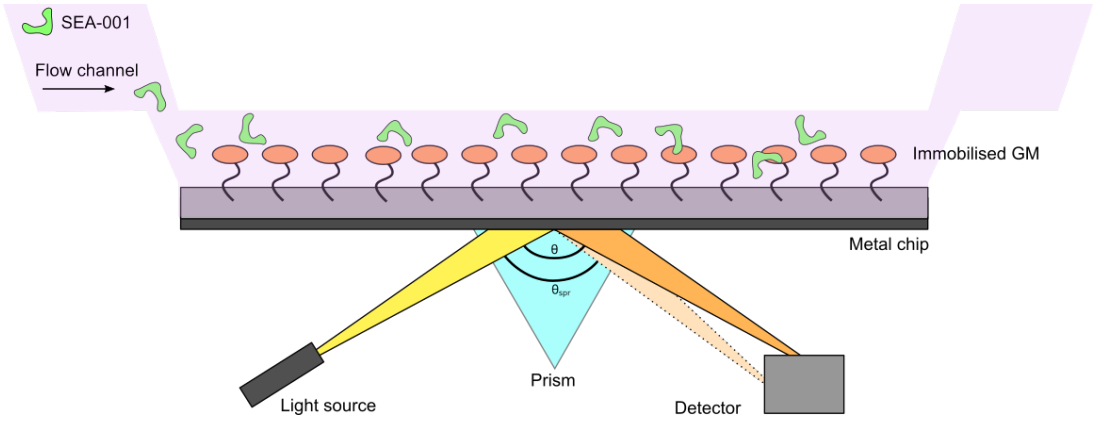


Figure 3.6 Outline of surface plasmon resonance

A flow channel contains protein or peptide (in this case, SEA-001) which is flowed over a chip with immobilised ligand (in this case, GM). Upon binding of protein, the refractive index (RI) of the surface is altered, resulting in a change in the angle of incidence light (θ) to give the resonance angle (θ_{spr}). This change is monitored *via* a photo-detector.

The change in θ_{spr} can be detected by the system using a photo-detection array and measured in resonance units (RU), where 1 RU is equal to an angle change of 1×10^{-4} degrees. Using these measurements, the association constant (k_{on} or k_a) and dissociation constant (k_{off} or k_d) can be obtained using Bioeval 3.0 (Biacore). These can be calculated over a range of concentrations using a dRU/dt versus RU plot, followed by a further plot of k_s versus C using equation 3.4, where k_s is the slope and C is the concentration of ligand (M) in solution.[191]

Equation 3.4 $$k_s = k_{on}C + k_{off}$$

From this second plot, the y-intercept gives a value for k_{off} and the slope gives a value for k_{on}. Using these values, K_D and subsequently, K_A can be calculated using equations 3.5 and 3.6.

Equation 3.5 $$K_D = \frac{k_{off}}{k_{on}}$$

Equation 3.6 $$K_A = \frac{1}{K_D}$$

Affinity can also be expression in terms of Gibbs Free Energy (ΔG^{o}) using equation 3.7, where C^{o} is the standard state concentration. Over a range of temperatures, this can also be used to calculate enthalpy (ΔH) and entropy values (ΔS).

$$Equation\ 3.7 \qquad \Delta G° = ln\frac{K_D}{C°}$$

3.2.2.2 SPR binding studies using immobilised SEA-001

Due to the associated costs of the biotinylated compounds, SPR was first carried out using biotinylated SEA-001 as the immobilised species. This meant that the expected change in RI upon binding of geldanamycin to the peptide would be smaller than if the GM was the immobilised species. Immobilisation of approximately 250 RU of biotinylated SEA-001, gave a chip concentration of 250 pg/mm^2 and a volume concentration of approximately 100 nM, as calculated using equation 3.8, approximating a 100 nm thickness for the pre-existing layering on the chip.[192]

$$Equation\ 3.8 \qquad C_{immobilised}\ (M\ L^{-1}) = \frac{C_{immobilised}\ (RU)}{100\ x\ M_r}$$

Using equation 3.9, this predicts and R_{max} of approximately 40 RU, where R_{max} defines the increase in surface thickness upon binding of 100% of geldanamycin binding to the immobilised SEA-001, taking into account their difference in mass.

$$Equation\ 3.9 \qquad R_{max} = \frac{M_r(free\ species)}{M_r(immobilised\ species)}\ x\ C_{immobilised}\ (RU)$$

Earlier runs of the experiment observed increases of approximately 50 RU when compared to the blank control flow cell. However, closer analysis and subsequent experimental repeats revealed the interference of dimethyl sulfoxide (DMSO), present in the buffer, with the observed signals. Large spikes were produced at points when flow solutions were changed, likely due to a minor mismatch between DMSO concentration between the buffer and geldanamycin solution or the poor solubility of geldanamycin. As a result of these spikes, negative responses were also observed. Though significant efforts were made to match the two concentrations, due to the

sensitive nature of the technique spikes still occurred, rendering the software-predicted association and dissociation curves inadequate for predicting k_{on} and k_{off} rates.

An additional experiment monitoring the effects of a serial dilution of DMSO in buffer also revealed that a response was observed between DMSO and the chip and/or immobilised species. This resulted in signals of up to 25 RU, similar to those observed in experiments with geldanamycin (figure 3.7). As there was no significant difference between the background DMSO signals and the R_{max} of the bound GM, it was determined that immobilisation of SEA-001 onto the flow cell would be insufficient to obtain reliable binding curves. Unfortunately, due to the poor solubility of GM in water, it was also not possible to remove DMSO from the experiment.

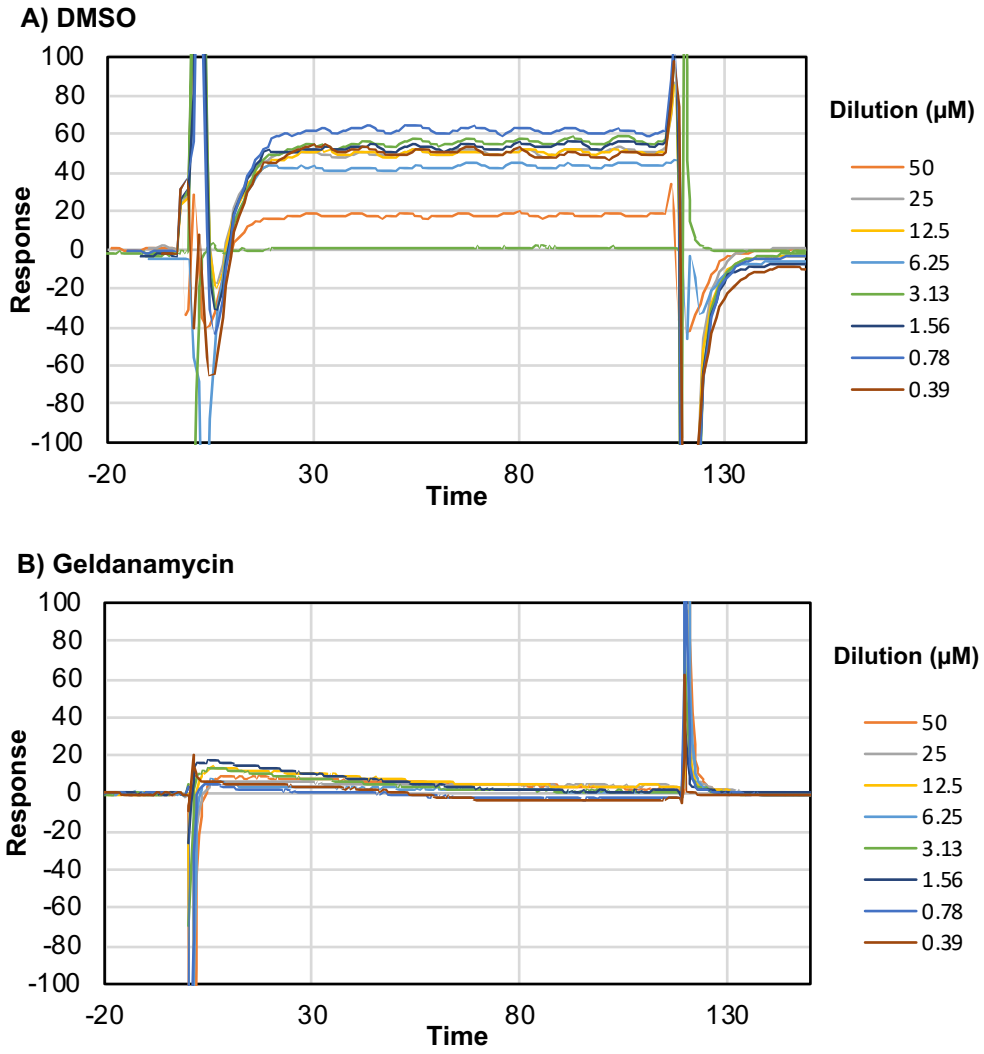


Figure 3.7 Comparison of raw SPR data for DMSO and GM dilution series

(A) Raw SPR data yielded large spikes at points corresponding to a change of flow solution and also yielded a response of up to 25 RU. The signals observed were comparable in response to those observed for geldanamycin (B). Note: the dilution range of DMSO quoted corresponds to the amount matching what would be present in corresponding micromolar concentrations of GM samples.

3.2.2.3 SPR binding studies using immobilised geldanamycin

As there appeared to be a background interference of up to 25 RU from the DMSO present in the buffer and analyte solutions, it was determined that a higher R_{max} would be needed to obtain more reliable results. The easiest way to achieve this, without using excessive amounts of reagents, was to immobilise the geldanamycin on the chip instead of the peptide. As predicted by equation 3.9 (above), the

immobilisation of 100 RU (100 nM) of biotinylated GM predicted an R_{max} of approximately 280 RU, which was expected to be high enough to observed any interaction between peptide and GM despite the background interference of DMSO. Earlier experiments appeared to observe some interaction between the two species, in that initial solutions of lower peptide concentrations appeared to show and incremental increase in RU with increasing concentration. However, as runs continued, there appeared to be no obvious link between the concentration of the peptide and the observed RU, with buffers also yielding signals (figure 3.8).

All curves appeared to follow a similar association and dissociation trend, including those corresponding to buffer only, indicating that the buffer used was not suitable for these experiments and was likely the reason for the observed change in RU. This theory was reinforced by the observation that all curves seemed to lie between 0 and 50 RU, which was significantly lower than the predicted R_{max} for binding between the two species.

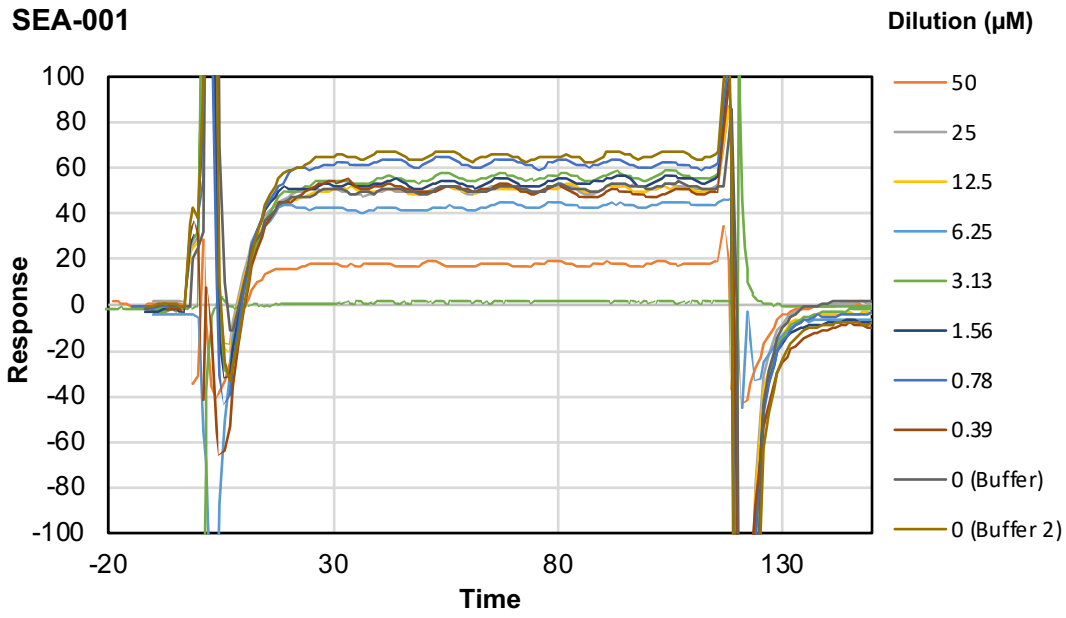


Figure 3.8 Raw SPR data for SEA-001 dilution series

Response signals were observed up to approximately 70 RU. However, increase in response did not relate to an in increase in concentration of SEA-001, with responses also being observed from buffer only samples.

3.2.2.4 Conclusion from SPR experiments with geldanamycin and SEA-001

Throughout all the SPR runs, a very fast off rate (k_{off}) was observed, usually indicative of transient binding. Upon further evaluation, it was concluded that this rapid off rate was a result of the change in DMSO concentration between the injected analyte and the running buffer. As a result, no significant binding could be observed between the species using this technique. As with the fluorescence studies, it is possible that the presence of the additional tag (in this case, biotin) interferes with binding. However, in this case, there is no conclusive way to know whether the two species bind due to the number of other factors contributing the increase in RU associated with the observed results.

3.2.3 Studying the binding between SEA-001 and geldanamycin using ITC

Isothermal titration calorimetry (ITC) was performed using an iTC200 (Malvern) accessible through the Astbury Centre for Structural Molecular Biology. Unlike previous binding techniques, additional tags such as FITC and biotin were not needed, eliminating them as a cause of limited binding.

3.2.3.1 Isothermal Titration Calorimetry

ITC is a technique used to observe the direct binding between two species.[193] One advantage of ITC over other binding techniques is that it does not rely on the use of fluorescent or chemical labels, such as those needed for fluorescence experiments and SPR. In addition to measuring binding, it is also capable of measuring stoichiometry and changes in enthalpy (ΔH) and entropy (ΔS).[194] ITC measures the energy associated with binding through measurement of the amount of heat released or absorbed upon the addition, and subsequent binding of an analyte to a protein,[195] or in this case, peptide.

The ITC instrument itself (figure 3.9) measures the amount of power that is required to maintain a constant temperature difference (as close to zero as possible) between a sample cell and reference cell upon addition of a titrant to the sample.[196] The temperature difference is maintained through the use of a thermostatted heat jacket within the machine, and the power is measured in μcal/sec. The reference cell usually contains water or a buffer matching that of the sample and titrant. Generally, the protein is contained in the cell and the ligand is titrated into the protein in small increments using a syringe.

Once the full volume of the syringe has been added to the cell, heat (Q) in kcal/mol of titrant versus $[titrant]_T/[cell]_T$ (molar ratio) can be plotted from the results; where $[titrant]_T$ and $[cell]_T$ are the total concentrations of species in the syringe and cell, respectively. This plot can usually be processed directly using ITC software, in this case a combination of Nitpic,[197] SEDPHAT[198] and GUSSI were used. From the generated plots, enthalpy (ΔH), dissociation constant (K_D) and stoichiometry (n) can be obtained (figure 3.10).

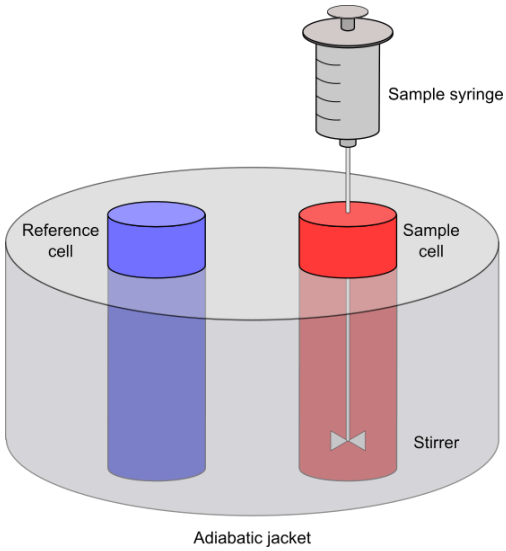


Figure 3.9 Outline of isothermal titration calorimetry

Sample (usually ligand) is titrated into the sample cell (usually containing protein) over the course of 20 injections with constant stirring. The amount of power needed to maintain a constant temperature difference ($\Delta T = 0^oC$) between the sample cell and reference cell (usually containing water) is measured.

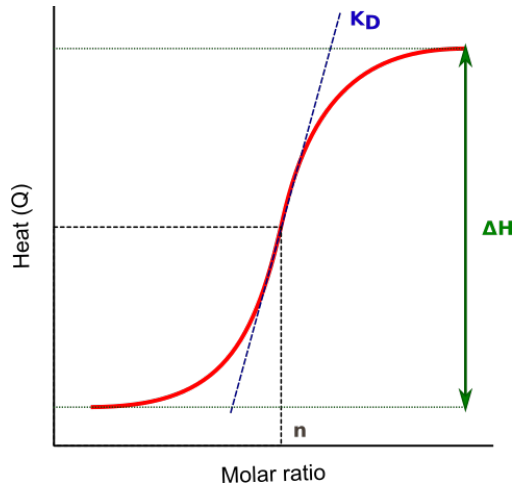


Figure 3.10 Example of plotted ITC results

A plot of heat (Q) against the molar ratio ($[titrant]_T/[cell]_T$) gives an s-shaped curve with a gradient of the dissociation constant, K_D. Stoichiometry, n, can be taken from the curve's midpoint and enthalpy, ΔH, from the difference in heat.

As the ΔH and K_D (and therefore K_A; see equation 3.6) are known, entropy (ΔS) and Gibbs energy (ΔG) can also be calculated using equation 3.10, where T is temperature and R is the universal gas constant.

Equation 3.10 $$\Delta G(T) = \Delta H(T) - T\Delta S(T) = -RTln(K_A)$$

3.2.3.2 Titration of geldanamycin into SEA-001

Geldanamycin (1 mM) was titrated into SEA-001 (100 µM) *via* a series of 20 injections over 50 minutes. This gave an initial readout of approximately 1.80 µcal/sec which gradually decreased to approximately 1.20 µcal/sec with sequential titrations. This gave an overall power change of 0.6 µcal/sec over the course of the experiment. However, the fitting of these changes resulted in a straight-line graph (figure 3.11) instead of the normal S-shaped curve obtained from ITC experiments. This straight line is usually attributed to a heat of dilution due to buffer mismatch, rather than a binding interaction. However, the pulses observed upon each injection of the experiment was much higher (1.20 – 1.80µcal/sec) than the pulses observed from injections in the control experiments; buffer into buffer (0.05 µcal/sec) and GM into buffer (0.04 µcal/sec). In addition to this, all samples were dialysed and diluted in the same buffer; PBS, 0.01% TCEP and 5.0% DMSO. As a result, it seemed unlikely that the results could be a result of such buffer mismatch.

It is possible that, as in SPR, the DMSO in the experiment had an unforeseen effect. Though GM needs the DMSO to solubilise, it is possible that the peptide is destabilised or precipitates over time at this concentration. In contrast, it may be that the GM, which is insoluble in water, precipitates out due to the low concentration of DMSO throughout the course of experiment and as a result it is not a heat of dilution, but a heat of precipitation effect that is observed.

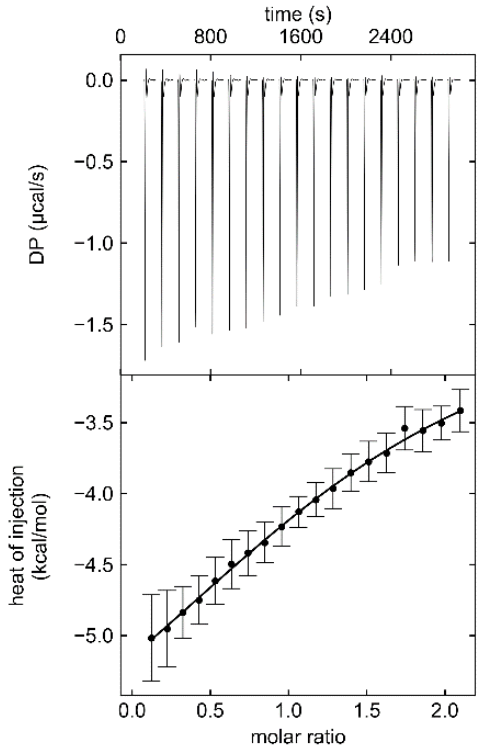


Figure 3.11 ITC plot of geldanamycin (1 mM) titrated in SEA-001 (100 µM)
Titration resulted in the observation of a partial curve indicating a potential binding interaction between the two species. Heat pulses were ten times higher than those observed in background experiments indicating that the observed change was not a result of heat of dilution. Due to problems with the solubility of geldanamycin, a full curve could not be obtained.

Another interpretation is that only the beginning, or end, of an S-shaped curve (as seen in fluorescence experiments) is being observed and, by optimising reagent concentrations, a full curve could be obtained. However, due to the solubility issues and the time and cost associated with continuing this work, other avenues were explored.

3.2.3.3 Titration of SEA-001 into geldanamycin

It was established that the solubility of GM was increased at lower concentrations and as such, it was less likely to precipitate during the ITC run. However, in order to facilitate the change in concentration, it was necessary to swap the reagents of the cell and syringe over. Hence, the syringe now held SEA-001 (360 μM) and the cell now contained GM (50 μM). As for previous titrations, a series of 20 injections occurred over the course of 50 minutes. Injections gave heat pulses of approximately 2.0 μcal/sec per injection, which was once again much higher than that control readings of 0.05 and 0.10 μcal/sec for buffer into buffer and SEA-001 into buffer, respectively. However, this power input remained constant throughout the experiment, which showed no real trend upon plotting (figure 3.12). Once again, this may be attributed to either heat of dilution or precipitation and are unlikely to be attributed to a real binding interaction.

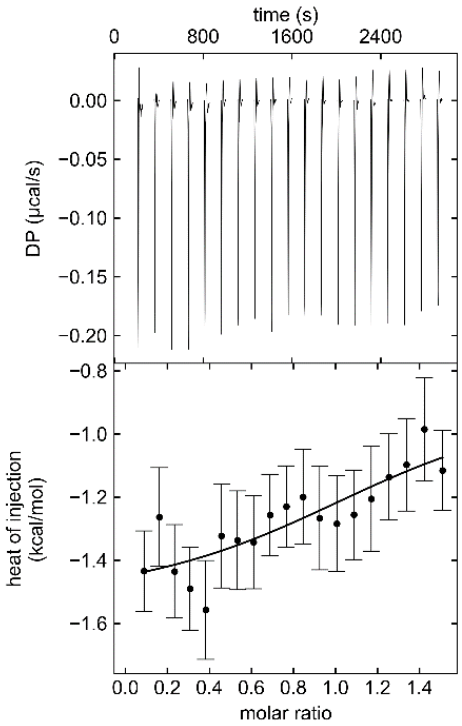


Figure 3.12 ITC plot from titration of SEA-001 (360 µM) into GM (50 µM)

Titration of peptide into geldanamycin showed no binding interaction. Though heat pulses ten times larger than those observed in control experiments were observed, there was no trend between concentration of titrant and the power needed to maintain a constant temperature. As a result, heat pulses were assumed to be attributed to heat of precipitation due to the poor solubility of GM.

3.2.3.4 Conclusions from ITC with SEA-001 and geldanamycin

Results from ITC experiments remained somewhat inconclusive, showing a potential partial binding interaction. However, due to the poor solubility of geldanamycin, it is possible that such results arise from reagent precipitation rather that a binding interaction. As mentioned previously, it is possible that the unconstrained nature of the free peptide SEA-001 does not mimic the phage-displayed sequence which showed binding during Magic Tag® experiments.

3.3 Exploring the binding interaction between MBP-constrained SEA-001 and geldanamycin

The SEA-001 peptide sequence was expressed as part of the *E. coli* protein, maltose binding protein (MBP). Several variants were produced in order to cover several different sites within the protein. More information on this can be found in Chapter 4.

Three variants, His_6-MBP-G6, His_6-MBP-D178 and His_6-MBP-T367, were used in a geldanamycin pull-down assay so as to probe the binding between the constrained SEA-001 peptide and geldanamycin. Variants represented N-terminal, internal and C-terminal insertions, respectively.

3.3.1 Geldanamycin pull-down assay with SEA-001 constrained within histidine-tagged MBP variants

The pull-down assay exploited the strong binding interaction between NeutrAvidin™, a deglycosylated avidin derivative, and biotin. NeutrAvidin™ agarose resin (Pierce) was used in conjunction with a gravity flow column. 0.5 mL of resin allowed for a maximum immobilisation of 33.5 μg or biotinylated geldanamycin. Assuming 100% immobilisation of biotin-GM, a maximum binding of 2.6 mg of H_6-MBP-D178 was calculated. A standard dilution/wash buffer; 0.1 M phosphate, 0.15 M NaCl, 5% DMSO, pH 7.2, was used to accommodate the solubility of both species. A total of 1 mg protein was loaded onto the column, giving a maximum binding capacity of 1 mg. Step-wise binding, washing and elution with 8.0 M Urea steps were monitored by SDS-PAGE (figure 3.13).

It was observed that the protein was washed from the column before the elution step and therefore, showed no binding to the immobilised geldanamycin. The geldanamycin was known to be bound to the column due to the presence of the pinkish hue of the compound.

The assay was repeated under the same conditions for H_6-MBP-G6 and H_6-MBP-T367, yielding the same result. As the pull-down assays were performed much later than the original binding assays, the geldanamycin was checked for degradation *via* ^{1}H NMR and was determined to be intact and suitable for use in the assay.

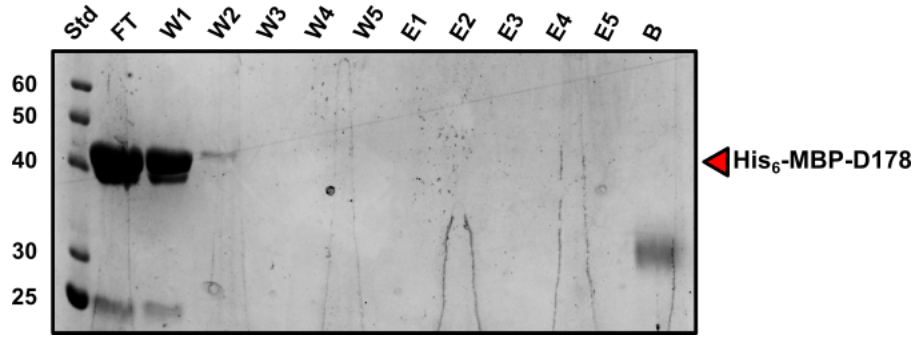


Figure 3.13 SDS-PAGE analysis of GM pull-down with H_6-MBP-D178

All protein (~44 kDa) was observed to leave the column in flow-through (FT) and wash (W) phases. No protein was observed in elution (E) phases. Post elution, beads were boiled in SDS-PAGE buffer; no H_6-MBP-D178 remained bound to beads.

3.3.2 Conclusions for geldanamycin pull-down experiment with MBP-constrained SEA-001

Though the presence of a biotin tag on the geldanamycin still remains a possible, reason for the lack of observed binding, it is much more likely that the binding interaction between geldanamycin and SEA-001 is either non-existent or too weak to be observed *via* this method. Immobilisation of biotinylated GM and subsequent analysis of MBP-constrained SEA-001 showed no observable interaction between the two species.

3.4 Conclusions on the binding interaction between SEA-001 and geldanamycin

The binding interaction of SEA-001 and geldanamycin was studied in an attempt to mimic the binding interaction between a phage-expressed *Alu* sequence and immobilised geldanamycin during a Magic Tag® experiment at the University of Warwick. Though initial fluorescence anisotropy results showed hope, the binding interaction appeared to be too weak to be studied by this method due to the cost impact of the large quantities of reagents needed. Fluorescence intensity experiments with immobilised SEA-001 showed no binding interaction. Though SPR experiments were attempted, the DMSO needed to counteract the poor solubility of geldanamycin interfered with the results. Nonetheless, although no plottable data was obtained, raw data indicated no, or extremely weak binding

between the species. ITC experiments showed a potential weak interaction between the two species, however the resultant heat spikes from titration of GM into SEA-001 may have been the result of reagent precipitation as opposed to a real binding interaction. A geldanamycin pull-down experiment using NeutrAvidin™ resin, biotinylated GM and MBP-constrained SEA-001 showed no binding interaction between the two species.

In the cases of fluorescence experiments, it remains true that the presence of the bulky FITC label may interfere with the binding interaction of the two species, especially if the binding interaction was not particularly strong to begin with. In SPR and geldanamycin pull-down experiments, it is also possible that the biotin label interferes with the binding interaction. However, this possibility seems unlikely due to the use of the similar immobilisation techniques in the Magic Tag® experiments. However, a consistent concern with any immobilisation technique is that there may be binding to the linker and/or tag instead of the desired immobilised species. It remains that poor solubility of geldanamycin made experiments difficult and therefore, could not be properly representative of the behaviour of the species in physiological conditions. However, the combination of the above results shows that the interaction observed in the Magic Tag® experiment could not be replicated through a broad range of binding experiments.

It possible that the differences in one or more of five residues between the Magic Tag® hit and the SEA-001 sequence may have been a major contributor to the binding interaction (see 3.1.2).[199] It is possible that the cysteine (C) from serine (S) substitution would have affected binding as although C and S share a similar arrangement in chemical space, C has the capacity to form disulfide bonds and cannot act as a hydrogen bond donor as S can. In addition, substitution of leucine (L) for proline (P) may cause a 'kink' in the spacial arrangement of the sequence as well as introducing a potential hydrogen bond acceptor. In the case of the substitution of isoleucine (I) for threonine (T), a hydrophobic residue has been substituted for a polar residue which may again have an effect. The histidine (H) from aspartic acid (D) results in a swap from positive to negative charge so it is possible that this may also affect binding. The substitution of asparagine (N) for lysine (K) may also have an impact as the changing of these residues would result in a change in charge from neutral to positive.

Though no defined binding interaction between SEA-001 and geldanamycin was observed, it is still possible that *Alu* elements could be exploited in drug discovery *via* other methods.

In retrospect, and if given more time, it would have been interesting to observe whether the binding interaction between GM and the *Alu*-derived sequence could be observed with a peptide matching the exact Magic Tag sequence. As we know that several hydrogen bonding interactions are involved in the binding of HSP90 to GM, it is possible that the substitution of K for N may have affected the binding interaction. However, it is also possible that the reading frame that we have used for studying the Magic Tag hit, as directed by bioinformatic analyses, does not match the reading frame that was expressed on the surface of the phage in the Magic Tag experiments.

Chapter 4
Translated *Alu* elements in human proteins

Overview

The extent of the effect of translated *Alu* elements in proteins is largely unknown, with reports of both advantageous and disadvantageous protein mutations.[200] However, it is known that they produce a source of protein variability through the expression of alternative protein isoforms as a result of alternative splicing.[201] Five target human recombinant proteins were chosen for expression trials in *E. coli*; ZMAT1, NEK4, BCAS4, PPP5D1 and ASCC1. Human proteins were cloned *via* PCR into a series of vectors for overexpression in *E. coli.* A number of different solubility tags and purification methods were performed in an attempt to obtain naturally occurring *Alu*-containing proteins.

4.1 Overexpression of naturally occurring *Alu*-containing human recombinant proteins

Generally, it is often difficult to express human recombinant protein in non-mammalian cell cultures such as *E. coli*. Many proteins fold incorrectly either due to degradation or accumulation to form an insoluble inclusion body.[202] Even if a protein can be expressed within the system, it is still possible that the protein will be inactive.[203] *E. coli* is often used for protein expression as large quantities of protein can be produced quickly at a relatively low cost.[204] However, there are a number of reasons that the expression of human recombinant proteins from *E. coli* can be challenging.

The first is that human genes contain rare codons which cannot be recognised by *E. coli* transfer RNAs (tRNAs).[205] This problem can be easily overcome by optimising genes for expression in *E. coli* and/or using specialised cell lines such as Rosetta™ (DE3). Secondly, expression of the foreign protein may be toxic to the host system and may slow or kill the *E. coli* by interfering with normal proliferation and homeostatic function.[206] The human body is a complex system and as such, the

normal expression of proteins within it may require the aid of co-factors and chaperones. Therefore, it may not be possible to express the protein of choice without co-expression with additional proteins which aid folding.[207] Additionally, the protein of choice may require post-translational modifications (PTMs) in order to fold correctly or maintain activity. For example, *E. coli* expression systems do not have the ability to perform PTMs such as O- or N-linked glycosylation, hydroxylation or sulfation.[208] As a result, the expression of human *Alu*-containing in *E. coli* was expected to be challenging.

4.1.1 Selection of human proteins for overexpression in *E. coli*

Of the 46 *Alu*-containing hits identified during bioinformatic analysis (outlined in chapter 2), only a select few were chosen for overexpression in *E. coli*. One of the main questions concerning translated *Alu* elements in this project was whether or not they were linked with human disease, in particular, cancer. As such, the main focus of the proteins chosen was based upon previous reports of their implications in cancer or as precursors to cancer. At the beginning of the project, the main protein of interest was M4K1 (Mitogen-activated protein kinase kinase kinase kinase 1) due to reports linking it to cancer.[209] However, due to the large size of the gene (2,631 bp) and no reports of previous overexpression in any system, it was determined that cloning (*via* PCR), and subsequent overexpression and purification, of the 91.2 kDa (canonical isoform; 833 residues) recombinant protein from *E. coli* was likely to be challenging. Research into other protein hits found four proteins with cancer implications which showed promise; BCAS4, NEK4, ASCC1 and ZMAT1. The protein structures for these proteins, as predicted by I-TASSER, can be observed in figure 4.1.

BCAS4 (breast cancer amplified sequence 4), as stated by its name, is overexpressed in breast cancer.[209] NEK4 (never in mitosis gene A (NIMA)-related kinase 4), though not directly linked to cancer, has been linked to DNA repair. As a result, mutations or the expression of alternative isoforms may result in inefficient DNA repair and subsequent additional mutations could potentially lead to cancer.[210] ASCC1 (activating signal cointegrator 1 complex subunit 1) has been associated with Barrett's Esophagus (BE), a pre-cursor to esophageal adenocarcinoma (EAC), through germline gene mutations.[211] ZMAT1 (zinc finger matrin-type protein 1) has been reported to serve as a prediction of poor prognosis in gastric cancer patients.[212]

All of the proteins have been reported to be translated and expressed in the cell as two or more different isoform variants with at least one *Alu*-containing isoform (AC) and one non-*Alu*-containing (nAC) isoform.

ZMAT1 and NEK4 were high risk proteins and were expected to have a lower chance of successful overexpression in *E. coli*. According to the Human Protein Atlas, NEK4 RNA is ubiquitously expressed; however, protein expression occurs within the cytoplasm and nucleus and is most abundant in the testes. Both the AC and nAC isoforms are predicted to express but no preference for the expression of one over the other is reported. NEK4 (figure 4.1A) had been reported to have been overexpressed before; however, this was from human HEK293T cells, not *E. coli*.[213] This, in addition to its large size (841 residues), predicted that overexpression in *E. coli* may prove challenging. However, NEK7, another kinase which shares 40% sequence similarity to NEK4, can be overexpressed in *E. coli*.[214] It should be noted that NEK7 is a much smaller protein than NEK4 at only 34.5 kDa. ZMAT1 RNA is detected in all tissue types, but protein expression is observed only in the cytoplasm of the epididymis, adrenal gland and testes. Expression of the protein is localised to the nucleoplasm. ZMAT1 RNA is detected in multiple cancers but shows low cancer specificity. Both the AC (figure 4.1B) and nAC isoforms of ZMAT1 are predicted to be expressed with no recorded bias towards expression of one over the other. Though expression protocols could not be found, can be purchased (MyBioSource) and has been quoted to be purified from multiple systems including *E. coli*. Due to the presence of four zinc fingers contained within the structure of ZMAT1, it was predicted that purification may be more of an issue in this case. Each zinc finger in ZMAT1 contains two histidine residues and two cysteine residues, each making a possible contribution to the binding to Ni-NTA (Nickel-nitrilotriacetic acid). Conversely, the polyhistidine purification tag may interact with the Zn^{2+} ions present in the buffer reducing binding to Ni-NTA.[215] Therefore, purification of such a protein may become difficult using polyhistidine tags. However, purification of zinc fingers using polyhistidine tags has been achieved.[216]

The Human Protein Atlas reports that BCAS4 RNA expression in enhanced in blood and lymphoid tissue, and protein expression generally occurs in the cytoplasm. RNA is also enhanced in memory and naive B-cells. Both AC and nAC isoforms of BCAS4 are expressed, but no preference is reported. BCAS4 (figure 4.1C) is a much smaller gene (1,404 bp) making it a good candidate for restriction free cloning.

Though BCAS4 itself has not been reported to have been expressed before, its much larger predicted functional partner BCAS3 can be purchased (MyBioSource). Due to its small size (canonical isoform; 22.8 kDa; 211 residues) and prominence in breast cancer cell lines, BCAS4 remained a protein of interest for overexpression in *E. coli*.

ASCC1 RNA is reported to be ubiquitous and protein expression occurs in the cytoplasm and nucleus of all tissues. ASCC1 has low cell line specificity. No preference for the expression of the AC or nAC isoform is reported. ASCC1 (figure 4.1D) is a 45.5 kDa protein (canonical isoform; 400 residues) which can be purchased (MyBioSource) as an overexpression product from *E. coli*. However, the protein isoform that can be purchased is the non-*Alu* (nAC) variant. It was predicted that ASCC1 may be the easiest to express of the four cancer-associated proteins due to one of its isoforms already having been overexpressed (though no published protocols were available). However, this prediction assumed that the translated *Alu* did not destabilise the *Alu*-containing (AC) isoform.

Though it has no known association with cancer, PPP5D1 (PPP5 TPR repeat domain-containing protein 1) was also chosen for overexpression due to its small size (19.6 kDa; 171 residues) and cytoplasmic expression in most tissues, including most cancerous tissues. PPP5D1 has only one described isoform but has two computationally mapped potential isoforms.[217]

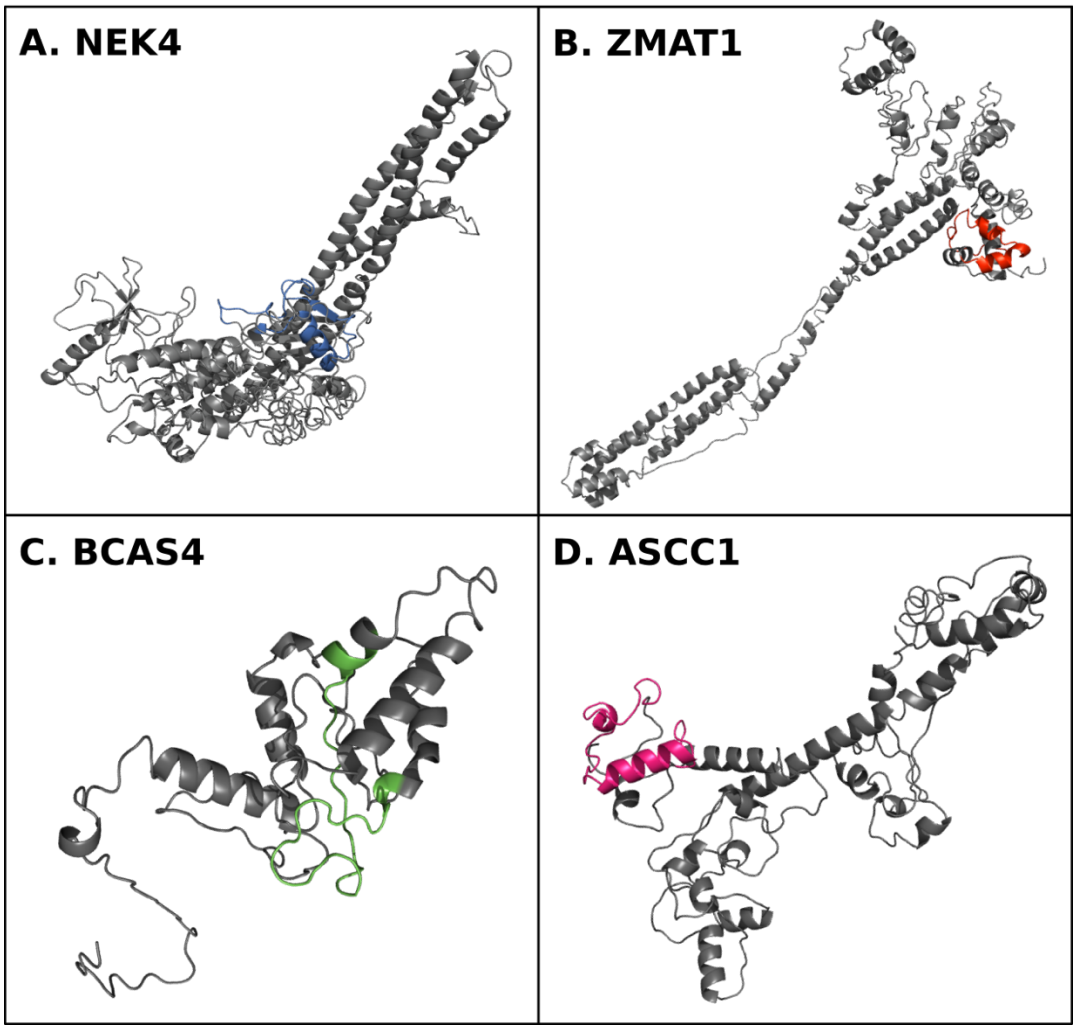


Figure 4.1 I-TASSER structure predictions of human proteins

The structures of *Alu*-containing isoforms of NEK4, ZMAT1, BCAS4 and ASCC1 were predicted using I-TASSER. *Alu* insertions are represented in colour. In most cases, the element appears to have either coiled or α-helical structure. All four proteins are predicted to be largely made up of α-helical character. *Alu* insertions in NEK4 and BCAS4 are fairly central to the structure, whereas they lie on the outside of ZMAT1 and ASCC1.

4.2 Cloning of human genes into *E. coli* expression vectors

The *Alu*-containing variants of human genes, NEK4, ASCC1, ZMAT1 and BCAS4 and the non-*Alu*-containing variant of ASCC1 were purchased in TrueORF® pCMV6-entry vectors from OriGene. It should be noted that these genes were not codon-optimised for overexpression in *E. coli*. Bacterial expression vectors for all proteins were generated *via* a combination of purchase and restriction-free (RF) cloning. The five gene variants listed above were subcloned into either pET-28a or pET-SUMO-28a (figure 4.2) to generate a N-terminal His_6 tag or His_6-SUMO fusion proteins. PPP5D1 was purchased (GenScript) as a custom, codon-optimised gene in pGEX-4T-1 to generate a GST fusion protein.

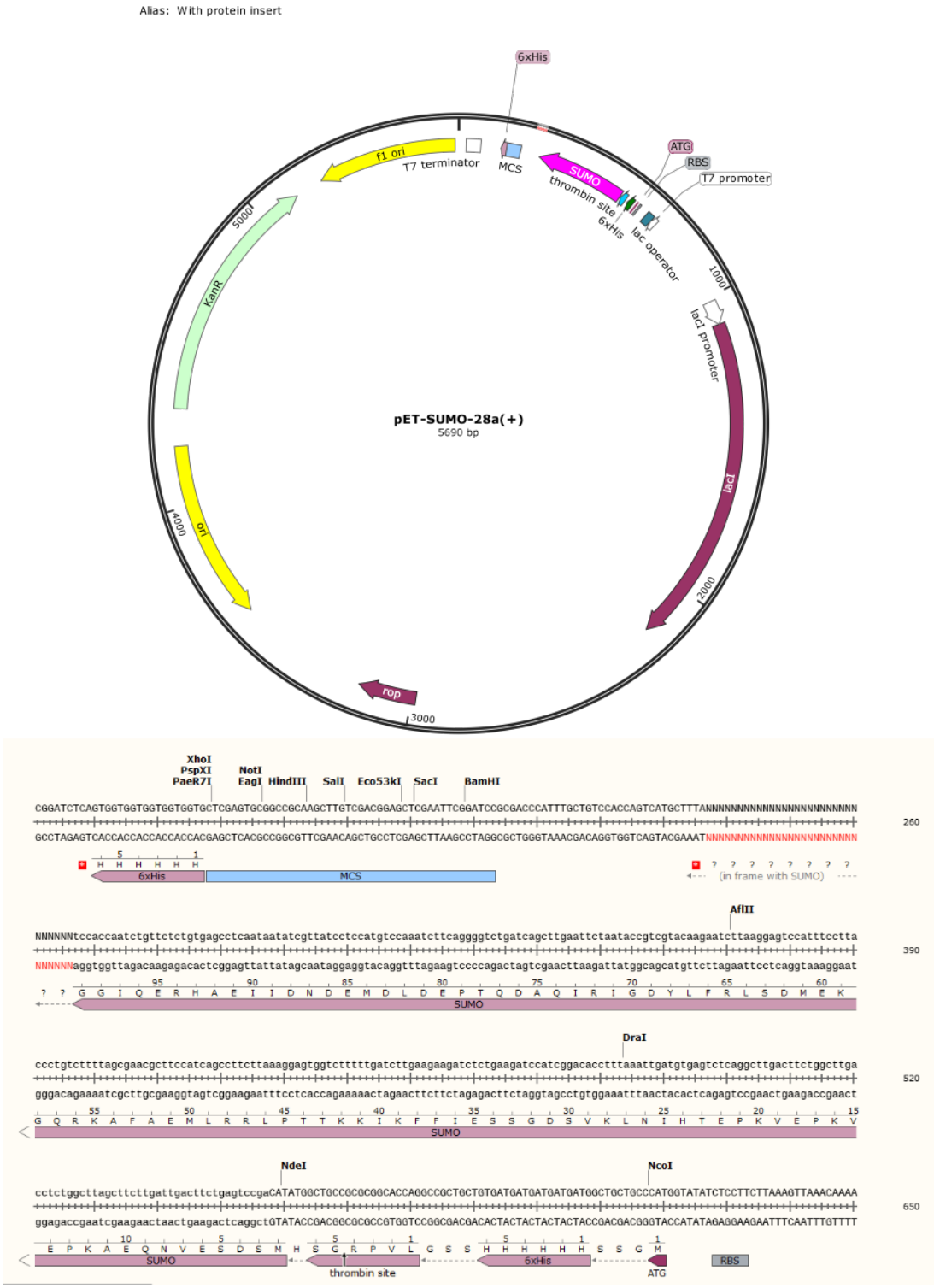


Figure 4.2 pET-SUMO-28a vector

pET-SUMO-28a was obtained from the Edwards Group (University of Leeds). Genes that were subcloned (red) into pET-SUMO-28a were inserted directly after the SUMO gene and ended in a terminated in a STOP codon. Proteins inserted into pET-28a were inserted in the same way without the presence of the SUMO gene.

4.2.1 4.2.1 Restriction-free cloning

Restriction-free (RF) cloning is a two-step polymerase chain reaction (PCR) cloning method which allows for the insertion of a gene into any region of a target vector without the use of restriction enzymes or DNA ligases.[218] The technique relies on the design of specific primers which contain regions matching both the target gene and the desired insertion location within the destination vector. The forward primer is composed of two parts; the desired point of insertion in the destination vector (ca. 21 bp) followed by the start of the gene of interest (ca. 21 bp) beginning with ATG. The reverse primer is a reverse complement composed of three parts; the end of the gene of interest (ca. 21 bp) followed by a stop codon and the desired point of insertion in the destination vector.

The designed primers were used in an initial PCR amplification round (RF_1) in which they annealed to the target gene and were amplified to give a 'megaprimer'. This megaprimer was comprised of the target gene flanked by sequences complementary to the desired insertion point in the destination vector. The megaprimer was gel extracted and used in a second PCR amplification round (RF_2) in which the flanking ends annealed to the destination vector and formed a gene-containing 'loop'. The primer then extended around the outside of the destination vector to give a final PCR product of the target gene contained within the destination vector (figure 4.4).[219] Dpn1 treatment was used to digest any methylated parental plasmid. The reaction mixture was used directly in standard transformation in *E. coli*. Following the isolation of individual colonies, inserted sequences were confirmed by sequencing (GeneWiz) across the whole gene and can be found in Appendix 2. A list of cloned constructs is shown in table 4.1.

Figure 4.3 RF_1 amplification of NEK4

Observed bands corresponded to the NEK4 'megaprimer' and were excised and purified for use in amplification round 2 (RF_2). The two lanes represent two identical reactions that took place side by side.

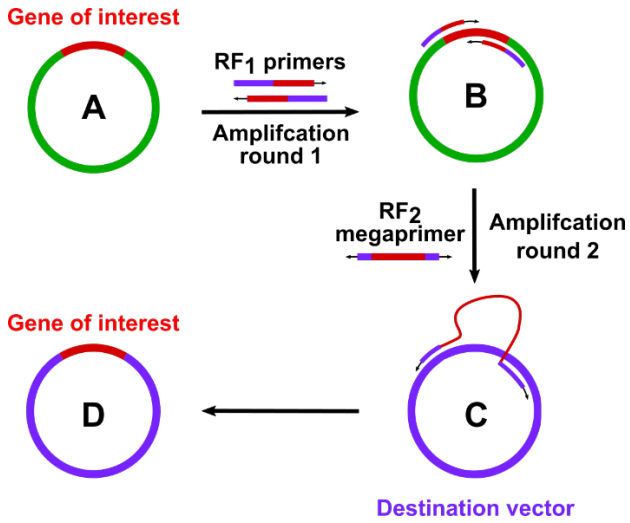


Figure 4.4 Restriction-free (RF) cloning

The gene of interest, confined within an entry vector (A), is amplified using specifically designed primers in amplification round 1 (RF_1: B). The resulting 'megaprimer' is used in a second amplification round (RF_2) which extends around the whole destination vector (C). The result is the gene of interest contained within the desired vector for use in immediate transformation, with no ligation steps required.

Gene	Variant	Vector	Predicted protein construct	Mass (kDa)
ASCC1	*Alu*	pET-SUMO-28a	His_6-SUMO-ASCC1	58.9
ASCC1	*Alu*	pET-28a	His_6-ASCC1	48.1
ASCC1	Non-*Alu*	pET-28a	His_6-ASCC1	43.8
BCAS4	*Alu*	pET-28a	His_6-BCAS4	24.9
NEK4	*Alu*	pET-SUMO-28a	His_6-SUMO-NEK4	108.0
PPP5D1	*Alu*	pGEX-4T-1	GST-PPP5D1	45.9
ZMAT1	*Alu*	pET-SUMO-28a	His_6-SUMO-ZMAT1	88.2

Table 4.1 Human gene constructs

Further information on sequences of plasmids and proteins can be found in Appendix 2.

4.3 Trial expression of human recombinant proteins

4.3.1 Fusion partners and purification tags

Genes cloned into the pET-28a vector (BCAS4) generated an N-terminal His_6 tag. The use of a polyhistidine tag is a widely-used purification method based on its interaction with Ni^{2+} metal ions immobilised on Ni-NTA (nickel-nitrilotriacetic acid) resin.[220] Due to the small size and charge of the polyhistidine tag relative to the attached protein, it was unlikely that it would interfere with protein activity and thus, it was not necessary to cleave it post-purification. However, had the

polyhistidine tag needed removed, it could be cleaved from the protein product *via* a thrombin cleavage site (LVPRGS).

Genes cloned into the pET-SUMO-28a vector (ZMAT1, NEK4, ASCC1) generated His_6-SUMO fusion proteins. SUMO (small ubiquitin-like modifier) family proteins are approximately 97 residues in length, giving a molecular weight of approximately 11 kDa, and have approximately 18% sequence similarity with ubiquitin (Ub).[221] Unlike Ub, the surface charge topology of SUMO has very distinct positively and negatively charged regions.[222] The fusion of SUMO proteins to partner proteins has been reported to improve their expression and solubility.[223] The expression of human proteins in *E. coli* expression systems is often challenging; as such, it was proposed that overexpression of the desired protein as part of a larger SUMO fusion product would improve solubility. As SUMO is generally used to improve solubility and not to aid purification, a polyhistidine tag was also incorporated at the N-terminus to enable protein purification using nickel affinity chromatography. SUMO has a C-terminal Gly-Gly motif through which cleavage could be achieved using Ubiquitin-like-specific protease 1 (Ulp1).[224]

PPP5D1 was overexpressed as a glutathione S-transferase (GST) fusion protein. GST is a eukaryotic protein which has a molecular weight approximately 26 kDa.[225] As well as aiding in solubilisation of its fusion partner, GST can also act as a purification tag through utilisation of its natural binding affinity for glutathione. As a result, no polyhistidine tag was required for purification of GST fusion proteins. GST could be easily cleaved from its fusion partner *via* a thrombin cleavage site.

4.3.2 Overexpression of His_6-SUMO-NEK4

NIMA (Never in Mitosis A) related kinase 4 is a serine/threonine protein kinase which is part of a larger protein family termed NIMA-related kinases (Nrks) which constitutes approximately 2% of the entire human kinome[††] (NEK1 through to NEK11).[226] Despite sharing approximately 40 - 45% sequence identity with NimA, a protein involved in mitotic entry through catalytic kinase domains at their N-termini, their sequences vary greatly elsewhere, in particular at their C-termini. Variations at the non-catalytic C-terminus are believed to be the cause of possible varied functionalities of Nrks which are not under mitotic control.[227] Relatively little

[††] The kinome refers to all of the protein kinases encoded by the genome.

is known about the exact function and structure of NEK4, though more research has been performed on other members of the Nrk family. However, it has been observed the NEK4 is present in most primary carcinomas.[228] There is limited structural data on NEK4 and previously reported expression has only been achieved in HEK293 cells.

The *Alu*-containing (AC) NEK4 transcript (variant 1, OriGene) was cloned into pET-SUMO-28a. The gene-containing plasmid was transformed into *E. coli* Rosetta™ (DE3) competent cells. Initial expression tests showed no overexpression of the desired protein product (figure 4.5).

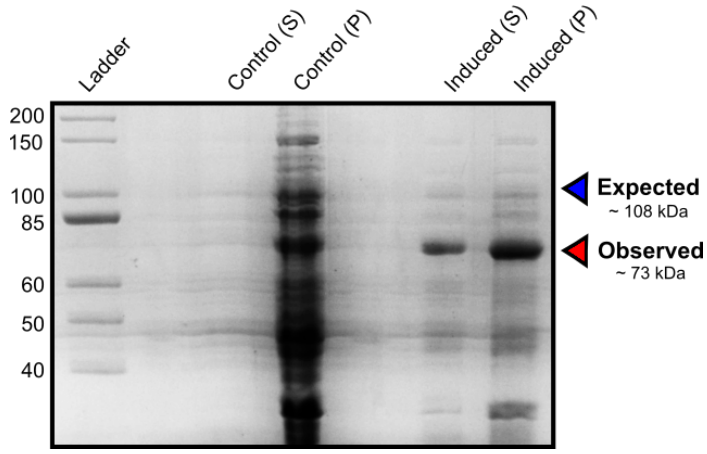


Figure 4.5 SDS PAGE analysis of His_6-SUMO-NEK4 overexpression

Polyhistidine-tagged SUMO-NEK4 fusion protein (ca. 108 kDa) test expression was performed in LB media with IPTG induction. SDS-PAGE analysis showed no overexpression of a protein product of the expected mass. Strong bands were observed at approximately 73 kDa in both the induced supernatant (Induced (S)) and induced pellet (Induced (P)); however, this corresponded to neither His_6-SUMO nor NEK4 alone and was also present in the control pellet (Control (P)).

The polyhistidine-tagged SUMO-NEK4 fusion protein had a calculated molecular weight of approximately 108 kDa. No protein of this size was observed upon SDS-PAGE analysis. Prominent bands could be observed at approximately 73 kDa which correlated to neither NEK4 alone (ca. 94.5 kDa) nor polyhistidine-tagged SUMO protein (ca. 13.3 kDa). This band could also be observed in the control pellet (no IPTG induction). Unfortunately, no MS analysis was obtained to verify the absence of the desired protein.

Due to the large size of NEK4 and the even larger size of the SUMO-NEK4 fusion protein, it is possible that this construct cannot be overexpressed from an *E. coli* expression system. Generally, it is difficult to express proteins over 100 kDa in *E. coli*. Though genes were not optimised for overexpression in *E. coli* through the substitution of rare codons, Rosetta™ cells were used. It is possible that NEK4 is glycosylated when expressed in humans. *E. coli* is limited in that it does not have the machinery to efficiently carry out this post-translational modification.[229]

Note: Data for His_6-SUMO-NEK4 was revisited towards the end of the PhD project and it was determined that we cannot say for certain that this protein was not overexpressed. It is possible that a truncated protein product was formed or that the presence of charged residues (constituting approximately 1/7 of the protein) affected mobility in SDS-PAGE. Given more time, large scale overexpression, purification and MS analysis would have been carried out to confirm the identity of the protein product.

4.3.3 Overexpression of His_6-SUMO-ZMAT1

Relatively little is known about the structure and function of zinc finger matrin-type protein 1 (ZMAT1); however, it is known that it contains four Cys2-His2 (C2H2)-type zinc fingers and is localised to the nucleus. The STRING database (ELIXIR)[230] predicts eight protein binding partners for the protein but generally the presence of C2H2 zinc finger motifs in the structure predicts DNA binding activity.[231] In terms of disease, the majority of reports relate ZMAT1 to gastric cancer,[232, 233] though this is usually in reference to the non-*Alu* long non-coding ZMAT1 RNA transcript.

The *Alu*-containing (AC) ZMAT1 transcript (variant 1, OriGene) was cloned into pET-SUMO-28a. The gene-containing plasmid was transformed in *E. coli* Rosetta™ (DE3) competent cells for overexpression. Overexpression of the polyhistidine-tagged SUMO-ZMAT1 protein (ca. 88 kDa) was tested *via* IPTG induction and auto-induction (figure 4.6), both of which resulted in the overexpression of an insoluble inclusion body.

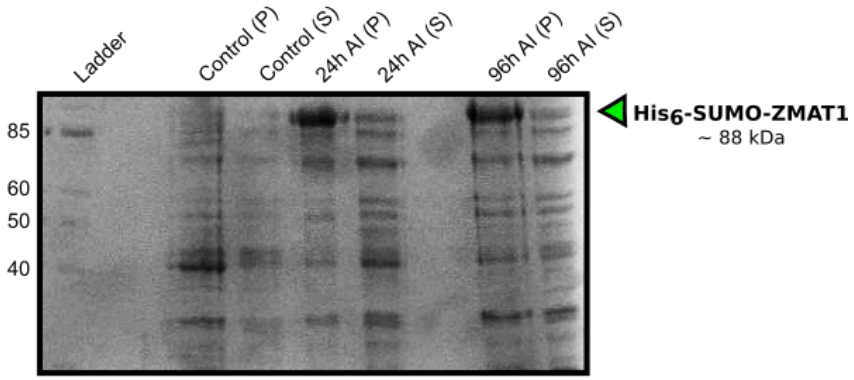


Figure 4.6 SDS-PAGE analysis of His_6-SUMO-ZMAT1 auto-induction
Auto-induced overexpression of polyhistidine-tagged SUMO-ZMAT1 fusion protein (ca. 88 kDa) was performed in LB auto-induction media over the course of 96 hours. Though overexpression was observed in auto-induced (AI) samples and not control samples, the protein appeared to be overexpressed in the cell pellet (P), not the supernatant (S). As such, the protein product was an insoluble inclusion body.

Due to the insolubility of the protein, it was not possible to purify it *via* normal methods*, i.e.* purification of solubilised protein *via* nickel affinity chromatography and subsequent size exclusion chromatography (SEC). Instead, purification was attempted through solubilisation as a result of denaturation with urea, purification *via* nickel affinity chromatography and subsequent refolding with L-arginine. Protein solubilisation and purification was successful (figure 4.7A). Unfortunately, upon dialysis of the refolded protein to remove urea, the majority of the protein precipitated (figure 4.7B), despite the presence of Zn^{2+} ions in all buffers. For the small amount of soluble protein that was recovered, cleavage of ZMAT1 from SUMO with Ulp1 was tested. Cleavage resulted in insoluble ZMAT1 protein product. No further avenues were explored for the purification of this protein.

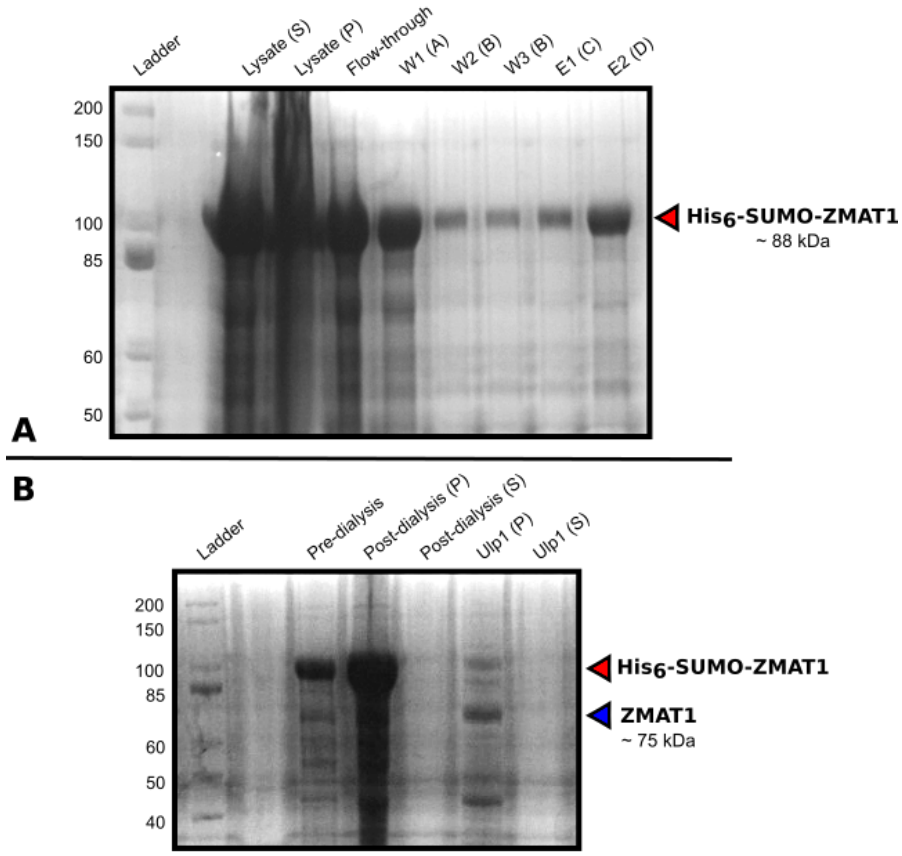


Figure 4.7 SDS-PAGE analysis of denaturation, refolding and Ulp1 cleavage of His_6-SUMO-ZMAT1

(A) Polyhistidine-tagged SUMO-ZMAT1 protein (ca. 88 kDa) was denatured in 8 M urea and purified *via* nickel affinity chromatography. Soluble protein was successfully obtained from column elution. (B) Polyhistidine-tagged SUMO-ZMAT1 (ca. 88 kDa; red) was refolded with L-arginine and dialysed to remove urea. Dialysis resulted in the majority of protein precipitating (Post-dialysis (P)). A small amount of soluble fusion protein was obtained; however, cleavage with Ulp1 resulted in the precipitation of ZMAT1 (Ulp1 (P); ca. 75 kDa; blue).

4.3.4 Overexpression of GST-PPP5D1

PPP5 tetratricopeptide repeat domain containing 1 (PPP5D1) is a small protein with partial sequence similarity with the serine/threonine phosphatase, PPP5C. However, unlike PPP5C, PPP5D1 does not contain a catalytic pseudo-phosphatase domain. As for previous proteins, there is very limited information available on the structure and function of PPP5D1.

Codon-optimised PPP5D1 in pGEX-4T-1 (GenScript) was transformed in *E. coli* Rosetta™ (DE3) supercompetent cells and overexpressed using IPTG induction. Cell lysate was purified *via* chromatography with Pierce ™ glutathione agarose (Thermo Scientific) which utilises the binding interaction between GST and glutathione to purify GST fusion proteins. SDS-PAGE analysis (figure 4.8) showed that a small amount of GST-PPP5D1 (ca. 46 kDa) was eluted with reduced-glutathione. However, the bulk of eluted protein was approximately 26 kDa which likely corresponded to the overexpression of GST alone rather than as part of the fusion protein. Bands of 46 kDa were also observed in the flow-through and wash phases. This could be attributed to one of two things; either there was another protein present in large amounts with a similar mass to the target protein, or the fusion protein was misfolded and therefore, could not bind to the column. No further avenues were explored for the expression and purification of GST-PPP5D1.

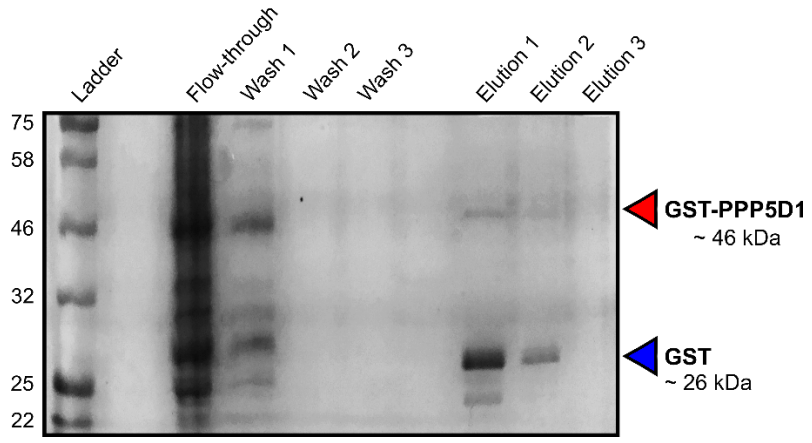


Figure 4.8 SDS-PAGE analysis of GST-PPP5D1 purification

A small amount of GST-PPP5D1 (ca. 46 kDa; red) appeared to elute from the column; however, the bulk of eluted protein was approximately 26 kDa (blue), likely corresponding to lone GST as opposed to the GST-PPP5D1 fusion protein. Bands at approximately 46 kDa in the flow-through and wash phases may indicate misfolding of the fusion protein.

4.3.5 Overexpression of His_6-BCAS4

Breast carcinoma amplified sequence 4 (BCAS4) is, again, a protein about which relatively little is known, other than its reported overexpression in a large number of breast cancer cell lines. In addition to this, it has been reported that, in some cell lines, the BCAS4 gene fuses with that of BCAS3 as a result of chromosome

rearrangement.[234, 235] The function of BCAS4, BCAS3 and the BCAS3/4 fusion protein is still something to be speculated about as homology studies find no obvious functional domains.

The *Alu*-containing (AC) BCAS4 transcript (variant 1, OriGene) was cloned into pET-28a. The gene-containing plasmid was transformed in *E. coli* Rosetta™ (DE3) competent cells for overexpression. Overexpression in LB media with IPTG induction yielded an insoluble inclusion body (figure 4.9). A series of lysis buffers were tested in an attempt to solubilise the protein, however, each resulted in an insoluble product.

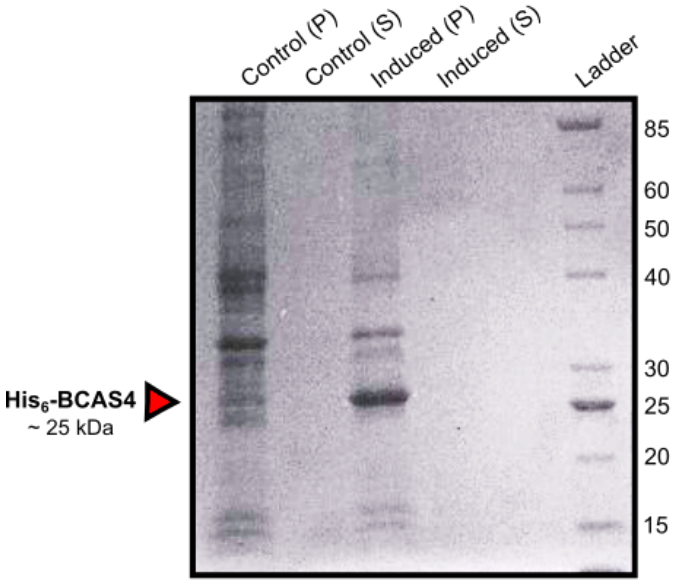


Figure 4.9 SDS-PAGE analysis of His_6-BCAS4 overexpression in *E. coli*
Overexpression of polyhistidine-tagged BCAS4 (ca. 25 kDa) in *E. coli* with IPTG induction resulted in the formation of the protein as an insoluble inclusion body (Induced (P)). No soluble protein was observed (Induced (S)).

As with His_6-SUMO-ZMAT1, polyhistidine-tagged BCAS4 was solubilised *via* denaturation with urea and purified using nickel affinity chromatography. SDS-PAGE analysis confirmed elution of the protein product. His_6-BCAS4 was diluted in refolding buffer containing L-arginine. However, upon dialysis to remove urea, the protein precipitated and therefore, could not be carried forward. The absence of soluble protein was confirmed *via* mass spectrometry. No further avenues were explored for the purification of His_6-BCAS4.

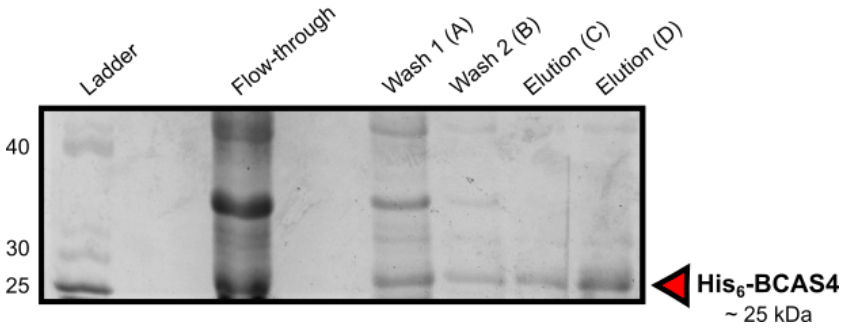


Figure 4.10 SDS-PAGE analysis of nickel affinity purification of denatured His_6-BCAS4

Urea-denatured polyhistidine-tagged BCAS4 (ca. 25 kDa) was purified *via* nickel affinity chromatography and eluted in a pH-dependent manner. SDS-PAGE analysis confirmed elution of the desired protein product. Some protein was observed to be washed from the column prior to elution due to column overloading.

4.3.6 Overexpression of His_6-SUMO-ASCC1

Activating signal cointegrator 1 complex subunit 1 (ASCC1) is probably the most studied protein of the selection explored in this project. Located in the nucleus, it is part of the activating signal cointegrator 1 (ASC-1) complex which is composed of three other protein subunits in addition to ASCC1; thyroid hormone receptor interactor 4 (TRIP4) and activating signal cointegrator 1 complex subunits 2 and 3 (ASCC2 & ASCC3).[236] Though the complex remains quite poorly understood, it has been reported to have links with rheumatoid arthritis (RA) through inhibition of the protein complex NF-κB (nuclear factor kappa-light-chain-enhancer of activated B-cells),[237] links with neuromuscular degeneration through truncation of ASCC1 variants,[238] and links with Barrett Esophagus (BE) and through germline mutations in the ASCC1 gene.

The *Alu*-containing (AC) ASCC1 transcript (variant 1, OriGene) was cloned into pET-SUMO-28a. The gene-containing plasmid was transformed in Rosetta™ (DE3) competent cells for expression from *E. coli*. Overexpression in LB media with IPTG induction gave an insoluble protein product (figure 4.11).

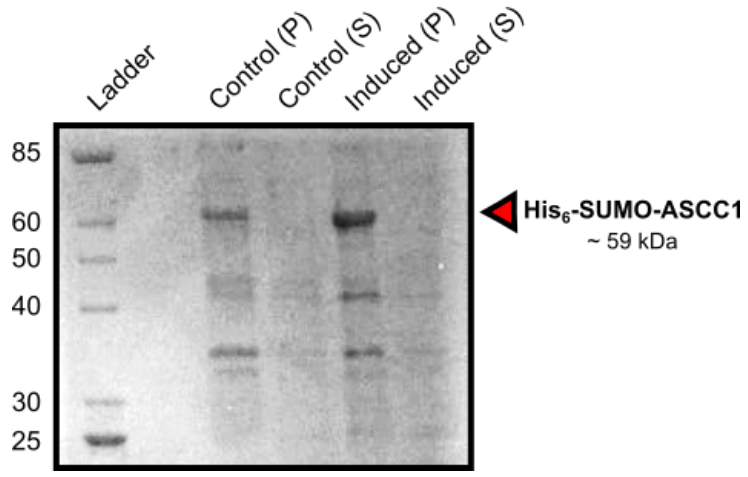


Figure 4.11 SDS-PAGE analysis of His_6-SUMO-ASCC1 overexpression with IPTG induction

Analysis showed a band at approximately 60 kDa corresponding to the poly-histidine tagged SUMO-ASCC1 fusion protein (ca. 59 kDa) as an insoluble inclusion body (Induced (P)). No soluble protein was observed.

As with previous insoluble protein products, purification was attempted *via* denaturation with urea and subsequent nickel affinity chromatography. Though purified His_6-SUMO-ASCC1 remained stable through refolding with L-arginine and subsequent dialysis to remove urea, cleavage of the SUMO tag Ulp1 yielded insoluble protein product. The product produced also appeared to be of a higher molecular weight (ca. 54 kDa) than that of the expected cleavage products; polyhistidine-tagged SUMO (ca. 13 kDa) and ASCC1 (ca. 47.5 kDa). Though ASCC1 could be purified as a SUMO fusion protein, for the type of studies we wished to perform, cleavage was necessary to ensure that folding of ASCC1 was correct and not influenced by or a result of SUMO fusion.

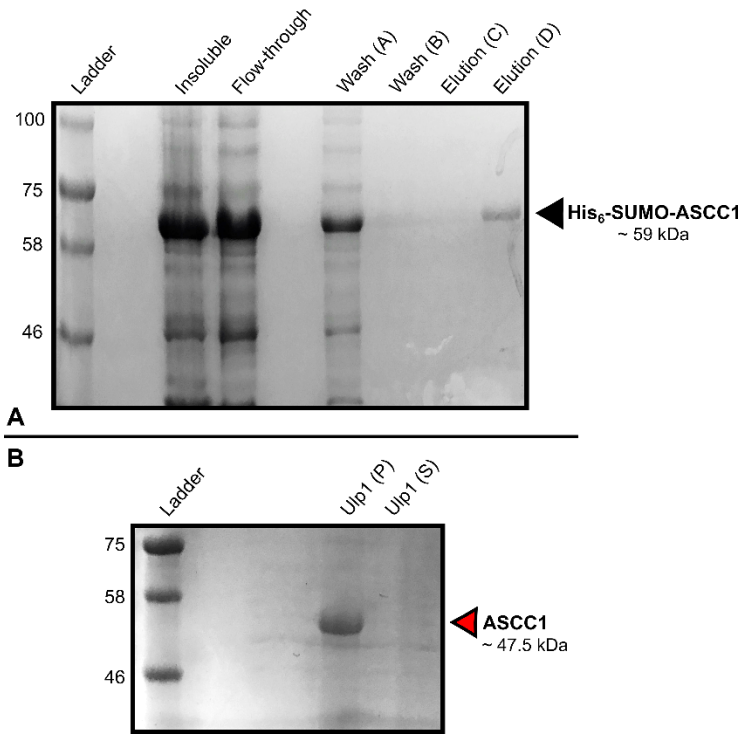
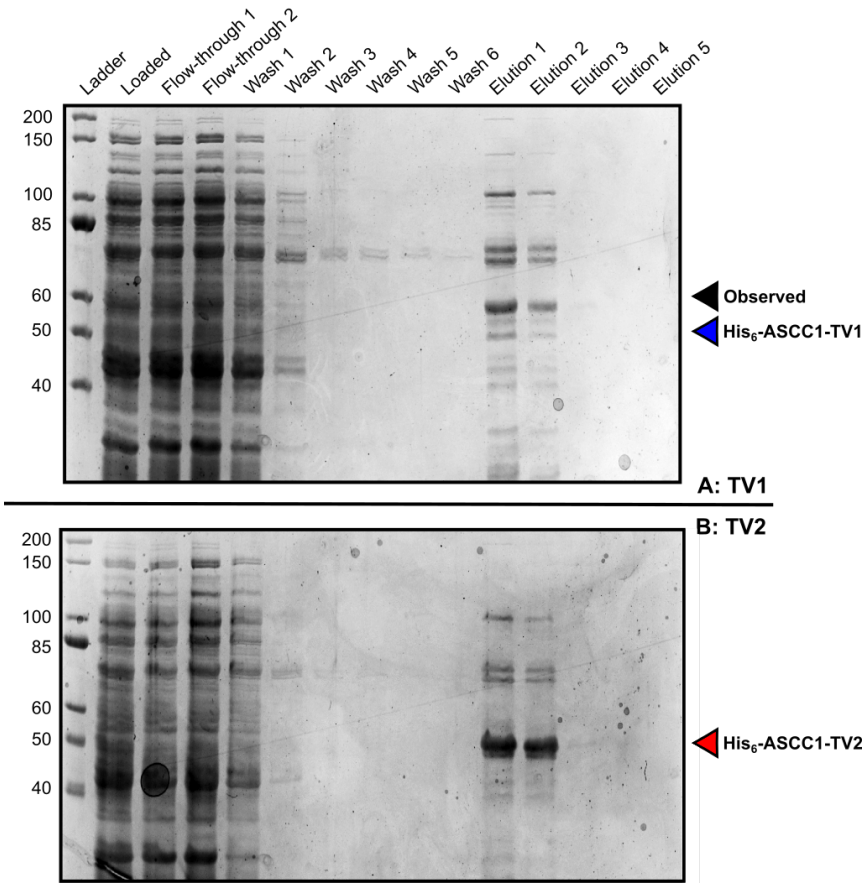


Figure 4.12 SDS-PAGE analysis of purification and subsequent cleavage of His_6-SUMO-ASCC1

(A) Denaturation of His_6-SUMO-ASCC1 (ca. 59 kDa) with 8 M urea and subsequent purification using nickel affinity chromatography yielded purified protein product which remained stable through re-folding and dialysis to remove urea. (B) Cleavage of the His_6-SUMO tag (ca. 13 kDa) from ASCC1 (ca. 47.5 kDa) resulted in the formation of an insoluble cleavage product of approximately 54 kDa (red) which corresponded to neither of the expected cleavage products.

After the observation that the non-*Alu*-containing (nAC) ASCC1 variant could be purchased as a polyhistidine-tagged protein overexpressed in *E. coli*, both the AC ASCC1 transcript and the nAC ASCC1 transcript (variant 3, OriGene) were cloned into pET-28a so as to contain a similar linker to the purchasable variant of the protein. This linker introduced a TEV (tobacco etch virus) cut site for cleavage of the polyhistidine tag. The gene-containing plasmids were transformed in *E. coli* Rosetta™ (DE3) competent cells for overexpression. Expression tests for both protein variants were carried out with both auto-induction and IPTG induction. The optimal conditions of those tested were determined to be 48 hour auto-induction in terrific broth (TB) with lysis *via* sonication in a high-salt phosphate buffer

containing dithiothreitol (DTT). Protein was purified *via* nickel affinity chromatography (figure 4.13) and subsequent size exclusion chromatography.



4.13 SDS-PAGE analysis of ASCC1 isoforms purified by nickel affinity chromatography

(A) Purification of ASCC1 transcript variant 1 (TV1; *Alu* containing) yielded a range of protein products upon column elution. The most prominent protein product was approximately 60 kDa (black) which did not correspond to His_6-ASCC1 TV1 (ca. 48 kDa). A band corresponding to this molecular weight (blue) can be observed but it was considerably fainter. (B) Purification of ASCC1 transcript variant 2 (TV2; non-*Alu*-containing) again saw a range of eluted products. A very prominent band can be observed at approximately 50 kDa however, this does not correspond to His_6-ASCC1 TV2 (ca. 43.7 kDa).

SDS-PAGE analysis showed that multiple protein products were eluted *via* nickel affinity chromatography. Interestingly, for both protein variants, the most prominent protein purified was approximately 10 kDa above the mass of the expected protein product. In the case of His_6-ASCC1 transcript variant 1 (TV1; *Alu* containing), faint

bands were observed at approximately 50 kDa which were assumed to correspond to the desired protein product (ca. 48 kDa). However, the main product was observed to be approximately 60 kDa. For His_6-ASCC1 transcript variant 2 (TV2; non-*Alu*-containing), the main protein product was observed to be approximately 50 kDa. Though faint bands were observed between 40 and 45 kDa which had the potential to correspond to the desired product (ca. 43.7 kDa), the bands could also be observed in elution products of TV1 lysate. Sequencing of the expression plasmids showed no sequence errors which would result in protein elongation and therefore, larger protein products that expected.

Elution products for both ASCC1 protein variants were further purified by size exclusion chromatography (SEC) and analysed *via* mass spectrometry (MS) for traces of the desired protein product. The trace obtained from size exclusion of His_6-ASCC1 TV1, as expected from SDS-PAGE analysis of the nickel elution product, showed multiple protein products (figure 4.14). MS analysis of products confirmed the absence of desired product in all elution fractions as no protein was observed in the range of 40 and 45 kDa.

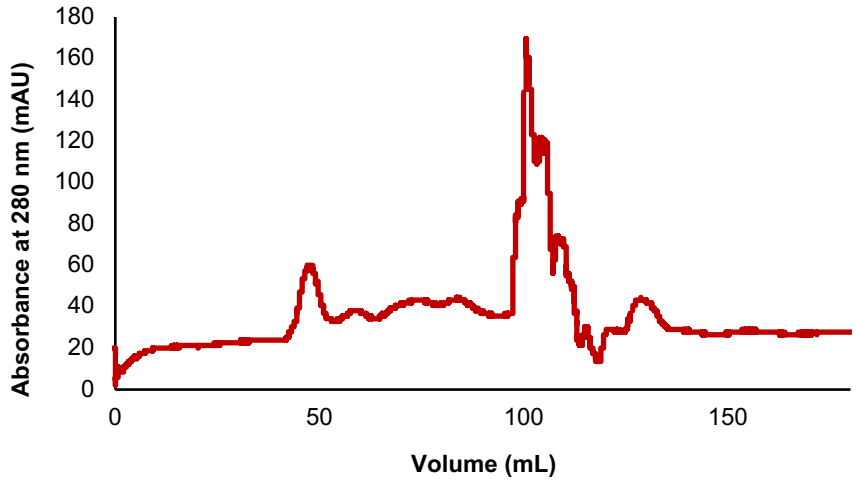
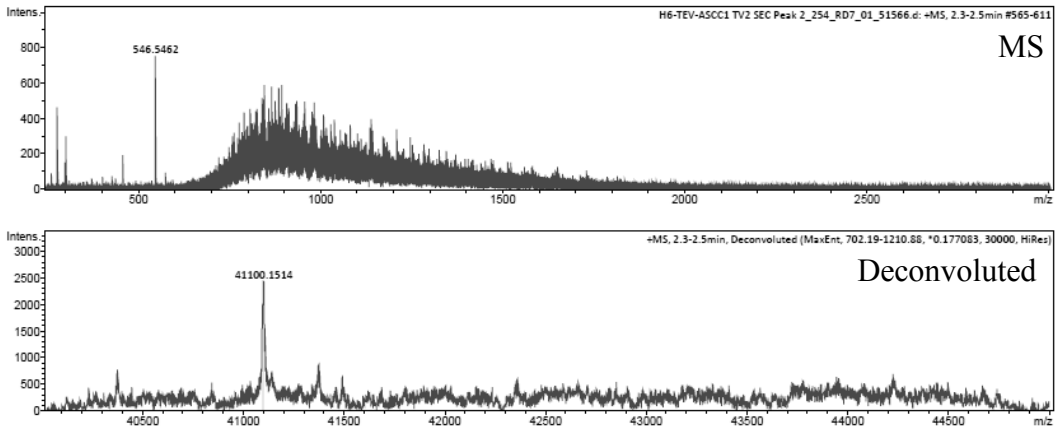


Figure 4.14 SEC trace for His_6-ASCC1 TV1

Size exclusion chromatography showed multiple protein products after purification *via* nickel affinity chromatography. Protein products were analysed *via* mass spectrometry to confirm the absence of desired protein product in all fractions.

Size exclusion of His_6-ASCC1 TV2 also showed the presence of multiple protein products. MS analysis of product peaks revealed a protein product with a molecular weight of 41100.1 Da (figure 4.15). The molecular weight of the polyhistidine-tagged ASCC1 TV2 was calculated to be 43771.5 Da, and the TEV-cleavage product had a calculated molecular weight of 41560.1 Da, neither of which corresponded to the observed mass. It was determined that the desired protein products for ASCC1 variants had not be obtained and no further avenues were explored for their expression and purification.



4.15 MS analysis of His_6-TEV-ASCC1 TV2

The expected protein mass for His_6-TEV-ASCC1 TV2 was 43771.5 Da; however, the observed protein mass = 41100.15 Da (deconvoluted) indicating the overexpression and purification of an incorrect protein product.

4.4 Conclusions on the overexpression of *Alu*-containing recombinant proteins

Five human genes were selected for cloning and overexpression in *E. coli*; NEK4, ZMAT1, PPP5D1, BCAS4 and ASCC1. Targets were selected dependent on the availability of the gene for purchase, the gene/protein's reported implications in disease, in particular cancer, and the predicted ease of expression of the protein.

Each gene was cloned into an *E. coli* expression vector using restriction free cloning aside from PPP5D1 which was purchased in a pGEX-4T-1 *E. coli* expression vector. For NEK4, ZMAT1, PPP5D1 and BCAS4, only the *Alu* containing (AC) transcript variants were cloned. In the case of ASCC1, both the AC and non-*Alu*-containing (nAC) transcript variants were cloned.

PPP5D1 was overexpressed as a fusion protein with GST and purified using glutathione agarose resin. Protein was observed to elute from the column in the wash phase indicating that the protein product was misfolded. SDS-PAGE analysis revealed a protein of approximately 26 kDa to be eluted from the column, likely corresponding to lone GST.

Expression tests of NEK4 in *E. coli* showed overexpression of a protein product of approximately 73 kDa *via* SDS-PAGE analysis. Initially, this was assumed not to be His_6-SUMO-NEK4 which had a calculated mass of approximately 108 kDa. Unfortunately, no MS data was obtained to determine the identity of the observed protein product. In hindsight, it is possible that the protein observed *via* SDS-PAGE may have been due to truncated product formation or unforeseen effects of charged residues on protein mobility. Overexpression of ZMAT1 as the fusion protein, His_6-SUMO-ZMAT1 gave an insoluble protein product when auto-induced for 96 hours. The protein was denatured with 8.0 M urea for purification *via* nickel affinity chromatography. SDS-PAGE analysis revealed elution of the desired product; however, re-folding with L-arginine, dialysis to remove urea and subsequent cleavage of the SUMO fusion tag with Ulp1 resulted in protein precipitation. His_6-BCAS4 was purified in the same way as His_6-SUMO-ZMAT1; however, refolding with L-arginine and dialysis to remove excess urea again resulted in the precipitation of the protein product.

Initially, expression tests for ASCC1 were attempted by which the AC variant of ASCC1 was over expressed as part of a fusion protein with polyhistidine-tagged SUMO. Overexpression resulted in the formation of an insoluble protein product and so purification *via* denaturation with urea was performed. Though the protein remained soluble through re-folding with L-arginine and dialysis to remove excess urea, cleavage of the His_6-SUMO tag resulted in protein precipitation. After observing that the nAC isoform of ASCC1 could be purchased as a recombinant protein from *E. coli*, both the AC and nAC variants were re-transformed into a pET-28a vector and mutated to contain a TEV cut-site to mimic the purchasable variant. Both variants were overexpressed and purified *via* nickel affinity chromatography followed by size exclusion chromatography. SDS-PAGE analysis after nickel elution showed the main elution products in both cases to be of a higher molecular weight than the expected product. In the case of His_6-ASCC1 TV1, a faint band at approximately 50 kDa was believed to be the desired product; however, size

exclusion and subsequent MS analysis of eluted proteins showed no trace of the desired product. MS analysis of protein products obtained from overexpression of His_6-ASCC1 TV2, revealed a product at 41100 kDa. Unfortunately, this corresponded neither to the polyhistidine-tagged protein (ca. 43.7 kDa) nor the TEV-cleavage product (41.5 kDa). Plasmid sequencing for both constructs revealed no mutations which would account for protein elongation.

For the most part, only *Alu*-containing isoforms of protein were used in expression trials, with the exception of ASCC1 for which overexpression of both AC and nAC variants was attempted. As previously stated, only the nAC variant of ASCC1 is commercially available as a purification product from *E. coli*. It is possible that the presence of the *Alu* in protein isoforms may affect the protein in such a way that it destabilises the protein sufficiently to make purification difficult. In most cases, overexpression in an *E. coli* expression system was successful and most difficulties lay in the purification of the selected proteins. Proteins precipitated at various steps of purification and as a result, final purified protein products could not be produced. Due to the difficulty in purification of the chosen human targets, it was determined that a different route would be taken to study the effect of *Alu* elements on the structure and function of proteins. As a result, no further work was carried out on the overexpression of human proteins.

Chapter 5
MBP as a model system for *Alu* expression

Overview

Due to the difficulties arising from the overexpression and purification of naturally occurring *Alu*-containing proteins, maltose binding protein (MBP) was used as a model system. The SEA-001 *Alu* sequence (discussed in Chapter 2) was cloned into eight different regions within MBP and studied for its effect on the protein folding and subsequent binding to ligands; D-(+)-maltose, maltotriose and β-cyclodextrin. The effect of translated *Alu* elements on the structure and function of MBP was studied through the use of protein expression studies, circular dichroism, differential scanning calorimetry and amylose purification studies.

5.1 Overexpression of MBP-*Alu* protein mutants

As an alternative to the expression of naturally-occurring *Alu*-containing proteins from *E. coli*, a well-studied and easy to express protein could be mutated to contain the *Alu* insertion as identified by bioinformatics. The gene for maltose binding protein (MBP) was mutated to give eight different variants, each containing the *Alu* sequence, SEA-001, at a different location within its sequence. Three additional mutants were cloned to contain a 'scrambled' variant of the *Alu* element. Each variant, in addition to the wild-type, was overexpressed and purified for use in folding and binding studies.

5.1.1 MBP as a model system for investigating translated *Alu* elements

Maltose binding protein (MBP) is an approximately 43 kDa periplasmic protein expressed from the *malE* gene of *E. coli*. It is involved in the transport of maltose in *E. coli* and as such, it has several potential high affinity ligands including maltose, maltotriose and β-cyclodextrin.[239] It is often utilised as a fusion partner in the overexpression of recombinant proteins due to its high solubility.[240] Due to ease of overexpression at high concentrations in *E. coli*, multiple binding partners[241] and good solubility, MBP (figure 5.1) was chosen as a model protein to study the effects of translated *Alu* elements in proteins.

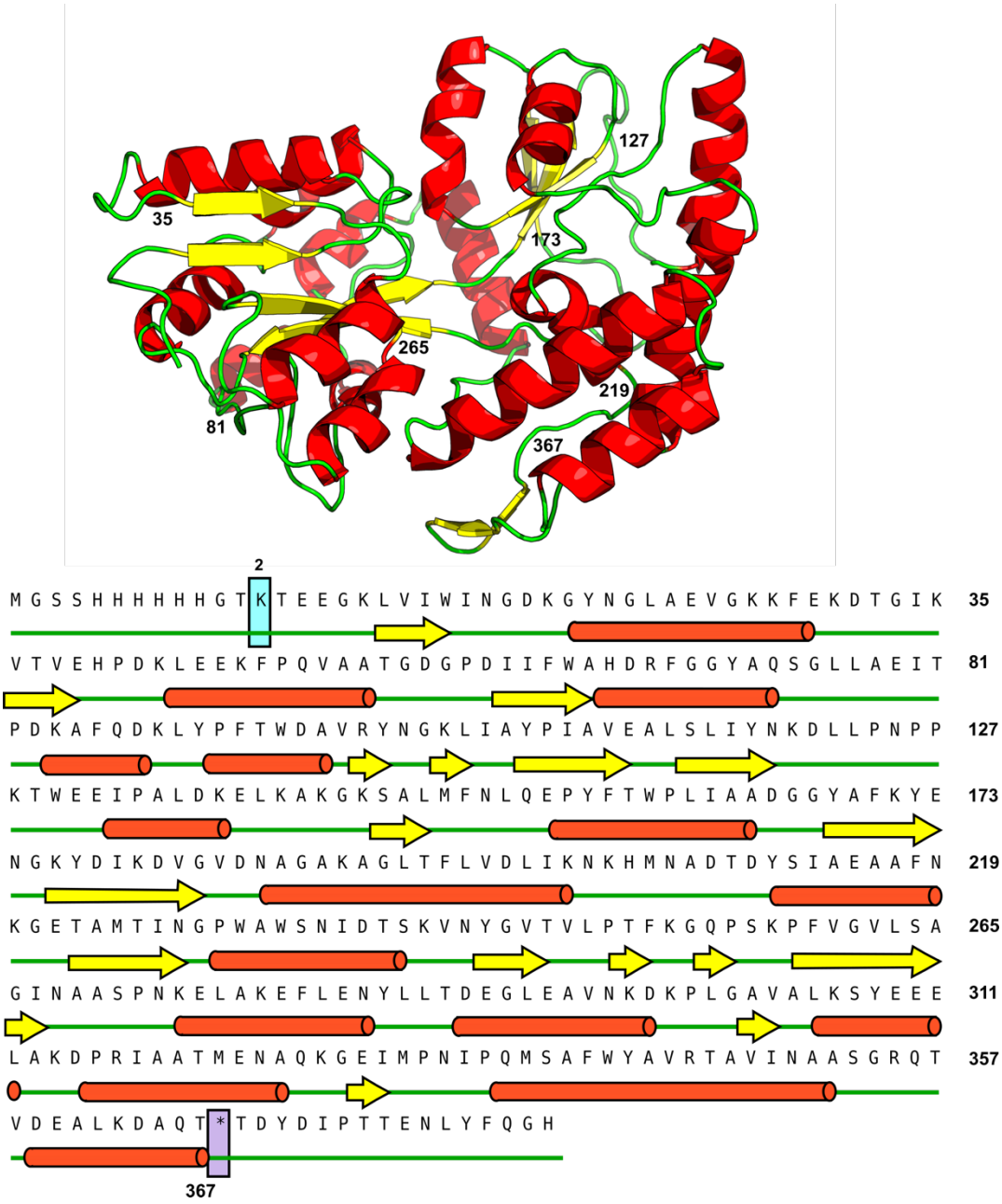


Figure 5.1 Structure of MBP

The crystal structure of maltose binding protein (MBP) when bound to maltose determined by crystallography (1ANF)[242] and breakdown of secondary structure. For this study an N-terminal His_6 tag was used; therefore, MBP begins at the Lys2 (K2) residue. A C-terminal stop codon was also introduced.

5.2 Site-directed mutagenesis of MBP-*Alu* constructs

Site-directed mutagenesis (SDM) was performed to introduce a C-terminal STOP codon into a commercially available pDB.His.MBP vector (DNASU; ID:

eVNO00085130) so as to encode a polyhistidine-tagged MBP variant. Further polymerase chain reactions (PCR) were performed on the resulting plasmid to introduce *Alu* sequences.

Alu insertions were introduced into the MBP gene *via* PCR-based SDM, derived from an inverse PCR technique.[243] Eight mutants were cloned to contain the desired *Alu* sequence, SEA-001. Three mutants were cloned to contain a scrambled variant of the desired *Alu* sequence, SEA-002.

SEA-001	L E C S G A I S A H C N L R L L G S S D S P A S A S R V A G I T G
SEA-002	I A R L H G P S A S N G T S S S T C A P D L G V G E S A L C I S R

A total of twelve constructs were subcloned including the wild-type (WT) MBP construct.

5.2.1.1 Site-directed mutagenesis

All site-directed mutagenesis was performed using a Quikchange mutagenesis kit (Agilent). Primers for introduction of a STOP codon into the pDB.His.MBP vector were designed so as to be complementary to one another and overlap the site of insertion (figure 5.2).[244] Following non-exponential amplification of the parental plasmid, template DNA was digested with Dpn1 and PCR products were transformed into *E. coli* XL1-Blue supercompetent cells. Plasmids were isolated and confirmed *via* sequencing of the whole gene.

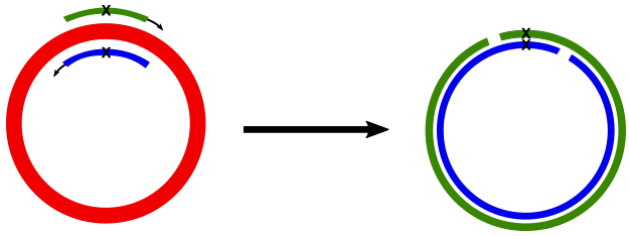


Figure 5.2 SDM to introduce a STOP codon into pDB.His.MBP

The three base STOP codon, TAG (**X**), was central to both the forward and reverse primers which were designed complementary to one another. PCR was performed using Quikchange mutagenesis with extension around the template plasmid to give the product, pDB.His.MBP.STOP, which required no ligation prior to transformation.

Introduction of *Alu* insertions was achieved *via* an adaption of inverse PCR using site-directed mutagenesis. All *Alu* and scrambled *Alu* insertions were introduced into the pDB.His.MBP.STOP vector. Primers pairs were designed so that one half of the desired 99 bp insertion sequence was contained within each of the forward and reverse primers, in addition to a sequence corresponding to the desired insertion site (figure 5.3).

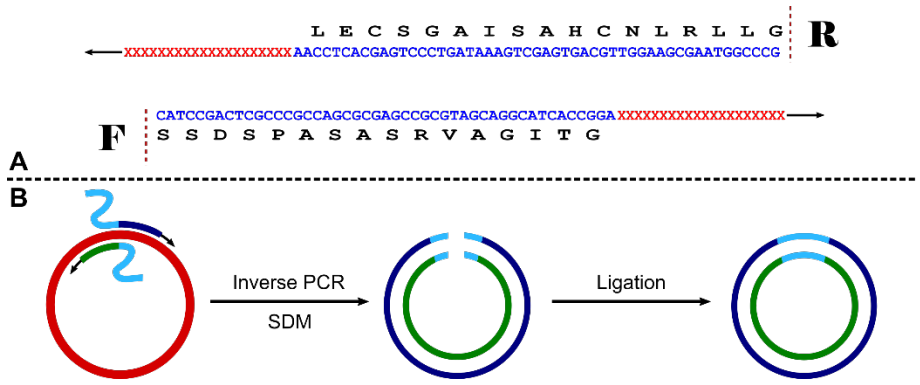


Figure 5.3 Outline of SDM as an adaption of inverse PCR

Alu insertions were introduced into pDB.His.MBP.STOP *via* an adapted inverse PCR method. (A) Forward and reverse primers were designed to each contain half of the desired 99 bp insertion sequence and a sequence matching the desired insertion site and (B) were annealed, extended around the outside of the vector template and amplified to give a PCR product containing the full insertion. The PCR product was annealed with T4 ligase and treated with Dpn1 to remove parental plasmid prior to transformation.

PCR was performed *via* mutagenesis by which primers were extended around the vector template to give a construct which was treated with T4 polynucleotide kinase (PNK) and ligated prior to Dpn1 digestion and subsequent transformation. Prior to sequencing, constructs were checked for the desired insertion *via* enzyme double digests. Samples were cut with restriction enzymes and analysed on a 1% agarose gel to observe digestion products of successful and unsuccessful mutagenesis reactions. A difference of 99 bp was observed between digestion products (figure 5.4). Samples which showed the correct digestion product were sequenced for confirmation of successful mutagenesis.

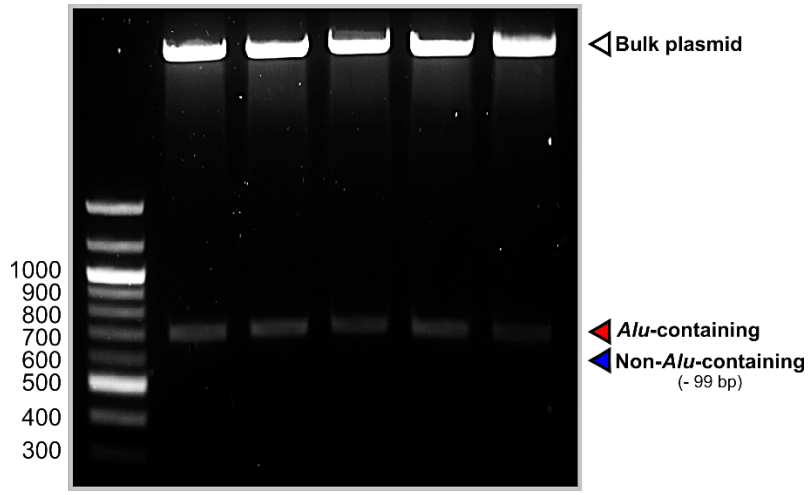


Figure 5.4 Example of enzyme double digest (His_6-MBP-T81)

Double digest analysis of *Alu* insertions in MBP constructs were performed with enzymes corresponding to the site of insertion. Successful mutagenesis reactions gave rise to a digestion product 99 bp larger than those that were unsuccessful. Note: Non-*Alu* refers to where the band would be observed – all reactions were successful.

5.2.1.2 Cloned MBP-*Alu* constructs

A total of twelve MBP constructs (table 5.1) were cloned *via* site-directed mutagenesis. A STOP codon was introduced into pDB.His.MBP (figure 5.4) which served as the expression vector for wild-type (WT) MBP.

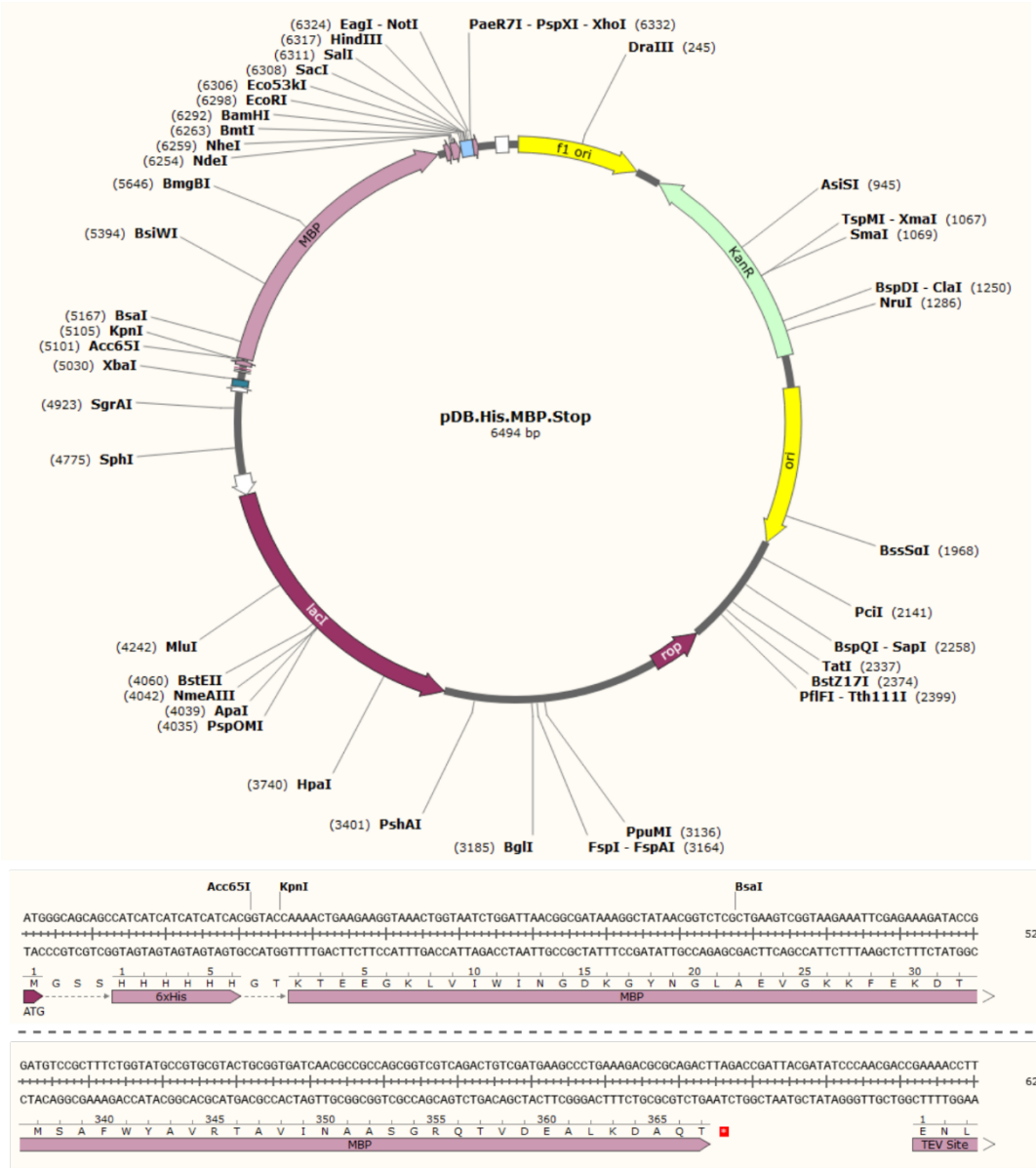


Figure 5.5 Vector map for pDB.His.MBP.STOP

The plasmid encodes a His_6 purification tag prior to the *malE* (MBP encoding) gene. A C-terminal stop codon was introduced at the end of the gene.

Insertion Location	Insertion Type	Predicted Protein
None	None	His_6-MBP
G6	*Alu*	His_6-MBP-G6
T81	*Alu*	His_6-MBP-T81
P126	*Alu*	His_6-MBP-P126
D178	*Alu*	His_6-MBP-D178
G253	*Alu*	His_6-MBP-G253
A293	*Alu*	His_6-MBP-A293
N333	*Alu*	His_6-MBP-N333
T367	*Alu*	His_6-MBP-T367
D178*	Scrambled *Alu*	His_6-MBP-D178*
G253*	Scrambled *Alu*	His_6-MBP-G253*
N333*	Scrambled *Alu*	His_6-MBP-N333*

Table 5.1 Cloned MBP constructs

Constructs were cloned *via* site directed mutagenesis. The WT construct was cloned through introduction of a STOP codon to pDB.His.MBP. The resulting vector (pDB.His.MBP.STOP) was mutated to contain *Alu* insertions at eight different sites. Three of the same sites were also used to introduce scrambled *Alu* insertions. Note: * refers to scrambled *Alu* sequences.

Eight *Alu* insertions of LECSGAISAHCNLRLLGSSDSPASASRVAGITG (SEA-001) and three scrambled *Alu* insertions of the sequence IARLHGPSASNGTSSSTCAPDLGVGESALCISR (SEA-002) were cloned into the pDB.His.MBP.STOP vector. Constructs were named after the residue of MBP encoded directly after the insertion sequence. In the case of pDB.His.MBP.STOP, the polyhistidine insert, GSSHHHHHHGT, does not contribute to the numbering of residues. Insertions were generally confined to parts of the gene which resulted in expression as part of a loop within the proteins secondary structure, with the exception of D178/D178*, which was located within a β-sheet and A293 which was located within an α-helix (figure 5.6).

Retrospective analysis of the tolerance of protein secondary structure in Chapter 2 revealed that though the majority of Alu insertions did arise in coiled regions, many were tolerated within alpha helices and some were even tolerated within beta strands. This practical work discussed in the Chapter worked with MBP constructs in which Alu insertions were placed in coiled regions of the protein in most cases (exceptions: A293, D178) as it was theorised that these would be less likely to be detrimental to protein expression. With the observed tolerance of insertions in mind, it may have been possible to place more insertions in alpha helices, and maybe some beta strands,

within the protein structure and still have obtained successful expression. Bioinformatic analysis shows that *Alu* insertions never arose near known structural domains of protein, such as binding sites, and as such, it is likely that any insertion in the binding site would not be tolerated, as originally predicted.

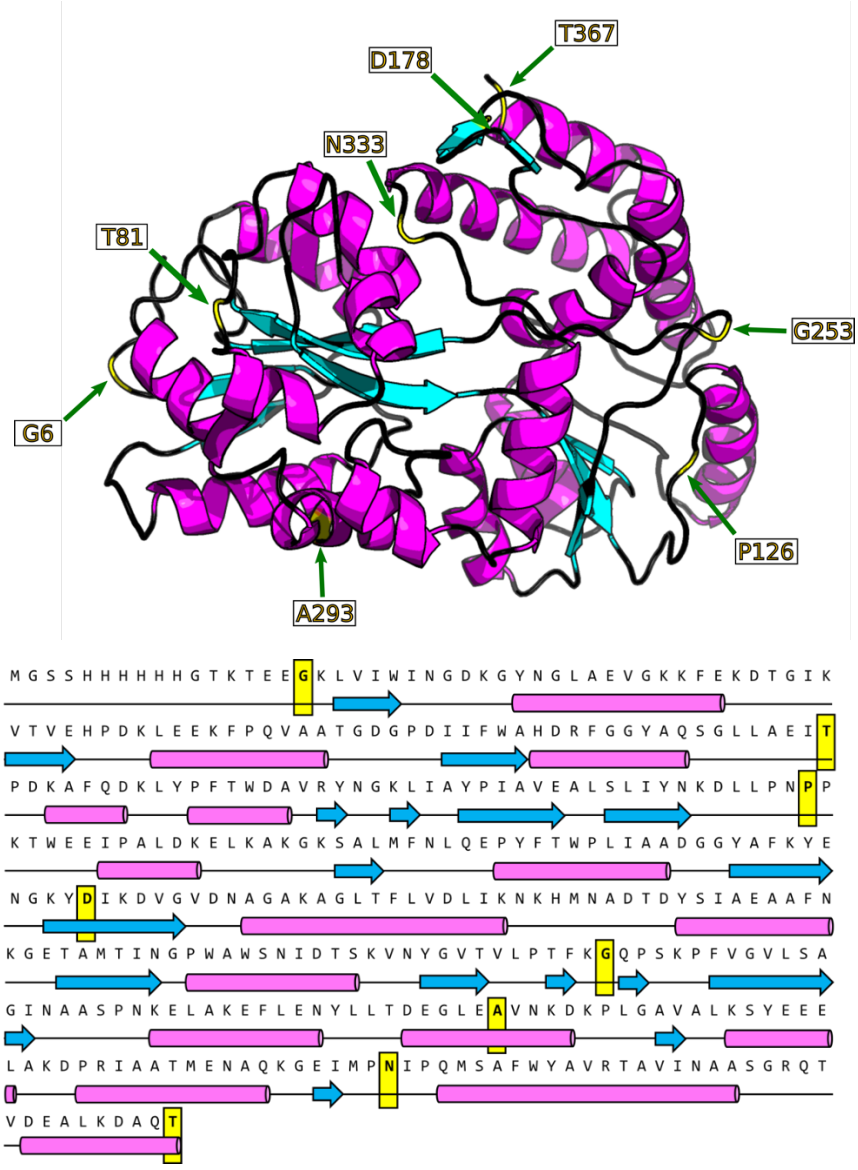


Figure 5.6 Locations of *Alu* insertions in the secondary structure of MBP

A total of eight locations within MBP were chosen to host *Alu* insertions. Most were located within protein loops, with the exception of D178, which was located within a β-sheet and A293, which was located within an α-helix. Scrambled *Alu* sequences were also inserted at D178, G253 and N333 locations. (Note: numbering refers to the original MBP structure without a polyhistidine tag - the first lysine residue observed is K2).

5.3 Overexpression and purification of MBP-*Alu* protein mutants

Wild-type MBP and MBP-*Alu* mutants were overexpressed in *E. coli* Rosetta cells *via* auto-induction at 25 ºC for approximately 48 hours. Cells were lysed *via* sonication and purified either by manual nickel affinity chromatography coupled with automated size-exclusion chromatography (SEC), or by fast protein liquid chromatography (FPLC) using a double column (Crude HisTrap followed by SEC) method. Protein products were analysed by SDS-PAGE (figure 5.7) and MS (figure 5.8) for purity and confirmation of the correct protein product.

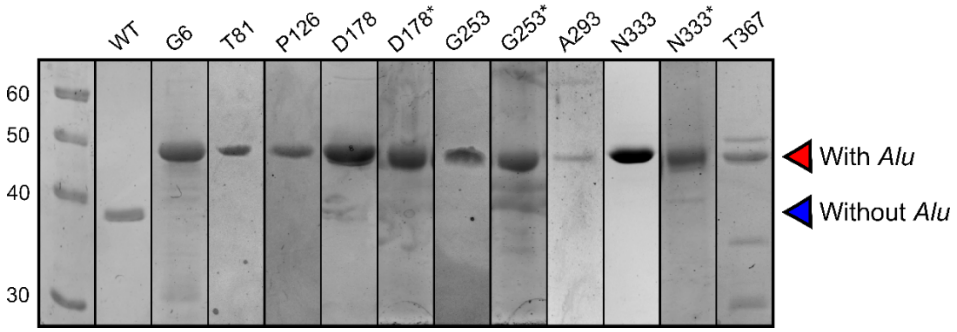


Figure 5.7 SDS-PAGE analysis of purified MBP variants

Proteins were purified *via* nickel affinity and size exclusion chromatography and analysed *via* SDS-PAGE. *Alu*-containing MBP constructs were approximately 44.6 kDa and wild-type MBP was approximately 41.4 kDa. Note: structures contained scrambled A*lu* sequences are marked with an Asterix (*).

Analysis by mass spectrometry observed the correct masses (ca. 44.6 kDa) for MBP variants G6 through to N333* when taking into account cleavage of the N-terminal formylmethionine (- 160 Da). Oddly, wild-type (WT) MBP and the T367 variant of MBP had an observed mass approximately 112 Da lower than their calculated masses of 41.4 and 44.6 kDa, respectively, when taking into account the cleavage of the N-formylmethionine. As plasmid sequencing had confirmed the absence of any unwanted mutations, the loss in mass was attributed to small truncations. For the wild-type, this corresponded to the loss of an N-terminal glycine and a C-terminal threonine. For the T367 variant, the loss corresponded to a loss of the final Thr-Gly residues of the *Alu* insertion at the C-terminus. As the truncations were small, protein products were carried through to further analyses.

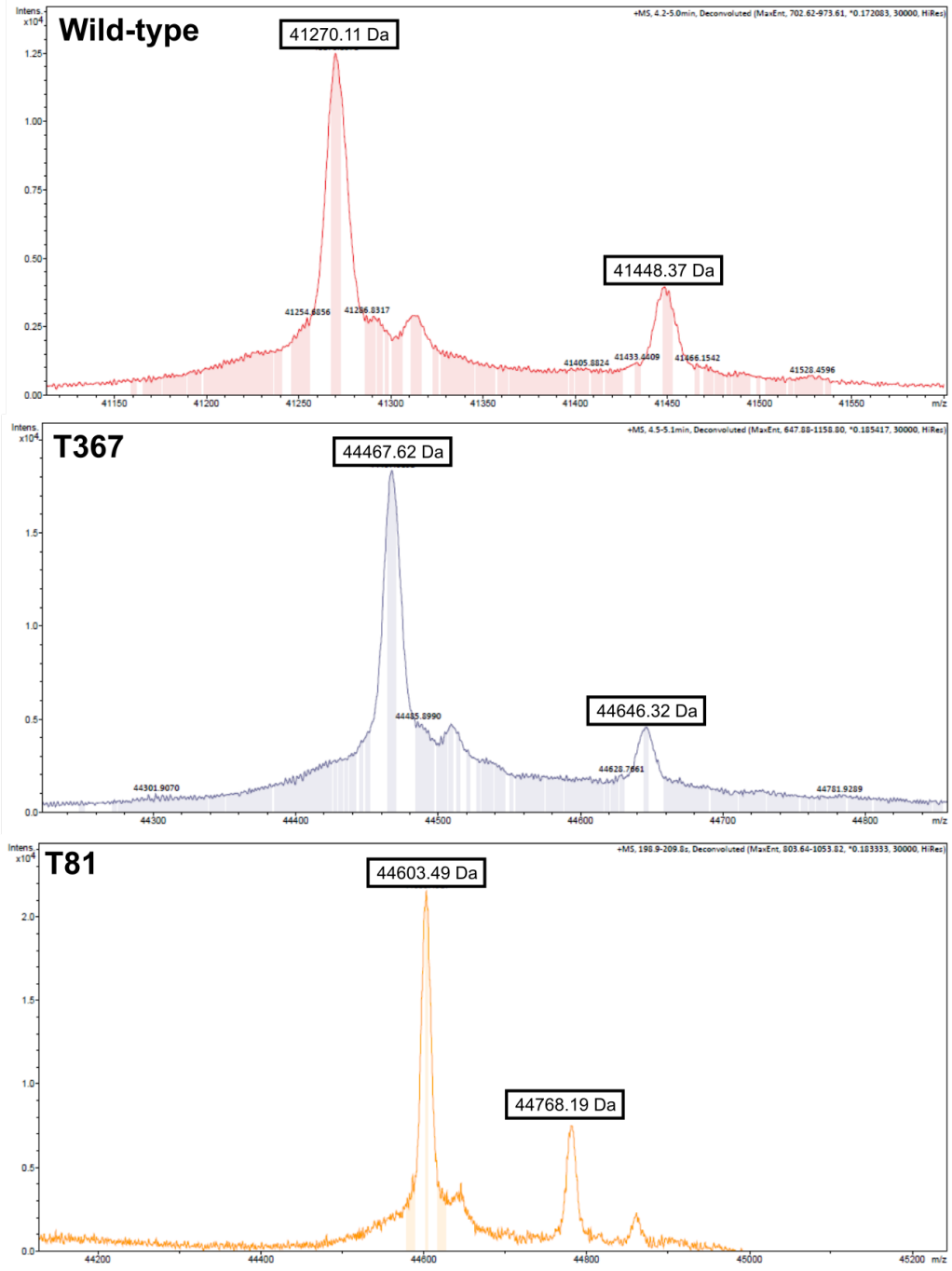


Figure 5.8 MS analysis of MBP constructs

Wild-type MBP showed a mass of 41,270 kDa, 112 kDa less than the predicted 41.4 kDa, corresponding to a loss of an N-terminal glycine and a C-terminal threonine. Similarly, the 112 kDa loss in mass observed for the T367 variant corresponded to a loss of the final Thr-Gly residues of the C-terminal *Alu* insertion. MS for all other *Alu*-MBP constructs yielded the correct mass and looked similar to the spectra observed for the T81 variant. Full MS data for all variants can be found in Appendix 4.

5.4 The effect of a translated *Alu* on MBP overexpression, folding and stability

5.4.1 Effect of *Alu* insertions on MBP overexpression

Expression tests were performed in order to determine whether the introduction of an *Alu* insertion, and the respective site of *Alu* insertion, affected the overexpression levels on MBP variants. Cells were grown in LB, 5 mL of each cell solution was adjusted to OD_{600} = 0.75 ± 0.05 and cells were induced with IPTG and incubated overnight. Cells were pelleted, lysed and analysed *via* SDS-PAGE. SDS-PAGE samples were prepared with equal amounts of loading buffer and lysate then equal volumes of each sample were loaded on to the gel to maintain consistency between samples. SDS-PAGE gels were stained with Coomassie blue and analysed *via* densitometry (figure 5.9).

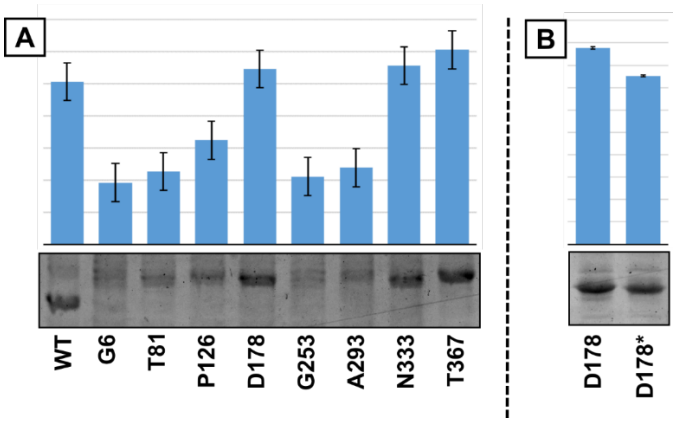


Figure 5.9 Densitometric analysis of MBP variant overexpression

(A) D178, N333 and T367 variants showed similar overexpression to wild-type MBP. All other variants were overexpressed at lower levels. (B) There was no difference in the overexpression of scrambled and non-scrambled *Alu* variants of the D178 construct. Error bars are calculated from the standard deviation of three biological repeats. No statistical difference was found between any of the MBP variants and the control ($p < 0.05$). Note: constructs marked with an Asterix (*) contain scrambled *Alu* sequences.

Insertions at D178, N333 and T367 sites made very little difference to the overexpression of MBP. Insertions at all other sites led to reduced overexpression. In five of eight cases, significant reduction in overexpression was observed. This result was further confirmed during large scale over expression (1 L) by which large

amounts of WT, D178 and T367 variants were obtained at a concentration of > 1.0 mM at volumes exceeding 1 mL. 1 L overexpression of other MBP variants yielded purified protein products of < 500 μM at lower volumes. In particular, only 100 μL of 25 μM A293 was obtained from 1 L of media. Reduced overexpression with the A293 variant was expected due to the insertion being located in an α-helix. As a reduction in overexpression was not observed in all cases, it was likely that the difference was due to the location of the insertion and was independent of the insertion sequence. In order to confirm this hypothesis, scrambled variants of D178, G253, and N333 were analysed. Unfortunately, overexpression of G253 and N333 variants was too weak to be studied under these conditions. However, minimal change in overexpression between D178 and D178* MBP variants confirmed that changes in overexpression were sequence independent.

5.4.2 Folding and stability analysis through circular dichroism

Circular dichroism is a biophysical technique which can be used to predict the secondary structure of proteins through the linear combination of absorption measurements between 180 and 260 nm wavelengths. In combination with a temperature gradient, it can also be used to predict the melting temperature (T_m) of a given protein sample and hence, predict thermal stability. MBP variants were analysed *via* CD to primarily, predict any distinct secondary structure attributed to *Alu* insertions and secondarily, to determine the effect of *Alu* insertions on the thermal stability of proteins.

5.4.2.1 Circular dichroism

Circular dichroism (CD) is an established method for the prediction of protein conformation and associated conformational changes in solution.[245] It refers to differential absorption of the left-handed (counter-clockwise, L) and right-handed (clockwise, R) circularly polarised components of plane polarised light.[246] Normally, the magnitudes of L and R circularly polarised light are equal. However, upon passage of light through a sample which possesses chirality (e.g. a protein), the absorbance magnitudes of L and R are no longer equal, and a CD signal is observed.[247]

CD spectra are obtained by plotting the ellipticity (θ, º) as a function of wavelength (λ, nm) which generally spans a range of approximately 180 – 260 nm, where possible.[248] The CD spectra obtained from protein samples primarily arise from the

far-UV absorption of the amide backbone.[249] Three distinct regions in the CD spectra of proteins can be attributed to secondary structure.[250] Two clear negative absorptions can be observed at approximately 222 nm and 208 nm, which correspond to α-helix folding. These arise due to $n \rightarrow \pi^*$ electron transitions of the amide bond. The third absorption is an intense positive absorption at approximately 192 nm, arising as a result of $\pi \rightarrow \pi^*$ electron transitions and corresponding to β-sheet folding.[251] These β-sheet-associated absorptions also overlap with negative random coil-associated absorptions at approximately 198 nm.

Protein samples consist of a mixture of all three structural elements, the ratio of which varies between proteins and protein states (e.g. ligand-bound). CD data analysis assumes that the CD spectrum of a protein is a linear combination of each of its secondary structural elements. As a result, three main regions of a proteins CD spectra are studied in relation to one another to predict the percentage of α-helical, β-sheet and random coil character (figure 5.10).

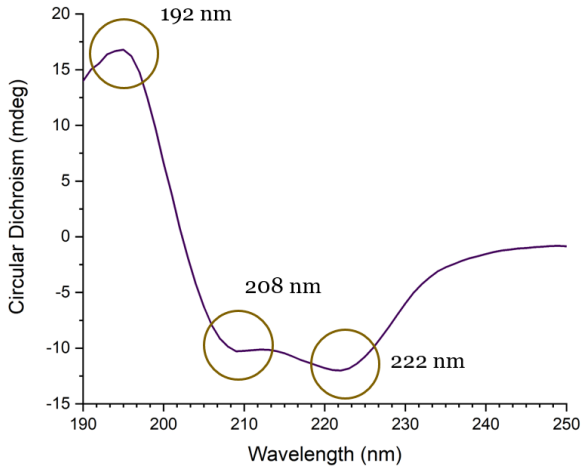


Figure 5.10 The linear combination of secondary elements in protein CD

Protein CD assumes that the spectra of a given protein sample is the linear combination of the protein's secondary structural elements; α-helix, β-sheet and random coil. The comparison of three distinct absorptions (192 nm, 208 nm and 222 nm) with respect to one another allows for prediction of the secondary structure of a sample.

Degradation of a protein sample results in a decrease in absorption. Generally, this leads to a decrease in α-helical and β-sheet CD signals. The introduction of a temperature gradient to CD experiments and the subsequent measurement of the change in secondary character can be used to predict thermal stability of a protein sample.[252] A plot of the concentration of folded versus unfolded protein as a function of temperature reveals the calculated melting point (T_m, °C) of a given sample and as a result, gives an indication of its thermal stability.

5.4.2.2 The effect of a translated *Alu* on the secondary structure of MBP

CD spectra of wild-type MBP (WT), *Alu*-containing variants and scrambled *Alu* variants were obtained using a Chirascan™ CD spectrometer (Applied Photophysics) available through the Astbury Centre of Structural Molecular Biology (University of Leeds). Protein samples (0.2 mg/mL) in 50 mM sodium phosphate (pH 7.5) were scanned over a wavelength range of 180 – 260 nm. Comparison of the resultant CD spectra for each variant showed minimal difference from the wild-type spectra. This indicated no overt effect of the *Alu* insertion on the secondary structure of MBP and also indicated that the *Alu* sequence itself was unlikely to present its own distinct secondary structure. It also confirmed correct protein folding of each of the MBP-*Alu* variants in most cases.

In general, little to no change in $\theta_{222}/\theta_{208}$ ratio was observed for most *Alu* variants. The A293 α-helix appeared to have the most effect on secondary character. A slight decrease in the $\theta_{222}/\theta_{208}$ ratio from 1.22 (WT) to 1.05 (A293) was observed, which would be expected upon direct disruption of α-helical character (i.e. insertion of an predominantly unstructured *Alu* sequence into the centre of an α-helix). Interestingly, insertion of an *Alu* sequence at the G253 position also showed a decrease in $\theta_{222}/\theta_{208}$ from 1.22 to 1.05, similar to that observed for the A293 α-helical insertion. However, introduction of a scrambled *Alu* at G253 positions seemingly led to a further reduction in $\theta_{222}/\theta_{208}$ to 1.00 (figure 5.11). This indicated that simple insertion of a sequence at this position was enough to partially disrupt its α-helical character. It should be noted that the effect was not sequence specific.

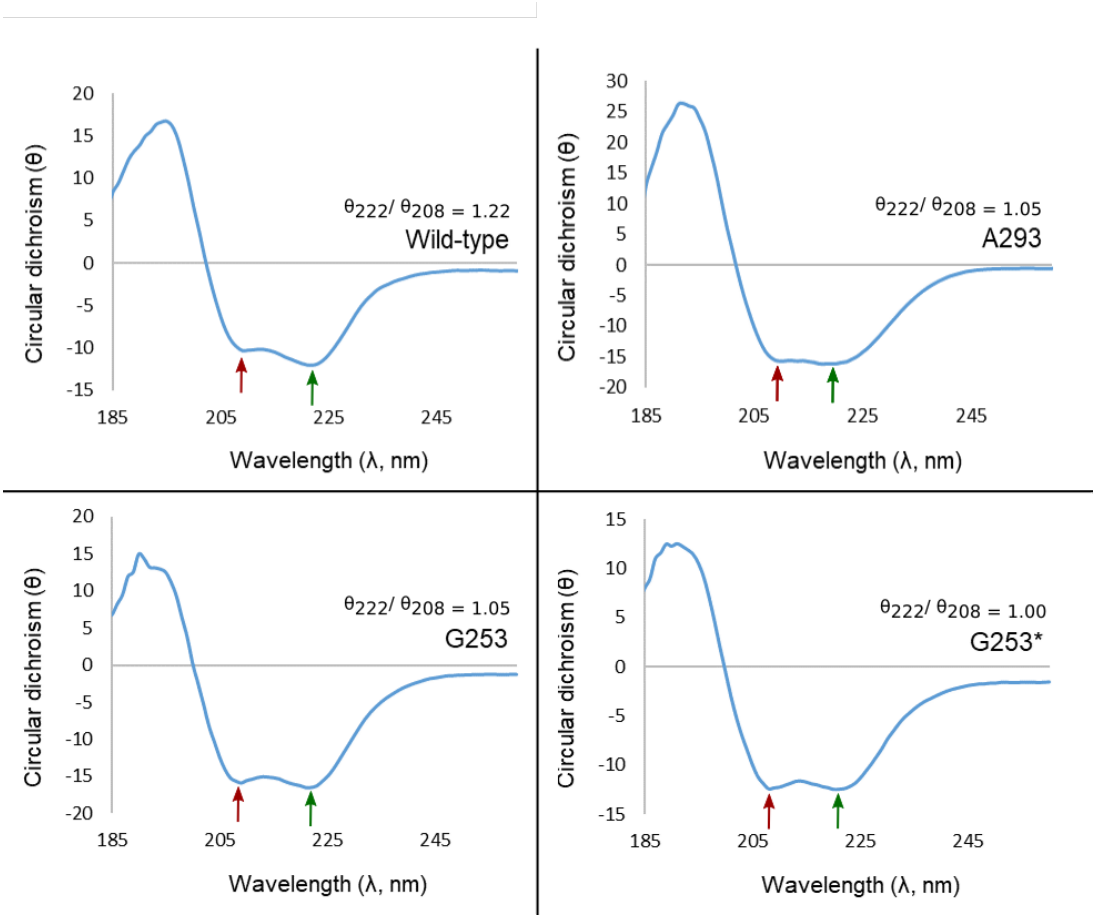


Figure 5.11 Observed changes in $\theta_{222}/\theta_{208}$ for MBP variants A293, G253 and G253*

Generally, most CD spectra observed were similar to that of wild-type MBP with minimal changes of up to 0.08 in $\theta_{222}/\theta_{208}$ ratio. Insertion of an *Alu* sequence at position A293, located within an α-helix, showed an expected change in α-helical character. Insertions of both *Alu* and scrambled *Alu* sequences at the G253 position result in a similar change in $\theta_{222}/\theta_{208}$. Note: constructs marked with an Asterix (*) contained scrambled *Alu* sequences.

Though CDNN (Circular Dichroism analysis using Neural Networks; Böhm 1997) and DichroWeb software[253] were used in an attempt to predict secondary structure, both programmes provided inaccurate analysis of α-helical content of MBP variants, including wild-type MBP, when compared with structural data recorded in the Protein Data Bank (PDB). As a result, any predictions made using either software were disregarded.

5.4.3 The effect of a translated *Alu* on the thermal stability on MBP

Two methods were used to determine the thermal stability of MBP variants and, by extension, any changes in stability upon insertion of *Alu* sequences. Melting points of all variants were determined using CD. Differential scanning calorimetry (DSC) of the wild-type and D178 variant was performed as a means to determine the accuracy of CD T_m calculations.

5.4.3.1 Melting point analysis using CD

Melting point analysis of MBP variants was performed using Global3 software (Applied Photophysics). CD spectra of protein samples were obtained over a temperature gradient of 20 – 70 °C. Subsequently, folded (%) and unfolded (%) protein was plotted against temperature to give an intercept equivalent to the T_m of the sample. However, the T_m of all *Alu*-containing MBP variants were determined to be higher than that of the wild-type protein. Further examination of CD data at 20 °C and 70 °C showed that even at 70 °C, proteins were not fully unfolded (figure 5.12). Even with the A293 MBP variant, which would be expected to be destabilised due to the location of the insertion within a protein α-helix, α-helical character is still observed at 70 °C. As Global3 assumes that the end point of the temperature gradient, in this case 70 °C, corresponds to a fully unfolded protein (with no α-helical character), T_m values calculated using this software were dismissed.

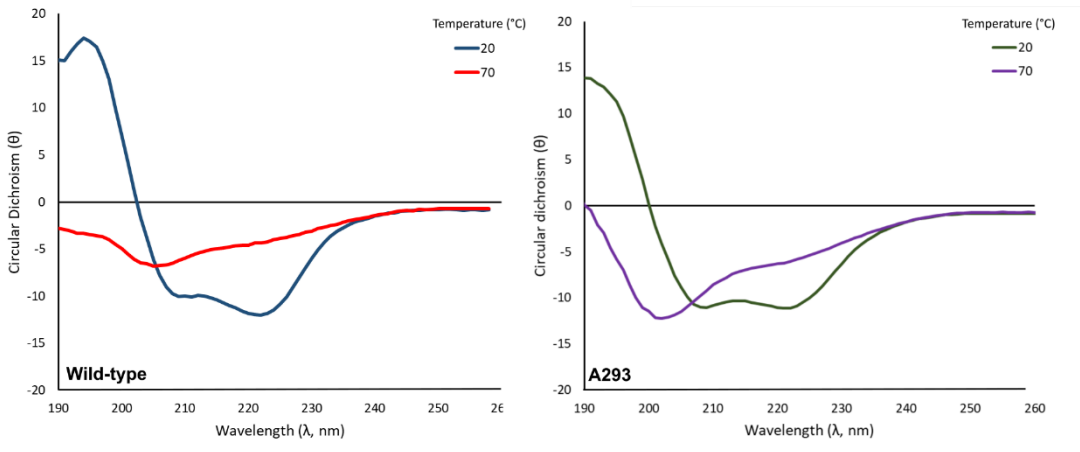


Figure 5.12 CD spectra for folded and unfolded WT and A293 variants

At 70 °C, proteins are still observed to have some α-helical character and thus, accurate T_m determination, which assumes full unfolding at the end of the temperature gradient, was not possible with CD data.

5.4.3.2 Thermal stability determination using DSC

Since CD data was insufficient to determine melting point, WT and D178 variants were analysed *via* differential scanning calorimetry (DSC) as a means to measure any difference in thermal stability upon *Alu* insertion. D178 was used as a comparison to wild-type MBP as this insertion lies at the end of a distinct secondary structure (β-sheet) within MBP and as such, would be expected to exhibit a decrease in thermal stability.

DSC is a biophysical method which can be used to assess the thermal stability of proteins upon mutation,[254] ligand-binding[255] and folding.[256] It measures the heat flow applied to a sample at constant $\Delta T/\Delta t$. Upon melting, a sample requires the application of more heat in order to maintain constant $\Delta T/\Delta t$ and to accommodate the endothermic state change and as a result, the heat capacity (Cp) is higher at this point. Measuring Cp as a function of temperature results in a sharp peak at the samples melting point. For a more thermally stable protein, the maximum of this peak would be observed at a higher temperature than a less thermally stable sample. In contrast, the exothermic aggregation of protein results in a dip in Cp.[257]

DSC was performed on 1.0 mg/mL protein samples in a buffer of 15 mM phosphate, 50 mM NaCl at pH 7.2. The predicted result was that wild-type MBP should yield a higher T_m and also a more stable re-folding pattern when cooled before the point of aggregation, when compared to D178. Heating of samples from 10 – 90 ºC, past the point of aggregation, revealed a slightly lower T_m for D178 of approximately 50 °C when compared to wild-type MBP with a T_m of approximately 53 ºC. However, both variants had a similar aggregation temperature at around 75 ºC (figure 5.13).

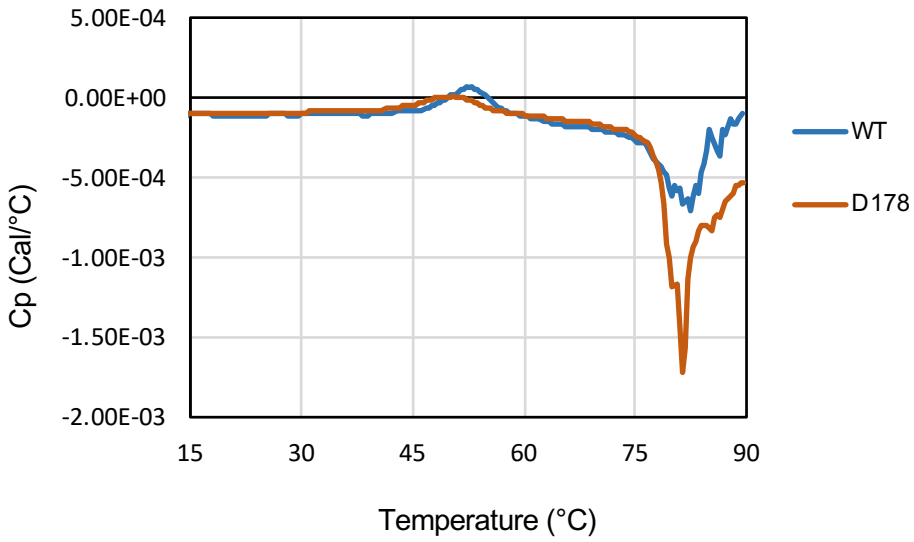


Figure 5.13 DSC analysis of WT and D178 MBP variants

Scanning between 10 and 90 °C, past the point of protein aggregation showed D178 to have a slightly lower T_m (50 °C) than wild-type MBP (53 °C). Both protein variants showed a similar aggregation temperature at approximately 75 °C.

Melting points obtained through DSC lay closer to published T_m values for MBP.[258, 259] It should be noted that the quoted T_m for maltose binding protein is approximately 63.4 °C and the values measured in this study are approximately 10 °C less than this. However, this may be due to the introduction of the cleavable N-terminal polyhistidine tag into the structure which has been observed to decrease thermal stability with other proteins.[260]

A second experiment was performed to monitor how well each protein re-folded after melting. This involved the heating and cooling of each sample between 10 and 70 °C with three repetitions. The highest temperature was 70 °C so as to refold the protein prior to protein aggregation. With each round of refolding a protein tends to become more destabilised and hence, it unfolds more easily requiring less energy to maintain constant $\Delta T/\Delta t$. Figure 5.18 shows that the D178 scaffold refolds less efficiently than the wild-type. This indicates that the *Alu* thermally destabilises the protein, once again contradicting the T_m values obtained from CD and following the expected observation for the insertion. As DSC was only performed on two MBP variants and not the scrambled D178* variant, it cannot be determined whether it is simply the insertion the destabilises the protein or the actual sequence that has an

effect. However, these results did lead to the dismissal of the conclusions drawn from T_m data obtained through CD.

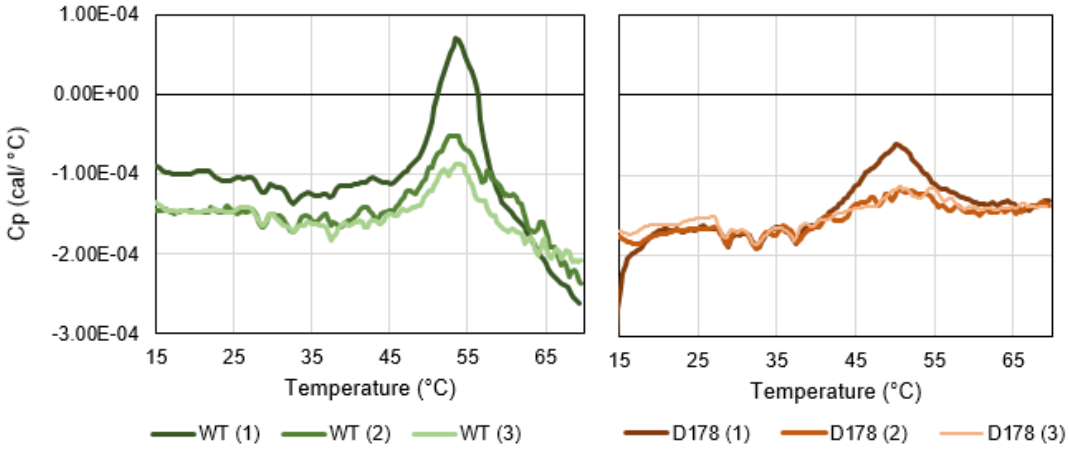


Figure 5.14 DSC analysis of wild-type and D178 re-folding

Protein samples were heated and cooled between 10 and 70 °C three times. A higher C_p was observed for wild-type MBP over D178 in addition to a higher T_m indicating better thermal stability. The wild-type protein was also observed to refold more efficiently.

5.4.3.3 Conclusions on *Alu* effect on the expression, folding and stability of MBP

The effect of *Alu* insertions on the thermal stability of MBP was probed using CD and DSC. Analysis of melting points (T_m) *via* CD were dismissed as T_m calculation relied on the assumption that full unfolding occurs at the top end (70 °C) of the temperature gradient, which was not achieved. DSC was performed on wild-type MBP and the D178 variant. As the *Alu* insertion in the D178 variant lies at the edge of a β-sheet, this would be likely to destabilise the protein and thus, have a lower T_m. Through DSC analysis, a slight decrease in T_m of approximately 3 °C was observed; a result which agreed much more with what was expected. A predicted T_m value for wild-type MBP, only 10 °C lower than those quoted in the literature, was also observed which could be partially attributed to the presence of a cleavable polyhistidine tag at the N-terminus of our MBP scaffold. Upon thermal unfolding and re-folding of the protein, more efficient refolding was also observed for the wild-type, further indicating that the *Alu* insertion slightly decreases thermal stability.

DSC does seem to indicate that a slight decrease in thermal stability is observed upon insertion of an *Alu* sequence into MBP; however, it is possible that this is due to the size and/or locations of the insertion as opposed to the *Alu* sequence itself.

5.5 Functional consequences of *Alu* insertions in MBP

5.5.1 Using amylose purification to predict correct protein folding

MBP has a binding affinity for amylose,[261] and a higher binding affinity for maltose, as indicated by its name. 1 mg of each purified protein variant (Ni-NTA/SEC), was loaded onto an amylose column. Flow-through and wash fractions were collected prior to elution in buffer containing 10 mM maltose. The addition of maltose displaces amylose in the binding site of MBP and allows elution of bound protein. Elution fractions were collected, and all fractions were analysed *via* SDS-PAGE (figure 5.15). Correct folding of MBP variants would expect to see minimal/no protein elution in the flow-through and wash stages and elution of protein only upon the addition of maltose. Protein unfolding, or partial unfolding, would result in protein product observed in the wash phases of chromatography.

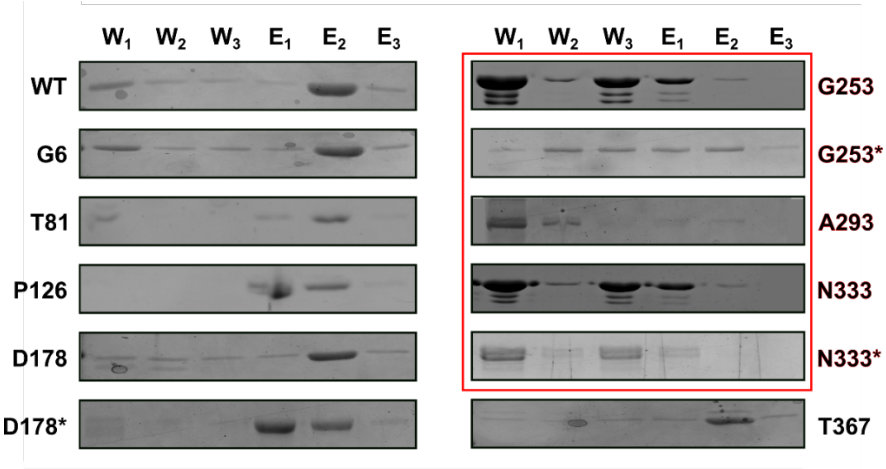


Figure 5.15 Amylose purification of MBP variants

MBP variants were purified *via* amylose column chromatography, utilising the binding interaction between amylose and MBP to predict correct protein folding. Bulk elution in wash phases (W) indicated that incorrect or disrupted protein folding had occurred, as observed by insertions at positions G253/*, A293 and N333/* (red). Bulk elution in the elution phases (E) indicated ligand binding and hence, correct protein folding as observed by wild-type MBP and insertions at G6 through D178* and T367. Note: constructs labelled with an Asterix (*) contained scrambled *Alu* sequences.

In most cases, proteins were eluted from the column upon addition of maltose and therefore, the bulk of the protein sample was observed in E_1/E_2 fractions. In the case of WT, G6, T81 and D178 insertions minor elution in the first wash (W_1) was observed; however, this is likely as result of oversaturation of the amylose resin. Interestingly, the insertion at D178, located at the end of a β-sheet, does not seem to have any effect on protein folding. However, the insertion at A293, within an α-helix, does result in protein misfolding. In addition to this, insertions at the G253 and N333 positions also disrupts amylose binding function, which may indicate incorrect folding.

Previous studies, from other groups, into the structure of maltose-binding protein[262, 263, 264] revealed that its structure consists of two domains (I and II), attributed to the proteins N- and C- terminus, respectively, bridged by three short loops (3 – 5 residues) which act as a hinge (figure 5.16).

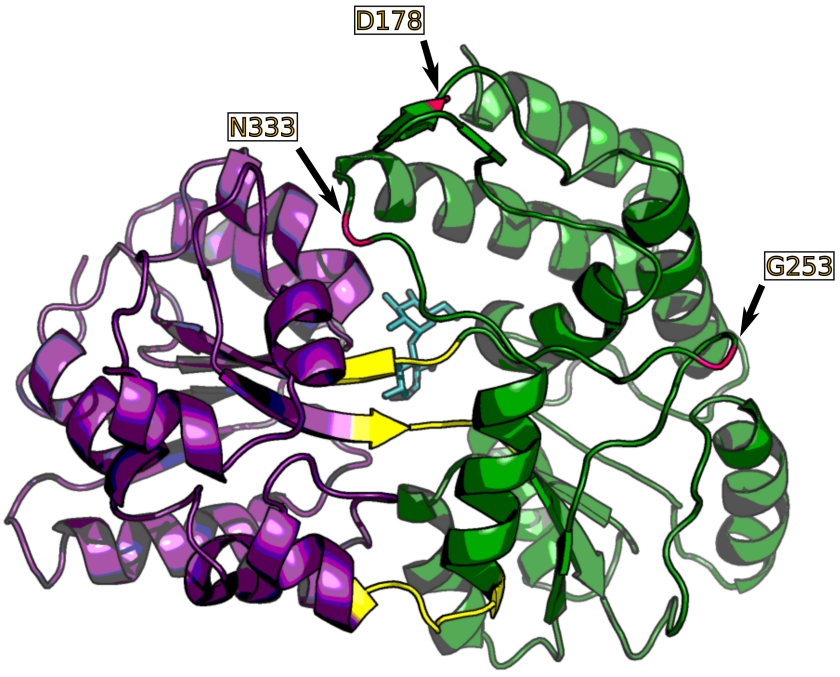


Figure 5.16 Insertion sites with respect to MBP domains and hinge

The two domains of MBP, I (N-terminal; purple) and II, (C-terminal; green) are hinged by three short loops ranging from 3 – 5 residues in length (yellow). The hinges contribute to the binding of MBP to ligands, such as maltose (light blue), and insertions affecting these loops may contribute to loss of binding affinity. Affected insertion sites are highlighted in pink. Structure adapted from PDB (1ANF).

As these loops contribute to the binding of MBP to its ligands, it is feasible to assume that insertions affecting this hinge would affect the binding of MBP variants to amylose. From this perspective, judging from the locations of our insertions, the only ones that may interfere with this hinge are N333/*, which lie above the loops and close to the maltose binding site. It has previously been reported than mutations at N333 may disrupt the packing of the interface between MBP domains I and II which is formed when MBP is in its open conformation.[265] As a result, the lack of binding observed for N333/* variants to amylose resin is likely attributed to the simple presence of an insertion at this site and not the nature of the insertion. Additionally, the lack of binding observed for the insertion at A293, within a protein α-helix, is unsurprising as this insertion was predicted to disrupt function.

Interest lies in the result that the insertion at G253, which lies in an outer loop of domain II of MBP, does not bind directly to amylose and therefore, loss of binding affinity indicates incorrect protein folding. This is concurrent with the previously observed reduction in protein overexpression and shift in α-helical character. It is yet more interesting that the insertion of a scrambled *Alu* at the same site, seems to slightly restore binding. For the G253 *Alu* variant, bulk protein is observed to elute from amylose resin almost immediately in protein wash phases. In contrast, for the G253* scrambled variants, the protein elutes evenly throughout wash and elution phases indicating that some weak binding may have occurred, though binding has not been fully restored. This result implies that the *Alu* sequence itself may have a small effect on protein folding in combination with insertion location.

In most cases, the insertion of an *Alu* sequence in MBP does not appear to have an effect on the binding of the protein to amylose. This is surprising in the case of the D178 insertion, located at the end of a β-sheet, which would be expected to disrupt folding. In contrast, the A293 insertion, within an α-helix, disrupts protein folding as expected.

Two other insertion sites were observed to disrupt protein folding for both the *Alu* and scrambled variants; G253 and N333. The N333 site has previously been reported to be involved in forming the open conformation of MBP and hence, it is not surprising that an insertion at this site, especially such a large one, would affect binding to amylose. In addition, this insertion lies close to the N333 binding site and so it is possible that the large insertion loop may cause steric hindrance at the binding site. There was no observed improvement to binding when substituting the *Alu* for

its scrambled variant (N333*). An interesting result was observed for insertions at G253, which lies on the outside on domain II of MBP. An *Alu* insertion at this site results in no binding to the amylose column and hence, predicts protein misfolding. More interestingly, substitution of the *Alu* sequence for its scrambled variant (G253*) resulted in a slight improvement in binding, though binding was not restored in full. This indicated that though insertion site may be the primary factor contributing to effect of *Alu* insertions, the sequence itself may also contribute to changes in protein folding and function.

5.5.2 Analysis of protein function using ITC

To further probe the effect of *Alu* insertions on protein function, the binding of wild-type MBP and four variants; G6 (N-terminal insertion), D178/* (internal, edge of β-sheet insertion) and T367 (C-terminal insertion), was analysed *via* isothermal titration calorimetry (ITC; discussed in Chapter 3). Binding interactions between protein variants and three MBP substrates; D-(+)-maltose, maltotriose and β-cyclodextrin (figure 5.17) were analysed. Ideally, ITC experiments with G253/* and N333/* would have been performed; however, due to problems which arose in the overexpression and purification of sufficient amounts of these MBP variants for ITC analysis, this was not possible.

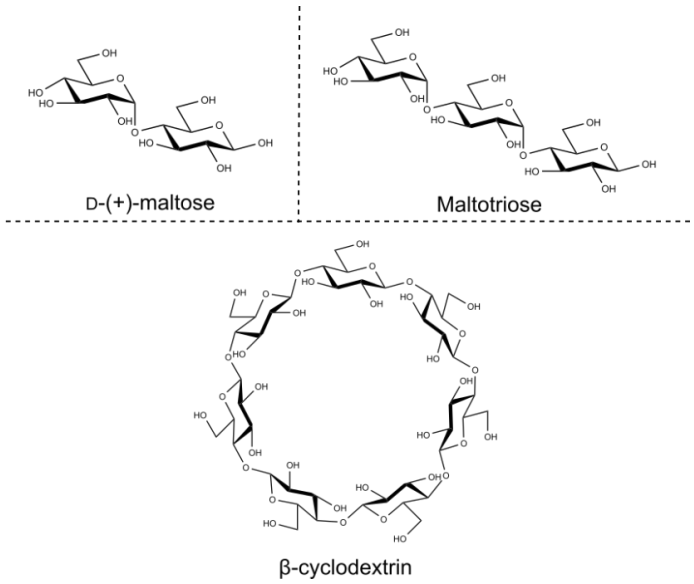


Figure 5.17 Ligands of MBP

Three ligands of MBP were used in binding studies; D-(+)-maltose, maltotriose and β-cyclodextrin.

5.5.2.1 Investigating the effect of a translated *Alu* on the binding of MBP to its ligands

Binding curves for our wild-type MBP variant with D-(+)-maltose, maltotriose and β-cyclodextrin were obtained so as to observe whether our variant gave K_d values similar to those in literature and hence, could be used as a suitable comparison. It should be noted that literature values for K_d of MBP to maltose ranges between 2.0 and 4.0 μM.[266, 267] The K_d for maltotriose is quoted as 0.15 – 0.4 μM[268] and 1.8 μM for β-cyclodextrin. Our ITC experiments observed values for binding of ligands to wild-type MBP of 8.6 ± 3.8 μM, 5.6 ± 1.3 μM and 4.0 ± 0.3 μM, respectively (figure 5.18). Though our values differ slightly from those quoted in literature, differences are relatively small and be attributed to a difference in buffer systems.

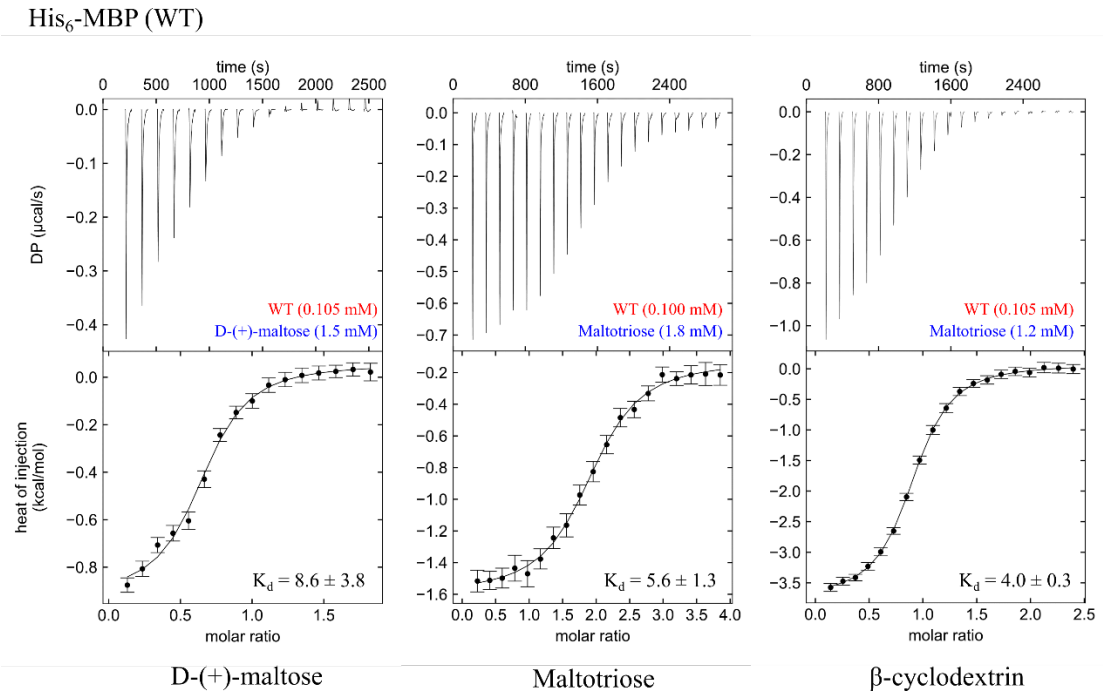


Figure 5.18 ITC curves for wild-type MBP and sugars

ITC curves gave K_d values for wild-type MBP with D-(+)-maltose, maltotriose, β-cyclodextrin of 8.6 μM, 5.6 μM and 4.0 μM, respectively. These values are in concurrence with those quoted in literature.

In the same way, binding affinity (K_d) of sugars to G6, D178 and T367 variants was measured as a way to represent N-terminal, internal and C-terminal insertions, respectively (table 5.2). For the most part, no significant change in K_d was observed for variants, with most differences being no more than 5.0 μM away from the wild-

type value. In addition, there was no change in the stoichiometry of binding; N = 1 for D-(+)-maltose and β-cyclodextrin and N = 2 for maltotriose. One exception to this lay in the binding of T367 to D-(+)-maltose with a K_d of 23.0 μM, compared to 1.6 μM when bound with wild-type. However, the large error on the fitting of this data is likely to be cause of this rather than a real change in binding affinity due to the insertion. The same can be said for the binding of maltotriose to the D178 variant from which outliers provided an error in the plotting of the curve in Sedphat. This change is slightly over a 10-fold decrease in binding. Insufficient data for D-(+)-maltose to the G6 variant was obtained; however, as minimal changes in binding affinity were observed for this variant with maltotriose and β-cyclodextrin, the same result was assumed for binding to D-(+)-maltose. Binding curves for all ITC experiments can be found in Appendix 5.

MBP Variants	K_d (μM)		
	D-(+)-maltose	Maltotriose	β-cyclodextrin
WT	8.6 ± 3.8	5.3 ± 1.3	4.0 ± 0.3
G6	-	2.9 ± 1.4	6.2 ± 2.1
D178	4.6 ± 1.2	27.5 ± 21.9	5.2, 4.8
T367	23.0 ± 15.4	2.8 ± 1.2	4.9 ± 0.8

Table 5.2 K_d values for binding of MBP variants to sugars

K_d values for binding of MBP variants to sugars calculated from ITC. The result highlighted in orange shows a change from exothermic to endothermic binding, resulting in two K_d values.

A second, more interesting, change is observed in the D178 insertion. Instead of the usual S-shaped exothermic binding curve observed for binding to β-cyclodextrin, an endothermic curve is observed (figure 5.19). Usually, this kind of binding curve is attributed to a conformational change of the protein in order to accommodate ligand binding. This insertion does lie relatively close to the ligand binding site of MBP and as such it is understandable that such a large insertion would affect binding. Due to the nature of the observed curve and the decrease in binding with increasing ligand size, it is proposed that an insertion at this location favours the closed conformation. Thus, the more open conformation required for the binding of larger ligands is unfavourable and an endothermic, two-step interaction is observed. ITC with the scrambled *Alu* variant, D178*, and β-cyclodextrin was performed in order to analyse

whether it was the insertion site or the insertion sequence itself that resulted in this change in binding. The scrambled *Alu* variant yielded a similar binding curve indicating that the change in binding was sequence independent and was solely related to the site at which the insertion occurred.

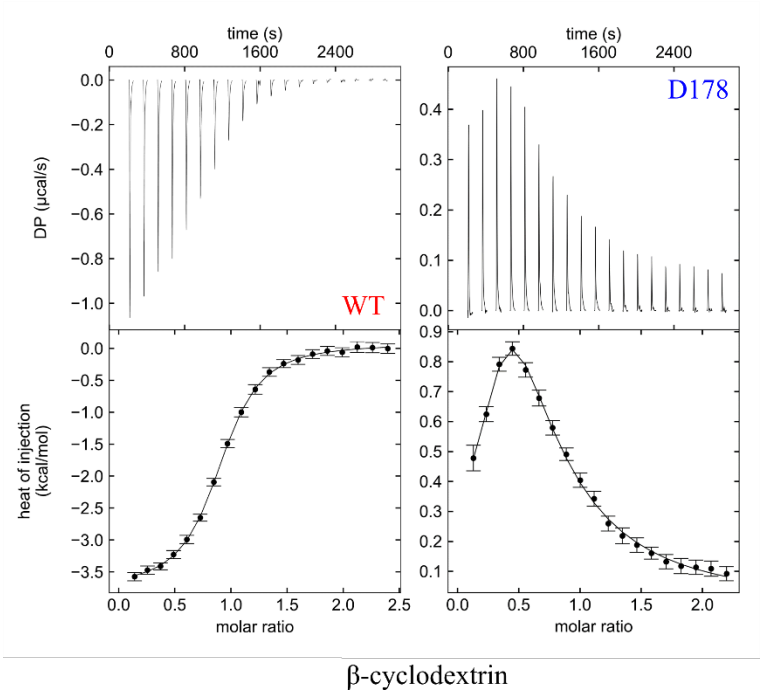


Figure 5.19 ITC binding curve for β-cyclodextrin with His_6-MBP-WT and His_6-MBP-D178

The curve showed an endothermic binding very dissimilar for the exothermic S-shaped curve observed for wild-type MBP with β-cyclodextrin. The switch in binding was likely attributed to a conformational change which is undergone in order for the protein variant to accommodate ligand binding.

5.5.2.2 Conclusions from ITC

Overall, ITC revealed that the insertion of *Alu* elements into MBP had minimal effect on ligand binding with D-(+)-maltose, maltotriose and β-cyclodextrin. An exception to this was observed with the D178 insertion site with β-cyclodextrin, which resulted in a switch from exothermic to endothermic binding, usually attributed to conformational change; however, the same switch was observed with the scrambled *Alu* variant, D178*. This indicated that it was the insertion site that was the primary cause of the change in binding and the sequence of insertion was relatively unimportant. Unfortunately, the most interesting insertions at G253 and

N333 locations could not be studied by ITC due to problems in overexpressing and purifying sufficient amounts of protein.

As bioinformatic analysis showed that *Alu* insertions generally do not occur naturally near known binding sites of structural motifs, it is unsurprising that the insertion at N333, near the hinge of MBP was disruptive to binding.

5.6 Conclusions on the effect of translated *Alu* elements in MBP

Wild-type MBP, eight *Alu*-containing MBP variants and three scrambled *Alu* MBP variants were sub-cloned and successfully overexpressed and purified from *E. coli.* A summary of the effects of *Alu* insertions on the overexpression, structure and function of MBP can be found in table 5.2.

Variant	Insert location	Effect on expression	Effect on binding	Quantitative binding
WT	None	None	None	N/A
G6	N-terminal	Reduction	None	Minimal changes to K_d (all ligands)
T81	Loop	Reduction	None	
P126	Loop	None	None	
D178	β-strand	None	None	Higher K_d (D-(+)-maltose and maltriose) Endothermic binding (β-cyclodextrin)
D178*	β -strand	Reduction	None	Endothermic binding (β-cyclodextrin)
G253	Loop	Reduction	Loss of binding	N/A
G253*	Loop	Reduction	Loss of binding	N/A
A293	α-helix	Reduction	Loss of binding	N/A
N333	Potential hinge region	Reduction	Loss of binding	N/A
N333*	Potential hinge region	Reduction	Loss of binding	N/A
T367	C-terminal	None	None	Minimal changes to K_d (all ligands)

Table 5.2 Summary of changes to MBP overexpression and binding upon *Alu* insertion

Effects on expression were determined *via* densitometry experiments. Effects on binding were determined by analysis of binding to an amylose column, with quantitative analysis of some constructs performed *via* ITC. Note: constructs marked with an Asterix (*) contain scrambled *Alu* sequences.

Small scale overexpression tests were performed which showed that insertion of *Alu* sequences within MBP generally lowered overexpression levels in *E. coli*. However, replacement of *Alu* sequences with scrambled sequences did not restore levels of overexpression and therefore, any loss was attributed to insertion site and not insertion sequence. A reduced level of overexpression was also observed using 1 L media, with much lower protein yields being obtained in the majority of cases (excluding G6, D178 and T367 insertions).

Quick analysis of protein binding to amylose resin showed that most *Alu*-containing MBP variants folded correctly and hence, bound to amylose. Interestingly, this also included the D178 insertion which was located at the edge of a β-sheet. The A293 insertion, located within an α-helix, did not bind to amylose and, as predicted, was not correctly folded. In addition to A293, insertions at G253 and N333 also showed no binding. Insertion of a scrambled *Alu* at the N333 location did not result in restoration of binding. As previous reports suggested that the N333 location may affect the three-loop hinge associated with open conformation of MBP, this result was not surprising. However, substitution of the *Alu* sequence for a scrambled variant at G253 seemed to partially, though not fully, restore binding. As this insertion was located on the outside of the protein and not in the vicinity of the binding site, it was interesting that this would have such a large effect and that the sequence itself seemed to contribute to that effect as opposed to just insertion location.

Study of MBP variants by CD revealed that there was no definitive secondary structure to the *Alu* insertion. Information on thermal stability of MBP variants obtained from CD was inconclusive; however, results from DSC saw a slight decrease in thermal stability when comparing the D178 insertion with wild-type MBP. Due to the insertion size (33 residues) this is not unexpected and cannot be attributed to the *Alu* sequence itself.

Isothermal titration calorimetry of wild-type MBP and G6, D178/* and T367 variants revealed that, in most cases, minimal changes to the binding of D-(+)-maltose, maltotriose and β-cyclodextrin were observed in the presence of *Alu* insertions. The exception to this was the binding of the D178 variant which had a lower affinity for all ligands, except maltose. Titration with β-cyclodextrin resulted in a switch from exothermic to endothermic binding, most likely attributed to the open conformation becoming unfavourable with this insertion site. The same switch

was observed upon the analysis of β-cyclodextrin binding to the scrambled *Alu* variant, D178*, indicating that the location of the insertion was the primary cause of the switch and that the observed result was sequence independent.

Overall, it appears that the sequence of the *Alu* insertion has minimal to no effect on the structure and function of proteins and any effects observed lies primarily with the location of the insertion independent of the insertion sequence. As such, it is likely that any effect that *Alu* sequences have upon proteins will be as a secondary effect in combination with a detrimental insertion site. It is likely that in most cases, *Alu* insertions in proteins will be mostly harmless unless the site in which they are inserted has an important role in structure or function. This is not to say that the insertion would not result in a disease-causing protein isoform; however, this probably occurs in the minority of cases.

Chapter 6
Assessing translation of *Alu* mRNAs in human cell lines and primary cells by polysome profiling

Overview

A combination of polysome profiling, reverse transcription and qPCR was used to assess the number of ribosomes attached to mRNAs of interest to determine the distribution of individual transcripts across the different translation complexes. *Alu*-containing and non-*Alu*-containing transcripts of BCAS4 and NEK4 were analysed and compared between two cell types; one cancer cell line, SH-SY5Y, and one non-cancer set of primary cells, NP-1. A second set of primary cancer cells was also analysed *via* polysome profiling, but transcript analysis was not performed.

This work had two mains aims: The first was to determine whether a difference in translation could be observed between *Alu*- and non-*Alu*-containing transcripts in a single cell type. The second was to observe whether a difference in translation was observed between 'cancer' and 'non-cancer' cell types indicating translation bias of one transcript over the other.

6.1 Polysome profiling of human cells

6.1.1 Polysome profiling

Modulation of protein levels occurs through a number of different mechanisms including, but not limited to, mRNA splicing, transport and translation.[269] Of these, mRNA translation is the most energy consuming and uses over 50% of a cell's energy.[270] Translation, or more importantly, the dysregulation of translation has been associated with cancer[271] and neurodegenerative disease[272] and as such, it is an important research area in further understanding both the causes and treatments of such diseases.

Polysome profiling (figure 5.1) is a molecular biology technique used to assess translation levels and global translation, *i.e.* all the mRNA translated at a single time within a cell.[273] It provides a means to measure the extent of active translation of transcripts of interest through the separation of ribosome-associated mRNA *via* sedimentation through a sucrose gradient.[274]

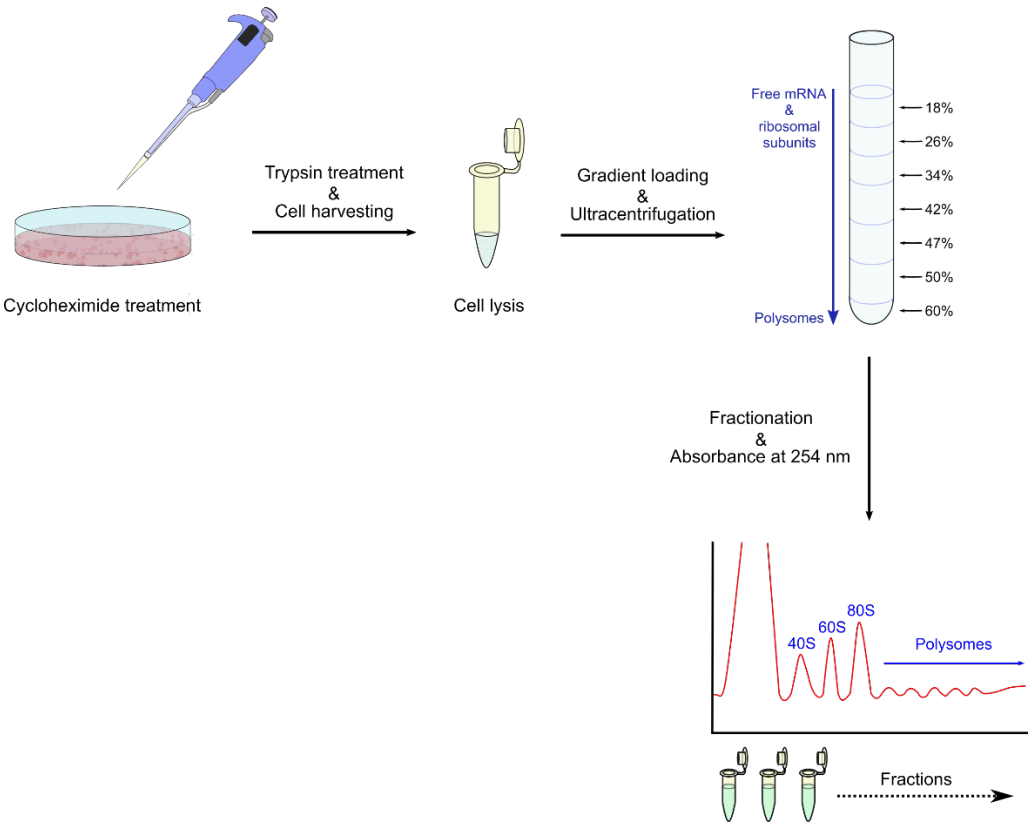


Figure 6.1 Outline of polysome profiling

Cell were treated with cycloheximide to inhibit the elongation stage of translation, then treated with trypsin to enable harvesting. Cells were lysed in the presence of $MgCl_2$ and RNase inhibitor to maintain RNA complexes and prevent RNA degradation (rRNA and mRNA), respectively. Cytoplasmic lysate was loaded onto a sucrose gradient (18 – 60%)[275] prior to ultracentrifugation, which separated mRNAs according to the number of ribosomes bound. Gradients were fractionated and a polysome profile was generated from absorbance readings at 254 nm.

The technique was performed by first inhibiting the second stage of translation, termed elongation, by treatment of cells with cycloheximide which binds to the 60S ribosomal subunit and effectively 'freezes' translation at its elongation step.[276, 277] It is important to halt the process at the elongation step as this allows for mRNA splicing to have occurred. Cells were lysed and RNA-ribosome complexes were separated by ultracentrifugation through a sucrose gradient based on their sedimentation.[278] Subsequent fractionation of centrifuged gradients collected mRNA according to its ribosomal-association, with free mRNA and ribosomal subunits present in lower percentage sucrose fractions and increasingly larger subunits present in higher percentage sucrose fractions. This was mapped through absorbance at 254 nm, generating a polysome profile.[279]

In these experiments, RNA fractions were precipitated, DNase treated and purified, in this case *via* phenol/chloroform extraction and ethanol precipitation, prior to analysis. Reverse-transcription was performed to obtain cDNA for use in real-time quantitative PCR (qPCR).

6.1.2 Cell lines

Translation of *Alu* and non-*Alu* mRNA was initially planned to be assessed in three cell types; SH-SY5Y, GMB1 and NP-1 to determine any difference in translation of each transcript in 'cancer' and 'non-cancerous' cells. SH-SY5Y, a neuroblastoma cell line derived from a bone marrow biopsy, was obtained from the Aspden group (Faculty of Biological Sciences, University of Leeds).[280] GBM1 (glioblastoma multiforme), primary glioblastoma cells obtained from patients undergoing surgery at Stanford Medical Centre,[281] and NP-1, primary brain cells obtained during surgery on epilepsy patients at Stanford Medical Centre, were obtained from the Wurdak group (St. James' Hospital, University of Leeds).[282]

For the purpose of this work, SH-SY5Y and GMB1 samples are referred to as 'cancerous' and NP-1 samples are referred to as 'non-cancerous' thus, giving a comparison between cancer and non-cancer cell-derived mRNA.

The minimum number of cells which could be used for reasonable detection was determined with SH-SY5Y cells. Samples containing 3×10^6, 5×10^6, 6×10^6, 7×10^6 and 9×10^6 were fractionated and analysed for suitable absorbance and polysome separation at a reasonable detection limit (above 0.1). All samples yielded good separation up to 6 + ribosomes. The resulting RNA obtained from fractionation of

the 5×10^6 cell sample was too minimal to be detected *via* absorbance at 260/280 nm. Therefore, it was determined that samples of more than 5×10^6 cells would be needed, and ideally 1×10^7 to 2×10^7 cells would be needed in order to perform analyses on multiple transcript targets *via* qPCR.

6.1.3 Target transcripts

Initially, two transcripts for each of ASCC1, BCAS4 and NEK4 were chosen as targets for qPCR analysis to compare the translation level of the *Alu* transcript of each gene in comparison to the non-*Alu* transcript.

As *Alu* and non-*Alu* variants of mRNA transcripts differ through their alternative splicing patterns, primers for BCAS4 and NEK4 were designed using Ensembl to match exons specific to the *Alu* and non-*Alu* transcripts so as to amplify 100 – 200 bp corresponding to each variant individually.

6.1.4 Polysome graphs for cells

Polysome profiling was performed on each cell type; SH-SY5Y, GBM1 and NP-1. For SH-SY5Y and NP-1 samples, 2×10^7 and 2.2×10^7 cells, respectively, were used for each of three replicate gradients. Each cell type yielded three near-identical polysome profiles. A representative profile for each can be observed in figure 6.2 – all polysome profiles obtained can be found in Appendix 6. SH-SY5Y appeared as previously observed by the Aspden group. To enable precise separation of ribosome-mRNA complexes across the gradient, 0.5 mL fractions were generated (table 6.1).

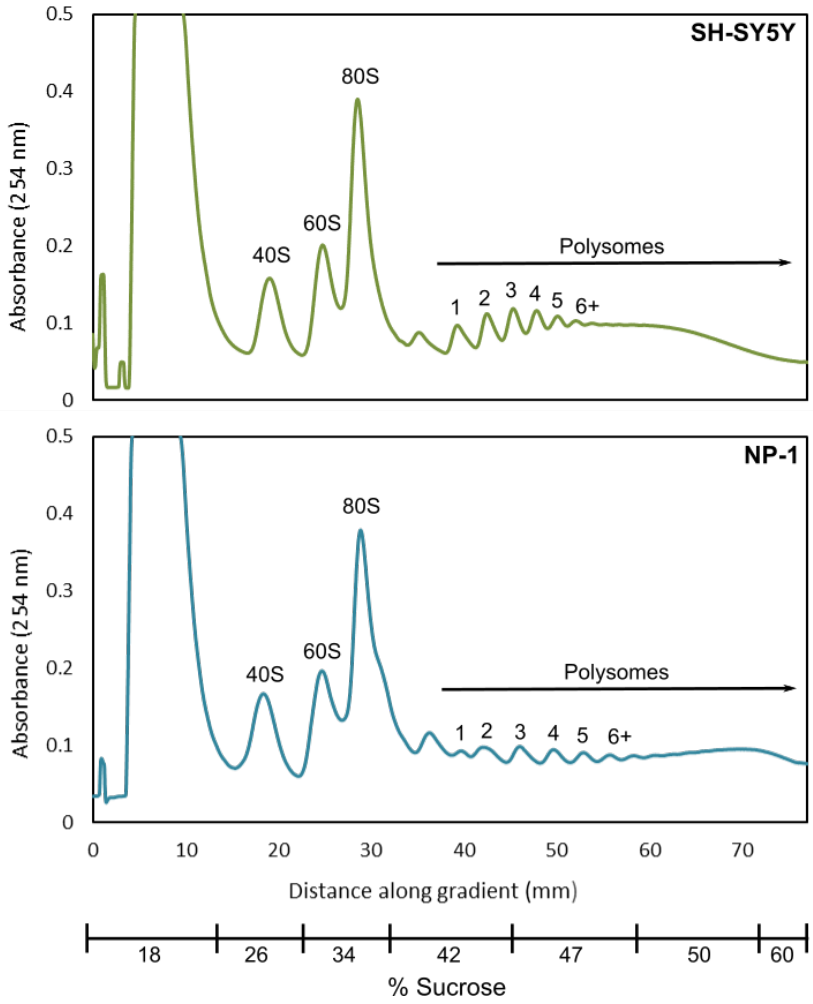


Figure 6.2 Polysome graphs for SH-SY5Y and NP-1 cells

Polysome graphs obtained for harvested SH-SY5Y and NP-1 cells. Polysome graphs were obtained in triplicate. All polysome graphs obtained can be found in Appendix 6. Ascending ribosome-association can be observed for both samples, beginning at approximately 38 mm along the gradient length.

As a result of the small fractionation volume, each fraction contained a majority of mRNA bound by a specific number of ribosomes (e.g. one ribosome, two ribosomes, etc.). This was true up to the association of approximately six ribosomes at which point peak absorbance became convoluted. Fractions past this point were labelled as being associated to '6 +' ribosomes.

Fraction	Distance (mm)	Association
1 – 4	0 – 14.2	Free mRNA
5 – 6	14.2 – 21.3	40S
7	21.3 – 24.8	60S
8	24.8 – 28.4	80S
9	28.4 – 31.9	< 1 ribosome
10	31.9 – 35.5	1 ribosome
11	35.5 – 39.0	2 ribosomes
12	39.0 – 42.6	3 ribosomes
13	42.6 – 46.1	4 ribosomes
14	46.1 – 49.7	5 ribosomes
15 - 21	49.7 – 78.6	6+ ribosomes

Table 6.1 Distribution of ribosome-bound mRNA across gradient fractions in SH-SY5Y and NP-1 samples

Unfortunately, the GBM1 cells did not reach suitable confluence during cell culture. Significant efforts were made to recover GBM1 cells to a healthy confluence, however, minimal progress was made. Nonetheless, approximately 0.7×10^6 cells were harvested, loaded onto a single gradient in a minimal amount of lysis buffer and fractionated, measuring absorbance at a more sensitive detection limit of 0.05 (figure 6.3).

The polysome graph for GBM1 cells required a more sensitive detection limit than those for SH-SY5Y and NP-1 due to the low number of cells used. The 80S peak (ca. 28 mm) was significantly smaller than for SH-SY5Y or NP1. Though clear polysome signals can be observed, no efforts were made to reverse transcribe GBM1 mRNA samples. Previous work had determined that sample sizes of less than 3×10^6 cells did not yield sufficient mRNA for qPCR analysis. However, mRNA was precipitated with isopropanol and stored for potential future work with higher sensitivity assays. Unfortunately, as no GBM1 mRNA samples were obtained, a direct comparison of 'cancerous' and 'non-cancerous' brain cells could not be made. The fact that both NP-1 and GBM1 cells were obtained from brain samples would have made a better comparison than with SH-SY5Y samples that were obtained from bone marrow. Nonetheless, a 'cancerous' and 'non-cancerous' sample was still obtained and analyses were carried out on these.

Though polysome profiling has been performed with primary tissue samples before,[283] there are no reports of use with these cells lines. The NP-1 cells used were culture as a 'healthy' cell culture by the Wurdak group. With these cell types, it is

possible to begin comparing translation between them; however, they are not ideal. For more accurate conclusions to be drawn, much more information on the cell origin would be required, such as age, gender and ethnicity. This work provides a basis for further research in this area.

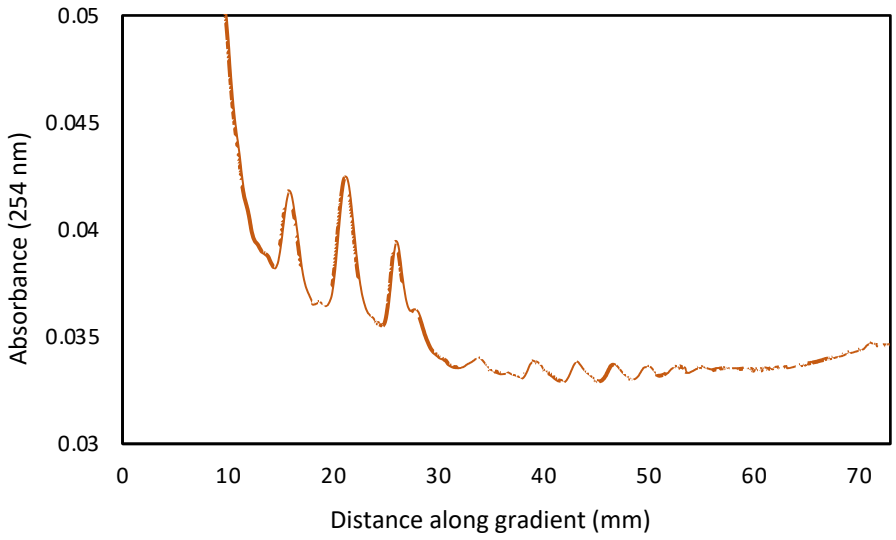


Figure 6.3 Polysome graph for GMB1

Polysomes can be observed and were fractionated accordingly. The absorbance detection limit for GBM1 cells was set at 10% of that used for NP1 and SH-SY5Y cells. Previous experiments indicated that insufficient mRNA would be obtained from this number of cells (0.7×10^6) for qPCR analysis.

Fractions from SH-SY5Y ('cancerous') and NP-1 ('non-cancerous') samples were DNase treated and RNA was extracted using acidic phenol/chloroform, followed by ethanol precipitation. Reverse transcription was performed to yielded cDNA for use in qPCR.

6.2 Comparing the translation of *Alu*: non-*Alu* transcripts within cells

For the comparison of *Alu* and non-*Alu* mRNA translation in NP-1 and SH-SY5Y cells, primers for *Alu* (AC) and non-*Alu*-containing (nAC) mRNAs for BCAS4 and NEK4 were designed. Translation of NEK4 nAC mRNA was insufficient to obtain a suitable standard curve. A suitable standard curve was defined as having an R^2 of 0.9 – 1.0 upon fitting and a primer efficiency of 100 – 112%. As a result, only

BCAS4 AC and nAC transcripts could be compared to one another at the polysome level in these cell samples. Standard curves for all qPCR analysis can be found in Appendix 6.

6.2.1 Translation of NEK4 AC mRNA in SH-SY5Y and NP-1 cells

Due to the low level of translation of NEK4 nAC mRNA in both NP-1 and SH-SY5Y samples, a suitable standard curve was not obtained to quantify translation of this transcript at the polysome level. However, quantification was possible for the AC transcript in both cell types. NEK4 AC translation in SH-SY5Y cells (figure 6.4) showed approximately 50% of mRNA bound to two or fewer ribosomes, indicating poor translation efficiency. 35% of mRNA was bound to three to five ribosomes and only 10% was bound to 6+ ribosomes. This in unsurprising, as publicly available RNA-Seq data (Human Protein Atlas) reported low NEK4 RNA expression in bone marrow.

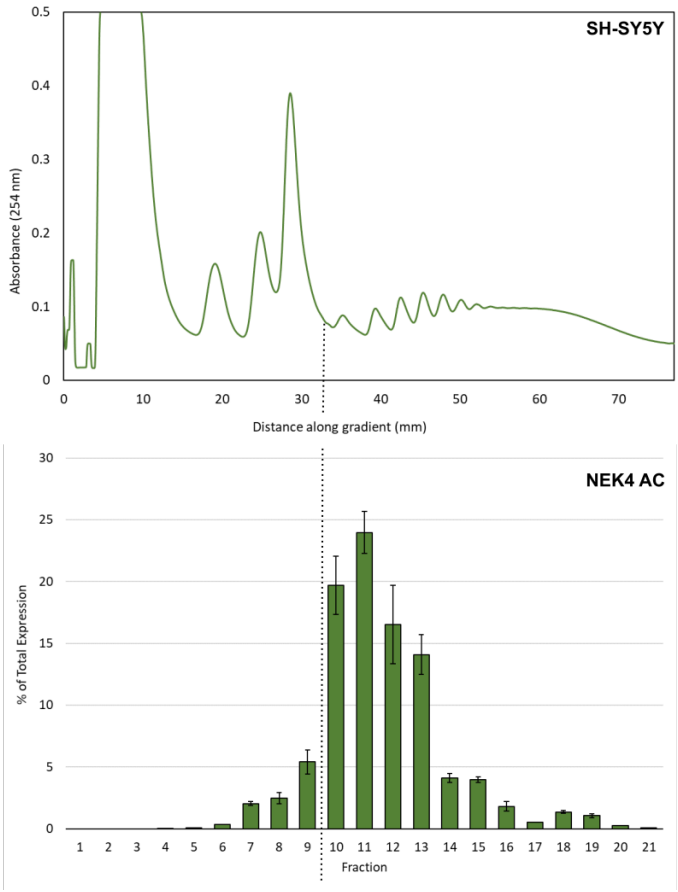


Figure 6.4 Ribosomal distribution of NEK4 AC mRNA in SH-SY5Y cells
50% of mRNA was bound to two or fewer ribosomes indicating poor translation efficiency. Ribosome-bound mRNA began in fraction 10 and increased in a fraction-based manner until approximately 6+ ribosomes at which point absorbance became convoluted. Error bars represent the standard error of three technical repeats.

In contrast, NEK4 AC mRNA has a higher translation efficiency in NP-1 cells (figure 6.5), with 39.5% of mRNA bound to 6+ ribosomes and only 18% of mRNA bound to two or fewer. This result is unexpected as publicly available RNA-Seq data suggests relatively low RNA expression in brain tissue.

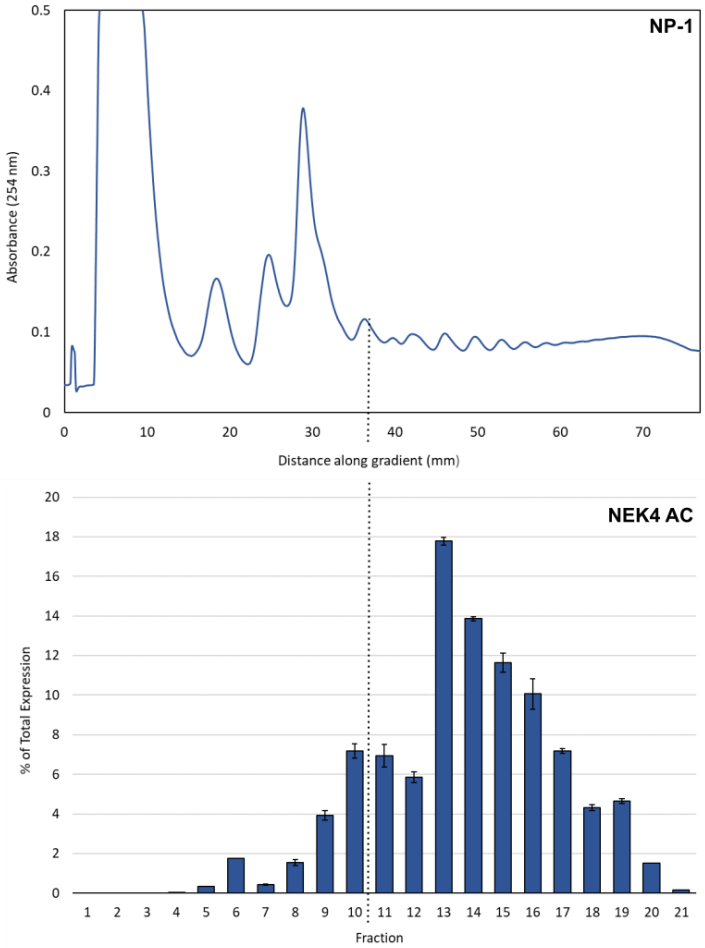


Figure 6.5 Ribosomal distribution of NEK4 AC mRNA in NP-1 cells

39.5% of mRNA was bound to six or more ribosomes indicating high translation efficiency. Only 18% of mRNA was bound to two or fewer ribosomes. Ribosome-bound mRNA began in fraction 10 and increased in a fraction-based manner until approximately 6+ ribosomes at which point absorbance became convoluted. Error bars represent the standard error of three technical repeats.

As translation of NEK4 nAC was insufficient to obtain quantitative data, and sufficient translation of NEK4 AC mRNA was possible, we can assume that the AC mRNA is the predominant transcript in both SH-SY5Y and NP-1 cell types.

6.2.2 The ratio of BCAS4 AC and nAC mRNA translation in SH-SY5Y and NP-1 cells

SH-SY5Y cells, as previously stated, are a neuroblastoma cell line derived from human bone marrow. There are three reported mRNA transcripts for BCAS4, two AC and one nAC. Due to the differences in transcripts, it was only possible to design primers to target both the AC transcripts at once.

For BCAS4, a similar ribosomal distribution of AC and nAC transcripts in polysomes was observed (figure 6.6) indicating similar translation levels when measured as a percentage of the all ribosomal mRNA. Ribosome-associated mRNA began from fraction 10; increasing ribosome count was consistent with increasing fraction number until there were over six ribosomes bound to mRNA.

It is reported in literature that mRNA bound to three or more ribosomes is generally translated efficiently in the cell. Both BCAS4 AC and nAC mRNAs were translated in SH-SY5Y neuroblastoma cells at a low level, with approximately 60% of mRNA bound to two or fewer ribosomes, respectively. Similar amounts of mRNA were observed to be bound by 3-5 ribosomes (26% and 30%, respectively) and 6+ ribosomes (5% and 4%, respectively). Performance of a Student's T-test* gave *p* values of 0.74, 0.63, and 0.37 for < 2, 3-5 and 6+ ribosomes, respectively, confirming no significant difference in translation of BCAS4 AC and nAC mRNAs in SH-SY5Y cells. Poor translation efficiency of BCAS4 mRNA in SH-SY5Y was unsurprising as publicly available RNA-Seq data reported low BCAS4 RNA expression in bone marrow.

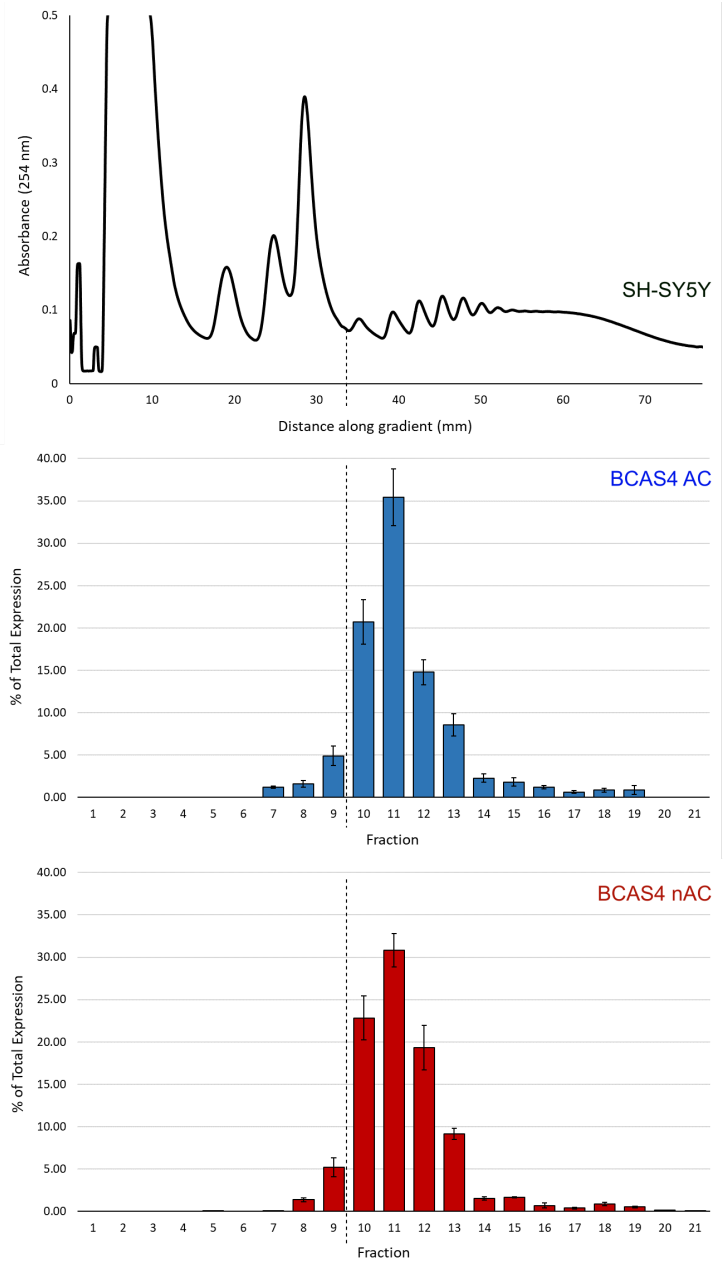


Figure 6.6 Ribosomal distribution of BCAS4 AC and nAC mRNAs in SH-SY5Y

A similar ribosomal distribution was observed in AC and nAC BCAS4 mRNAs indicting similar translation levels of both mRNAs in SH-SY5Y cells ($p > 0.3$ in all cases.). Ribosome-bound mRNA began in fraction 10 and increased in a fraction-based manner until approximately 6+ ribosomes at which point absorbance became convoluted. Error bars represent the standard error of three technical repeats.

There is generally a higher translation of both AC and nAC BCAS4 mRNAs in NP-1 cells than in SH-SY5Y. This can be observed in figure 6.7, where 45% and 51% of mRNA is bound by 3-5 ribosomes, respectively. However, unlike in SH-SY5Y, a difference in mRNA bound by two or fewer or 6+ ribosomes was observed. A small difference is observed in the case of mRNA bound to two or fewer ribosomes, with 20% of AC mRNA bound and 25% of nAC bound. However, the AC has 30% of mRNA bound to 6+ ribosomes in comparison to only 18% for the nAC, indicating more efficient translation of the AC variant. Performance of a Student's T-Test gave *p* values of 0.53, 0.50, and 0.07 for < 2, 3-5 and 6+ ribosomes, respectively. Though this showed no significant difference in mRNA bound to 0-5 ribosomes, there may be a significant difference in amount of BCAS4 AC and nAC mRNA bound by 6+ ribosomes in NP-1 cells. The difference in size of mRNAs is less than 100 bases which would account for only one additional bound ribosome. Taking this into account, AC BCAS4 mRNA may be more efficiently translated in NP-1 cells. However, in order to determine this, three biological repeats would need to be performed rather than three technical repeats. Unfortunately, due to time constraints, three biological repeats were not possible.

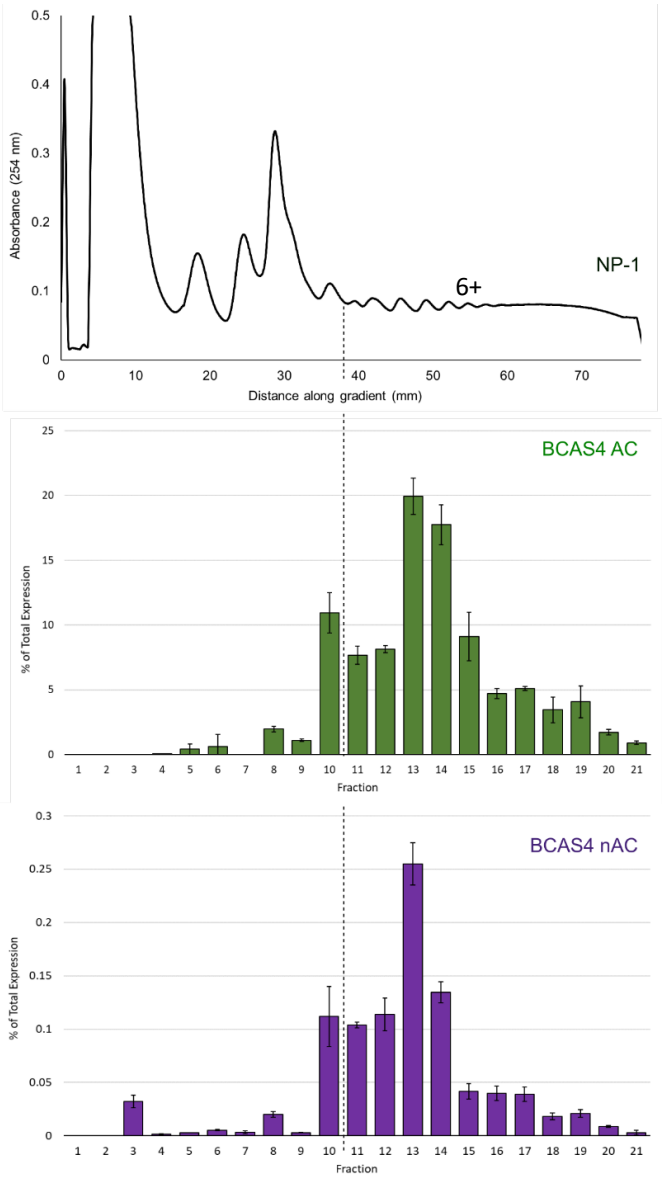


Figure 6.7 Ribosomal distribution of BCAS4 AC and nAC mRNAs in NP-1 cells

A different ribosomal distribution was observed for AC and nAC BCAS4 mRNAs indicating different translation efficiency of both mRNAs in NP-1 cells (p = 0.07 for 6+ ribosomes). Ribosome-bound mRNA began in fraction 10 and increased in a fraction-based manner until approximately 6+ ribosomes at which point absorbance became convoluted. Error bars represent the standard error of three technical repeats.

6.3 Conclusions on the translation of *Alu* and non-*Alu* mRNAs in cancerous and non-cancerous tissue samples

Three mRNA targets were initially chosen for polysome profiling by qPCR; ASCC1, BCAS4 and NEK4. Primers were designed to individually amplify *Alu*-containing (AC) or non-*Alu*-containing (nAC) isoforms *via* qPCR. mRNA trancript translation was measured in two human cell types: NP-1 (primary brain cells derived from epilepsy patients) and SH-SY5Y (stable neuroblastoma cell line derived from patient bone marrow). A third cell type, GBM1 (primary glioblastoma cells) was cultured and fractioned from a sucrose gradient. Due to the low number of cells obtained from fractionation, qPCR was not performed on these samples at this time. However, samples were stored for potential future use in higher sensitivity assays. Given more time, more cells would have been cultured and gradients would have been repeated. Both AC and nAC mRNAs for ASCC1 were insufficiently translated in both NP-1 and SH-SY5Y cells to obtain a standard curve for analysis of ribosomal distribution. This was surprising as publically available RNA-Seq data reported ASCC1 mRNA to be expressed in all tissues, and particularly well expressed in brain tissue (Human Protein Atlas). The same was true for the nAC variant of NEK4. As a result, it was confirmed that in both cell lines, the *Alu*-containing mRNA of NEK4 was predominantly translated. A translation of NEK4 AC mRNA was observed in NP-1 cells with the 39.5% of transcript mRNA bound to six or more ribosomes, compared to 9% in SH-SY5Y. Both AC and nAC BCAS4 mRNAs are translated at a similar level in SH-SY5Y cells, with approximately 60% of mRNA bound by two or fewer ribosomes, indicating low translation efficiency. However, in NP-1 cells, a higher proportion of *Alu*-containing mRNA isoform was observed to be bound by six or more ribosomes (39.5%) when compared than for the non-*Alu*-containing isoform (17.9%). Although, 3 + ribosomes is generally considered to represent a well translated transcript, the more associated ribosomes, the better the translation. As 6+ ribosomes is the point at with the polysome graphs in this project reach their upper limit is 6+ ribosomes, this is used as the upper limit for highly translated species. Unfortunately, due to time constraints, only three technical repeats could be obtained for BCAS4 results in a *p*-value of 0.07. In order to truly determine significant difference between AC and nAC BCAS4 transcripts, at least three bioligcal repeats would be required. This indicated higher translation level of the AC than the nAC

mRNA isoform. As the difference in transcript length is only 97 bp, only one additional ribosome can be attributed to the change in size. This indicated that as for BCAS4, the *Alu*-containing transcript is predominant in non-cancerous brain cells (NP-1).

Polysome profiling techniques coupled with qPCR provides a good basis for the study of alternative gene transcripts which arise as a result of *Alu* insertions. It was possible to study mRNA translation in different cell samples, including those that are patient-derived, providing that translation levels were high enough for detection by qPCR. Cell lines that can be cultured to a high cell concentration are more feasible for these studies. Unfortunately, in this case the GBM1 cell batch did not replicate efficiently and hence, an insufficient number of cells was obtained for analysis. However, cells were stored for potential use in future, higher-sensitivity assays. With more time and carefully chosen cell samples, it could be possible to directly compare any differences in translation of AC and nAC mRNAs from the same gene in different tissues. Ideally, cell samples would be patient-derived and would be matched dependent on host. With carefully chosen cells, it could be possible to compare the translation of AC and nAC transcripts between different cell types which not only relate to disease but also trends in gender, ethnicity and age could be studied. As the extent of *Alu*-containing protein-coding transcript still remains an understudied area within the field, it may be important to study the changes in *Alu* translation between different cell types/cell origins, not only to discover disease but also to understand the way their translation may change over time or with geographic location. This may provide more insight into their diversity in the human population. So far, no research has been reported which compares the translation efficiency of *Alu* and non-*Alu* transcripts which arise from protein-coding genes.

Chapter 7
Conclusions and future work

The aim of this work was to establish the extent of *Alu* exonisation and hence, explore the abundance and nature of translated *Alu* elements in the human proteome. The abundance of *Alu* exonisation and any related trends were studied *via* bioinformatic methods strongly based around local alignment of nucleotide and protein sequences. The nature and effect of translated *Alu* sequences on protein structure and function was explored *via* peptide binding studies, protein expression in *E. coli* and subsequent biophysical studies. Additionally, the changes in the expression of *Alu* and non-*Alu* mRNA in 'cancerous' and 'non-cancerous' human cell samples was studied using a combination of polysome profiling and quantitative PCR.

7.1 Individual conclusions

7.1.1 Bioinformatic analysis

Initial bioinformatic analyses compared nucleotide consensus sequences, and respective translated open reading frames (ORFs), of eight *Alu* subfamilies; J, Sx, Sp, Sq, Sc, Sb, Sb1 and Yb (formerly Sb2). Sequence alignment showed 80% conservation between subfamilies. Translation of each consensus sequence into its six possible ORFs and subsequent BLAST analysis (sequence identity > 60%, E-value < 1×10^{-8}) identified 46 human proteins containing an *Alu*-like insertion. Of the identified hits, 32 gave rise to both AC and non-*Alu*-containing (nAC) isoforms. Analysis of the locations of *Alu* insertions within protein hits revealed a bias towards terminal insertions over insertions internal to the protein sequence.

Further analyses using larger a dataset of 37 *Alu* consensus sequences, made available by the Dfam database, allowed for identification of the parent *Alu* of each insertion. Alignment of insertions with *Alu* sequences at the nucleotide level revealed that 88% of *Alu* insertions were derived from the *Alu* right arm. In addition to this, 83% of insertions were copies of the antisense (-) strand of the parental *Alu*.

Alignment of *Alu* insertions revealed a sequence of residues which was conserved between most protein hits:

LECS-X_1-GAISAHCNLRLLGSSD-X_2–PASASQ-X_3-AGITG

Direct alignment with protein hits confirmed conservation. This result was concurrent with previous observations that the majority of *Alu* insertions arose from the left arm and were copies of the antisense (-) strand of their parental *Alu*.
Following on from the work of Lev-Maor *et al*. which identified 3' AG splice sites at positions 290 and 286 of antisense *Alu* strands, an additional three new potential 3' AG splice sites were identified at positions 265, 258 and 116. Most splice sites lay relatively close together and therefore, accounted for the sequence conservation observed in *Alu* insertions. This bioinformatic work provides an origin for the extent and tolerance of *Alu* insertions within the human proteome. As a result, it may be easier to identify translated *Alu* sequences within proteins and predict their origin.

7.1.2 *Alu* binding with geldanamycin

The conserved *Alu* sequence identified during bioinformatic analysis was purchased as a peptide, SEA-001, and used in binding studies with geldanamycin, a potent anti-cancer molecule. Previous work at Warwick University had used phage display to identify a binding interaction between geldanamycin and a translated *Alu* sequence displayed on the surface of a phage. This work aimed to recreate the interaction and utilise it as a potential target for cancer therapeutics. Though several biophysical binding techniques (FA, FI, SPR, ITC) were used, the binding interaction between geldanamycin and the *Alu*-like sequence could not be recreated.

7.1.3 *Alu*-containing human proteins

Initially, five human proteins were chosen for overexpression in *E. coli*; NEK4 (Never in mitosis A (NIMA)-related kinase 4), ZMAT1 (Zinc finger matrin-type protein 1), PPP5D1 (PPP5 TPR repeat domain-containing protein 1), BCAS4 (breast cancer amplified sequence 4) and ASCC1 (activating signal cointegrator 1 complex subunit 1). Proteins were chosen dependent on availability of gene purchase, the association of the protein with disease and the predicted ease of protein expression.

Genes were cloned into a series of expression plasmids and transformed into *E. coli*. A number of different overexpression and purification methods were attempted to express the *Alu*-containing (AC) variant of each protein, however, no purified soluble protein products were obtained.

Unfortunately, as none of the above proteins could be produced, it was not possible to assess the effect of the *Alu* on their structure and function. Due to their previous links with disease, it would have been insightful to determine whether the *Alu* insertion contributed to faults in the protein.

7.1.4 Effect of *Alu* insertions on MBP

Due to the difficulties arising in the purification of human proteins from *E. coli*, maltodextrin binding protein (MBP) was used as a model system to study the effect of *Alu* elements on proteins. The peptide sequence, SEA-001, was inserted into seven different locations within MBP, labelled by the residue lying directly after the insertion: G6 (N-terminal), T81, P126, D178 (β-sheet), G253, A293 (α-helix), N333 and T367 (C-terminal). A second sequence, a scrambled variation of SEA-001, labelled SEA-002, was inserted at the D178, G253 and N333 locations to give a total of twelve constructs including wild-type MBP.

SEA-001	L E C S G A I S A H C N L R L L G S S D S P A S A S R V A G I T G
SEA-002	I A R L H G P S A S N G T S S S T C A P D L G V G E S A L C I S R

Circular dichroism (CD) revealed that the *Alu* insertion had no distinct secondary structure. Thermal stability studies using CD were inconclusive and further analysis using differential scanning calorimetry (DSC) showed minimal effect of the *Alu* on the thermal stability of MBP.

Quick purification of the MBP constructs *via* amylose affinity chromatography showed that most variants bound to maltose indicating correct protein folding. Exceptions to this were the α-helix insertion at position A293 which was expected to disrupt folding, and insertions at G253 and N333 positions. Previous reports have suggested that mutations at the N333 positions affect the hinge movement of MBP and therefore, the lack of binding with an insertion at this position was not surprising. Insertion of the scrambled *Alu* sequence (N333*) at the same position showed no restoration of binding. The insertion at G253 lies further from the binding

site and replacement of the *Alu* sequence with a scrambled variant (G253*) appeared to slightly restore binding.

Isothermal titration calorimetry (ITC) showed minimal changes in the binding of MBP with its ligands; D-(+)-maltose, maltotriose and β-cyclodextrin, upon the insertion of an *Alu* sequence at positions G6 and T367. The binding of β-cyclodextrin to both *Alu* and scrambled *Alu* (*) insertions at the D178 position observed an endothermic binding event which contradicted the usual exothermic curve. Due to the location of the insertion, relatively close to the binding site of ligands, it was concluded that insertions at this position appeared to favour a closed protein conformation and as such, the binding of larger ligands such as β-cyclodextrin required an unfavourable conformational change. As the same effect was observed for the scrambled variant, it was determined that this change was a result of the insertion site and was independent of the *Alu* sequence.

This work provides an insight into the effect of *Alu* elements on human proteins and hence begins to predict how such insertions may influence protein structure and function. Combined with the list of known *Alu* insertions in human proteins identified *via* bioinformatics, and *Alu* insertions which may be identified in other work, it may be possible to predict whether their presence may disrupt the function alternate protein isoforms, and therefore, could be used as a means to begin to predict faulty proteins or disease.

7.1.5 Polysome profiling

Polysome profiling provided a way in which to separate ribosome-bound mRNA *via* sedimentation through a sucrose gradient and subsequent fractionation. Analysis of this mRNA using qPCR, with primers for selected targets, allowed for determination of *Alu* and non-*Alu* transcript expression within different cell lines.

Three targets were chosen for analysis by qPCR; ASCC1, BCAS4, NEK4. Primers were designed so as to selectively target AC and nAC mRNAs for each gene. Three cell samples, SH-SY5Y, NP-1 and GBM1, were chosen for this work. qPCR analysis of mRNAs from GMB1 cells was not performed due to low cell count, which yielded too little mRNA for subsequent analysis, according to previous experiments.

AC and nAC mRNAs of ASCC1 were not studied as translation levels were too low to obtain a sufficient standard curve for quantification. Similarly, a suitable standard curve for NEK4 nAC was not obtained. However, as the AC mRNA was translation

to a sufficient level, it can be concluded that, in both SH-SY5Y and NP-1 cells, the AC NEK4 transcript is predominantly translated. A higher translation efficiency of *Alu*-containing NEK4 was observed in NP-1 cells (brain tissue) than in SH-SY5Y (bone marrow), with the highest proportion of mRNA being bound by six or more ribosomes.

Both BCAS4 AC and nAC mRNAs were translated at a similar level in SH-SY5Y cells, with 60% of mRNA bound to two or fewer ribosomes in both cases. In NP-1 cells, more efficient translation of AC mRNA is seen compared to the nAC mRNA with 30% of mRNA bound to six or more ribosomes, compared to 18%, respectively. This indicated that in non-cancer brain cells, the *Alu*-containing variant of BCAS4 is predominately translated.

Polysome gradienting coupled with qPCR provided a novel way to compare the study of alternative mRNAs of *Alu*-associated genes, whereby the translation efficiency of AC and nAC transcripts can be compared within a single cell line, provided that translation is of a high enough level to obtain a suitable standard curve for quantification. In addition to this, the study can also be applied to *Alu*-associated mRNAs within patient-derived cell lines allowing for the potential examination of *Alu* elements in a wide variety of diseases, including cancer. Unfortunately, in this case, the cancerous brain cell line, GBM1, was insufficient to make a direct study of the change, if any, of AC and nAC translation levels between 'healthy' and 'cancerous' brain cells. However, by choosing carefully selection patient-derived cell samples, it should be possible to perform such comparisons.

7.2 Overall conclusions

Using bioinformatic analysis, a well conserved *Alu* sequence that was incorporated into protein-coding regions of human DNA hits was identified. This conservation arose from the incorporation of 3' splices sites associated with the antisense *Alu* left arm which led to *Alu* exonisation. In addition to two 3' AG splice sites earlier identified by Lev-Maor *et al*. (positions 290 and 286), three other potential 3' AG splice sites were observed which could lead to *Alu* exonisation (265, 258 and 116). In many cases, the exonisation of *Alu* sequences led to an alternative splicing event which resulted in the expression of an *Alu*-containing (AC) protein isoform; 32 proteins were identified to have both AC and non-AC (nAC) isoforms in this study. Though the effect of natural *Alu* insertions in human proteins could not be studied due to problems arising in protein overexpression and purification, the *Alu* sequence SEA-001 appeared to have minimal effect when inserted into multiple different sites within maltodextrin binding protein (MBP). Sites which appeared most effected were N333, which is reported to affect the hinge motion of MBPs open binding conformation, G253 and D178. Insertions at D178 showed hindrance to MBP's binding of β-cyclodextrin and were assumed to contribute to a unfavourable open binding conformation which was required to accommodate the large ligand. However, this interaction was sequence-independent and arose due to the position of the *Alu* insertion as opposed to the *Alu* sequence itself. Ligand binding at position G253 was also abolished upon *Alu* insertion and appeared to be minorly restored upon insertion of a scrambled sequence at the same position. However, in all cases, the location of the insertion had a predominant contribution to changes in protein function and any minor contributions from the *Alu* sequence occurred as a secondary factor.

In most cases, the presence of an *Alu* sequence in a protein is likely to have a minimal effect as most insertions occur at protein termini and hence, are unlikely to be in locations that greatly contribute to protein function. However, in the small number of cases were insertions occur in an internal protein region, it is possible that the insertion could lead to a disease-causing isoform with limited function. It should be noted, that any functional changes that may arise would likely be as a result of the insertion location not the sequence. Nonetheless, if the sequence can be recognised

as an *Alu* and is not present in an alternative functional protein isoform, it could be speculated that the *Alu* sequence could still be used as a drug target.

Using a combination of polysome profiling, reverse transcription and qPCR, it was possible to selectively target AC and nAC mRNAs arising from the same gene and hence, compare their translation levels in both primary cells and stable cell lines. Through the application of this to carefully selected patient-derived cells, it should be possible to measure difference in translation of such transcripts in a range of disease-associated cells and their 'healthy' counterparts. This would give scope for the prediction of whether increased or decreased translation efficiency of *Alu*-associated mRNAs could be an indicator of certain diseases.

In general, the extent of *Alu* translation in human proteins is understudied. This project has provided a basis to probe the extent and effect of translated *Alu* elements on the human proteome. This work builds on previous work to provide insight as to how *Alu* elements can lead to the production of alternate transcripts and are translated into alternate protein isoforms. Furthermore, it begins to predict how such isoforms may be affected by the insertion in terms of structure and functionality. This could be used as a way to predict how *Alu* insertions, identified in disease-related proteins might contribute to the disease. Through the study of translational efficiency of AC and nAC transcripts using polysome-profiling, this project has provided a toolkit to study *Alu* transcripts and their translation in comparison to non-*Alu* transcripts arising from the same gene. Through careful selection of cell samples, it could be possible to not only study changes in the translation of *Alu* transcripts in relation to disease, but it may also be possible to study trends in their translation with respect to gender, age or ethnicity. Overall, the project provides a deeper understanding of *Alu* elements and their importance to the human genome and proteome.

7.3 Future work

The bioinformatic work associated with this project, in combination with other reported computational analyses, has provided a basis for the study of *Alu* effects on the human proteome. As studies into the physical effects of *Alu* transcripts are still limited, this leaves a large area still to be explored. Ideally, AC and nAC isoforms of naturally occurring human proteins would be overexpressed and studied for changes in their structure and function, in particular, those known to be disease-related. However, it is likely that a different expression system such as insect cells or human cells would be required to obtain a purified protein product for many of the hits observed in bioinformatic analysis. This would yield lower concentrations of protein. In addition, expanding polysome profiling techniques to a variety of cell types and *Alu*-associated targets would not only highlight potentially disease-associated *Alu* insertions, but may also provide a means to direct which proteins should be studied.

Chapter 8
Bioinformatic Methods

8.1 Building and refinement of an *Alu* database

The bioinformatic methods used in this research were a series of informative techniques by which comparisons could be made between nucleotide or protein sequences. Sequences, or partial sequences, were either be directly compared to another sequence, or multiple sequences; or screened against a known database of sequences. Bioinformatic analyses were carried out using NCBI BLAST, ExPASy Translate and ProtParam, Clustal Omega, JackHMMER, Dfam and MEGA7. The eight *Alu* consensus sequences initially used to identify *Alu*-containing proteins were obtained from the NCBI database. This was later expanded to 37 subfamily sequences from the Dfam database.

8.1.1 BLAST analysis of translated *Alu* sequences

Eight *Alu* consensus sequences (J, Sx, Sp, Sq, Sc, Sb, Sb1 and Sb2) were obtained from the National Centre for Biotechnology Information (NCBI) database. Each *Alu* consensus was translated into its corresponding six open reading frames (ORFs) using ExPASy Translate. This gave a total of 48 translated *Alu* sequences. Individual ORFs for each *Alu* consensus were screened against the UniProt/Swiss-Prot database[3] of known human proteins using NCBI protein BLAST (Basic Local Alignment Search Tool). Full BLAST parameters are outlined in table 8.1. Protein 'hits' were recorded in a Microsoft Excel database. Putative and uncharacterised proteins were dismissed. Proteins with an identity match of less than 68% were also omitted. For each protein hit, the following information was recorded in the database; protein name, protein accession number, the number of protein isoforms and the differences between them. For isoforms matching translated *Alu* sequences, the following was also recorded; expect value, location of the match in both the protein hit and the translated *Alu* sequence, the percentage identity match and the number of bases matched. 57 protein matches were identified. An example database entry is shown in figure 8.2.

Search Set		
Database	UniProt/Swiss-Prot (swissprot)	
Organism	Humans (taxid: 9605)	
Program selection		
Algorithm	blastp (protein-protein blast)	
General parameters		
Max. target sequences	100	
Expect threshold	10	
Word size	6	
Max. matches in query range	0	
Scoring parameters		
Matrix	BLOSUM62	
Gap costs	Existence: 11	Extension: 1
Compositional adjustments	Conditional compositional score matrix adjustment	
Filters and masking	None	

Table 8.1 NCBI BLAST search parameters

The parameters used for initial screening of translated *Alu* ORFs to identify potential proteins containing regions encoded by *Alu*-like sequences.

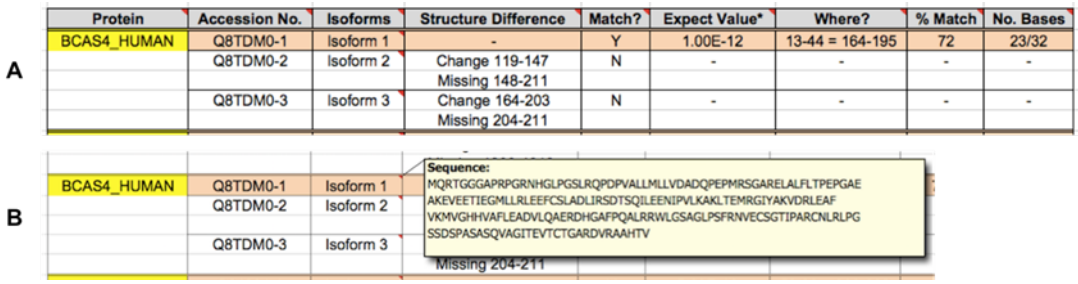


Figure 8.1 Example of a database entry from refined search results

Entries showed the name and accession number of hits as well as information on different isoforms. Red markers next to isoforms indicated where protein sequences have been recorded (B). AC isoforms also had details listed about the matched region. Red markers next to column headings indicated where additional information (e.g. heading definition) could be observed as a 'pop-up' in the database.

8.1.2 Database refinement

Translated *Alu* sequences were individually aligned with each isoform of their corresponding protein hits to confirm matched regions. In addition to the original percentage identity threshold of > 68%, an additional threshold of an E-value of less than 1×10^{8} was used. This additional threshold reduced hits from 57 to 46. A secondary database was built for each *Alu* consensus to contain the refined data, as well as additional information on possible protein function, the tissue in which it is expressed and any disease association (information obtained from UniProt).

8.2 Alignments to determine *Alu* locations in proteins

Proteins were defined as three distinct regions; N-terminal (0 – 20% of residues), internal (20 – 80%) and C-terminal (80 – 100%). Translated *Alu* sequences and individual protein hits were aligned with one another using the NCBI protein BLAST parameters outlined in table 7.1. The resulting alignments defined start and end regions (residue numbers) of each match. Using the insertion mid-point and the total length of the protein hit (obtained from NCBI database and UniProt), the location of the insertion in the protein could be calculated as a decimal (equation 8.1).

$$Equation\ 8.1 \quad Insertion\ location\ in\ protein = \frac{Insertion\ midpoint\ (Residue\ no.)}{Total\ length\ of\ protein\ (AA)}$$

Calculated values below 0.20 (< 20%) were labelled as N-terminal insertions. Those between 0.21 and 0.79 (21 – 79%) were labelled as internal and those lying above 0.80 (> 80%) were classed as c-terminal.

8.3 Alignments to determine which *Alu* region leads to insertions

Bioinformatic analysis was performed on nucleotide sequences of identified hits. For each hit, the cDNA sequence was obtained from the NCBI database. Through direct alignment with the corresponding *Alu* consensus sequence using NCBI nucleotide BLAST, the region of the *Alu* from which the insertion arose could be

identified. As well as identifying the matched region, it was also important to note the strand orientation (sense/antisense).

The nucleotide at the mid-point of the matched region was compared to the total length of the *Alu* consensus to give a ratio between 0 and 1 (equation 8.2)

$$Equation\ 8.2 \qquad Location\ in\ Alu = \frac{Midpoint\ of\ matched\ region}{Total\ length\ of\ Alu\ (bp)}$$

Mid-points giving a ratio of 0 – 0.5 in the sense direction, or 0.5 – 1.0 in the antisense direction, corresponded in insertions from the *Alu* left arm. Mid-points giving a ratio of 0 – 0.5 in the antisense direction, or 0.5 – 1.0 in the sense direction, corresponded in insertions from the *Alu* right arm.

8.4 8.4 Alignments to identify sequence conservation between hits

NCBI alignments of translated *Alu* sequences and protein hits gave the location of matched residues within the protein. Using ProtParam to view the protein sequence, the exact matched region was procured for each hit. Matched regions for each protein were aligned using JackHMMER. For proteins with multiple isoforms of the same matched region, only one copy of the sequence was analysed to give a total of 46 insertion sequences. Model positions were generated using JackHMMER to show sequence conservation.

Further sequence conservation analysis was performed using only hits obtained from primary reading frames, totaling to a JackHMMER alignment of 26 insertion sequences. A generally well conserved sequence was obtained through analysis of model positions generated through JackHMMER.

8.5 Comparison of 3’ splice sites

8.5.1 Examining 3’ splice sites in Dfam consensus sequences

50 *Alu* subfamily consensus sequences were obtained from the Dfam database. For each sequence, the reverse complement was obtained to give the antisense strand sequence. *Alu* sequences were aligned using MEGA7 software and examined for AG splice sites matching those reported by Lev Maor et al (see chapter 2). 3’ AG splice sites were identified manually through visualisation in Microsoft Excel.

8.5.2 Examination of 3' splices sites in matched genes

Nucleotide sequences for matches genes were compared to the Dfam database using their online search tool, resulting in the identification of the parent *Alu* for each insertion and the direction of the parent *Alu* strand (sense/antisense). Genes with insertions corresponding with antisense strand insertions were directly aligned with their parent *Alu* consensus sequence using MEGA7 software. AG splice sites were identified manually through visualisation in Microsoft Excel. New 3' AG splice sites were identified through direct comparison of the parent *Alu* sequence and the AG doublets located around the site of insertion in the gene.

Chapter 9
Biological materials and methods

9.1 Cloning methods

9.1.1 Site-directed mutagenesis of pDB.His.MBP

The pDB.His.MBP plasmid was mutated using a Quikchange II site-directed mutagenesis (SDM) protocol to introduce a c-terminal STOP codon in the MBP-encoding gene. The resulting plasmid will from hereon be referred to as pDB.His.MBP.Stop.

9.1.1.1 Parental plasmid

The parental pDB.His.MBP plasmid (Clone ID: EvNO00085130) was purchased from DNASU plasmid repository.

9.1.1.2 Primer design

Primers SA128 and SA129 were designed in accordance with the Quikchange II SDM protocol. Primers were made complementary to one another and were 47 bp in length with a melting temperature (T_m) of 67.7 ºC and a GC content of 49%. The desired mutation was contained in the middle of the primer flanked by 15 or more bases of correct sequence on both sides.

SA128:

TCG ATG AAG CCC TGA AAG ACG CGC AGA CTT AGA CCG ATT ACG ATA TC

SA129:

GAT ATC GTA ATC GGT CTA AGT CTG CGC GTC TTT CAG GGC TTC ATC GA

9.1.1.3 Polymerase chain reaction (PCR)

PCR reactions were performed using a Techne™ TC-512 Gradient Thermal Cycler (Bibby Scientific). Site-directed mutagenesis was performed using the polymerase chain reaction (PCR) parameters outlined below.

Component	Amount
10 × reaction buffer	5.0 μL
Primer A: SA128 (10 μM)	2.5 μL
Primer B: SA129 (10 μM)	2.5 μL
Plasmid DNA (pDB.His.MBP)	10 - 20 ng
PfuUltra High Fidelity DNA polymerase (2.5 U/μL)	1.0 μL
Deionised water	Make up to 50 μL

Table 9.1 Composition of Quikchange II site-directed mutagenesis (SDM) mixture

Step	Temperature	Time	Cycle
Initial denaturation	95 °C	30 sec	1
Denaturation	95 °C	30 sec	18
Annealing	55 (± 10) °C	1 m 30 sec	
Extension	68 °C	8 m 00 sec	
Storage	4 °C	-	-

Table 9.2 PCR cycling parameters for Quikchange II site-directed mutagenesis (SDM)

9.1.1.4 PCR analysis

PCR products were treated with DpnI (1 μL; 10 U) and incubated at 37 °C for 2 hours to remove parental plasmid DNA. PCR product (5 μL) was mixed with 6 × purple loading dye (NEB) and separated on a 1% agarose (in 1 × TAE buffer) gel treated with SYBR safe gel stain (1:100) at 100 V for 30 minutes. Gels were analysed under UV light using a ChemiDoc™ imaging system (BioRad).

9.1.1.5 Transformation of PCR product in XL1-Blue Supercompetent cells

PCR product (5 μL) was added to XL1-Blue Supercompetent cells (50 μL; Agilent Technologies) and cooled on ice for 30 minutes. Cells were heat-shocked in a 42 °C water bath for 45 seconds then cooled on ice for a further 5 minutes. LB (300 μL) was added and cells were incubated at 37 °C for 1 hour, with shaking. Cells were centrifuged at 17,000 × g for 1 minute. Supernatant (200 μL) was removed and cells were re-suspended in the remaining volume. Cells were spread on an LB agar plate treated with kanamycin (50 μg/mL) and incubated at 37 °C overnight. Plates were stored at 4 °C.

9.1.1.6 DNA purification

Single colonies were picked and added to LB (10 mL) treated with kanamycin (50 μg/mL). Mini-cultures were incubated at 37 °C overnight, with shaking. Overnight mini-culture (4 mL) was centrifuged at 4500 × g and the supernatant discarded. DNA was extracted using a QIAprep® Spin Miniprep kit (Qiagen) following the manufacturer's instructions. DNA was eluted in buffer EB (30 μL) and stored at -20 °C.

9.1.1.7 Glycerol stocks

Single colonies were picked, added to LB (10 mL) treated with kanamycin (50 μg/mL) and mini-cultures were incubated at 37 °C overnight, with shaking. Overnight mini-culture (1 mL) was added to 50% glycerol solution in water (0.5 mL) and stored at -80 °C.

9.1.2 Cloning of MBP-*Alu* constructs

Alu regions were inserted into the pDB.His.MBP.Stop plasmid using site-directed mutagenesis via inverse PCR. PCR was performed following a Phusion® High-Fidelity Master Mix (NEB) protocol. Resulting MBP-*Alu* constructs are labelled via the encoded amino acid directly after the point of *Alu* insertion, with the exception of the c-terminal *Alu* insertion which is labelled by the threonine (T367) at the c-terminus of MBP.

9.1.2.1 Primer design

Primers SA152 – SA167 were designed to meet the criteria outlined below. The forward primer was comprised of the second half of the *Alu* sequence (48 bp) followed by 18 – 20 bp of correct plasmid sequence that lay directly after the desired site of insertion (shown in red). The reverse primer was the reverse compliment of 18 – 20 bp of correct plasmid sequence that lay directly before the desired site of insertion (shown in blue) followed by the first half of the *Alu* sequence (51 bp). Examples of forward and reverse primers are shown below:

Forward primer (e.g. SA152):

TCA TCC GAC TCG CCC GCC AGC GCG AGC CGC GTA GCA GGC ATC ACC GGA GAC ATT AAA GAC GTG GGC

Reverse primer (e.g. SA153):

GCC CGG TAA GCG AAG GTT GCA GTG AGC TGA AAT AGT CCC TGA GCA CTC CAA GTA CTT GCC GTT TTC ATA C

Annealing primer ends (i.e. those matching the destination plasmid) terminated in one or more C/G bases and had a melting temperature (T_m) between 52 and 69 °C, with no more than a 6 °C difference between the T_m of primer pairs.

9.1.2.2 Polymerase chain reaction (PCR)

PCR reactions were performed using a Techne™ TC-512 Gradient Thermal Cycler (Bibby Scientific). Site-directed mutagenesis via inverse PCR was performed using the parameters outlined below.

Component	Amount
2 × Phusion® Master Mix	25.0 µL
Forward primer (10 µM)	2.5 µL
Reverse primer (10 µM)	2.5 µL
Plasmid DNA (pDB.His.MBP)	10 - 20 ng
DMSO	1.5 µL
Deionised water	Make up to 50 µL

Table 9.3 Composition of Phusion® High-Fidelity (HF) Master Mix PCR mixture

Step	Temperature	Time	Cycle
Initial denaturation	95 °C	30 sec	1
Denaturation	95 °C	10 sec	35
Annealing	55 °C	30 sec	
Extension	72 °C	8m 00 sec	
Storage	4 °C	-	-

Table 9.4 PCR cycling parameters for site-directed mutagenesis (SDM) via inverse PCR using Phusion® HF Master Mix.

9.1.2.3 PCR analysis

PCR product was treated with DpnI (1 µl; 10 U) and incubated at 37 °C for one hour to remove parental plasmid. PCR product (50 µL) was mixed with 6 × purple loading dye (NEB) and separated on a 1% agarose (in 1 × TAE buffer) gel treated with SYBR safe gel stain (1:100) at 100 V for 30 minutes. Gels were analysed under UV light using a ChemiDoc™ imaging system (BioRad). Successful PCR products were excised and DNA was extracted using a QIAquick gel extraction kit (Qiagen) following the manufacturer's instructions and eluted in buffer EB (30 µL).

9.1.2.4 Phosphorylation and ligation of PCR product

PCR product (approx. 200 ng) was phosphorylated through the addition of T4 ligation buffer (0.5 µL) and T4 polynucleotide kinase (PNK: 0.5 µL; 5 U) at 37 °C for 30 minutes. The reaction was cooled to room temperature and T4 ligation buffer (0.5 µL), T4 DNA ligase (0.5 µL; 200 U) and deionised water (3.0 µL) were added. Ligation occurred for 1 hour at room temperature.

9.1.2.5 Transformation of ligation product in XL1-Blue Supercompetent cells

Ligation product (10 µL) was added to XL1-Blue Supercompetent cells (100 µL; Agilent Technologies) and cooled on ice for 30 minutes. Cells were heat-shocked in a 42 °C water bath for 45 seconds then cooled on ice for a further 5 minutes. LB (300 µL) was added and cells were incubated at 37 °C for 1 hour, with shaking. Cells were centrifuged at 17,000 × g for 1 minute. Supernatant (200 µL) was removed and cells were re-suspended in the remaining volume. Cells were spread on an LB agar plate treated with kanamycin (50 µg/mL) and incubated at 37 °C overnight. Plates were stored at 4 °C.

9.1.2.6 Cloning analysis

Single colonies were randomly selected and added to LB (10 mL) treated with kanamycin (50 µg/mL). Mini-cultures were incubated at 37 °C overnight, with shaking. Overnight mini-culture (4 mL) was centrifuged at 4500 × g and supernatant was discarded. DNA was extracted using a QIAprep Spin Miniprep kit (Qiagen) following the manufacturer's instructions. DNA was eluted in buffer EB (30 µL).

9.1.2.7 Glycerol stocks

Single colonies were picked, added to LB (10 mL) treated with kanamycin (50 µg/mL) and mini-cultures were incubated at 37 °C overnight, with shaking. Overnight mini-culture (1 mL) was added to 50% glycerol solution in water (0.5 mL) and stored at -80 °C.

9.1.3 Restriction-free cloning of human genes into *E. coli* compatible plasmids

Human genes were cloned into plasmids for expression in E. coli using restriction-free (RF) cloning. The BCAS4 gene was cloned into a pET-28a plasmid and ZMAT1 and ASCC1 genes were cloned into a modified pET-SUMO-28a plasmid.

9.1.3.1 Primer design

Forward and reverse primers were designed so as to be no longer than 50 bp in length. The forward primer contained two regions; 20-26 bp matching the desired insertion site of with the cloning plasmid (red) followed by 24 bp in length, matching the beginning of the target gene, starting with ATG (blue).

The reverse primer incorporated three regions to give the reverse complement of the last 24 bp of the target gene (red), followed by a STOP codon (black) and 20-23 bp matching the desired insertion site within the cloning plasmid (blue).

Examples for BCAS4 in pET-28a are shown below:

SA030 (Forward):

GCT CAC AGA GAA CAG ATT GTT GGA ATG CAG CGG ACC GGG GGC GGG GCT

SA031 (Reverse):

TCG ACG GAG TCT GAA TTC GGA TTA TAC AGT GTG GGC AGC TCG TAC

9.1.3.2 PCR amplification

Restriction-free (RF) cloning consists of two rounds of PCR amplification, separated by a PCR purification step prior to Dpn1 treatment and transformation. PCR reactions were performed using a Techne™ TC-512 Gradient Thermal Cycler (Bibby Scientific). PCR amplification round 1 (RF_1) was performed using the parameters outlines in tables 9.5 and 9.6.

Component	Amount
2X Phusion® Master Mix	25.0 µL
Forward primer (10 µM)	2.5 µL
Reverse primer (10 µM)	2.5 µL
Template DNA*	2.5 µL
DMSO	1.5 µL
Deionised water	Make up to 50 µL

Table 9.5 Composition of Phusion® High-Fidelity (HF) Master Mix PCR mixture for RF amplification round 1 (RF_1)

*Template DNA as made up to 20 ng/µL

Step	Temperature	Time	Cycle
Initial denaturation	98 °C	30 sec	1
Denaturation	98 °C	10 sec	35
Annealing	68 °C	30 sec	
Extension	72 °C	30 sec/kb template DNA	
Final extension	72 °C	5 minutes	1
Storage	4 °C	-	-

Table 9.6 PCR cycling parameters for RF amplification round 1 (RF_1)

RF_1 amplification product was analysed by electrophoresis on a 1% agarose gel. Successfully amplified samples were purified using a QIAquick PCR purification kit (Qiagen).

In RF amplification round 2, the product from RF_1 acts a mega primer for the reaction. PCR amplification round 2 (RF_2) was performed using the parameters outlines in tables 9.7 and 9.8.

Component	Amount
2X Phusion® Master Mix	25.0 µL
RF_1 amplification product	Variable*
Destination plasmid	1.0 µL
DMSO	1.5 µL
Deionised water	Make up to 50 µL

Table 9.7 Composition of Phusion® High-Fidelity (HF) Master Mix PCR mixture for RF amplification round 2 (RF_2)

*RF_1 amplification product concentration was measured using a Nanodrop 2000 (Thermo Scientific) and the volume used was calculated at a molar ratio of 20:1 of insert to plasmid with 20 ng of parental plasmid starting material.

Step	Temperature	Time	Cycle
Initial denaturation	98 °C	30 sec	1
Denaturation	98 °C	10 sec	20
Annealing	68 °C	30 sec	
Extension	72 °C	2 min/kb destination plasmid	
Final extension	72 °C	5 minutes	1
Storage	4 °C	-	-

Table 9.8 PCR cycling parameters for RF amplification round 2 (RF_2)

9.1.3.3 Transformation of RF_2 amplification product in XL1-Blue Supercompetent cells

Dpn1 (1 µL) was added to RF_2 amplification product (10 µL) and incubated at 37 ºC for 1 hour, with shaking. The resulting solution was added to XL1-Blue Supercompetent cells (100 µL) and cooled on ice for 30 minutes. Cells were heat-shocked at 42 ºC for 45 seconds and cooled on ice for a further 5 minutes. LB media (300 µL) was added and cells were incubated at 37 ºC for 1 hour, with shaking. Cells were centrifuged at 17,000 × g for 1 minute. Supernatant (200 µL) was removed and cells were re-suspended in the remaining volume. Cells were spread on an LB agar plate treated with kanamycin (50 µg/mL) and incubated at 37 °C overnight. Plates were stored at 4 °C.

9.2 Protein Expression

9.2.1 Expression of His-tagged MBP-*Alu* proteins

All His-tagged MBP-*Alu* variants were transformed into Rosetta 2(DE3) competent cells (Novagen), grown in standard LB media (see 9.6.3 for recipe), and proteins were expressed using IPTG induction.

9.2.1.1 Transformation of plasmid DNA in Rosetta 2(DE3) Competent cells

Plasmid DNA (1 μL) was added to Rosetta 2(DE3) competent cells (50 μL) and cooled on ice for 30 minutes. Cells were heat-shocked in a 42 °C water bath for 45 seconds then cooled on ice for a further 5 minutes. LB (300 μL) was added and cells were incubated at 37 °C for 1 hour, with shaking. Cells were centrifuged at 17,000 × g for 1 minute. Supernatant (200 μL) was removed and cells were re-suspended in the remaining volume. Cells were spread on an LB agar plate treated with kanamycin (50 μg/mL) and chloramphenicol (25 μg/mL) and incubated at 37 °C overnight. Plates were stored at 4 °C.

9.2.1.2 Glycerol stocks

Single colonies were picked, added to LB (10 mL) treated with kanamycin (50 μg/mL) and chloramphenicol (25 μg/mL) and mini-cultures were incubated at 37 °C overnight, with shaking. Overnight mini-culture (1 mL) was added to 50% glycerol solution in water (0.5 mL) and stored at -80 °C.

9.2.1.3 Cell growth and protein expression

A small amount of glycerol stock (or single plated colony) was added to LB (10 mL) treated with kanamycin (50 μg/mL) and chloramphenicol (25 μg/mL) and mini-cultures were incubated at 37 °C overnight, with shaking. Overnight mini-culture was added to LB (1 L) treated with kanamycin (50 μg/mL) and chloramphenicol (25 μg/mL). Large cultures were incubated at 37 °C, with shaking until OD_{600} was between 0.6 and 0.8. Cultures were cooled to 18 °C, induced with IPTG (0.5 mM) and incubated at 18 °C overnight, with shaking. Cultures were centrifuged at 10,000 × g for 20 minutes and supernatant was discarded. Pellets not undergoing immediate lysis and subsequent purification were stored at -80 °C.

9.2.2 Differential expression of MBP constructs

A small amount of glycerol stock was added to LB media (5 mL) treated with kanamycin (50 μg/mL) and chloramphenicol (25 μg/mL) and mini-cultures were incubated at 37 °C overnight, with shaking. Overnight mini-culture (150 μL) was added to auto-induction LB (15 mL) and incubated for 24 hours at 37 °C with shaking. OD_{600} for each sample was corrected to 1.0 ± 0.05. OD_{600}-corrected cultures (1.5 mL) were centrifuged at 13,000 × g for 10 minutes. Supernatant was removed and cells were resuspended in lysis buffer (20 mM Tris HCl, 150 mM NaCl, 2 mM $MgCl_2$, 1 mg/mL lysozyme, 3 mM deoxycholic acid, 16 U DNAse I, pH 7.5) and incubated at 37 °C for 1 hour, with shaking. Cells were centrifuged at 13, 000 × g for 10 minutes. Lysate (50 μL) was added to loading buffer (30 μL) and boiled at 95 °C for 3 minutes. SDS-PAGE samples (20 μL) were loaded on a 10% acrylamide gel, and run at 180 V for 1 hour. Gel was stained in InstantBlue™ Coomassie stain and imaged using a ChemiDoc™ MP imaging system (Bio-Rad).

9.3 Protein purification

Proteins were purified in one of two ways. Lysate was either directly purified using a double column (HisTrap and SEC) *via* FPLC. Or lysate was first purified manually *via* nickel affinity chromatography prior to further purification *via* size exclusion chromatography.

9.3.1 Purification *via* FPLC: double column (HisTrap and SEC)

Cell pellets were suspended in lysis buffer (approx. 30 mL: 20 mM Tris HCl, 150 mM NaCl, Pierce™ EDTA-free protease inhibitor (Thermo Fisher Scientific); pH 7.5). Cells were lysed via sonication on ice (60% power, 3 × 2 minutes) and centrifuged at 16,000 × g for 40 minutes. Supernatant was sterile filtered through a Minisart 0.45 μM syringe filter (Sartorius UK). Protein was purified from cell lysate *via* FPLC on an NGC™ Chromatography system (BioRad) using a HisTrap™ FF column (5 mL; GE Healthcare) and a HiLoad™ 16/60 Superdex™ 200 Prep Grade column (GE Healthcare). The following FPLC buffers were used: running buffer A (20 mM Tris HCl, 150 mM NaCl, pH 7.5) and elution buffer B (20 mM Tris HCl, 150 mM NaCl, 500 mM imidazole, pH 7.5). The resultant protein was concentrated, aliquoted and stored at -80 °C.

9.3.2 Purification *via* manual nickel affinity column prior to SEC

Cell pellets were suspended in lysis buffer (approx. 30 mL: 20 mM Tris HCl, 150 mM NaCl, Pierce™ EDTA-free protease inhibitor (Thermo Fisher Scientific); pH 7.5). Cells were lysed via sonication on ice (60% power, 3 × 2 minutes) and centrifuged at 16,000 × g for 40 minutes. Lysate was loaded onto Ni-NTA agarose resin (Generon) and washed with 10 × CV wash buffer (20 mM Tris HCl, 150 mM NaCl, 20 mM imidazole, pH 7.5). Protein was eluted with 1.5 × CV elution buffer (20 mM Tris HCl, 150 mM NaCl, 300 mM imidazole, pH 7.5). Eluted protein was immediately purified in wash buffer *via* FPLC on an NGC™ Chromatography system (BioRad) using a HiLoad™ 16/60 Superdex™ 200 Prep Grade column (GE Healthcare). The resultant protein was concentrated, aliquoted and stored at -80 °C.

9.3.3 Purification of MBP constructs *via* amylose affinity chromatography (as a prediction of MBP folding)

Amylose binding resin was used to predict whether *Alu* insertions into MBP affected the folding and subsequent ligand binding of the protein. Constructs containing the *Alu* insertion were compared with wild-type MBP.

Protein (1.0 mg) was loaded on equilibrated amylose resin (1.0 mL; New England Biolabs) and flow-through (FT) was collected. Resin was washed with three column volumes of column buffer CB (20 mM Tris HCl, 150 mM NaCl, 1 mM EDTA, pH 7.2) and washes were collected as 1 mL fractions. Bound protein was eluted with three column volumes of elution buffer (20 mM Tris HCl, 150 mM NaCl, 1 mM EDTA, 10 mM D-(+)-maltose, pH 7.2) and eluent was collected as 1 mL fractions. Fractions were analysed *via* SDS-PAGE.

9.4 Biophysical methods

Biophysical methods were performed using FITC- and biotin-labelled derivatives of geldanamycin and the SEA-001 peptide as well as the unlabelled species. The structures of the GM derivatives are shown in figure 9.1.

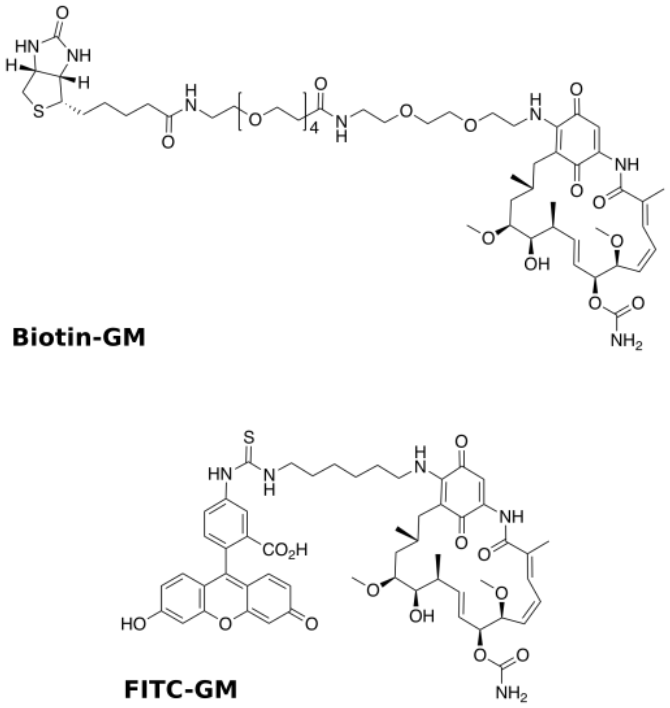


Figure 9.1 Structures of FITC-labelled and biotin-labelled GM

9.4.1 Fluorescence assays

Fluorescence assays were performed using an Envision Multiplate reader (Perkin Elmer) accessible from the School of Chemistry (University of Leeds). Fluoroscein isothiocyanate (FITC) was used as a fluorescent marker giving and excitation wavelength of 490 nm and an emission wavelength of 525 nm.

9.4.1.1 Fluorescence anisotropy

Fluorescence anisotropy experiments were performed in standard 96-well black polypropylene plates (Greiner Bio-One). Stock solutions of FITC-GM (1 mM) and SEA-001 (1 mM) were prepared in DMSO. Serial dilutions of SEA-001 were performed in the plate using buffer (PBS, 0.1 mg/mL BSA, 1 mM EDTA, 1 mM DTT). Earlier experiments gave a concentration range of 0.95 nM to 500 μM. Later experiments gave a concentration range of 1.47 nM to 775 μM. Three serial dilution repeats were performed per plate. FITC-GM was diluted to give a final concentration of 33 nM in each well. The following controls were used: buffer only, FITC-GM with buffer (no SEA-001) and a serial dilution of SEA-001 with buffer (no FITC-GM). Plates were read for fluorescence anisotropy immediately, after 1 hour and the next morning.

9.4.1.2 Fluorescence intensity

Fluorescence intensity experiments were performed in streptavidin-coated 96-well black polypropylene plates (Sigma Aldrich). Stock solution of biotinylated SEA-001 (1 mM) and FITC-GM (1 mM) were prepared in DMSO. Biotinylated SEA-001 was diluted to a final concentration of 100 μM in PBS. A serial dilution of FITC-GM in PBS buffer was made to give a range of concentrations from 0.24 – 750 μM. Streptavidin-coated plated were washed three times with blocking buffer (PBS, 0.05% Tween-20, 0.05% BSA) and three times with PBS. Biotinylated SEA-001 (100 μM) was added to well and allowed to bind at room temperature for 1 hour. Unbound SEA-001 was washed from the plate three times with PBS. FITC-GM dilutions were added to wells to give three dilution replicates. The plate was incubated at room temperature for three hours to allow for binding. Unbound FITC-GM was washed from the plate three time with PBS. The plate was read for fluorescence.

9.4.2 Surface plasmon resonance

SPR experiments were performed using a Biacore 3000 (Biacore). For each experimental run, one flow cell remained blank to allow for measurements of background response and any non-specific binding to the chip.

Stock solutions of GM (1 mM), biotinylated GM (1 mM), SEA-001 (1 mM) and biotinylated SEA-001 were prepared in DMSO. All buffers and samples were prepared to match the following buffer conditions; PBS, 0.1 mg/mL BSA, 0.01% Tween-20, 5% DMSO.

9.4.2.1 Immobilised SEA-001

Biotinylated SEA-001 (10 nM) was immobilised on a sensor chip SA (GE Healthcare) at a flow rate of 5 μL/min to give a response of approximately 250 RU. Injections of GM (ascending serial dilution; 0.39 μM – 10 μM) were passed over the flow cells at a flow rate of 40 μL/min for 2 minutes, followed by a 3 minute dissociation, with an injection of buffer before and after each series replicate. Injection series were repeated three times.

9.4.2.2 DMSO only controls

Biotinylated SEA-001 (10 nM) was immobilised on a sensor chip SA (GE Healthcare) at a flow rate of 5 μL/min to give a response of approximately 250 RU.

Injections of DMSO matching the DMSO concentrations present in a serial dilution of GM (0.39 – 10 μM) were passed over the flow cells at a flow rate of 40 μL/min for 2 minutes, followed by a 3 minute dissociation, with an injection of buffer before and after each series replicate. Injection series were repeated three times.

9.4.2.3 Immobilised GM

Biotinylated GM (100 nM) was immobilised on a sensor chip SA at a flow rate of 5 μL/min to a response of approximately 100 RU. Injections of SEA-001 (ascending serial dilution; 0.39 – 50 μM) were passed over flow cells at a flow rate of 40 μL/min, with an injection of buffer before and after each series replicate. Injection series were repeated three times.

9.4.3 Isothermal titration calorimetry

All ITC experiments were performed using a MicroCal iTC200 (Malvern). Prior to experimental runs, syringe, sample cell and reference cell were rinsed with surfactant, several cell volumes of water and finally buffer. The reference cell was filled with degassed water. Initial control runs: buffer into buffer and ligand into buffer, were performed to ensure that any background was kept to a minimum. All samples and buffers were matched through overnight dialysis prior to experimental runs.

9.4.3.1 ITC analysis of binding between geldanamycin and free *Alu* peptide, SA001

ITC was used to detect the presence of a potential binding interaction between the free *Alu* peptide, SA001, and geldanamycin (GM). Samples were dialysed in PBS buffer (PBS tablets; Sigma Aldrich) containing 5% DMSO and 0.01% TCEP.

The sample cell was loaded with 200 μL of peptide at a concentration of approximately 100 μM. Excess liquid was removed from around the cell. The syringe was filled with 40 μL of geldanamycin at a concentration of 1 mM. Cells were pre-heated to 25 ºC before lowering of syringe into sample then maintained at 25 ºC whilst DP (data signal) was measured. Ligand was titrated into protein over 20 × 4 s injections (1 × 0.5 μL sacrificial injection then 19 × 2 μL injections). There was a 120 s recovery time between injections. All data was fitted and analysed using NITPIC and SEDPHAT and results were displayed using GUSSI (NIH).

9.4.3.2 ITC analysis of binding between MBP constructs and native MBP ligands

Isothermal titration calorimetry (ITC) was used to study the effect of *Alu* insertion on the binding affinity of MBP to its native binding partners; D-(+)-maltose, β-cyclodextrin and maltotriose.[250] Mutated MBP-*Alu* constructs were compared with wild-type MBP. Samples were dialysed in a buffer containing 15 mM sodium phosphate and 50 mM NaCl at pH 7.2.

The sample cell was loaded with 200 μL of protein at a concentration of approximately 200 μM. Excess liquid was removed from around the cell. The syringe was filled with 40 μL of ligand at a concentration dependent on each ligand: D-(+)maltose (3.0 mM), β-cyclodextrin (2.4 mM) and maltotriose (3.6 mM). Cells were pre-heated to 25 ºC before lowering of syringe into sample then maintained at 25 ºC whilst DP (data signal) was measured. Ligand was titrated into protein over 20 × 4 s injections (1 × 0.5 μL sacrificial injection then 19 × 2 μL injections). There was a 120 s recovery time between injections. All data was fitted and analysed using NITPIC and SEDPHAT and results were displayed using GUSSI (NIH).

9.4.4 Circular dichroism

Circular dichroism (CD) was measured using a Chirascan™ Circular Dichroism spectrometer (Applied Photophysics) made available by the Astbury Centre of Structural Molecular Biology. CD data was analysed using CDNN and Global3

software (Applied Photophysics) and DichroWeb. Each sample was provided as 200 µL of ca. 0.2 mg/mL protein in 50 mM sodium phosphate.

CD for each construct was measured using Pro-Data Chirascan (Applied Photophysics) as the difference in absorbance of right-handed circularly polarised light and left-handed circularly polarised light between 180.0 and 260.0 nm at a bandwidth of 2.0 nm. Two repeats were performed per construct at a temperature of 20 °C. Secondary structure was then analysed using CDNN and DichroWeb software.

Time-dependent CD was performed between 180.0 nm and 260.0 nm at a bandwidth of 2.0 nm and a path length of 1.0 cm. A step-wise temperature ramp was performed at a range of 20 – 90 °C increasing at 1 °C/min. Temperature was cooled back to 20 °C between samples. The melting temperature (T_m) of each sample was calculated using Global3 software.

9.5 Polysome profiling

Polysome profiling was used to determine whether there was a change in transcript number of *Alu* and non-*Alu* transcripts between cancerous and non-cancerous cell lines. The following genes were studied; ASCC1, NEK4, BCAS4 and RPGRIP1L. Polysome profiling was performed in collaboration with the Aspden Group (University of Leeds), with cells provided by the Wurdak Group (St. James' Hospital, Leeds). All centrifugation was carried out at 4 ºC and after the harvesting of cells, all samples were kept on ice unless otherwise stated. Approximately 10 × 10^6 cells were used for each polysome gradient.

9.5.1 Cell preparation

9.5.1.1 SH-SY5Y cells

SH-SY5Y neuroblastoma cells were provided by the Aspden group (University of Leeds). Cells were grown in Dulbeco's modified eagle medium (DMEM) supplemented with 10% fetal bovine serum (FBS) and 1% penicillin streptomycin (Pen-Strep). All incubation periods were performed at a maintained temperature of 37 ºC.

Cells were thawed and added to pre-warmed growth media (12 mL) in a vented T75 flask. Cells were allowed to attach for 2-3 hours. Media was removed and fresh media (12 mL) was added. Prior to splitting, media was removed and 1 × trypsin (3

mL; Lonza) was added to cells and incubated for 5 minutes at 37 °C. Media (5 – 6 mL) was added to quench trypsin. Cells were split at a ratio of 1:11 every 5-7 days, allowing for no more than four passages (P_4) before plating.

Three days prior to polysome gradienting, 1.0 – 1.5 × 10^6 cells were added to medium (12 mL) in 10 cm plates.

9.5.1.2 GBM1 Cells

GBM1 glioblastoma cells were provided by the Wurdak group (University of Leeds). All GBM1 cells were grown on coated flasks prepared in the following way. Poly-l-ornithine (10 mg/mL) was added to T75 flasks at a dilution of 1:2000 (10 mL) and incubated at room temperature for 1 hour. Flasks were washed with sterile water. Laminin at a dilution of 1:500 (10 mL) was added and cells were incubated at room temperature overnight. Flasks that were not for immediate use were stored at -20 °C for up to two months. Prior to cell seeding, flasks were washed with PBS.

GBM1 cells were grown in a Neurobasal+ medium containing the following; neurobasal medium (Thermo Fisher Scientific) supplemented with 40 ng/mL recombinant human EGF protein (EGF; R&D Systems), 40 ng/mL human recombinant FGF-basic (FGF; Thermo Fisher Scientific), 0.5 × B-27 supplement (Thermo Fisher Scientific) and 0.5 × N-2 supplement (Thermo Fisher Scientific).

Cells were thawed, added to media (5 mL) and centrifuged at 300 × g for 5 minutes. Pellet was resuspended in fresh media (12 mL) and added to a vented T75 flask. Cells were incubated at 37 °C. Prior to splitting, media was removed and 1 × trypsin (3 mL; Lonza) was added to cells and incubated for 5 minutes at 37 °C. Media (5 – 6 mL) was added and cells were centrifuged at 300 × g for 5 minutes. Media was discarded and cells were resuspended in fresh media. Cells were split at a ratio of 1:11 every 5-7 days, allowing for no more than four passages (P_4) before plating.

9.5.1.3 NP1 Cells

NP1 brain cells were provided by the Wurdak group (University of Leeds). All NP1 cells were grown on coated flasks prepared in the following way. Poly-l-ornithine (10 mg/mL) was added to T75 flasks at a dilution of 1:2000 (10 mL) and incubated at room temperature for 1 hour. Flasks were washed with sterile water. Laminin at a dilution of 1:500 (10 mL) was added and cells were incubated at room temperature overnight. Flasks that were not for immediate use were stored at -20 °C for up to two months. Prior to cell seeding, flasks were washed with PBS.

Cells were grown in NP medium containing the following; DMEM-F12 (Thermo Fisher Scientific) supplemented with 20 ng/mL FGF, 20 ng/mL EGF, 0.5 × B-27 supplement, 0.5 × N-2 supplement, 1 × Glutamax-1 supplement (Thermo Fisher Scientific) and 5% FBS.

Cells were thawed, added to media (5 mL) and centrifuged at 300 × g for 5 minutes. Pellet was resuspended in fresh media (12 mL) and added to a vented T75 flask. Cells were incubated at 37 °C. Prior to splitting, media was removed and 1 × trypsin (3 mL; Lonza) was added to cells and incubated for 5 minutes at 37 °C. Media (5 – 6 mL) was added and cells were centrifuged at 300 × g for 5 minutes. Media was discarded and cells were resuspended in fresh media. Cells were split at a ratio of 1:11 every 5-7 days, allowing for no more than four passages (P_4) before plating.

9.5.2 Gradient preparation

Sucrose stock solutions were prepared at the following sucrose percentages (w/v); 18%, 26%, 34%, 42%, 47%, 50% and 60%. Stock solutions also contained 50 mM Tris HCl pH 8.0, 150 mM NaCl and 10 mM $MgCl_2$. Stock solutions were stored at 4 ºC for up to one month.

The day prior to polysome gradienting, sucrose aliquots were prepared. To the necessary amount of stock solution for the number of gradients prepared, the following was added; cycloheximide (100 µg/mL), DTT (1 mM) and 1 × cOmplete™ protease inhibitor cocktail (Roche). Gradients of decreasing sucrose concentration (bottom to top) were made to give a total of 11.7 mL per tube. The composition of the gradient is outline in table 9.9.

Aliquot	**60%**	**50%**	**47%**	**42%**	**34%**	**26%**	**18%**
mL	0.5	2.0	2.0	2.0	1.4	1.4	1.4

Table 9.9 Composition of sucrose gradients for polysome profiling.

Layers were flash frozen using liquid nitrogen between the addition of each aliquot. Gradients were stored carefully at 4 ºC overnight.

9.5.3 Polysome gradients

9.5.3.1 Cell harvesting and lysis

All wash steps were carried out using PBS supplemented with cycloheximide (100µg/mL) but will be referred to as PBS. Fresh lysis buffer (10 mM Tris HCl pH 8.0, 150 mM NaCl, 10 mM $MgCl_2$, 1 mM DTT, 1% IGPAL, 100 ug/mL cycloheximide, 24 U/mL Turbo™ DNase, 90 U RNasin® Plus, 1 × cOmplete™ protease inhibitor cocktail) was prepared and kept at 4 ºC.

Cells were treated with cycloheximide (100 µg/mL) at 37 ºC for 3 minutes. Medium was removed and plates were washed with PBS (2 - 3 mL). 1 × trypsin (2 – 3 mL) was added as cells were incubated at 37 ºC for 5 minutes. Media (5 – 6 mL) was added to quench trypsin and cells were harvested. Cells were pelleted at 800 x g for 8 minutes and media was discarded. Cells were washed with PBS and pelleted as before. PBS was discarded and cells were re-suspended in lysis buffer (300 µL) and incubated on ice for 40 minutes. Cell debris was pelleted at 13,000 × g for 5 minutes. Lysate (300 µL) was loaded onto each gradient and centrifuge tubes were balanced using 18% sucrose solution (from aliquots). Gradients were centrifuged at 31,000 rpm for 3.5 hours at 4 ºC.

9.5.3.2 Gradient fractionation

All gradients were fractionated using a Piston Gradient Fractionater™ (BioComp) coupled with an automated fraction collector and an EM-1 Econo UV monitor (Bio-Rad). Profiles were generated using Gradient Profiler software (BioComp).

Fractionator was rinsed thoroughly with RNAse-free water prior to each gradient. Gradients were fractionated into 0.5 mL aliquots. Gradient remnants were collected as an additional fraction, and the fractionator was aired to collect any solution remaining in the machine. “Air” was collected as the final fraction.

9.5.3.3 9.5.3.3 RNA precipitation

To each gradient fraction, an equal volume (0.5 mL) was added. NaCl solution was added to a final concentration of 0.3 M, followed by GlycoBlue™ Coprecipitant (1 µL; Invitrogen). RNA was precipitated at -80 ºC for at least 12 hours.

Samples were thawed and RNA was pelleted at 13,000 x g for 30 minutes. Supernatant was removed and pellets were washed twice with 70% ethanol *via* centrifugation. Pellets were air-dried for ca. 10 minutes. Pellets were resuspended in RNAse-free water (30 µL) and concentrations were measured using a Nanodrop™

8000 (Thermo Fisher Scientific). To each sample, an appropriate amount of Turbo™ DNAse (1 µL per 10 µg RNA) was added. Turbo™ DNAse reaction buffer (1 ×) and water were added to a final volume of 100 µL. Solution was mixed well and incubated at 37 ºC for 30 minutes.

Sample was diluted to 200 µL with RNAse-free water. Acidic phenol/chloroform (200 µL) was added and samples were mixed thoroughly. Samples were centrifuged at 13,000 × g for 5 minutes. Aqueous phase was added to 100% ethanol (500 µL). NaCl solution was added to a final concentration of 0.3 M followed by addition of GlycoBlue™ Coprecipitant (1 µL; Invitrogen). RNA was precipitated at -80 ºC for at least 12 hours.

Samples were thawed and RNA was pelleted at 13,000 × g for 30 minutes. Supernatant was removed and pellets were washed twice with 70% ethanol *via* centrifugation. Pellets were air-dried for ca. 10 minutes then resuspended in RNAse-free water (30 µL). Concentrations were measured using a Nanodrop™ 8000 (Thermo Fisher Scientific). Purified RNA was stored at -80 ºC.

9.5.4 Reverse transcription PCR

Reverse transcription PCR (RT-PCR) was carried out using a qScript™ cDNA synthesis kit (QuantaBio) in a T100™ Thermal Cycler (Bio-Rad).

RT-PCR was performed using the parameters outlined below.

Component	Amount
RNA	1 µg - 10 pg *
Nuclease-free water	Make up to 20 µL
5X qScript™ reaction buffer	4.0 µL
qScript™ RT	1.0 µL

Table 9.10 Composition of qScript™ RT-PCR reaction mixture

*Equal volumes of RNA were added to each reaction so as to fit a range of 1 µg to 10 pg across all fractions.

Cycle	Temperature (°C)	Time
1	22	5 min
1	42	30 min
1	85	5 min
Hold	4	-

Table 9.11 PCR cycling parameters to RT-PCR using qScript™ reverse transcriptase

The resulting cDNA concentration was measured using a Nanodrop™ 8000 (Thermo Fisher Scientific) and stored at -20 °C.

9.5.5 Quantitative PCR

Quantitative PCR (qPCR) was performed in a 96-well PCR reaction plate using a CFX Connect™ Real-Time PCR Detection System (Bio-Rad). qPCR was run and analysed using Bio-Rad CFX Manager™ software.

9.5.5.1 Primer design

Primers were designed to individually target *Alu* and non-*Alu* transcripts of NEK4, BCAS4, ASCC1 and RPGRIP1L. Exon-exon junctions specific to each transcript type were identified using Ensemble and primers were designed to span the appropriate junctions accordingly. Primers were 18-25 nt in length with a GC content of 40-60%. Primer melting point (T_m) was between 50 and 70 °C, with no more than 5°C difference between primer pairs. The predicted transcript produced by primer pairs spanned from 100 – 150 nt. Examples of forward and reverse primers are shown below:

SAQ001: BCAS4 AC Forward

Sequence corresponding to part of exon 5

```
GGG TTC AAG TGA TTC TCC TGC
```

SAQ002: BCAS4 AC Reverse

Sequence corresponding to reverse complement of part of exon 6

```
CTA TAC AGT GTG GGC AGC TC
```

Predicted transcripts were checked for undesired amplification products using basic local alignment tool (BLAT) analysis available from the UCSC Genome Browser. Primers were ordered from Integrated DNA Technologies (IDT).

9.5.5.2 qPCR setup

Prior to qPCR, cDNA samples were diluted equally so that the highest sample concentration was approximately 2.5 ng/µL. A sample of combined cDNA was prepared by combining a small amount of each fraction cDNA in a 1/5 dilution. Serial dilutions of 1/50, 1/500 and 1/5000 were then made. RNA corresponding to each cDNA sample was diluted to the same concentration as samples. A reaction master mix was made up so as to meet the reaction components outlined in table 9.12.

Component	Amount
2X PowerUp™ SYBR™ Green Master Mix	10 µL
Forward primer (10 µM)	0.6 µL
Reverse primer (10 µM)	0.6 µL
Deionised water	3.8 µL

Table 9.12 Reaction master mix components per well for qPCR using PowerUp™ SYBR™ Green

In each well, 15 µL of master mix was added to 5 µL of samples and mixed well. The plate was filled so as to accommodate the samples outlined in table 9.13.

Type	Sample	No. Wells
Standard curve	1/2.5 dilution	3
Standard curve	1/12.5 dilution	3
Standard curve	1/62.5 dilution	3
Standard curve	1/312.5 dilution	3
Samples	cDNA for each fraction	3
No RT control (NRT)	RNA for each fraction	1
No template control (NTC)	Water	1

Table 9.13 Overview of 96-well PCR plate for qPCR with PowerUp™ SYBR™ Green.

Plate was covered and centrifuged at < 1000 rpm for 30 seconds. qPCR was carried out using the parameters outlined in table 9.10.

Step	Temperature	Time	Cycle
UDG activation	50 °C	2 minutes	1
Dual-Lock™ DNA polymerase	95 °C	2 minutes	1
Denaturation	95 °C	15 seconds	40
Annealing	55-60 °C*	15 seconds	
Extension	72 °C	1 minute	

Table 9.14 Cycling conditions of real-time quantitative PCR (RT-qPCR)

*Annealing temperature should be set to the melting temperature of the primer used.

9.5.5.3 qPCR data analysis

Data obtained from qPCR was analysed using Microsoft Excel. Standard curves were generated by plotting log (starting quantity) against quantification cycle (Cq). Trend lines were fitted and standard curves were excepted for an R^2 value between 0.9 and 1.0. Primer efficiency was calculated using equation 9.1. Primers with an efficiency between 100 and 110% we accepted.

$$Equation\ 9.1 \qquad Efficiency\ (\%) = (10^{(-1/slope)} - 1) \times 100$$

Starting quantities (SQ) for fractions was calculated through the rearrangement of the equation associated with the standard curve trend line (equation 9.2).

$$Equation\ 9.2 \qquad SQ = 10^{(Cq - y_{int}/slope)}$$

SQ mean and standard error were calculated from fraction triplicates. Results were plotted as histograms of fraction versus SQ mean. Secondary histograms were generated showing transcript expression per fraction as a percentage of total expression.

Data for fractions 9 – 10, representative of polysome population and discounting free mRNA and ribosomal subunits, was plotted as a percentage of total polysomal RNA. Ascending and descending cumulative percentages were plotted on a x-axis of ribosome number and the intercept between the two plots represented the mean ribosomal distribution of each target transcript.

9.6 General recipes

9.6.1 Agarose (1%) gel

Agarose (0.5 g; Thermo Fisher Scientific) was added to deionised water (50 mL) and microwaved for 1 minute 30 seconds. SYBR™ Safe gel stain (0.5 μL; Invitrogen) was added and gel was poured into mould (BioRad) and allowed to set at room temperature.

9.6.2 Tris acetate EDTA (TAE) buffer

A 50 × TAE buffer stock solution was prepared by adding Tris (244 g), acetate (57.1 mL) and 0.5 M Na_4EDTA (100 mL) to deionised water (up to 1L). 1 × TAE buffer was prepared freshly for use.

9.6.3 Luria-Bertani (LB) media

LB Broth – Miller (25 g; Formedia) was added to deionised water (1 L) and autoclaved. The amount of powder was scaled for the amount of media required.

9.6.4 LB agar

LB Broth – Miller (2.5 g; Formedia) and agar (1.2 g; Formedia) were added to deionised water (100 mL) and autoclaved. Autoclaved LB agar was cooled until touchable by hand. Necessary antibiotics were added and LB agar was poured into sterile petri dishes and allowed to set at room temperature. Plates were seal with parafilm and stored top-down at 4 °C.

9.6.5 SDS-PAGE (10% acrylamide) gel

SDS-PAGE gel solutions were prepared as outlined in table 9.15.

	Separating gel	Stacking gel
Deionised water	5.0 mL	6.26 mL
Tris HCl pH 6.8	-	2.5 mL
Tris HCl pH 8.8	2.5 mL	-
40% acrylamide solution	2.5 mL	1.24 mL
10% SDS	100 μL	100 μL
10% APS	100 μL	100 μL

Table 9.15 Components for SDS-PAGE gel (10% acrylamide) solutions

TEMED (10 μL) was added to separating gel solution, mixed and loaded in a 1.0 mm plates using a gel casting kit (BioRad). Isopropanol was added to maintain an even gel. Gel was allowed to set at room temperature. Isopropanol was removed. TEMED (10 μL) was added to stacking gel solution and mixed. Stacking gel was loaded, well comb added and gel was allowed to set at room temperature.

9.6.6 SDS-PAGE running buffer

A 5 × SDS-PAGE buffer was prepared by adding Tris (15.1 g), glycine (94 g) and sodium dodecyl sulfate (SDS; 5 g) to deionised water (1 L). 1 × buffer was prepared freshly prior to gel runs.

9.6.7 SDS loading buffer (2 ×)

Loading buffer was made up according to table 9.16.

Component	Amount
Tris HCl	100 mM
SDS	4%
Bromophenol blue	0.2%
Glycerol	20%
DTT	20 mM

Table 9.16 Components of SDS-PAGE loading buffer

9.6.8 Antibiotic stocks

Antibiotic stocks (1000 ×) were made up to according to table 9.17.

Antibiotic	Concentration	Solvent
Kanamycin	50 mg/mL	Water
Chloramphenicol	25 mg/mL	Ethanol
Ampicillin	125 mg/mL	Water

Table 9.17 Antibiotic stock concentrations

Appendix

1 Bioinformatic data

Subfamily	Consensus sequence
J	GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGG CGGGAGGATCACTTGAGCCCAGGAGTTCGAGACCAGCCTGGGCAACATAGTG AAACCCCGTCTCTACAAAAAATACAAAAATTAGCCGGGCGTGGTGGCGCGCG CCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGGATCGCTTGAGCCCG GGAGGTCGAGGCTGCAGTGAGCCGTGATCGCGCCACTGCACTCCAGCCTGGG CGACAGAGCGAGACCCTGTCTCAAAAAAAA
Sx	GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGG CGGGCGGATCACCTGAGGTCAGGAGTTCGAGACCAGCCTGGCCAACATGGTG AAACCCCGTCTCTACTAAAAATACAAAAATTAGCCGGGCGTGGTGGCGCGCG CCTGTAATCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATCGCTTGAACCCG GGAGGCGGAGGTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGG CGACAGAGCGAGACTCCGTCTCAAAAAAAA
Sp	GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGG CGGGCGGATCACCTGAGGTCGGGAGTTCGAGACCAGCCTGACCAACATGGAG AAACCCCGTCTCTACTAAAAATACAAAAATTAGCCGGGCGTGGTGGCGCATG CCTGTAATCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATCGCTTGAACCCG GGAGGCGGAGGTTGCGGTGAGCCGAGATCGCGCCATTGCACTCCAGCCTGGG CAACAAGAGCGAAACTCCGTCTCAAAAAAAA
Sq	GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGG CGGGTGGATCACCTGAGGTCAGGAGTTCGAGACCAGCCTGGCCAACATGGTG AAACCCCGTCTCTACTAAAAATACAAAAATTAGCCGGGCGTGGTGGCGGGCG CCTGTAATCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATCGCTTGAACCCG GGAGGCGGAGGTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGG CAACAAGAGCGAAACTCCGTCTCAAAAAAAA
Sc	GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGG CGGGCGGATCACGAGGTCAAGAGATCGAGACCATCCTGGCCAACATGGTGAA ACCCCGTCTCTACTAAAAATACAAAAATTAGCTGGGCGTGGTGGCGCGCGCC TGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATCGCTTGAACCCGGG AGGCGGAGGTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGCGA CAGAGCGAGACTCCGTCTCAAAAAAAA

Subfamily	Consensus sequence
Sb	GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGG CGGGCGGATCACGAGGTCAGGAGATCGAGACCATCCTGGCTAACACGGTGAA ACCCCGTCTCTACTAAAAATACAAAAATTAGCCGGGCGTGGTGGCGGGCGCC TGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGGG AGGCGGAGCTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCG ACAGAGCGAGACTCCGTCTCAAAAAAAA
Sb1	GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGG CGGGCGGATCACGAGGTCAGGAGATCGAGACCATCCCGGCTAAAACGGTGAA ACCCCGTCTCTACTAAAAATACAAAAATTAGCCGGGCGTAGTGGCGGGCGCC TGTAGTCCCAGCTACTTGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGGG AGGCGGAGCTTGCAGTGAGCCGAGATCCCGCCACTGCACTCCAGCCTGGGCG ACAGAGCGAGACTCCGTCTCAAAAAAAA
Yb	GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGG CGGGTGGATCATGAGGTCAGGAGATCGAGACCATCCTGGCTAACAAGGTGAA ACCCCGTCTCTACTAAAAATACAAAAAATTAGCCGGGCGCGGTGGCGGGCGC CTGTAGTCCCAGCTACTGGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGG GAAGCGGAGCTTGCAGTGAGCCGAGATTGCGCCACTGCAGTCCGCAGTCCGG CCTGGGCGACAGAGCGAGACTCCGTCTCAAAAAAAA

A1.1 Consensus sequences for NCBI database *Alu* subfamilies

Class	Subfamilies analysed
Alu J	Jb, Jo, Jr, Jr4
Alu S	Sc, Sc5, Sc8, Sg, Sg4, Sg7, Sp, Sq, Sq2, Sq4, Sq10, Sx, Sx1, Sx3, Sx4, Sz, Sz6
Alu Y	Y, Ya5, Ya8, Yb8, Yb9, Yc, Yc3, Yd8, Ye5, Yf1, Yg6, Yh9, Yk4, Yk11, Yk12

A1.2 List of Dfam *Alu* sequences used in analysis

ID	Protein	Full name	Match (%)	Isoforms	Strand	Location
Q7Z5D8	NANGN	NANOG neighbour homeobox	97	1	+	N
Q99959	PKP2	Plakophilin-2	95	2	+	I
Q8NHA8	OR1FC	Olfactory receptor 1F12	95	1	+	C
Q86V71	ZN429	Zinc finger protein 429	95	1	+	C
Q9NRJ1	MOST1	Protein MOST-1	94	1	+	C
Q96ET8	TV23C	Golgi apparatus membrane protein TVP23 homolog C	94	3	N/A	I
Q58FG0	HS905	Putative heat shock protein HSP 90-alpha A5	94	1	-	N
Q9UJ41	RABX5	Rab5 GDP/ GTP exchange factor	93	4	N/A	I
Q9BUA6	MYL10	Myosin regulatory light chain 10	93	1	-	N
Q9NV72	ZN701	Zinc finger protein 701	92	2	-	N
Q9HC38	GLOD4	Glyoxalase domain-containing protein 4	92	3	-	N
Q8N7M2	ZN283	Zinc finger protein 283	92	1	-	N
Q09FC8	ZN415	Zinc finger protein 415	91	6	-	N
Q5TCQ9	MAGI3	Membrane-associated guanylate kinase, WW and PDZ domain containing protein 3	90	4	N/A	I
Q92918	M4K1	Mitogen-activated protein kinase kinase kinase kinase 1	89	2	-	C
Q05481	ZNF91	Zinc finger protein HPF7	88	2	-	C
Q8N9N2	ASCC1	Activating signal cointegrator complex subunit 1	88	2	-	C
O14628	ZN195	Zinc finger protein 195	85	8	-	N
Q96N38	ZN714	Zinc finger protein 714	84	3	+	C
Q8TDM0	BCAS4	Breast cancer amplified sequence 4	84	3	-	C
Q04864	REL	Proto-oncogene c-rel	84	2	-	I
Q8N7I0	GVQW1	GVQW motif-containing protein 1	84	1	-	I
P49796	RGS3	Regulator of G-protein signalling 3	84	9	+	N
Q5T7P6	TMM78	Transmembrane protein 78	83	1	-	C
Q5H5K5	ZMAT1	Zinc finger matrin-type protein 1	83	2	-	N
Q8IV13	CCNJL	Cyclin-J-like protein	83	2	-	I
Q8NEM8	CBPC3	Cytosolic carboxypeptidase 3	83	4	N/A	I
Q6P3R8	NEK5	Never in mitosis A-related kinase 5	83	1	-	I

ID	Protein	Full name	Match (%)	Isoforms	Strand	Location
Q9Y2ZO	SGT1	Protein SGT1 homolog	82	2	-	I
Q495B1	AKD1A	Ankyrin repeat and death domain-containing protein 1A	82	4	-	C
Q6UX73	CP089	UPF0764 protein C16orf89	81	2	-	C
Q96J02	ITCH	E3 ubiquitin-protein ligase Itchy homolog	81	3	-	N
P51957	NEK4	Never in mitosis A-related kinase 4	80	3	-	I
Q8WTZ3	YS049	Zinc finger protein ENSP00000375192	78	1	-	I
Q96M98	PACRG	Parkin coregulated gene protein	77	2	-	I
Q9BUB5	MKNK1	MAP kinase-interacting serine/ threonine-protein kinase 1	77	3	-	I
Q68CZ1	FTM	Protein fantom	77	2	-	C
P78312	F193A	Protein FAM193A	76	6	-	C
Q9NXG0	CNTLN	Centlein	75	3	N/A	C
O15488	GLYG2	Glycogenin	74	6	-	N
F2Z398	LMO7D	LMO7 downstream neighbour protein	74	1	-	I
Q96T75	DSCR8	Down syndrome critical region protein 8	73	4	-	I
O94966	UBP19	Ubiquitin carboxyl-hydrolase 19	71	7	-	N
E7EU14	PPP5D1	PPP5 TPR repeat domain-containing protein 1	70	1	-	C
Q8NDZ0	BEND2	BEN domain-containing protein 2	70	2	-	N
Q96ME1	FXL18	F-box/ LRR-repeat protein 18	70	4	-	C
Q6NY19	KANK3	KN motif and ankyrin repeat domain-containing protein 3	68	2	-	C

A1.3 Refined list of *Alu*-containing protein matches

List contains the *Alu* matches identified using a percentage identity of 68% or above and an E-value of 1×10^{8} or lower. Protein IDs are taken from the NCBI database. Stands are listed at sense (+) or antisense (+) dependent on the orientation of the parent *Alu* and location refers to the area of the protein in which the insertion was located; N-terminal (N), internal (I) or C-terminal (C).

Alignment	SID	E-Val.
AKD1A 495 GWSTMARSQLTATSASRVQMILVPQPPE 522 **AluSz6** 9 ...**AV**...**R**.........**A**..... 36	86%	1×10^{-14}
ASCC1 351 LLPRLEYNDAISAHCNLCLPGSSDSPASASQVAGITG 387 **AluSq** 6 ......**CSG**.....**R**.....**R**..... 42	86%	3×10^{-19}
BCAS4 164 VECSGTIPARCNLRLPGSSDSPASASQVAGIT 195 **AluSq** 10 **L**....**A**.**S**.**H**.....**R**..... 41	84%	3×10^{-16}
BEND2 80 GSGSVTQAGVQWHDHSSLQPQPLGLKQFFHLSLPSSWDDRRTPPCP 125 **AluJr** 4 ..**R**..**A**.....**R**.....**RTP**...**RSSR**.....**Y**..**A**..**R**. 49	72%	3×10^{-14}
CBPC3 704 AHCKLRLPGSRHSPASASRVAGTTGTRHHTWLIFVFLVEMG 744 **AluSc** 18 ...**N**.....**SD**.....**A**...**AQ**.....**T**. 58	83%	8×10^{-22}
CCNJL 94 NGVSLLSPRLKCSGMISAHCNLHLPGSSNSPASA 127 **AluSq** 1 **D**.....**L**...**E**...**A**.....**R**.....**D**..... 34	82%	4×10^{-16}
CNTLN 1373 QSLTLSPRLKCNGAIVAHQNLRLPDSSSS-ASAS 1405 **AluSx** 2 **R**..**A**.....**E**.**S**...**S**..**C**.....**G**..**D**.**P**.... 35	74%	6×10^{-12}
CP089 367 FYIFLVETGFHHVAHAGLELLISRDPPTSGSQSVGL 402 **AluSz** 50 **IFV**.....**GQ**.....**T**.**S**...**A**.**A**...**A**.**I** 85	69%	8×10^{-16}
DSCR8 28 LFLSPRLECSGSITDHCSLHLP 49 **ALUSz6** 4 .**A**.....**A**.**SA**..**N**.**R**.. 25	73%	5×10^{-10}
F193A 1167 DGVSLLLPSLGYNGAILAHCNLRLPGSSDCAASASQVVGIT 1207 **AluSq2** 1 ......**S**.**R**.**EC**....**S**.....**SP**....**R**.**A**... 41	78%	2×10^{-16}
FTM 1096 IKQSLALSPGLGCSSAISAHCNFRLPGSSDFPASASQVDGITGACHHTQ 1144 **AluSg4** 1 **LRR**.....**R**.**E**..**G**.....**L**.....**S**.....**R**.**A**.....**R**.**RAR** 49	71%	8×10^{-22}
FXL18 757 ETESHSVVQAGVQWRDLSSLQPLLSGLQ 784 **AluSz6** 1 ....**R**..**A**.....**G**....**PPP**.**FK** 28	71%	4×10^{-11}
GLOD4 30 KVESCSVARLECSGAISAHCS 50 **AluSc5** 1 **ET**..**R**.....**N** 21	81%	1×10^{-9}
GLYG2 3 ETEFHHGAQAGLELLRSSNSPTSASQSAGMT 33 **AluJb** 56 ..**G**...**V**.....**G**..**DP**.**A**.....**I**. 86	77%	9×10^{-12}
ITCH 159 NGVSLCLPRLECNSAISAHCNLCLPGLSDSPISASRVAGFTGASQN 204 **ALUSQ2** 1 **D**....**LS**.....**G**.....**R**...**S**....**A**.....**I**...**RHH** 46	76%	2×10^{-21}
M4K1 797 SPRLECSGTISPHCNLLLPGSSNSPASASRVAGITG 832 **AluSx** 7 .....**A**..**A**....**R**.....**D**..... 42	89%	2×10^{-18}
MKNK1 189 LGSSDPPTSASQVAGTTGIAHR 210 **AluJb** 70 .....**A**....**S**..**I**..**VS**.. 91	77%	7×10^{-9}

Alignment	SID	E-Val.
NEK4 456 QSLALSPKLECSGTILAHSNLRLLGSSDSPASASRVAGITGVCHHAQ 502 **AluSx** 2 **R**.....**R**.....**A**.**S**..**C**....**P**.....**AR**...**R** 48	81%	5×10^{-22}
PACRG 210 PRLECSGAIMARCNLDHLGSSDPPTSASQVA 240 **AluJo** 8 .........**I**.**H**.**S**.**EL**.......**A**...**R**.. 38	77%	3×10^{-15}
REL 308 VETGFRHVDQDGLELLTSGDPPTLASQSAGIT 339 **AluSz6** 55 .....**C**..**G**.**A**.....**A**.**S**...**AS**..... 86	78%	2×10^{-13}
RGS3 2 PVIPALWEVEMGRSQGQEIETILAN 26 **AluYk4** 8 ........**A**.**A**.**G**.**R**.......... 32	84%	3×10^{-11}
SGT1 110 IETGFHRVGQAGLQLLTSSDPPALDSQSAGITG 142 **AluSz** 55 **V**.....**H**......**E**.........**SA**..... 87	85%	2×10^{-17}
ZMAT1 2 ESCSVTRLECSGAISAHCSLHLPGSSDSPASASQIAGTTDA 42 **AluSc** 3 ..**R**..**A**.....**N**.**R**.....**RV**....**G**. 43	83%	2×10^{-20}
ZN195 76 EMGFHHATQACLELLGSSDLPASASQSAGITGVNHRAQ 113 **AluJb** 56 .**T**....**VA**..**G**.....**P**.....**S**...**R** 93	82%	1×10^{-18}
ZN415 61 RLECNGAISAHCNLRLPDSNDSPASASRVAGIT 93 **AluSq2** 9 .....**G**.**S**..... 41	94%	9×10^{-19}
ZN701 2 GFLHVGQDGLELPTSGDPPASASQSAGITGVSHRTQ 37 **AluSz** 58 ..**H**....**A**....**L**..**S**.....**AR** 93	83%	8×10^{-17}
ZN714 504 GMVAHACNPNTLRGLGEQIARSGVQDQPGQHGKTPSLLKIQKFAGCGGRRL 554 **AluSg** 2 .**A**.....**S**..**G**.**R**.**GR**.**T**....**R**.....**E**.....**L**..**R**..**A**.. 52	76%	4×10^{-24}

A1.4 Alignments of protein hits with a single ORF from the parental *Alu*

2 Human proteins DNA and protein data

Plasmid sequences for human proteins

Genes were cloned into plasmids followed by a stop codon with no addition mutations to the gene. In cases were plasmids encoded fusion proteins, the fusion partner interest is highlight in red.

pET.SUMO.28a.NEK4 transcript variant 1 sequence 5'-3'

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGTCG
GACTCAGAAGTCAATCAAGAAGCTAAGCCAGAGGTCAAGCCAGAAGTCAAGCCTGAGACTCACATC
AATTTAAAGGTGTCCGATGGATCTTCAGAGATCTTCTTCAAGATCAAAAAGACCACTCCTTTAAGA
AGGCTGATGGAAGCGTTCGCTAAAAGACAGGGTAAGGAAATGGACTCCTTAAGATTCTTGTACGAC
GGTATTAGAATTCAAGCTGATCAGACCCCTGAAGATTTGGACATGGAGGATAACGATATTATTGAG
GCTCACAGAGAACAGATTGGTGGAATGCCCCTGGCCGCCTACTGCTACCTGCGGGTCGTGGGCAAG
GGGAGCTATGGAGAGGTGACGCTTGTGAAGCACCGGCGGGACGGCAAGCAGTATGTCATCAAAAAA
CTGAACCTCCGAAATGCCTCTAGCCGAGAGCGGCGAGCTGCTGAACAGGAAGCCCAGCTCTTGTCT
CAGTTGAAGCATCCCAACATTGTCACCTACAAGGAGTCATGGGAAGGAGGAGATGGTCTGCTCTAC
ATTGTCATGGGCTTCTGTGAAGGAGGTGATTTGTACCGAAAGCTCAAGGAGCAGAAAGGGCAGCTT
CTGCCGGAGAATCAGGTGGTAGAGTGGTTTGTACAGATCGCCATGGCTTTGCAGTATTTACATGAA
AAACACATCCTTCATCGAGATCTGAAAACTCAAAATGTCTTCCTAACAAGAACAAACATCATCAAA
GTAGGGGACCTAGGAATTGCCCGAGTGTTGGAGAACCACTGTGACATGGCTAGCACCCTCATTGGC
ACACCCTACTACATGAGCCCTGAATTGTTCTCAAACAAACCCTACAACTATAAGTCTGATGTTTGG
GCTCTAGGATGCTGTGTCTATGAAATGGCCACCTTGAAGCATGCTTTCAATGCAAAAGATATGAAT
TCTTTAGTTTATCGGATTATTGAAGGAAAGCTGCCAGCAATGCCAAGAGATTACAGCCCAGAGCTG
GCAGAACTGATAAGAACAATGCTGAGCAAAAGGCCTGAAGAAAGGCCGTCTGTGAGGAGCATCCTG
AGGCAGCCTTATATAAAGCGGCAAATCTCCTTCTTTTTGGAGGCCACAAAGATAAAAACCTCCAAA
AATAACATTAAAAATGGTGACTCTCAATCCAAGCCTTTTGCTACAGTGGTTTCTGGAGAGGCAGAA
TCAAATCATGAAGTAATCCACCCCCAACCACTCTCTTCTGAGGGCTCCCAGACATATATAATGGGT
GAAGGCAAATGTTTGTCCCAGGAGAAACCCAGGGCCTCTGGTCTCTTGAAGTCACCTGCCAGTCTG
AAAGCCCATACCTGCAAACAGGACTTGAGCAATACCACAGAACTAGCCACAATCAGTAGCGTAAAT
ATTGACATCTTACCTGCAAAAGGGAGGGATTCAGTGAGTGATGGCTTTGTTCAGGAGAATCAGCCA
AGATATTTGGATGCCTCTAATGAGTTAGGAGGTATATGCAGTATTTCTCAAGTGGAAGAGGAGATG
CTGCAGGACAACACTAAATCCAGTGCCCAGCCTGAAAACCTGATTCCCATGTGGTCCTCTGACATT
GTCACTGGGGAAAAGAATGAACCAGTGAAGCCTCTGCAGCCCCTAATCAAAGAACAAAAGCCAAAG
GACCAGAGTCTTGCCCTGTCGCCCAAGCTGGAGTGCAGTGGCACAATCTTGGCTCACAGCAACCTC
CGCCTCCTGGGTTCAAGTGATTCTCCAGCCTCAGCCTCCCGAGTAGCTGGGATTACAGGCGTGTGC
CACCACGCCCAGGATCAAGTTGCTGGTGAATGTATTATAGAAAAACAGGGCAGAATCCACCCAGAT
TTACAGCCACACAACTCTGGGTCTGAACCTTCCCTGTCTCGACAGCGACGGCAAAAGAGGAGAGAA

CAGACTGAGCACAGAGGGGAAAAGAGACAGGTCCGCAGAGATCTCTTTGCTTTCCAAGAGTCGCCT
CCTCGATTTTTGCCTTCTCATCCCATTGTTGGGAAAGTGGATGTCACATCAACACAAAAAGAGGCT
GAAAACCAACGTAGAGTGGTCACTGGGTCTGTGAGCAGTTCAAGGAGCAGTGAGATGTCATCATCA
AAGGATCGACCATTATCAGCCAGAGAGAGGAGGCGACTAAAGCAGTCACAGGAAGAAATGTCCTCT
TCAGGCCCTTCAGTGAGGAAAGCGTCTCTGAGTGTAGCAGGGCCAGGAAAACCCCAGGAAGAAGAC
CAGCCCTTGCCTGCCCGACGGCTCTCCTCTGACTGCAGCGTCACTCAGGAAAGGAAACAGATTCAT
TGTCTGTCTGAGGATGAGTTAAGTTCTTCTACAAGTTCAACTGATAAGTCAGATGGGGATTACGGG
GAAGGGAAAGGTCAGACAAATGAAATTAATGCCTTGGTACAATTGATGACTCAGACCCTGAAACTG
GATTCTAAAGAGAGCTGTGAAGATGTCCCGGTAGCAAACCCAGTGTCAGAATTCAAACTTCATCGG
AAATATCGGGACACACTGATACTTCATGGGAAGGTTGCAGAAGAGGCAGAGGAAATCCATTTTAAA
GAGCTACCTTCAGCTATTATGCCAGGTTCTGAAAAGATCAGGAGACTAGTTGAAGTCTTGAGAACT
GATGTAATTCGTGGCCTGGGAGTTCAGCTTTTAGAGCAGGTGTATGATCTTTTGGAGGAGGAGGAT
GAATTTGATAGAGAGGTACGTTTGCGGGAGCACATGGGTGAAAAGTATACAACTTACAGTGTGAAA
GCTCGCCAGTTGAAATTTTTTGAAGAAAACATGAATTTT

pET.SUMO.28a.ZMAT1 transcript variant 1 sequence 5'-3'

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGTCG
GACTCAGAAGTCAATCAAGAAGCTAAGCCAGAGGTCAAGCCAGAAGTCAAGCCTGAGACTCACATC
AATTTAAAGGTGTCCGATGGATCTTCAGAGATCTTCTTCAAGATCAAAAAGACCACTCCTTTAAGA
AGGCTGATGGAAGCGTTCGCTAAAAGACAGGGTAAGGAAATGGACTCCTTAAGATTCTTGTACGAC
GGTATTAGAATTCAAGCTGATCAGACCCCTGAAGATTTGGACATGGAGGATAACGATATTATTGAG
GCTCACAGAGAACAGATTGGTGGAATGGAGTCTTGCTCTGTCACCAGGCTGGAGTGCAGTGGCGCA
ATCTCGGCTCACTGCAGCCTCCACCTCCCGGGTTCAAGCGATTCCCCTGCCTCAGCCTCCCAAATA
GCTGGGACTACAGACGCCATTTGGAATGAACAGGAAAAGGCTGAACTTTTTACAGATAAGTTTTGT
CAAGTATGTGGAGTGATGCTACAGTTTGAATCACAAAGAATTTCACATTATGAGGGTGAAAAACAT
GCTCAAAATGTTAGTTTTTATTTTCAAATGCATGGGGAACAAAATGAAGTGCCTGGTAAGAAAATG
AAGATGCATGTTGAGAATTTTCAGGTGCATAGGTATGAAGGAGTGGACAAAAACAAATTTTGTGAT
CTCTGCAACATGATGTTTAGCTCTCCACTTATTGCTCAGTCTCACTATGTGGGAAAGGTCCATGCT
AAAAAACTGAAGCAATTAATGGAGGAACATGATCAGGCATCTCCATCAGGATTTCAACCAGAGATG
GCATTTAGTATGAGAACCTATGTTTGCCATATTTGTAGTATTGCTTTTACATCTTTAGATATGTTC
CGGTCCCACATGCAAGGAAGTGAACATCAAATTAAAGAATCCATTGTTATCAATCTAGTGAAGAAT
TCAAGGAAGACACAAGACTCTTACCAAAATGAGTGTGCAGATTACATCAATGTGCAGAAAGCCAGA
GGACTAGAGGCCAAGACTTGTTTCAGAAAGATGGAAGAGAGTTCTTTGGAAACCCGTAGATACAGA
GAAGTGGTCGATTCCAGACCCAGACATAGAATGTTTGAACAAAGACTCCCATTTGAGACTTTCCGG
ACATACGCAGCACCATACAATATTTCACAAGCAATGGAAAAGCAGTTACCTCATTCAAAGAAGACA
TATGACTCTTTCCAAGATGAACTTGAAGATTACATCAAAGTACAGAAAGCCAGAGGACTAGATCCA
AAGACTTGTTTCAGAAAGATGAGAGAGAACTCTGTGGATACTCATGGGTACAGAGAAATGGTTGAT
TCTGGACCCAGATCAAGAATGTGTGAGCAAAGATTTTCACATGAGGCTTCCCAGACCTACCAACGA
CCATACCATATTTCACCAGTGGAAAGCCAGTTACCTCAGTGGCTACCAACCCATTCAAAGAGGACA
TATGATTCTTTCCAAGATGAACTTGAAGATTACATAAAAGTGCAGAAAGCCAGAGGACTAGAGCCA

AAAACTTGTTTCAGAAAGATAGGAGATAGCTCTGTAGAAACACACAGGAACAGAGAAATGGTTGAT
GTCAGACCCAGACATAGAATGTTGGAGCAAAAGCTCCCATGTGAGACTTTCCAGACCTATTCAGGA
CCATATAGTATTTCACAAGTAGTGGAAAACCAGTTACCTCATTGCTTACCAGCTCATGATAGCAAA
CAGAGACTAGATTCTATTAGCTACTGTCAACTCACCAGAGACTGTTTCCCAGAAAAACCAGTACCC
TTGAGCCTTAATCAGCAAGAAAATAACTCTGGCTCATACAGTGTAGAATCTGAAGTTTACAAGCAC
CTCTCTTCAGAAAACAATACTGCTGACCATCAAGCAGGTCATAAACGGAAACATCAGAAGAGAAAA
CGACACCTAGAAGAAGGCAAAGAAAGGCCAGAGAAAGAGCAGTCCAAGCATAAAAGGAAAAAGAGT
TATGAAGATACAGATTTAGACAAAGACAAGAGCATCAGACAAAGGAAAAGAGAGGAGGATAGAGTC
AAGGTCAGTTCAGGAAAGCTTAAGCATCGAAAAAAGAAAAAAAGCCATGATGTACCCTCCGAGAAA
GAAGAACGTAAGCACAGGAAAGAGAAAAAGAAATCTGTTGAAGAAAGGACAGAAGAGGAAATGCTT
TGGGATGAGTCTATTCTTGGATTT

pGEX.4T.1.PPP5D1 transcript sequence 5'-3'

ATGTCCCCTATACTAGGTTATTGGAAAATTAAGGGCCTTGTGCAACCCACTCGACTTCTTTTGGAA
TATCTTGAAGAAAAATATGAAGAGCATTTGTATGAGCGCGATGAAGGTGATAAATGGCGAAACAAA
AAGTTTGAATTGGGTTTGGAGTTTCCCAATCTTCCTTATTATATTGATGGTGATGTTAAATTAACA
CAGTCTATGGCCATCATACGTTATATAGCTGACAAGCACAACATGTTGGGTGGTTGTCCAAAAGAG
CGTGCAGAGATTTCAATGCTTGAAGGAGCGGTTTTGGATATTAGATACGGTGTTTCGAGAATTGCA
TATAGTAAAGACTTTGAAACTCTCAAAGTTGATTTTCTTAGCAAGCTACCTGAAATGCTGAAAATG
TTCGAAGATCGTTTATGTCATAAAACATATTTAAATGGTGATCATGTAACCCATCCTGACTTCATG
TTGTATGACGCTCTTGATGTTGTTTTATACATGGACCCAATGTGCCTGGATGCGTTCCCAAAATTA
GTTTGTTTTAAAAAACGTATTGAAGCTATCCCACAAATTGATAAGTACTTGAAATCCAGCAAGTAT
ATAGCATGGCCTTTGCAGGGCTGGCAAGCCACGTTTGGTGGTGGCGACCATCCTCCAAAATCGGAT
CTGGTTCCGCGTGGATCCATGGCGGAAATGAGAGCTTGGCGCCCATTGGTCCGACCTTCCCTGCAA
TGCGTCAAACTGGGGCGAGCCACTGCAAGGTGGTGGTGGGTGGTCAAGGTGAAGCCCCACGACAAG
GATGCCAAAATGGAATACCAGGAGTGCAACAAGATCGTGAAGCAGAAGGCCTTTGAGCGGGCCATC
GCAGGCGACGAGCACAAGCGCTCCGTGGTGGACTCGCTGGACATCGAGAGCATGACCATCGAGGGT
GAGTACAGCGGACCCAAGCTTGAGGACGACAAAGTGACAATCACCTTCATGAAGGGGCTCATGCAG
TGGTACAAGGACCAGAAGAAACTGCACCAGAAATGCGCCTACCAGGGTCTTGCTCTATCACCCAGG
CTGAAGTGCAGTGGTACGGTCACGGCTCACTGCAGCCTCAACCTCCTGGGCCCACGTGATCCTCCC
GCCTCAGCATCCCAAGTAGCTGTGACCGAGGGCATGCACCACCACACCTGGCTAATTTTTTTATTT
TTATAG

pET.28a.BCAS4 transcript variant 1 sequence 5'-3'

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGCAG
CGGACCGGGGGCGGGGCTCCGAGGCCCGGGCGCAACCACGGGCTCCCAGGCAGCCTCCGCCAGCCG
GACCCCGTCGCCCTCCTGATGCTGCTCGTGGACGCTGATCAGCCGGAGCCCATGCGCAGCGGGGCG
CGCGAGCTCGCGCTCTTCCTGACCCCCGAGCCTGGGGCCGAGGCGAAGGAGGTGGAGGAGACCATC
GAGGGCATGCTCCTCAGGCTGGAAGAGTTTTGCAGCCTGGCTGACCTGATCAGGAGTGATACTTCA

CAGATCCTGGAGGAAAACATCCCAGTCCTTAAGGCCAAACTGACAGAAATGCGTGGCATCTATGCC
AAAGTGGACCGGCTAGAGGCCTTCGTCAAGATGGTTGGACACCACGTCGCCTTCCTGGAAGCAGAC
GTGCTTCAGGCTGAGCGGGACCATGGGGCCTTCCCTCAGGCCCTGCGGAGGTGGCTGGGATCCGCA
GGGCTCCCCTCCTTCAGGAACGTGGAGTGCAGTGGCACAATCCCAGCTCGCTGCAACCTCCGCCTC
CCGGGTTCAAGTGATTCTCCTGCCTCCGCCTCCCAAGTAGCTGGGATTACAGAAGTCACCTGCACC
GGTGCCCGTGACGTACGAGCTGGCCACACTGTA

pET.SUMO.28a.ASCC1 transcript variant 1 sequence 5'-3'

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGTCG
GACTCAGAAGTCAATCAAGAAGCTAAGCCAGAGGTCAAGCCAGAAGTCAAGCCTGAGACTCACATC
AATTTAAAGGTGTCCGATGGATCTTCAGAGATCTTCTTCAAGATCAAAAAGACCACTCCTTTAAGA
AGGCTGATGGAAGCGTTCGCTAAAAGACAGGGTAAGGAAATGGACTCCTTAAGATTCTTGTACGAC
GGTATTAGAATTCAAGCTGATCAGACCCCTGAAGATTTGGACATGGAGGATAACGATATTATTGAG
GCTCACAGAGAACAGATTGGTGGAATGGAAGTTCTGCGTCCACAGCTTATAAGAATTGATGGCCGG
AATTACAGGAAGAATCCAGTCCAAGAACAGACCTATCAACATGAAGAAGATGAAGAGGACTTCTAT
CAAGGCTCCATGGAGTGTGCTGATGAGCCCTGTGATGCCTACGAGGTGGAGCAGACCCCACAAGGA
TTCCGGTCTACTTTGAGGGCCCCCAGCTTGCTCTATAATCTCATTCACTTGAACACATCAAACGAC
TGTGGGTTCCAGAAGATAACTTTGGATTGTCAGAATATTTATACTTGGAAGTCCAGGCATATAGTT
GGAAAGAGAGGGGACACTAGGAAGAAAATAGAAATGGAGACCAAAACTTCTATTAGCATTCCTAAA
CCTGGACAAGACGGGGAAATTGTAATCACTGGCCAGCATCGAAATGGTGTAATTTCAGCCCGAACA
CGGATTGATGTTCTTTTGGACACTTTTCGAAGAAAGCAGCCCTTCACTCACTTCCTTGCCTTTTTC
CTCAATGAAGTTGAGGTTCAGGAAGGATTCCTGAGATTCCAGGAGGAAGTACTGGCGAAGTGCTCC
ATGGATCATGGGGTTGACAGCAGCATTTTCCAGAATCCTAAAAAGCTTCATCTAACTATTGGGATG
TTGGTGCTTTTGAGTGAGGAAGAGATCCAGCAGACATGTGAGATGCTACAGCAGTGTAAAGAGGAA
TTCATTAATGATATTTCTGGGGGTAAACCCCTAGAAGTGGAGATGGCAGGGATAGAATACATGAAT
GATGATCCTGGCATGGTGGATGTTCTTTACGCCAAAGTCCATATGAAAGATGGCTCCAACAGGCTA
CAAGAATTAGTTGATCGAGTGCTGGAACGTTTTCAGGCATCTGGACTAATAGTGAAAGAGTGGAAT
AGTGTGAAACTGCATGCTACAGTTATGAATACACTATTCAGGAAAGACCCCAATGCTGAAGGCAGG
TACAATCTCTACACAGCGGAAGGCAAATATATCTTCAAGGAAAGAGAATCATTTGATGGCCGAAAT
ATTTTAAAGAGCTTTGCCTTGTTGCCCAGGCTGGAGTACAATGATGCAATCTCCGCTCACTGCAAC
CTGTGCCTCCCGGGTTCAAGTGATTCTCCTGCCTCAGCCTCCCAAGTAGCTGGGATTACAGGTGTC
TCTGATGCATATTCTCAGAGCCTACCAGGAAAATCC

pDB.His.ASCC1 transcript variants

ASCC1 transcript variant genes (Origene) were cloned into pDB.His.MBP.Stop using restriction free (RF) cloning. Primers were designed in such a way that the MBP encoding gene of pDB.His.MBP was 'overwritten' by the ASCC1 gene. A TEV (tobacco etch virus) cut site (shown in red) was then added between the 6X

histidine tag (shown in blue) and the protein start codon (shown in green) using Quikchange II site-directed mutagenesis (SDM).

pDB.His.TEV.ASCC1 transcript variant 1 sequence 5'-3'

ATGGGCAGCAGCCATCATCATCATCATCACTCCTCAGGTGAGAATCTGTATTTTCAGGGCATGGGA
TCGATGGAAGTTCTGCGTCCACAGCTTATAAGAATTGATGGCCGGAATTACAGGAAGAATCCAGTC
CAAGAACAGACCTATCAACATGAAGAAGATGAAGAGGACTTCTATCAAGGCTCCATGGAGTGTGCT
GATGAGCCCTGTGATGCCTACGAGGTGGAGCAGACCCCACAAGGATTCCGGTCTACTTTGAGGGCC
CCCAGCTTGCTCTATAATCTCATTCACTTGAACACATCAAACGACTGTGGGTTCCAGAAGATAACT
TTGGATTGTCAGAATATTTATACTTGGAAGTCCAGGCATATAGTTGGAAAGAGAGGGGACACTAGG
AAGAAAATAGAAATGGAGACCAAAACTTCTATTAGCATTCCTAAACCTGGACAAGACGGGGAAATT
GTAATCACTGGCCAGCATCGAAATGGTGTAATTTCAGCCCGAACACGGATTGATGTTCTTTTGGAC
ACTTTTCGAAGAAAGCAGCCCTTCACTCACTTCCTTGCCTTTTTCCTCAATGAAGTTGAGGTTCAG
GAAGGATTCCTGAGATTCCAGGAGGAAGTACTGGCGAAGTGCTCCATGGATCATGGGGTTGACAGC
AGCATTTTCCAGAATCCTAAAAAGCTTCATCTAACTATTGGGATGTTGGTGCTTTTGAGTGAGGAA
GAGATCCAGCAGACATGTGAGATGCTACAGCAGTGTAAAGAGGAATTCATTAATGATATTTCTGGG
GGTAAACCCCTAGAAGTGGAGATGGCAGGGATAGAATACATGAATGATGATCCTGGCATGGTGGAT
GTTCTTTACGCCAAAGTCCATATGAAAGATGGCTCCAACAGGCTACAAGAATTAGTTGATCGAGTG
CTGGAACGTTTTCAGGCATCTGGACTAATAGTGAAAGAGTGGAATAGTGTGAAACTGCATGCTACA
GTTATGAATACACTATTCAGGAAAGACCCCAATGCTGAAGGCAGGTACAATCTCTACACAGCGGAA
GGCAAATATATCTTCAAGGAAAGAGAATCATTTGATGGCCGAAATATTTTAAAGAGCTTTGCCTTG
TTGCCCAGGCTGGAGTACAATGATGCAATCTCCGCTCACTGCAACCTGTGCCTCCCGGGTTCAAGT
GATTCTCCTGCCTCAGCCTCCCAAGTAGCTGGGATTACAGGTGTCTCTGATGCATATTCTCAGAGC
CTACCAGGAAAATCC

pDB.His.TEV.ASCC1 transcript variant 2 sequence 5'-3'

ATGGGCAGCAGCCATCATCATCATCATCACTCCTCAGGTGAGAATCTGTATTTTCAGGGCATGGGA
TCGATGGAAGTTCTGCGTCCACAGCTTATAAGAATTGATGGCCGGAATTACAGGAAGAATCCAGTC
CAAGAACAGACCTATCAACATGAAGAAGATGAAGAGGACTTCTATCAAGGCTCCATGGAGTGTGCT
GATGAGCCCTGTGATGCCTACGAGGTGGAGCAGACCCCACAAGGATTCCGGTCTACTTTGAGGGCC
CCCAGCTTGCTCTATAAGCATATAGTTGGAAAGAGAGGGGACACTAGGAAGAAAATAGAAATGGAG
ACCAAAACTTCTATTAGCATTCCTAAACCTGGACAAGACGGGGAAATTGTAATCACTGGCCAGCAT
CGAAATGGTGTAATTTCAGCCCGAACACGGATTGATGTTCTTTTGGACACTTTTCGAAGAAAGCAG
CCCTTCACTCACTTCCTTGCCTTTTTCCTCAATGAAGTTGAGGTTCAGGAAGGATTCCTGAGATTC
CAGGAGGAAGTACTGGCGAAGTGCTCCATGGATCATGGGGTTGACAGCAGCATTTTCCAGAATCCT
AAAAAGCTTCATCTAACTATTGGGATGTTGGTGCTTTTGAGTGAGGAAGAGATCCAGCAGACATGT
GAGATGCTACAGCAGTGTAAAGAGGAATTCATTAATGATATTTCTGGGGGTAAACCCCTAGAAGTG
GAGATGGCAGGGATAGAATACATGAATGATGATCCTGGCATGGTGGATGTTCTTTACGCCAAAGTC
CATATGAAAGATGGCTCCAACAGGCTACAAGAATTAGTTGATCGAGTGCTGGAACGTTTTCAGGCA

TCTGGACTAATAGTGAAAGAGTGGAATAGTGTGAAACTGCATGCTACAGTTATGAATACACTATTC
AGGAAAGACCCCAATGCTGAAGGCAGGTACAATCTCTACACAGCGGAAGGCAAATATATCTTCAAG
GAAAGAGAATCATTTGATGGCCGAAATATTTTAAAGTTGTTTGAGAACTTCTACTTTGGCTCCCTA
AAGCTGAATTCAATTCACATCTCTCAGAGGTTCACCGTAGACAGCTTTGGAAACTACGCTTCCTGT
GGACAAATTGACTTCTCCTAA

Human protein sequences

All protein sequences are given in ProtParam format. Fusion partners and purification tags are highlighted in red.

His_6-SUMO-NEK4

```
    10         20         30         40         50         60
MGSSHHHHHH SSGLVPRGSH MSDSEVNQEA KPEVKPEVKP ETHINLKVSD GSSEIFFKIK

    70         80         90        100        110        120
KTTPLRRLME AFAKRQGKEM DSLRFLYDGI RIQADQTPED LDMEDNDIIE AHREQIGGMP

   130        140        150        160        170        180
LAAYCYLRVV GKGSYGEVTL VKHRRDGKQY VIKKLNLRNA SSRERRAAEQ EAQLLSQLKH

   190        200        210        220        230        240
PNIVTYKESW EGGDGLLYIV MGFCEGGDLY RKLKEQKGQL LPENQVVEWF VQIAMALQYL

   250        260        270        280        290        300
HEKHILHRDL KTQNVFLTRT NIIKVGDLGI ARVLENHCDM ASTLIGTPYY MSPELFSNKP

   310        320        330        340        350        360
YNYKSDVWAL GCCVYEMATL KHAFNAKDMN SLVYRIIEGK LPAMPRDYSP ELAELIRTML

   370        380        390        400        410        420
SKRPEERPSV RSILRQPYIK RQISFFLEAT KIKTSKNNIK NGDSQSKPFA TVVSGEAESN

   430        440        450        460        470        480
HEVIHPQPLS SEGSQTYIMG EGKCLSQEKP RASGLLKSPA SLKAHTCKQD LSNTTELATI

   490        500        510        520        530        540
SSVNIDILPA KGRDSVSDGF VQENQPRYLD ASNELGGICS ISQVEEEMLQ DNTKSSAQPE

   550        560        570        580        590        600
NLIPMWSSDI VTGEKNEPVK PLQPLIKEQK PKDQSLALSP KLECSGTILA HSNLRLLGSS
```

```
    610         620         630         640         650         660
DSPASASRVA GITGVCHHAQ DQVAGECIIE KQGRIHPDLQ PHNSGSEPSL SRQRRQKRRE

    670         680         690         700         710         720
QTEHRGEKRQ VRRDLFAFQE SPPRFLPSHP IVGKVDVTST QKEAENQRRV VTGSVSSSRS

    730         740         750         760         770         780
SEMSSSKDRP LSARERRRLK QSQEEMSSSG PSVRKASLSV AGPGKPQEED QPLPARRLSS

    790         800         810         820         830         840
DCSVTQERKQ IHCLSEDELS SSTSSTDKSD GDYGEGKGQT NEINALVQLM TQTLKLDSKE

    850         860         870         880         890         900
SCEDVPVANP VSEFKLHRKY RDTLILHGKV AEEAEEIHFK ELPSAIMPGS EKIRRLVEVL

    910           920           930           940           950
RTDVIRGLGV QLLEQVYDLL EEEDEFDREV RLREHMGEKY TTYSVKARQL KFFEENMNF
```

His_6-SUMO-ZMAT1

```
    10          20          30          40          50          60
MGSSHHHHHH SSGLVPRGSH MSDSEVNQEA KPEVKPEVKP ETHINLKVSD GSSEIFFKIK

    70          80          90         100         110         120
KTTPLRRLME AFAKRQGKEM DSLRFLYDGI RIQADQTPED LDMEDNDIIE AHREQIGGME

    130         140         150         160         170         180
SCSVTRLECS GAISAHCSLH LPGSSDSPAS ASQIAGTTDA IWNEQEKAEL FTDKFCQVCG

    190         200         210         220         230         240
VMLQFESQRI SHYEGEKHAQ NVSFYFQMHG EQNEVPGKKM KMHVENFQVH RYEGVDKNKF

    250         260         270         280         290         300
CDLCNMMFSS PLIAQSHYVG KVHAKKLKQL MEEHDQASPS GFQPEMAFSM RTYVCHICSI

    310         320         330         340         350         360
AFTSLDMFRS HMQGSEHQIK ESIVINLVKN SRKTQDSYQN ECADYINVQK ARGLEAKTCF

    370         380         390         400         410         420
RKMEESSLET RRYREVVDSR PRHRMFEQRL PFETFRTYAA PYNISQAMEK QLPHSKKTYD
```

```
    430         440         450         460         470         480
SFQDELEDYI  KVQKARGLDP  KTCFRKMREN  SVDTHGYREM  VDSGPRSRMC  EQRFSHEASQ

    490         500         510         520         530         540
TYQRPYHISP  VESQLPQWLP  THSKRTYDSF  QDELEDYIKV  QKARGLEPKT  CFRKIGDSSV

    550         560         570         580         590         600
ETHRNREMVD  VRPRHRMLEQ  KLPCETFQTY  SGPYSISQVV  ENQLPHCLPA  HDSKQRLDSI

    610         620         630         640         650         660
SYCQLTRDCF  PEKPVPLSLN  QQENNSGSYS  VESEVYKHLS  SENNTADHQA  GHKRKHQKRK

    670         680         690         700         710         720
RHLEEGKERP  EKEQSKHKRK  KSYEDTDLDK  DKSIRQRKRE  EDRVKVSSGK  LKHRKKKKSH

    730                             740                             750
DVPSEKEERK HRKEKKKSVE ERTEEEMLWD ESILGF
```

GST-PPP5D1

```
    10          20          30          40          50          60
MSPILGYWKI  KGLVQPTRLL  LEYLEEKYEE  HLYERDEGDK  WRNKKFELGL  EFPNLPYYID

    70          80          90          100         110         120
GDVKLTQSMA  IIRYIADKHN  MLGGCPKERA  EISMLEGAVL  DIRYGVSRIA  YSKDFETLKV

    130         140         150         160         170         180
DFLSKLPEML  KMFEDRLCHK  TYLNGDHVTH  PDFMLYDALD  VVLYMDPMCL  DAFPKLVCFK

    190         200         210         220         230         240
KRIEAIPQID  KYLKSSKYIA  WPLQGWQATF  GGGDHPPKSD  LVPRGSMAEM  RAWRPLVRPS

    250         260         270         280         290         300
LQCVKLGRAT  ARWWWVVKVK  PHDKDAKMEY  QECNKIVKQK  AFERAIAGDE  HKRSVVDSLD

    310         320         330         340         350         360
IESMTIEGEY  SGPKLEDDKV  TITFMKGLMQ  WYKDQKKLHQ  KCAYQGLALS  PRLKCSGTVT

    370                             380                             390
AHCSLNLLGP RDPPASASQV AVTEGMHHHT WLIFLFL
```

His_6-BCAS4

```
    10         20         30         40         50         60
MGSSHHHHHH SSGLVPRGSH MQRTGGGAPR PGRNHGLPGS LRQPDPVALL MLLVDADQPE

    70         80         90        100        110        120
PMRSGARELA LFLTPEPGAE AKEVEETIEG MLLRLEEFCS LADLIRSDTS QILEENIPVL

    130        140        150        160        170        180
KAKLTEMRGI YAKVDRLEAF VKMVGHHVAF LEADVLQAER DHGAFPQALR RWLGSAGLPS

    190            200            210            220            230
FRNVECSGTI PARCNLRLPG SSDSPASASQ VAGITEVTCT GARDVRAGHT V
```

His_6-SUMO-ASCC1

```
    10         20         30         40         50         60
MGSSHHHHHH SSGLVPRGSH MSDSEVNQEA KPEVKPEVKP ETHINLKVSD GSSEIFFKIK

    70         80         90        100        110        120
KTTPLRRLME AFAKRQGKEM DSLRFLYDGI RIQADQTPED LDMEDNDIIE AHREQIGGME

    130        140        150        160        170        180
VLRPQLIRID GRNYRKNPVQ EQTYQHEEDE EDFYQGSMEC ADEPCDAYEV EQTPQGFRST

    190        200        210        220        230        240
LRAPSLLYNL IHLNTSNDCG FQKITLDCQN IYTWKSRHIV GKRGDTRKKI EMETKTSISI

    250        260        270        280        290        300
PKPGQDGEIV ITGQHRNGVI SARTRIDVLL DTFRRKQPFT HFLAFFLNEV EVQEGFLRFQ

    310        320        330        340        350        360
EEVLAKCSMD HGVDSSIFQN PKKLHLTIGM LVLLSEEEIQ QTCEMLQQCK EEFINDISGG

    370        380        390        400        410        420
KPLEVEMAGI EYMNDDPGMV DVLYAKVHMK DGSNRLQELV DRVLERFQAS GLIVKEWNSV

    430        440        450        460        470        480
KLHATVMNTL FRKDPNAEGR YNLYTAEGKY IFKERESFDG RNILKSFALL PRLEYNDAIS
```

```
    490                             500                             510
AHCNLCLPGS SDSPASASQV AGITGVSDAY SQSLPGKS
```

His_6-TEV-ASCC1 Isoform 1 (*Alu*)

```
    10          20          30          40          50          60
MGSSHHHHHH SSGENLYFQG MGSMEVLRPQ LIRIDGRNYR KNPVQEQTYQ HEEDEEDFYQ

    70          80          90          100         110         120
GSMECADEPC DAYEVEQTPQ GFRSTLRAPS LLYNLIHLNT SNDCGFQKIT LDCQNIYTWK

    130         140         150         160         170         180
SRHIVGKRGD TRKKIEMETK TSISIPKPGQ DGEIVITGQH RNGVISARTR IDVLLDTFRR

    190         200         210         220         230         240
KQPFTHFLAF FLNEVEVQEG FLRFQEEVLA KCSMDHGVDS SIFQNPKKLH LTIGMLVLLS

    250         260         270         280         290         300
EEEIQQTCEM LQQCKEEFIN DISGGKPLEV EMAGIEYMND DPGMVDVLYA KVHMKDGSNR

    310         320         330         340         350         360
LQELVDRVLE RFQASGLIVK EWNSVKLHAT VMNTLFRKDP NAEGRYNLYT AEGKYIFKER

    370         380         390         400         410         420
ESFDGRNILK SFALLPRLEY NDAISAHCNL CLPGSSDSPA SASQVAGITG VSDAYSQSLP


GKS
```

His_6-TEV-ASCC1 Isoform 2 (Non-*Alu*)

```
    10          20          30          40          50          60
MGSSHHHHHH SSGENLYFQG MGSMEVLRPQ LIRIDGRNYR KNPVQEQTYQ HEEDEEDFYQ

    70          80          90          100         110         120
GSMECADEPC DAYEVEQTPQ GFRSTLRAPS LLYKHIVGKR GDTRKKIEME TKTSISIPKP

    130         140         150         160         170         180
GQDGEIVITG QHRNGVISAR TRIDVLLDTF RRKQPFTHFL AFFLNEVEVQ EGFLRFQEEV

    190         200         210         220         230         240
```

```
LAKCSMDHGV DSSIFQNPKK LHLTIGMLVL LSEEEIQQTC EMLQQCKEEF INDISGGKPL

    250         260         270         280         290         300
EVEMAGIEYM NDDPGMVDVL YAKVHMKDGS NRLQELVDRV LERFQASGLI VKEWNSVKLH

    310         320         330         340         350         360
ATVMNTLFRK DPNAEGRYNL YTAEGKYIFK ERESFDGRNI LKLFENFYFG SLKLNSIHIS

    370                                                         380
QRFTVDSFGN YASCGQIDFS
```

3 MBP DNA and protein sequences

MBP plasmid sequences

pDB.His.MBP.Stop

The pDB.His.MBP.Stop plasmid was created through site-directed mutagenesis of the pDB.His.MBP plasmid (Clone ID: EvNO00065130, DNASU) to introduce a STOP codon (shown in red) directly after the MBP-coding gene (shown in blue). The vector also encodes a polyhistidine (shown in green) at the beginning of the MBP coding gene.

The pDB.His.MBP.Stop plasmid and subsequent *Alu*-containing plasmids were used to express *Alu*-containing histidine-tagged MBP variants in *E.coli* cells.

Full pDB.His.MBP.Stop sequence 5'-3'

```
TGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGT
GACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCAC
GTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTT
ACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATA
GACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGG
AACAACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATAAGGGATTTTGCCGATTCGGCCTAT
TGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTTTAACAAAATATTAACGTTTACA
ATTTCAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACAT
TCAAATATGTATCCGCTCATGAATTAATTCTTAGAAAAACTCATCGAGCATCAAATGAAACTGCAA
TTTATTCATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTGTAATGAAGGAGAAAA
CTCACCGAGGCAGTTCCATAGGATGGCAAGATCCTGGTATCGGTCTGCGATTCCGACTCGTCCAAC
ATCAATACAACCTATTAATTTCCCCTCGTCAAAAATAAGGTTATCAAGTGAGAAATCACCATGAGT
GACGACTGAATCCGGTGAGAATGGCAAAAGTTTATGCATTTCTTTCCAGACTTGTTCAACAGGCCA
```

GCCATTACGCTCGTCATCAAAATCACTCGCATCAACCAAACCGTTATTCATTCGTGATTGCGCCTG
AGCGAGACGAAATACGCGATCGCTGTTAAAAGGACAATTACAAACAGGAATCGAATGCAACCGGCG
CAGGAACACTGCCAGCGCATCAACAATATTTTCACCTGAATCAGGATATTCTTCTAATACCTGGAA
TGCTGTTTTCCCGGGGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAATGCTT
GATGGTCGGAAGAGGCATAAATTCCGTCAGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATT
GGCAACGCTACCTTTGCCATGTTTCAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCGATA
GATTGTCGCACCTGATTGCCCGACATTATCGCGAGCCCATTTATACCCATATAAATCAGCATCCAT
GTTGGAATTTAATCGCGGCCTAGAGCAAGACGTTTCCCGTTGAATATGGCTCATAACACCCCTTGT
ATTACTGTTTATGTAAGCAGACAGTTTTATTGTTCATGACCAAAATCCCTTAACGTGAGTTTTCGT
TCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCG
TAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGC
TACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCCTTCTAG
TGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAA
TCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGAT
AGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGC
GAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAG
GGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTC
CAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGAT
TTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGT
TCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATA
ACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGT
CAGTGAGCGAGGAAGCGGAAGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTT
CACACCGCATATATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATAC
ACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCCCCGACACCCGCCAACACCCGCTGACGCG
CCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGC
ATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCTGCGGTAAAGCTCATCAGCG
TGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTCATCCGCGTCCAGCTCGTTGAGTTTCTCCAGA
AGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGTTTGGTCACT
GATGCCTCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGAGGATG
CTCACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACTG
GCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACA
GATGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCAG
GGCGCTGACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATTCATGTTGTTGCT
CAGGTCGCAGACGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCGGTGATTCATTCTGC
TAACCAGTAAGGCAACCCCGCCAGCCTAGCCGGGTCCTCAACGACAGGAGCACGATCATGCGCACC
CGTGGGGCCGCCATGCCGGCGATAATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTG
ACGAAGGCTTGAGCGAGGGCGTGCAAGATTCCGAATACCGCAAGCGACAGGCCGATCATCGTCGCG
CTCCAGCGAAAGCGGTCCTCGCCGAAAATGACCCAGAGCGCTGCCGGCACCTGTCCTACGAGTTGC
ATGATAAAGAAGACAGTCATAAGTGCGGCGACGATAGTCATGCCCCGCGCCCACCGGAAGGAGCTG
ACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGAGATCCCGGTGCCTAATGAGTGAGCTAACTTACA
TTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGA

ATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCCAGGGTGGTTTTTCTTTTCACCAG
TGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCAC
GCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCT
GTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAAT
GGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTC
ATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTAT
CGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGA
ACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAG
TCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAA
TAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTT
AATGATCAGCCCACTGACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCC
GCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGC
GACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTT
GCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTT
TTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGAC
ACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACTCTC
TTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCGCCATTCGATGGTGTCCGGGATCTCGAC
GCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGTAGGTTGAGGCCGTTGAGCACCG
CCGCCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCACGGGGCCTGCCA
CCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGA
TGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGG
CGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGA
TAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACCATGGGCAGCAGCC
ATCATCATCATCATCACGGTACCAAAACTGAAGAAGGTAAACTGGTAATCTGGATTAACGGCGATA
AAGGCTATAACGGTCTCGCTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGAATTAAAGTCACCG
TTGAGCATCCGGATAAACTGGAAGAGAAATTCCCACAGGTTGCGGCAACTGGCGATGGCCCTGACA
TTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTCAATCTGGCCTGTTGGCTGAAATCACCC
CGGACAAAGCGTTCCAGGACAAGCTGTATCCGTTTACCTGGGATGCCGTACGTTACAACGGCAAGC
TGATTGCTTACCCGATCGCTGTTGAAGCGTTATCGCTGATTTATAACAAAGATCTGCTGCCGAACC
CGCCAAAAACCTGGGAAGAGATCCCGGCGCTGGATAAAGAACTGAAAGCGAAAGGTAAGAGCGCGC
TGATGTTCAACCTGCAAGAACCGTACTTCACCTGGCCGCTGATTGCTGCTGACGGGGGTTATGCGT
TCAAGTATGAAAACGGCAAGTACGACATTAAAGACGTGGGCGTGGATAACGCTGGCGCGAAAGCGG
GTCTGACCTTCCTGGTTGACCTGATTAAAAACAAACACATGAATGCAGACACCGATTACTCCATCG
CAGAAGCTGCCTTTAATAAAGGCGAAACAGCGATGACCATCAACGGCCCGTGGGCATGGTCCAACA
TCGACACCAGCAAAGTGAATTATGGTGTAACGGTACTGCCGACCTTCAAGGGTCAACCATCCAAAC
CGTTCGTTGGCGTGCTGAGCGCAGGTATTAACGCCGCCAGTCCGAACAAAGAGCTGGCGAAAGAGT
TCCTCGAAAACTATCTGCTGACTGATGAAGGTCTGGAAGCGGTTAATAAAGACAAACCGCTGGGTG
CCGTAGCGCTGAAGTCTTACGAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACCATGGAAA
ACGCCCAGAAAGGTGAAATCATGCCGAACATCCCGCAGATGTCCGCTTTCTGGTATGCCGTGCGTA
CTGCGGTGATCAACGCCGCCAGCGGTCGTCAGACTGTCGATGAAGCCCTGAAAGACGCGCAGACTT
AGACCGATTACGATATCCCAACGACCGAAAACCTTTACTTCCAGGGCCATATGGCTAGCATGACTG

GTGGACAGCAAATGGGTCGCGGATCCGAATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAG
CACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCT
GCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTG
CTGAAAGGAGGAACTATATCCGGAT

pDB.His.MBP.Stop *Alu* sequences

The following codon-optimised *Alu* sequence (shown in red) was inserted into eight different locations within the MBP coding gene using site-directed mutagenesis via inverse PCR.

TTGGAGTGCTCAGGGACTATTTCAGCTCACTGCAACCTTCGCTTACCGGGCTCATCCGACTCGCCC
GCCAGCGCGAGCCGCGTAGCAGGCATCACCGGA

The following sequences show only the region of plasmid encoding the histidine-tagged MBP protein containing *Alu* insert. The remainder of the plasmid is identical to that pDB.His.MBP.Stop shown previously.

Sequences are labelled by the amino acid encoded directly after the site of *Alu* insertion, with the exception of T367 which corresponds to an *Alu* insertion at the end of the MBP coding gene directly before the STOP codon. (Note: amino acids are numbered from the beginning MBP, not from the beginning of the histidine tag.)

G6 *Alu* 5'-3' sequence

ATGGGCAGCAGCCATCATCATCATCATCACGGTACCAAAACTGAAGAATTGGAGTGCTCAGGGACT
ATTTCAGCTCACTGCAACCTTCGCTTACCGGGCTCATCCGACTCGCCCGCCAGCGCGAGCCGCGTA
GCAGGCATCACCGGAGGTAAACTGGTAATCTGGATTAACGGCGATAAAGGCTATAACGGTCTCGCT
GAAGTCGGTAAGAAATTCGAGAAAGATACCGGAATTAAAGTCACCGTTGAGCATCCGGATAAACTG
GAAGAGAAATTCCCACAGGTTGCGGCAACTGGCGATGGCCCTGACATTATCTTCTGGGCACACGAC
CGCTTTGGTGGCTACGCTCAATCTGGCCTGTTGGCTGAAATCACCCCGGACAAAGCGTTCCAGGAC
AAGCTGTATCCGTTTACCTGGGATGCCGTACGTTACAACGGCAAGCTGATTGCTTACCCGATCGCT
GTTGAAGCGTTATCGCTGATTTATAACAAAGATCTGCTGCCGAACCCGCCAAAAACCTGGGAAGAG
ATCCCGGCGCTGGATAAAGAACTGAAAGCGAAAGGTAAGAGCGCGCTGATGTTCAACCTGCAAGAA
CCGTACTTCACCTGGCCGCTGATTGCTGCTGACGGGGGTTATGCGTTCAAGTATGAAAACGGCAAG
TACGACATTAAAGACGTGGGCGTGGATAACGCTGGCGCGAAAGCGGGTCTGACCTTCCTGGTTGAC
CTGATTAAAAACAAACACATGAATGCAGACACCGATTACTCCATCGCAGAAGCTGCCTTTAATAAA
GGCGAAACAGCGATGACCATCAACGGCCCGTGGGCATGGTCCAACATCGACACCAGCAAAGTGAAT
TATGGTGTAACGGTACTGCCGACCTTCAAGGGTCAACCATCCAAACCGTTCGTTGGCGTGCTGAGC
GCAGGTATTAACGCCGCCAGTCCGAACAAAGAGCTGGCGAAAGAGTTCCTCGAAAACTATCTGCTG
ACTGATGAAGGTCTGGAAGCGGTTAATAAAGACAAACCGCTGGGTGCCGTAGCGCTGAAGTCTTAC
GAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACCATGGAAAACGCCCAGAAAGGTGAAATC

ATGCCGAACATCCCGCAGATGTCCGCTTTCTGGTATGCCGTGCGTACTGCGGTGATCAACGCCGCC
AGCGGTCGTCAGACTGTCGATGAAGCCCTGAAAGACGCGCAGACTTAG

T81 *Alu* 5'-3' sequence

ATGGGCAGCAGCCATCATCATCATCATCACGGTACCAAAACTGAAGAAGGTAAACTGGTAATCTGG
ATTAACGGCGATAAAGGCTATAACGGTCTCGCTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGA
ATTAAAGTCACCGTTGAGCATCCGGATAAACTGGAAGAGAAATTCCCACAGGTTGCGGCAACTGGC
GATGGCCCTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTCAATCTGGCCTGTTG
GCTGAAATCTTGGAGTGCTCAGGGACTATTTCAGCTCACTGCAACCTTCGCTTACCGGGCTCATCC
GACTCGCCCGCCAGCGCGAGCCGCGTAGCAGGCATCACCGGAACCCCGGACAAAGCGTTCCAGGAC
AAGCTGTATCCGTTTACCTGGGATGCCGTACGTTACAACGGCAAGCTGATTGCTTACCCGATCGCT
GTTGAAGCGTTATCGCTGATTTATAACAAAGATCTGCTGCCGAACCCGCCAAAAACCTGGGAAGAG
ATCCCGGCGCTGGATAAAGAACTGAAAGCGAAAGGTAAGAGCGCGCTGATGTTCAACCTGCAAGAA
CCGTACTTCACCTGGCCGCTGATTGCTGCTGACGGGGGTTATGCGTTCAAGTATGAAAACGGCAAG
TACGACATTAAAGACGTGGGCGTGGATAACGCTGGCGCGAAAGCGGGTCTGACCTTCCTGGTTGAC
CTGATTAAAAACAAACACATGAATGCAGACACCGATTACTCCATCGCAGAAGCTGCCTTTAATAAA
GGCGAAACAGCGATGACCATCAACGGCCCGTGGGCATGGTCCAACATCGACACCAGCAAAGTGAAT
TATGGTGTAACGGTACTGCCGACCTTCAAGGGTCAACCATCCAAACCGTTCGTTGGCGTGCTGAGC
GCAGGTATTAACGCCGCCAGTCCGAACAAAGAGCTGGCGAAAGAGTTCCTCGAAAACTATCTGCTG
ACTGATGAAGGTCTGGAAGCGGTTAATAAAGACAAACCGCTGGGTGCCGTAGCGCTGAAGTCTTAC
GAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACCATGGAAAACGCCCAGAAAGGTGAAATC
ATGCCGAACATCCCGCAGATGTCCGCTTTCTGGTATGCCGTGCGTACTGCGGTGATCAACGCCGCC
AGCGGTCGTCAGACTGTCGATGAAGCCCTGAAAGACGCGCAGACTTAG

P126 *Alu* 5'-3' sequence

ATGGGCAGCAGCCATCATCATCATCATCACGGTACCAAAACTGAAGAAGGTAAACTGGTAATCTGG
ATTAACGGCGATAAAGGCTATAACGGTCTCGCTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGA
ATTAAAGTCACCGTTGAGCATCCGGATAAACTGGAAGAGAAATTCCCACAGGTTGCGGCAACTGGC
GATGGCCCTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTCAATCTGGCCTGTTG
GCTGAAATCACCCCGGACAAAGCGTTCCAGGACAAGCTGTATCCGTTTACCTGGGATGCCGTACGT
TACAACGGCAAGCTGATTGCTTACCCGATCGCTGTTGAAGCGTTATCGCTGATTTATAACAAAGAT
CTGCTGCCGAACTTGGAGTGCTCAGGGACTATTTCAGCTCACTGCAACCTTCGCTTACCGGGCTCA
TCCGACTCGCCCGCCAGCGCGAGCCGCGTAGCAGGCATCACCGGACCGCCAAAAACCTGGGAAGAG
ATCCCGGCGCTGGATAAAGAACTGAAAGCGAAAGGTAAGAGCGCGCTGATGTTCAACCTGCAAGAA
CCGTACTTCACCTGGCCGCTGATTGCTGCTGACGGGGGTTATGCGTTCAAGTATGAAAACGGCAAG
TACGACATTAAAGACGTGGGCGTGGATAACGCTGGCGCGAAAGCGGGTCTGACCTTCCTGGTTGAC
CTGATTAAAAACAAACACATGAATGCAGACACCGATTACTCCATCGCAGAAGCTGCCTTTAATAAA
GGCGAAACAGCGATGACCATCAACGGCCCGTGGGCATGGTCCAACATCGACACCAGCAAAGTGAAT
TATGGTGTAACGGTACTGCCGACCTTCAAGGGTCAACCATCCAAACCGTTCGTTGGCGTGCTGAGC

GCAGGTATTAACGCCGCCAGTCCGAACAAAGAGCTGGCGAAAGAGTTCCTCGAAAACTATCTGCTG
ACTGATGAAGGTCTGGAAGCGGTTAATAAAGACAAACCGCTGGGTGCCGTAGCGCTGAAGTCTTAC
GAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACCATGGAAAACGCCCAGAAAGGTGAAATC
ATGCCGAACATCCCGCAGATGTCCGCTTTCTGGTATGCCGTGCGTACTGCGGTGATCAACGCCGCC
AGCGGTCGTCAGACTGTCGATGAAGCCCTGAAAGACGCGCAGACTTAG

D178 *Alu* 5'-3' sequence

ATGGGCAGCAGCCATCATCATCATCATCACGGTACCAAAACTGAAGAAGGTAAACTGGTAATCTGG
ATTAACGGCGATAAAGGCTATAACGGTCTCGCTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGA
ATTAAAGTCACCGTTGAGCATCCGGATAAACTGGAAGAGAAATTCCCACAGGTTGCGGCAACTGGC
GATGGCCCTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTCAATCTGGCCTGTTG
GCTGAAATCACCCCGGACAAAGCGTTCCAGGACAAGCTGTATCCGTTTACCTGGGATGCCGTACGT
TACAACGGCAAGCTGATTGCTTACCCGATCGCTGTTGAAGCGTTATCGCTGATTTATAACAAAGAT
CTGCTGCCGAACCCGCCAAAAACCTGGGAAGAGATCCCGGCGCTGGATAAAGAACTGAAAGCGAAA
GGTAAGAGCGCGCTGATGTTCAACCTGCAAGAACCGTACTTCACCTGGCCGCTGATTGCTGCTGAC
GGGGGTTATGCGTTCAAGTATGAAAACGGCAAGTACTTGGAGTGCTCAGGGACTATTTCAGCTCAC
TGCAACCTTCGCTTACCGGGCTCATCCGACTCGCCCGCCAGCGCGAGCCGCGTAGCAGGCATCACC
GGAGACATTAAAGACGTGGGCGTGGATAACGCTGGCGCGAAAGCGGGTCTGACCTTCCTGGTTGAC
CTGATTAAAAACAAACACATGAATGCAGACACCGATTACTCCATCGCAGAAGCTGCCTTTAATAAA
GGCGAAACAGCGATGACCATCAACGGCCCGTGGGCATGGTCCAACATCGACACCAGCAAAGTGAAT
TATGGTGTAACGGTACTGCCGACCTTCAAGGGTCAACCATCCAAACCGTTCGTTGGCGTGCTGAGC
GCAGGTATTAACGCCGCCAGTCCGAACAAAGAGCTGGCGAAAGAGTTCCTCGAAAACTATCTGCTG
ACTGATGAAGGTCTGGAAGCGGTTAATAAAGACAAACCGCTGGGTGCCGTAGCGCTGAAGTCTTAC
GAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACCATGGAAAACGCCCAGAAAGGTGAAATC
ATGCCGAACATCCCGCAGATGTCCGCTTTCTGGTATGCCGTGCGTACTGCGGTGATCAACGCCGCC
AGCGGTCGTCAGACTGTCGATGAAGCCCTGAAAGACGCGCAGACTTAG

G253 *Alu* 5'-3' sequence

ATGGGCAGCAGCCATCATCATCATCATCACGGTACCAAAACTGAAGAAGGTAAACTGGTAATCTGG
ATTAACGGCGATAAAGGCTATAACGGTCTCGCTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGA
ATTAAAGTCACCGTTGAGCATCCGGATAAACTGGAAGAGAAATTCCCACAGGTTGCGGCAACTGGC
GATGGCCCTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTCAATCTGGCCTGTTG
GCTGAAATCACCCCGGACAAAGCGTTCCAGGACAAGCTGTATCCGTTTACCTGGGATGCCGTACGT
TACAACGGCAAGCTGATTGCTTACCCGATCGCTGTTGAAGCGTTATCGCTGATTTATAACAAAGAT
CTGCTGCCGAACCCGCCAAAAACCTGGGAAGAGATCCCGGCGCTGGATAAAGAACTGAAAGCGAAA
GGTAAGAGCGCGCTGATGTTCAACCTGCAAGAACCGTACTTCACCTGGCCGCTGATTGCTGCTGAC
GGGGGTTATGCGTTCAAGTATGAAAACGGCAAGTACGACATTAAAGACGTGGGCGTGGATAACGCT
GGCGCGAAAGCGGGTCTGACCTTCCTGGTTGACCTGATTAAAAACAAACACATGAATGCAGACACC
GATTACTCCATCGCAGAAGCTGCCTTTAATAAAGGCGAAACAGCGATGACCATCAACGGCCCGTGG
GCATGGTCCAACATCGACACCAGCAAAGTGAATTATGGTGTAACGGTACTGCCGACCTTCAAGTTG

GAGTGCTCAGGGACTATTTCAGCTCACTGCAACCTTCGCTTACCGGGCTCATCCGACTCGCCCGCC
AGCGCGAGCCGCGTAGCAGGCATCACCGGAGGTCAACCATCCAAACCGTTCGTTGGCGTGCTGAGC
GCAGGTATTAACGCCGCCAGTCCGAACAAAGAGCTGGCGAAAGAGTTCCTCGAAAACTATCTGCTG
ACTGATGAAGGTCTGGAAGCGGTTAATAAAGACAAACCGCTGGGTGCCGTAGCGCTGAAGTCTTAC
GAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACCATGGAAAACGCCCAGAAAGGTGAAATC
ATGCCGAACATCCCGCAGATGTCCGCTTTCTGGTATGCCGTGCGTACTGCGGTGATCAACGCCGCC
AGCGGTCGTCAGACTGTCGATGAAGCCCTGAAAGACGCGCAGACTTAG

A293 *Alu* 5'-3' sequence

ATGGGCAGCAGCCATCATCATCATCATCACGGTACCAAAACTGAAGAAGGTAAACTGGTAATCTGG
ATTAACGGCGATAAAGGCTATAACGGTCTCGCTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGA
ATTAAAGTCACCGTTGAGCATCCGGATAAACTGGAAGAGAAATTCCCACAGGTTGCGGCAACTGGC
GATGGCCCTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTCAATCTGGCCTGTTG
GCTGAAATCACCCCGGACAAAGCGTTCCAGGACAAGCTGTATCCGTTTACCTGGGATGCCGTACGT
TACAACGGCAAGCTGATTGCTTACCCGATCGCTGTTGAAGCGTTATCGCTGATTTATAACAAAGAT
CTGCTGCCGAACCCGCCAAAAACCTGGGAAGAGATCCCGGCGCTGGATAAAGAACTGAAAGCGAAA
GGTAAGAGCGCGCTGATGTTCAACCTGCAAGAACCGTACTTCACCTGGCCGCTGATTGCTGCTGAC
GGGGGTTATGCGTTCAAGTATGAAAACGGCAAGTACGACATTAAAGACGTGGGCGTGGATAACGCT
GGCGCGAAAGCGGGTCTGACCTTCCTGGTTGACCTGATTAAAAACAAACACATGAATGCAGACACC
GATTACTCCATCGCAGAAGCTGCCTTTAATAAAGGCGAAACAGCGATGACCATCAACGGCCCGTGG
GCATGGTCCAACATCGACACCAGCAAAGTGAATTATGGTGTAACGGTACTGCCGACCTTCAAGGGT
CAACCATCCAAACCGTTCGTTGGCGTGCTGAGCGCAGGTATTAACGCCGCCAGTCCGAACAAAGAG
CTGGCGAAAGAGTTCCTCGAAAACTATCTGCTGACTGATGAAGGTCTGGAATTGGAGTGCTCAGGG
ACTATTTCAGCTCACTGCAACCTTCGCTTACCGGGCTCATCCGACTCGCCCGCCAGCGCGAGCCGC
GTAGCAGGCATCACCGGAGCGGTTAATAAAGACAAACCGCTGGGTGCCGTAGCGCTGAAGTCTTAC
GAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACCATGGAAAACGCCCAGAAAGGTGAAATC
ATGCCGAACATCCCGCAGATGTCCGCTTTCTGGTATGCCGTGCGTACTGCGGTGATCAACGCCGCC
AGCGGTCGTCAGACTGTCGATGAAGCCCTGAAAGACGCGCAGACTTAG

N333 *Alu* 5'-3' sequence

ATGGGCAGCAGCCATCATCATCATCATCACGGTACCAAAACTGAAGAAGGTAAACTGGTAATCTGG
ATTAACGGCGATAAAGGCTATAACGGTCTCGCTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGA
ATTAAAGTCACCGTTGAGCATCCGGATAAACTGGAAGAGAAATTCCCACAGGTTGCGGCAACTGGC
GATGGCCCTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTCAATCTGGCCTGTTG
GCTGAAATCACCCCGGACAAAGCGTTCCAGGACAAGCTGTATCCGTTTACCTGGGATGCCGTACGT
TACAACGGCAAGCTGATTGCTTACCCGATCGCTGTTGAAGCGTTATCGCTGATTTATAACAAAGAT
CTGCTGCCGAACCCGCCAAAAACCTGGGAAGAGATCCCGGCGCTGGATAAAGAACTGAAAGCGAAA
GGTAAGAGCGCGCTGATGTTCAACCTGCAAGAACCGTACTTCACCTGGCCGCTGATTGCTGCTGAC
GGGGGTTATGCGTTCAAGTATGAAAACGGCAAGTACGACATTAAAGACGTGGGCGTGGATAACGCT
GGCGCGAAAGCGGGTCTGACCTTCCTGGTTGACCTGATTAAAAACAAACACATGAATGCAGACACC

GATTACTCCATCGCAGAAGCTGCCTTTAATAAAGGCGAAACAGCGATGACCATCAACGGCCCGTGG
GCATGGTCCAACATCGACACCAGCAAAGTGAATTATGGTGTAACGGTACTGCCGACCTTCAAGGGT
CAACCATCCAAACCGTTCGTTGGCGTGCTGAGCGCAGGTATTAACGCCGCCAGTCCGAACAAAGAG
CTGGCGAAAGAGTTCCTCGAAAACTATCTGCTGACTGATGAAGGTCTGGAAGCGGTTAATAAAGAC
AAACCGCTGGGTGCCGTAGCGCTGAAGTCTTACGAGGAAGAGTTGGCGAAAGATCCACGTATTGCC
GCCACCATGGAAAACGCCCAGAAAGGTGAAATCATGCCGTTGGAGTGCTCAGGGACTATTTCAGCT
CACTGCAACCTTCGCTTACCGGGCTCATCCGACTCGCCCGCCAGCGCGAGCCGCGTAGCAGGCATC
ACCGGAAACATCCCGCAGATGTCCGCTTTCTGGTATGCCGTGCGTACTGCGGTGATCAACGCCGCC
AGCGGTCGTCAGACTGTCGATGAAGCCCTGAAAGACGCGCAGACTTAG

T367 *Alu* 5'-3' sequence

ATGGGCAGCAGCCATCATCATCATCATCACGGTACCAAAACTGAAGAAGGTAAACTGGTAATCTGG
ATTAACGGCGATAAAGGCTATAACGGTCTCGCTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGA
ATTAAAGTCACCGTTGAGCATCCGGATAAACTGGAAGAGAAATTCCCACAGGTTGCGGCAACTGGC
GATGGCCCTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTCAATCTGGCCTGTTG
GCTGAAATCACCCCGGACAAAGCGTTCCAGGACAAGCTGTATCCGTTTACCTGGGATGCCGTACGT
TACAACGGCAAGCTGATTGCTTACCCGATCGCTGTTGAAGCGTTATCGCTGATTTATAACAAAGAT
CTGCTGCCGAACCCGCCAAAAACCTGGGAAGAGATCCCGGCGCTGGATAAAGAACTGAAAGCGAAA
GGTAAGAGCGCGCTGATGTTCAACCTGCAAGAACCGTACTTCACCTGGCCGCTGATTGCTGCTGAC
GGGGGTTATGCGTTCAAGTATGAAAACGGCAAGTACGACATTAAAGACGTGGGCGTGGATAACGCT
GGCGCGAAAGCGGGTCTGACCTTCCTGGTTGACCTGATTAAAAACAAACACATGAATGCAGACACC
GATTACTCCATCGCAGAAGCTGCCTTTAATAAAGGCGAAACAGCGATGACCATCAACGGCCCGTGG
GCATGGTCCAACATCGACACCAGCAAAGTGAATTATGGTGTAACGGTACTGCCGACCTTCAAGGGT
CAACCATCCAAACCGTTCGTTGGCGTGCTGAGCGCAGGTATTAACGCCGCCAGTCCGAACAAAGAG
CTGGCGAAAGAGTTCCTCGAAAACTATCTGCTGACTGATGAAGGTCTGGAAGCGGTTAATAAAGAC
AAACCGCTGGGTGCCGTAGCGCTGAAGTCTTACGAGGAAGAGTTGGCGAAAGATCCACGTATTGCC
GCCACCATGGAAAACGCCCAGAAAGGTGAAATCATGCCGAACATCCCGCAGATGTCCGCTTTCTGG
TATGCCGTGCGTACTGCGGTGATCAACGCCGCCAGCGGTCGTCAGACTGTCGATGAAGCCCTGAAA
GACGCGCAGACTTTGGAGTGCTCAGGGACTATTTCAGCTCACTGCAACCTTCGCTTACCGGGCTCA
TCCGACTCGCCCGCCAGCGCGAGCCGCGTAGCAGGCATCACCGGATAG

pDB.His.MBP.Stop scrambled *Alu* sequences

The following codon-optimised *Alu* sequence (shown in blue) was inserted into eight different locations within the MBP coding gene using site-directed mutagenesis via inverse PCR.

ATAGCGCGTCTGCATGGTCCTTCCGCAAGTAATGGGACTTCCTCCTCTACTTGTGCGCCCGATCTT
GGCGTGGGGGAATCAGCCTTGTGTATTTCCCGC

The following sequences show only the region of plasmid encoding the histidine-tagged MBP protein containing *Alu* insert. The remainder of the plasmid is identical to that pDB.His.MBP.Stop shown previously.

D178* Scrambled *Alu* 5-3' Sequence

ATGGGCAGCAGCCATCATCATCATCATCACGGTACCAAAACTGAAGAAGGTAAACTGGTAATCTGG
ATTAACGGCGATAAAGGCTATAACGGTCTCGCTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGA
ATTAAAGTCACCGTTGAGCATCCGGATAAACTGGAAGAGAAATTCCCACAGGTTGCGGCAACTGGC
GATGGCCCTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTCAATCTGGCCTGTTG
GCTGAAATCACCCCGGACAAAGCGTTCCAGGACAAGCTGTATCCGTTTACCTGGGATGCCGTACGT
TACAACGGCAAGCTGATTGCTTACCCGATCGCTGTTGAAGCGTTATCGCTGATTTATAACAAAGAT
CTGCTGCCGAACCCGCCAAAAACCTGGGAAGAGATCCCGGCGCTGGATAAAGAACTGAAAGCGAAA
GGTAAGAGCGCGCTGATGTTCAACCTGCAAGAACCGTACTTCACCTGGCCGCTGATTGCTGCTGAC
GGGGGTTATGCGTTCAAGTATGAAAACGGCAAGTACATAGCGCGTCTGCATGGTCCTTCCGCAAGT
AATGGGACTTCCTCCTCTACTTGTGCGCCCGATCTTGGCGTGGGGGAATCAGCCTTGTGTATTTCC
CGCGACATTAAAGACGTGGGCGTGGATAACGCTGGCGCGAAAGCGGGTCTGACCTTCCTGGTTGAC
CTGATTAAAAACAAACACATGAATGCAGACACCGATTACTCCATCGCAGAAGCTGCCTTTAATAAA
GGCGAAACAGCGATGACCATCAACGGCCCGTGGGCATGGTCCAACATCGACACCAGCAAAGTGAAT
TATGGTGTAACGGTACTGCCGACCTTCAAGGGTCAACCATCCAAACCGTTCGTTGGCGTGCTGAGC
GCAGGTATTAACGCCGCCAGTCCGAACAAAGAGCTGGCGAAAGAGTTCCTCGAAAACTATCTGCTG
ACTGATGAAGGTCTGGAAGCGGTTAATAAAGACAAACCGCTGGGTGCCGTAGCGCTGAAGTCTTAC
GAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACCATGGAAAACGCCCAGAAAGGTGAAATC
ATGCCGAACATCCCGCAGATGTCCGCTTTCTGGTATGCCGTGCGTACTGCGGTGATCAACGCCGCC
AGCGGTCGTCAGACTGTCGATGAAGCCCTGAAAGACGCGCAGACTTAG

G253* Scrambled *Alu* 5-3' Sequence

ATGGGCAGCAGCCATCATCATCATCATCACGGTACCAAAACTGAAGAAGGTAAACTGGTAATCTGG
ATTAACGGCGATAAAGGCTATAACGGTCTCGCTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGA
ATTAAAGTCACCGTTGAGCATCCGGATAAACTGGAAGAGAAATTCCCACAGGTTGCGGCAACTGGC
GATGGCCCTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTCAATCTGGCCTGTTG
GCTGAAATCACCCCGGACAAAGCGTTCCAGGACAAGCTGTATCCGTTTACCTGGGATGCCGTACGT
TACAACGGCAAGCTGATTGCTTACCCGATCGCTGTTGAAGCGTTATCGCTGATTTATAACAAAGAT
CTGCTGCCGAACCCGCCAAAAACCTGGGAAGAGATCCCGGCGCTGGATAAAGAACTGAAAGCGAAA
GGTAAGAGCGCGCTGATGTTCAACCTGCAAGAACCGTACTTCACCTGGCCGCTGATTGCTGCTGAC
GGGGGTTATGCGTTCAAGTATGAAAACGGCAAGTACGACATTAAAGACGTGGGCGTGGATAACGCT
GGCGCGAAAGCGGGTCTGACCTTCCTGGTTGACCTGATTAAAAACAAACACATGAATGCAGACACC
GATTACTCCATCGCAGAAGCTGCCTTTAATAAAGGCGAAACAGCGATGACCATCAACGGCCCGTGG
GCATGGTCCAACATCGACACCAGCAAAGTGAATTATGGTGTAACGGTACTGCCGACCTTCAAGATA
GCGCGTCTGCATGGTCCTTCCGCAAGTAATGGGACTTCCTCCTCTACTTGTGCGCCCGATCTTGGC

GTGGGGGAATCAGCCTTGTGTATTTCCCGCGGTCAACCATCCAAACCGTTCGTTGGCGTGCTGAGC
GCAGGTATTAACGCCGCCAGTCCGAACAAAGAGCTGGCGAAAGAGTTCCTCGAAAACTATCTGCTG
ACTGATGAAGGTCTGGAAGCGGTTAATAAAGACAAACCGCTGGGTGCCGTAGCGCTGAAGTCTTAC
GAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACCATGGAAAACGCCCAGAAAGGTGAAATC
ATGCCGAACATCCCGCAGATGTCCGCTTTCTGGTATGCCGTGCGTACTGCGGTGATCAACGCCGCC
AGCGGTCGTCAGACTGTCGATGAAGCCCTGAAAGACGCGCAGACTTAG

N333* Scrambled *Alu* 5-3' Sequence

ATGGGCAGCAGCCATCATCATCATCATCACGGTACCAAAACTGAAGAAGGTAAACTGGTAATCTGG
ATTAACGGCGATAAAGGCTATAACGGTCTCGCTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGA
ATTAAAGTCACCGTTGAGCATCCGGATAAACTGGAAGAGAAATTCCCACAGGTTGCGGCAACTGGC
GATGGCCCTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTCAATCTGGCCTGTTG
GCTGAAATCACCCCGGACAAAGCGTTCCAGGACAAGCTGTATCCGTTTACCTGGGATGCCGTACGT
TACAACGGCAAGCTGATTGCTTACCCGATCGCTGTTGAAGCGTTATCGCTGATTTATAACAAAGAT
CTGCTGCCGAACCCGCCAAAAACCTGGGAAGAGATCCCGGCGCTGGATAAAGAACTGAAAGCGAAA
GGTAAGAGCGCGCTGATGTTCAACCTGCAAGAACCGTACTTCACCTGGCCGCTGATTGCTGCTGAC
GGGGGTTATGCGTTCAAGTATGAAAACGGCAAGTACGACATTAAAGACGTGGGCGTGGATAACGCT
GGCGCGAAAGCGGGTCTGACCTTCCTGGTTGACCTGATTAAAAACAAACACATGAATGCAGACACC
GATTACTCCATCGCAGAAGCTGCCTTTAATAAAGGCGAAACAGCGATGACCATCAACGGCCCGTGG
GCATGGTCCAACATCGACACCAGCAAAGTGAATTATGGTGTAACGGTACTGCCGACCTTCAAGGGT
CAACCATCCAAACCGTTCGTTGGCGTGCTGAGCGCAGGTATTAACGCCGCCAGTCCGAACAAAGAG
CTGGCGAAAGAGTTCCTCGAAAACTATCTGCTGACTGATGAAGGTCTGGAAGCGGTTAATAAAGAC
AAACCGCTGGGTGCCGTAGCGCTGAAGTCTTACGAGGAAGAGTTGGCGAAAGATCCACGTATTGCC
GCCACCATGGAAAACGCCCAGAAAGGTGAAATCATGCCGATAGCGCGTCTGCATGGTCCTTCCGCA
AGTAATGGGACTTCCTCCTCTACTTGTGCGCCCGATCTTGGCGTGGGGGAATCAGCCTTGTGTATT
TCCCGCAACATCCCGCAGATGTCCGCTTTCTGGTATGCCGTGCGTACTGCGGTGATCAACGCCGCC
AGCGGTCGTCAGACTGTCGATGAAGCCCTGAAAGACGCGCAGAC

TTAG

MBP Protein sequences

All protein sequences are given in ProtParam format. Protein regions encoded by *Alu* sequences are shown in red, and scrambled *Alu* insertions are shown in blue.

MBP-*Alu* Constructs

His_6-MBP

```
    10         20         30         40         50         60
MGSSHHHHHH GTKTEEGKLV IWINGDKGYN GLAEVGKKFE KDTGIKVTVE HPDKLEEKFP

    70         80         90        100        110        120
QVAATGDGPD IIFWAHDRFG GYAQSGLLAE ITPDKAFQDK LYPFTWDAVR YNGKLIAYPI

    130        140        150        160        170        180
AVEALSLIYN KDLLPNPPKT WEEIPALDKE LKAKGKSALM FNLQEPYFTW PLIAADGGYA

    190        200        210        220        230        240
FKYENGKYDI KDVGVDNAGA KAGLTFLVDL IKNKHMNADT DYSIAEAAFN KGETAMTING

    250        260        270        280        290        300
PWAWSNIDTS KVNYGVTVLP TFKGQPSKPF VGVLSAGINA ASPNKELAKE FLENYLLTDE

    310        320        330        340        350        360
GLEAVNKDKP LGAVALKSYE EELAKDPRIA ATMENAQKGE IMPNIPQMSA FWYAVRTAVI

    370
NAASGRQTVD EALKDAQT
```

His_6-MBP-G6

```
    10         20         30         40         50         60
MGSSHHHHHH GTKTEELECS GTISAHCNLR LPGSSDSPAS ASRVAGITGG KLVIWINGDK

    70         80         90        100        110        120
GYNGLAEVGK KFEKDTGIKV TVEHPDKLEE KFPQVAATGD GPDIIFWAHD RFGGYAQSGL

    130        140        150        160        170        180
LAEITPDKAF QDKLYPFTWD AVRYNGKLIA YPIAVEALSL IYNKDLLPNP PKTWEEIPAL

    190        200        210        220        230        240
DKELKAKGKS ALMFNLQEPY FTWPLIAADG GYAFKYENGK YDIKDVGVDN AGAKAGLTFL

    250        260        270        280        290        300
VDLIKNKHMN ADTDYSIAEA AFNKGETAMT INGPWAWSNI DTSKVNYGVT VLPTFKGQPS

    310        320        330        340        350        360
KPFVGVLSAG INAASPNKEL AKEFLENYLL TDEGLEAVNK DKPLGAVALK SYEEELAKDP
```

```
    370           380           390           400           410
RIAATMENAQ KGEIMPNIPQ MSAFWYAVRT AVINAASGRQ TVDEALKDAQ T
```

His_6-MBP-T81

```
    10          20          30          40          50          60
MGSSHHHHHH GTKTEEGKLV IWINGDKGYN GLAEVGKKFE KDTGIKVTVE HPDKLEEKFP

    70          80          90          100         110         120
QVAATGDGPD IIFWAHDRFG GYAQSGLLAE ILECSGTISA HCNLRLPGSS DSPASASRVA

    130         140         150         160         170         180
GITGTPDKAF QDKLYPFTWD AVRYNGKLIA YPIAVEALSL IYNKDLLPNP PKTWEEIPAL

    190         200         210         220         230         240
DKELKAKGKS ALMFNLQEPY FTWPLIAADG GYAFKYENGK YDIKDVGVDN AGAKAGLTFL

    250         260         270         280         290         300
VDLIKNKHMN ADTDYSIAEA AFNKGETAMT INGPWAWSNI DTSKVNYGVT VLPTFKGQPS

    310         320         330         340         350         360
KPFVGVLSAG INAASPNKEL AKEFLENYLL TDEGLEAVNK DKPLGAVALK SYEEELAKDP

    370           380           390           400           410
RIAATMENAQ KGEIMPNIPQ MSAFWYAVRT AVINAASGRQ TVDEALKDAQ T
```

His_6-MBP-P126

```
    10          20          30          40          50          60
MGSSHHHHHH GTKTEEGKLV IWINGDKGYN GLAEVGKKFE KDTGIKVTVE HPDKLEEKFP

    70          80          90          100         110         120
QVAATGDGPD IIFWAHDRFG GYAQSGLLAE ITPDKAFQDK LYPFTWDAVR YNGKLIAYPI

    130         140         150         160         170         180
AVEALSLIYN KDLLPNLECS GTISAHCNLR LPGSSDSPAS ASRVAGITGP PKTWEEIPAL

    190         200         210         220         230         240
DKELKAKGKS ALMFNLQEPY FTWPLIAADG GYAFKYENGK YDIKDVGVDN AGAKAGLTFL

    250         260         270         280         290         300
```

```
VDLIKNKHMN ADTDYSIAEA AFNKGETAMT INGPWAWSNI DTSKVNYGVT VLPTFKGQPS

    310        320        330        340        350        360
KPFVGVLSAG INAASPNKEL AKEFLENYLL TDEGLEAVNK DKPLGAVALK SYEEELAKDP

    370            380            390            400            410
RIAATMENAQ KGEIMPNIPQ MSAFWYAVRT AVINAASGRQ TVDEALKDAQ T
```

His_6-MBP-D178

```
    10         20         30         40         50         60
MGSSHHHHHH GTKTEEGKLV IWINGDKGYN GLAEVGKKFE KDTGIKVTVE HPDKLEEKFP

    70         80         90        100        110        120
QVAATGDGPD IIFWAHDRFG GYAQSGLLAE ITPDKAFQDK LYPFTWDAVR YNGKLIAYPI

    130        140        150        160        170        180
AVEALSLIYN KDLLPNPPKT WEEIPALDKE LKAKGKSALM FNLQEPYFTW PLIAADGGYA

    190        200        210        220        230        240
FKYENGKYLE CSGTISAHCN LRLPGSSDSP ASASRVAGIT GDIKDVGVDN AGAKAGLTFL

    250        260        270        280        290        300
VDLIKNKHMN ADTDYSIAEA AFNKGETAMT INGPWAWSNI DTSKVNYGVT VLPTFKGQPS

    310        320        330        340        350        360
KPFVGVLSAG INAASPNKEL AKEFLENYLL TDEGLEAVNK DKPLGAVALK SYEEELAKDP

    370            380            390            400            410
RIAATMENAQ KGEIMPNIPQ MSAFWYAVRT AVINAASGRQ TVDEALKDAQ T
```

His_6-MBP-G253

```
    10         20         30         40         50         60
MGSSHHHHHH GTKTEEGKLV IWINGDKGYN GLAEVGKKFE KDTGIKVTVE HPDKLEEKFP

    70         80         90        100        110        120
QVAATGDGPD IIFWAHDRFG GYAQSGLLAE ITPDKAFQDK LYPFTWDAVR YNGKLIAYPI

    130        140        150        160        170        180
AVEALSLIYN KDLLPNPPKT WEEIPALDKE LKAKGKSALM FNLQEPYFTW PLIAADGGYA
```

```
    190         200         210         220         230         240
FKYENGKYDI  KDVGVDNAGA  KAGLTFLVDL  IKNKHMNADT  DYSIAEAAFN  KGETAMTING

    250         260         270         280         290         300
PWAWSNIDTS  KVNYGVTVLP  TFKLECSGTI  SAHCNLRLPG  SSDSPASASR  VAGITGGQPS

    310         320         330         340         350         360
KPFVGVLSAG  INAASPNKEL  AKEFLENYLL  TDEGLEAVNK  DKPLGAVALK  SYEEELAKDP

    370             380             390             400             410
RIAATMENAQ  KGEIMPNIPQ  MSAFWYAVRT  AVINAASGRQ  TVDEALKDAQ  T
```

H_6-MBP-A293

```
    10          20          30          40          50          60
MGSSHHHHHH  GTKTEEGKLV  IWINGDKGYN  GLAEVGKKFE  KDTGIKVTVE  HPDKLEEKFP

    70          80          90          100         110         120
QVAATGDGPD  IIFWAHDRFG  GYAQSGLLAE  ITPDKAFQDK  LYPFTWDAVR  YNGKLIAYPI

    130         140         150         160         170         180
AVEALSLIYN  KDLLPNPPKT  WEEIPALDKE  LKAKGKSALM  FNLQEPYFTW  PLIAADGGYA

    190         200         210         220         230         240
FKYENGKYDI  KDVGVDNAGA  KAGLTFLVDL  IKNKHMNADT  DYSIAEAAFN  KGETAMTING

    250         260         270         280         290         300
PWAWSNIDTS  KVNYGVTVLP  TFKGQPSKPF  VGVLSAGINA  ASPNKELAKE  FLENYLLTDE

    310         320         330         340         350         360
GLELECSGTI  SAHCNLRLPG  SSDSPASASR  VAGITGAVNK  DKPLGAVALK  SYEEELAKDP

    370             380             390             400             410
RIAATMENAQ  KGEIMPNIPQ  MSAFWYAVRT  AVINAASGRQ  TVDEALKDAQ  T
```

H_6-MBP-N333

```
    10          20          30          40          50          60
MGSSHHHHHH  GTKTEEGKLV  IWINGDKGYN  GLAEVGKKFE  KDTGIKVTVE  HPDKLEEKFP

    70          80          90          100         110         120
```

```
QVAATGDGPD IIFWAHDRFG GYAQSGLLAE ITPDKAFQDK LYPFTWDAVR YNGKLIAYPI

    130        140        150        160        170        180
AVEALSLIYN KDLLPNPPKT WEEIPALDKE LKAKGKSALM FNLQEPYFTW PLIAADGGYA

    190        200        210        220        230        240
FKYENGKYDI KDVGVDNAGA KAGLTFLVDL IKNKHMNADT DYSIAEAAFN KGETAMTING

    250        260        270        280        290        300
PWAWSNIDTS KVNYGVTVLP TFKGQPSKPF VGVLSAGINA ASPNKELAKE FLENYLLTDE

    310        320        330        340        350        360
GLEAVNKDKP LGAVALKSYE EELAKDPRIA ATMENAQKGE IMPLECSGTI SAHCNLRLPG

    370            380            390            400            410
SSDSPASASR VAGITGNIPQ MSAFWYAVRT AVINAASGRQ TVDEALKDAQ T
```

His_6-MBP-T367

```
    10         20         30         40         50         60
MGSSHHHHHH GTKTEEGKLV IWINGDKGYN GLAEVGKKFE KDTGIKVTVE HPDKLEEKFP

    70         80         90        100        110        120
QVAATGDGPD IIFWAHDRFG GYAQSGLLAE ITPDKAFQDK LYPFTWDAVR YNGKLIAYPI

    130        140        150        160        170        180
AVEALSLIYN KDLLPNPPKT WEEIPALDKE LKAKGKSALM FNLQEPYFTW PLIAADGGYA

    190        200        210        220        230        240
FKYENGKYDI KDVGVDNAGA KAGLTFLVDL IKNKHMNADT DYSIAEAAFN KGETAMTING

    250        260        270        280        290        300
PWAWSNIDTS KVNYGVTVLP TFKGQPSKPF VGVLSAGINA ASPNKELAKE FLENYLLTDE

    310        320        330        340        350        360
GLEAVNKDKP LGAVALKSYE EELAKDPRIA ATMENAQKGE IMPNIPQMSA FWYAVRTAVI

    370            380            390            400            410
NAASGRQTVD EALKDAQTLE CSGTISAHCN LRLPGSSDSP ASASRVAGIT G
```

Scrambled MBP-*Alu* Constructs

His_6-MBP-D178*

```
    10          20          30          40          50          60
MGSSHHHHHH GTKTEEGKLV IWINGDKGYN GLAEVGKKFE KDTGIKVTVE HPDKLEEKFP

    70          80          90         100         110         120
QVAATGDGPD IIFWAHDRFG GYAQSGLLAE ITPDKAFQDK LYPFTWDAVR YNGKLIAYPI

   130         140         150         160         170         180
AVEALSLIYN KDLLPNPPKT WEEIPALDKE LKAKGKSALM FNLQEPYFTW PLIAADGGYA

   190         200         210         220         230         240
FKYENGKYIA RLHGPSASNG TSSSTCAPDL GVGESALCIS RDIKDVGVDN AGAKAGLTFL

   250         260         270         280         290         300
VDLIKNKHMN ADTDYSIAEA AFNKGETAMT INGPWAWSNI DTSKVNYGVT VLPTFKGQPS

   310         320         330         340         350         360
KPFVGVLSAG INAASPNKEL AKEFLENYLL TDEGLEAVNK DKPLGAVALK SYEEELAKDP

   370          380          390          400          410
RIAATMENAQ KGEIMPNIPQ MSAFWYAVRT AVINAASGRQ TVDEALKDAQ T
```

His_6-MBP-G253*

```
    10          20          30          40          50          60
MGSSHHHHHH GTKTEEGKLV IWINGDKGYN GLAEVGKKFE KDTGIKVTVE HPDKLEEKFP

    70          80          90         100         110         120
QVAATGDGPD IIFWAHDRFG GYAQSGLLAE ITPDKAFQDK LYPFTWDAVR YNGKLIAYPI

   130         140         150         160         170         180
AVEALSLIYN KDLLPNPPKT WEEIPALDKE LKAKGKSALM FNLQEPYFTW PLIAADGGYA

   190         200         210         220         230         240
FKYENGKYDI KDVGVDNAGA KAGLTFLVDL IKNKHMNADT DYSIAEAAFN KGETAMTING

   250         260         270         280         290         300
```

PWAWSNIDTS KVNYGVTVLP TFKIARLHGP SASNGTSSST CAPDLGVGES ALCISRGQPS

310 320 330 340 350 360
KPFVGVLSAG INAASPNKEL AKEFLENYLL TDEGLEAVNK DKPLGAVALK SYEEELAKDP

370 380 390 400 410
RIAATMENAQ KGEIMPNIPQ MSAFWYAVRT AVINAASGRQ TVDEALKDAQ T

His_6-MBP-N333*

10 20 30 40 50 60
MGSSHHHHHH GTKTEEGKLV IWINGDKGYN GLAEVGKKFE KDTGIKVTVE HPDKLEEKFP

70 80 90 100 110 120
QVAATGDGPD IIFWAHDRFG GYAQSGLLAE ITPDKAFQDK LYPFTWDAVR YNGKLIAYPI

130 140 150 160 170 180
AVEALSLIYN KDLLPNPPKT WEEIPALDKE LKAKGKSALM FNLQEPYFTW PLIAADGGYA

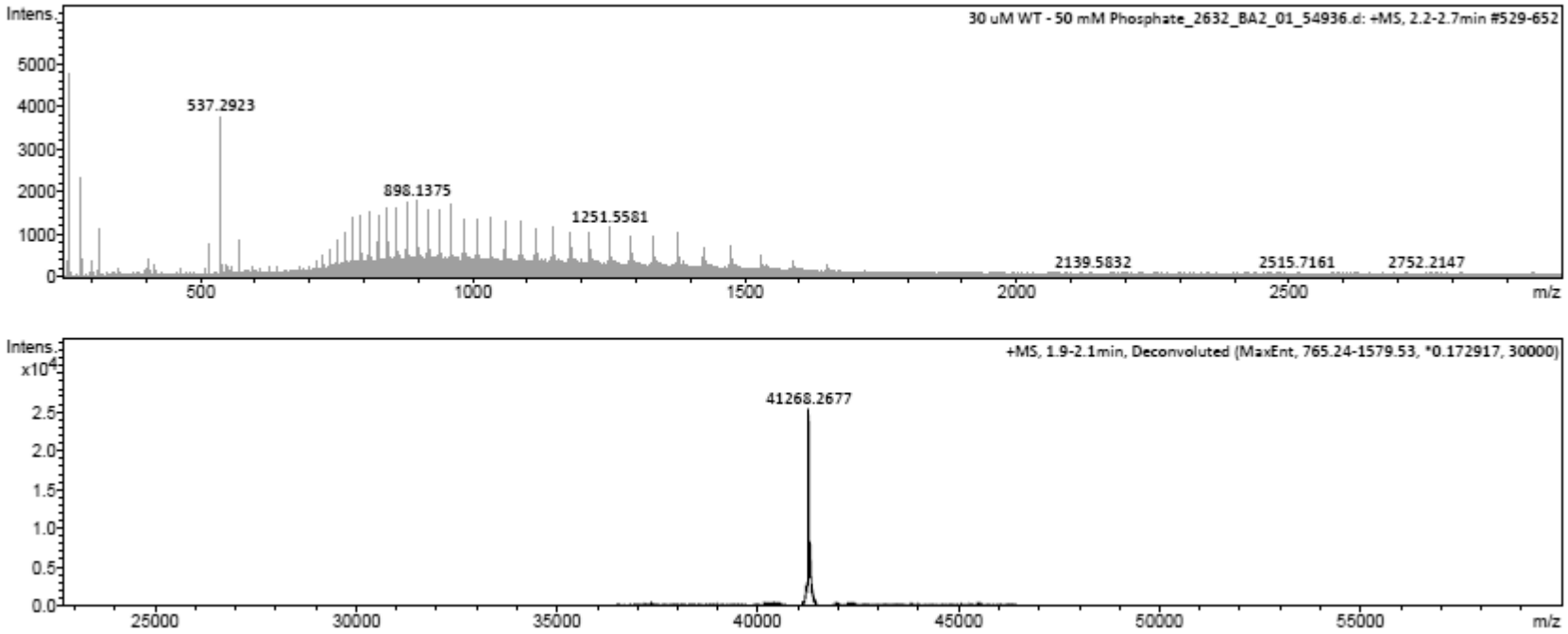
190 200 210 220 230 240
FKYENGKYDI KDVGVDNAGA KAGLTFLVDL IKNKHMNADT DYSIAEAAFN KGETAMTING

250 260 270 280 290 300
PWAWSNIDTS KVNYGVTVLP TFKGQPSKPF VGVLSAGINA ASPNKELAKE FLENYLLTDE

310 320 330 340 350 360
GLEAVNKDKP LGAVALKSYE EELAKDPRIA ATMENAQKGE IMPIARLHGP SASNGTSSST

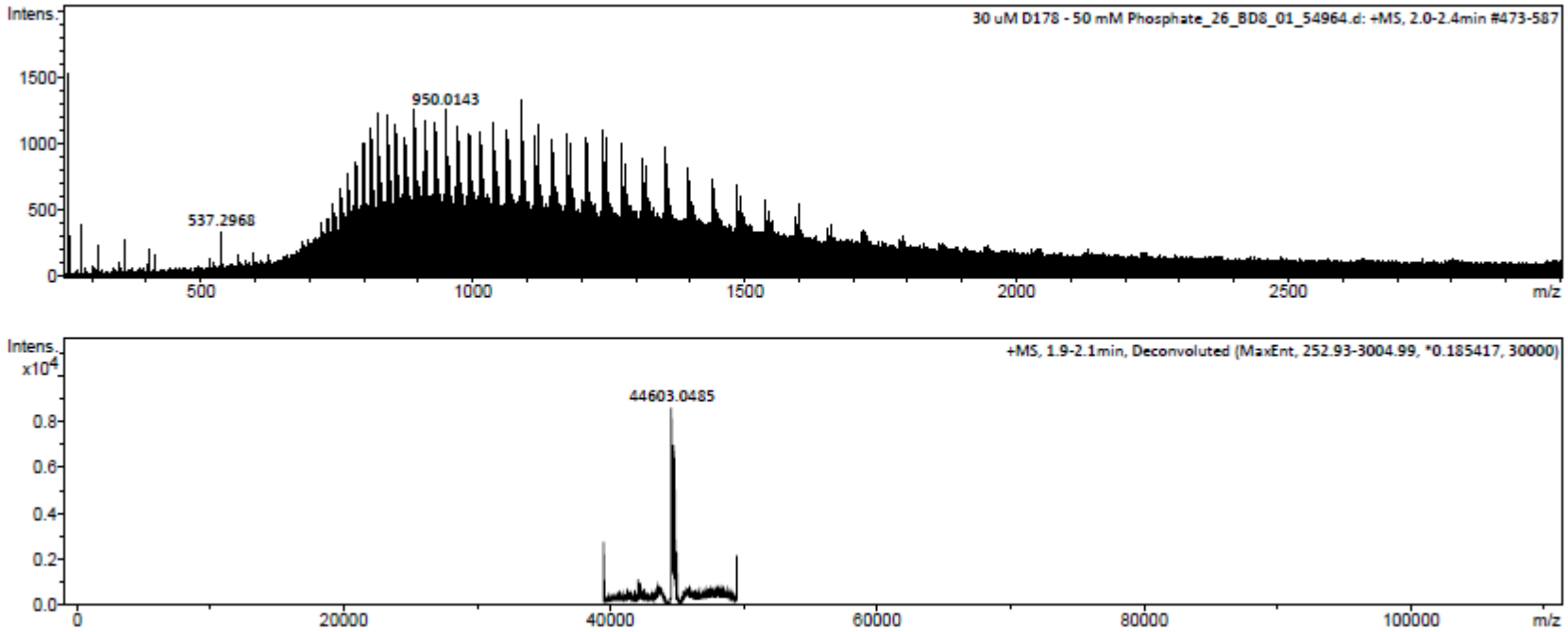
370 380 390 400 410
CAPDLGVGES ALCISRNIPQ MSAFWYAVRT AVINAASGRQ TVDEALKDAQ T

4 High resolution mass spectrometry



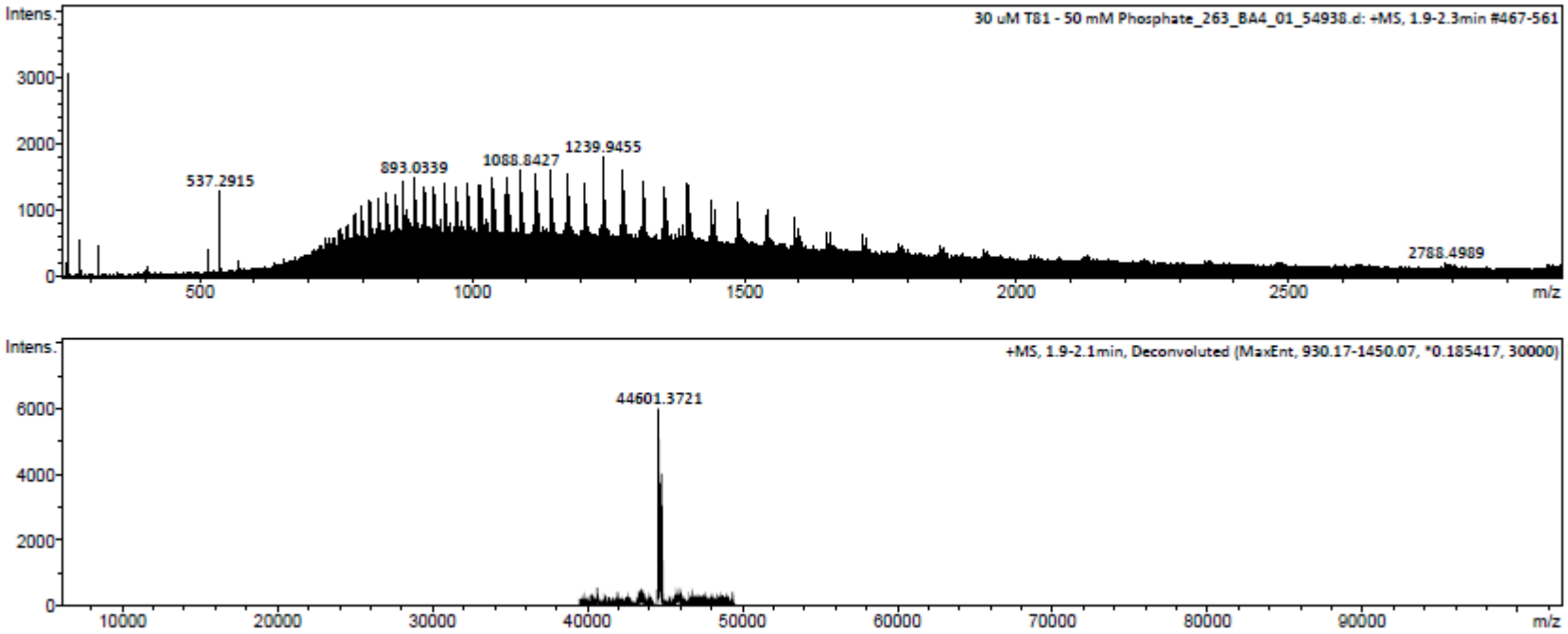
A4.1 Mass Spectrometry – His_6-MBP (WT)

Expected protein mass = 41,541.0 Da; observed protein mass = 41268.27 Da



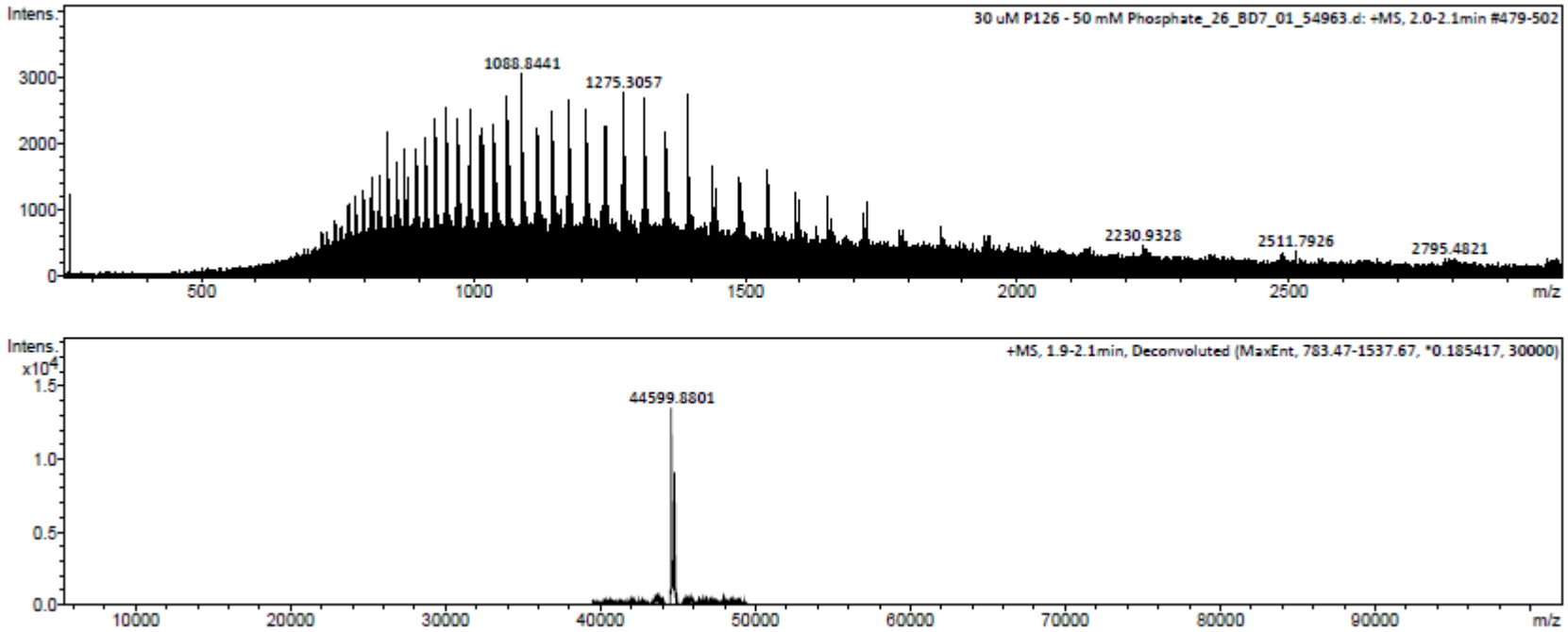
A4.2 Mass Spectrometry – His_6-MBP-G6

Expected protein mass = 44737.6 Da; observed protein mass = 44603.05 Da



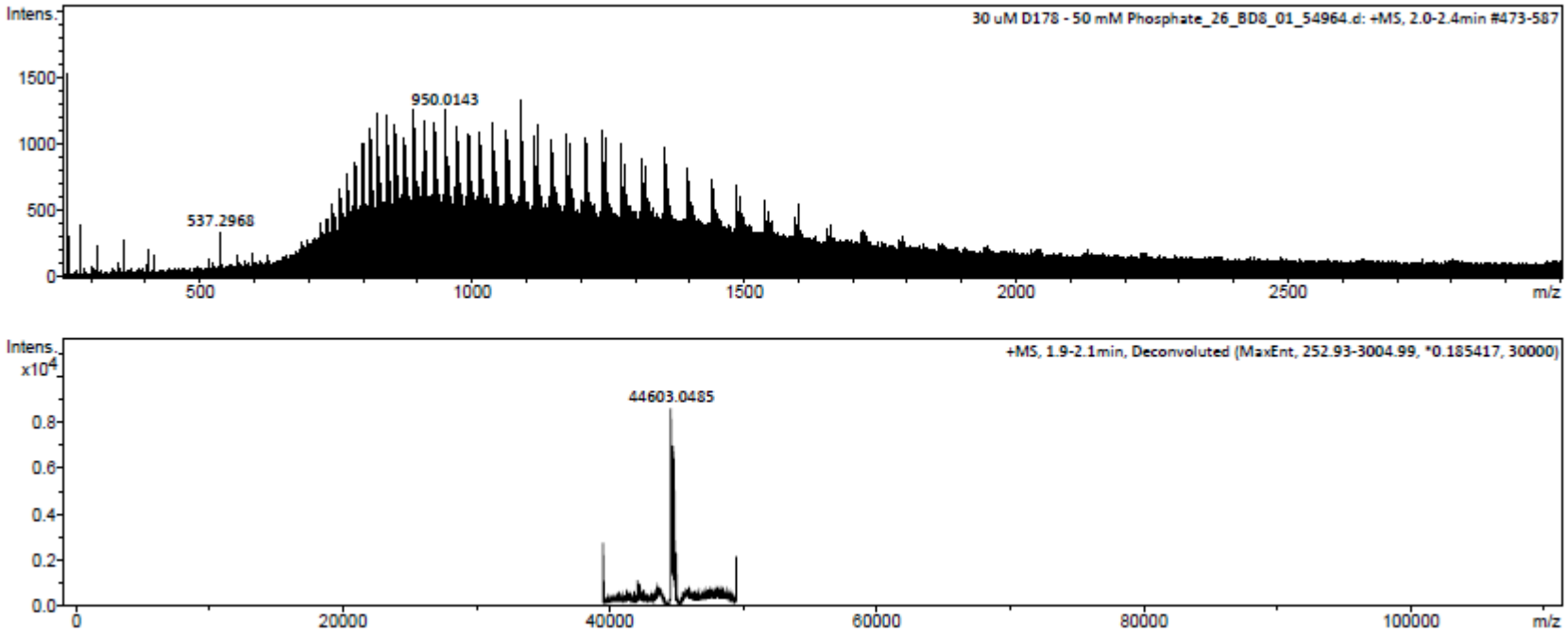
A4.3 Mass Spectrometry – His_6-MBP-T81

Expected protein mass = 44737.6 Da; observed protein mass = 44601.37 Da



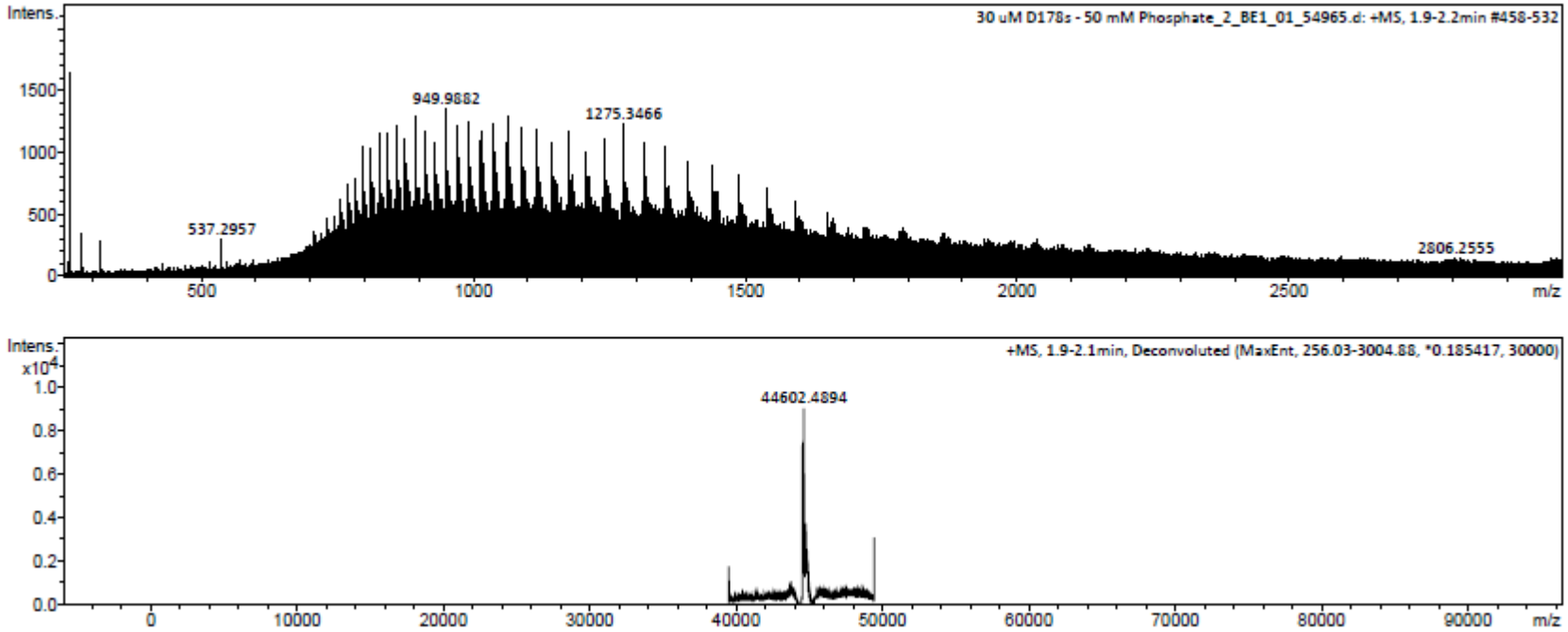
A4.4 Mass Spectrometry – His_6-MBP-P126

Expected protein mass = 44737.6 Da; observed protein mass = 44599.88 Da



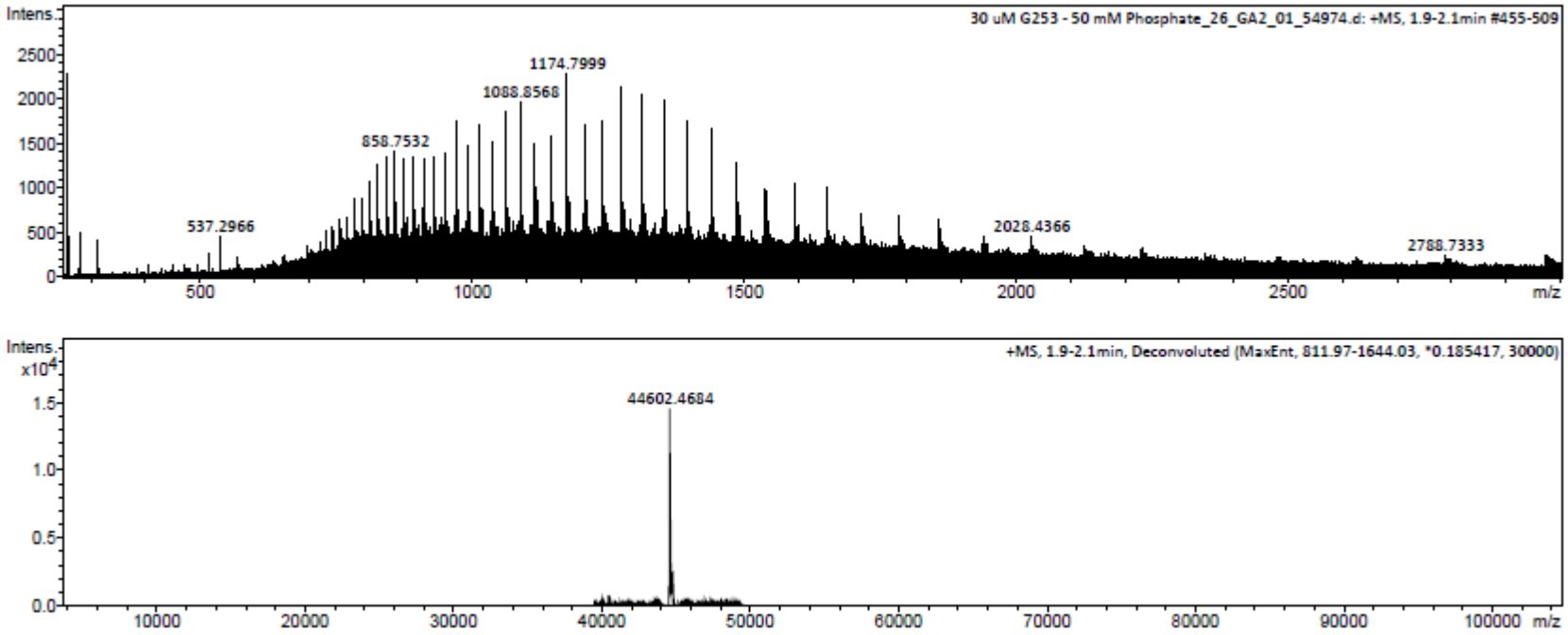
A4.5 Mass Spectrometry – His_6-MBP-D178

Expected protein mass = 44737.6 Da; observed protein mass = 44603.05 Da



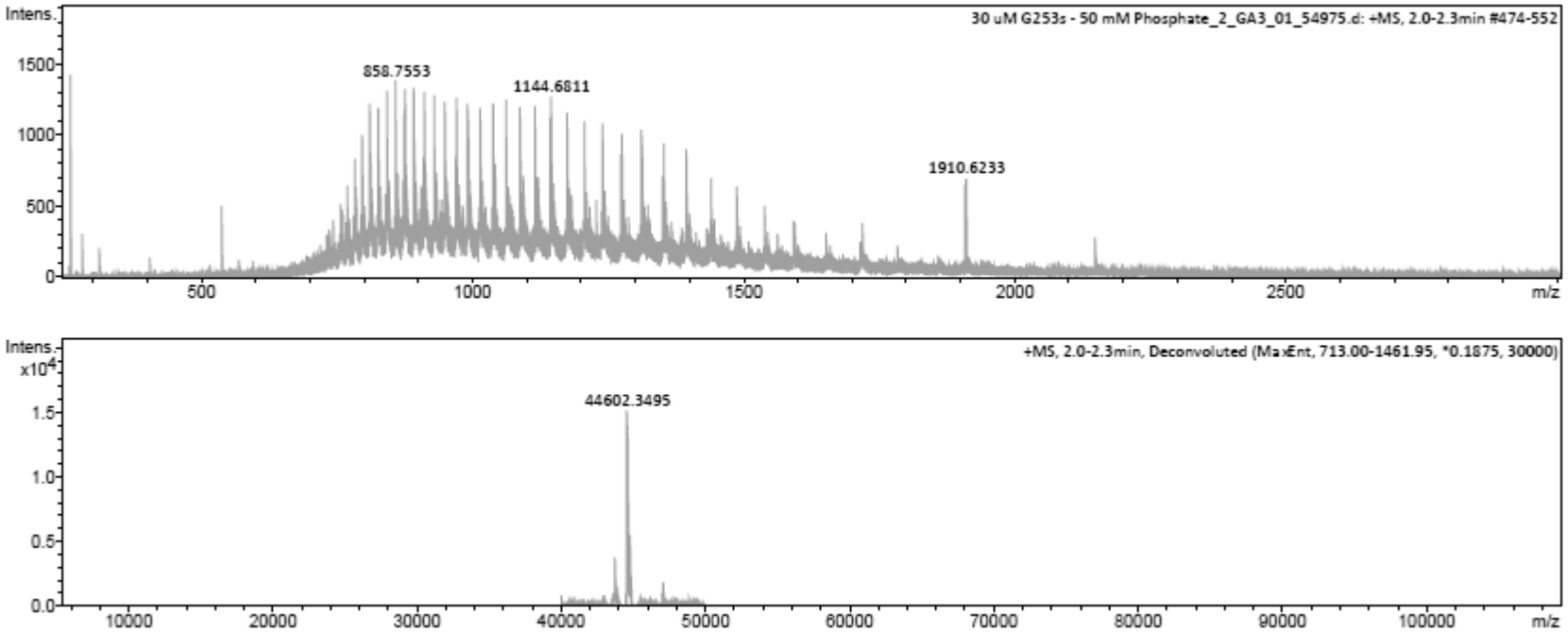
A4.6 Mass Spectrometry – His_6-MBP-D178*

Expected protein mass = 44737.6 Da; observed protein mass = 44602.49 Da



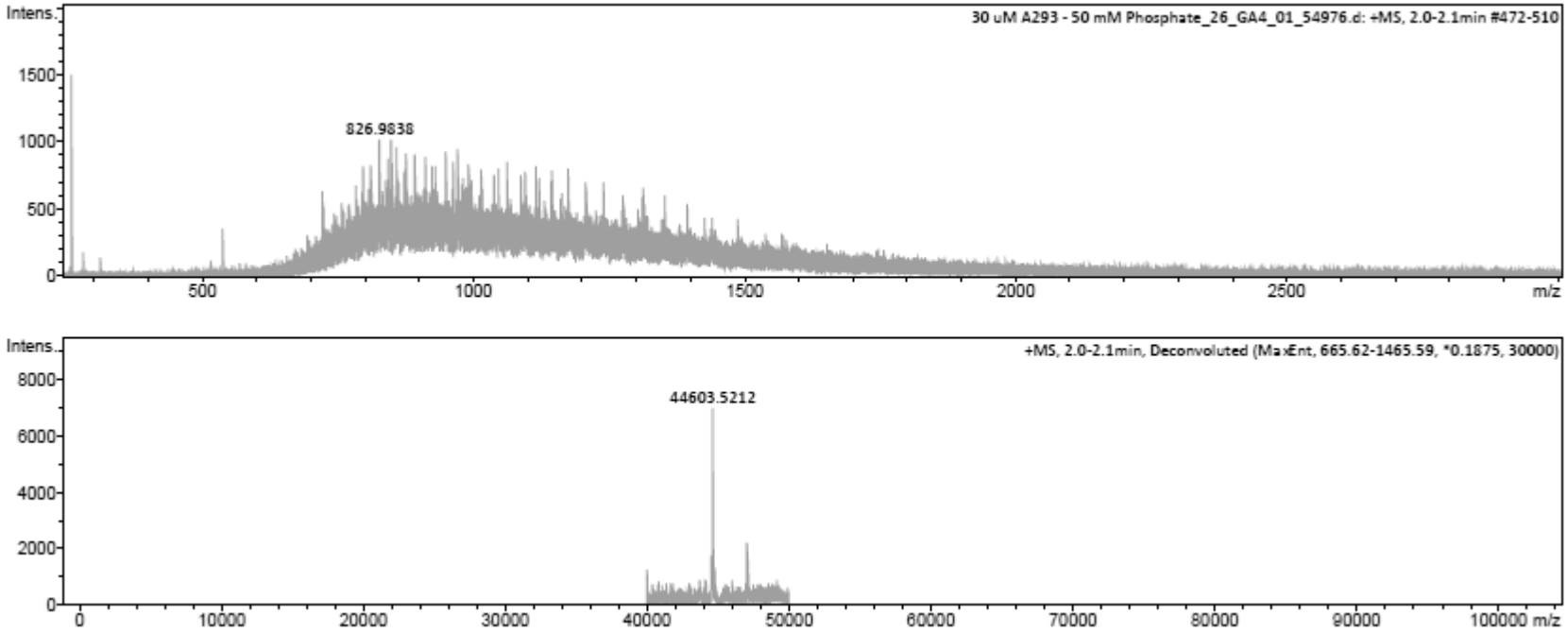
A4.7 Mass Spectrometry – His_6-MBP-G253

Expected protein mass = 44737.6 Da; observed protein mass = 44602.47 Da



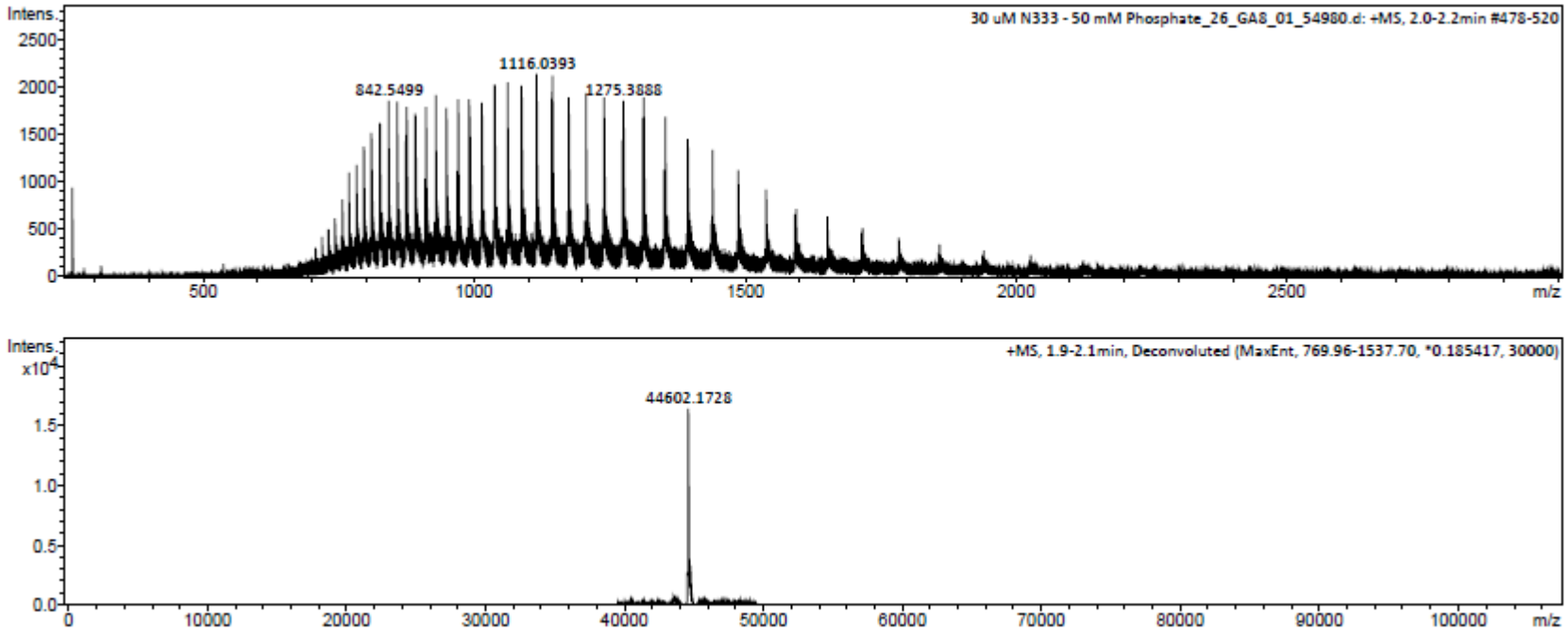
A4.8 Mass Spectrometry – His_6-MBP-G253*

Expected protein mass = 44737.6 Da; observed protein mass = 44602.35 Da



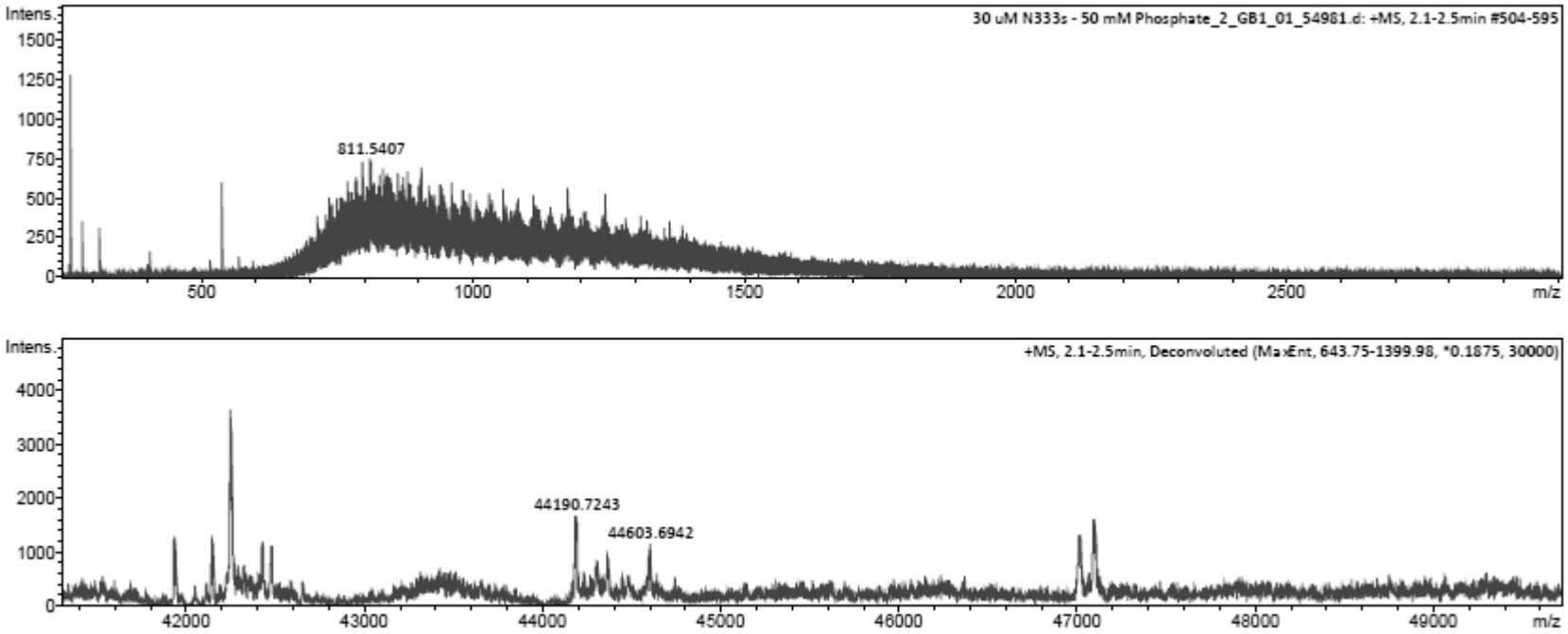
A4.9 Mass Spectrometry – His_6-MBP-A293

Expected protein mass = 44737.6 Da; observed protein mass = 44603.52 Da



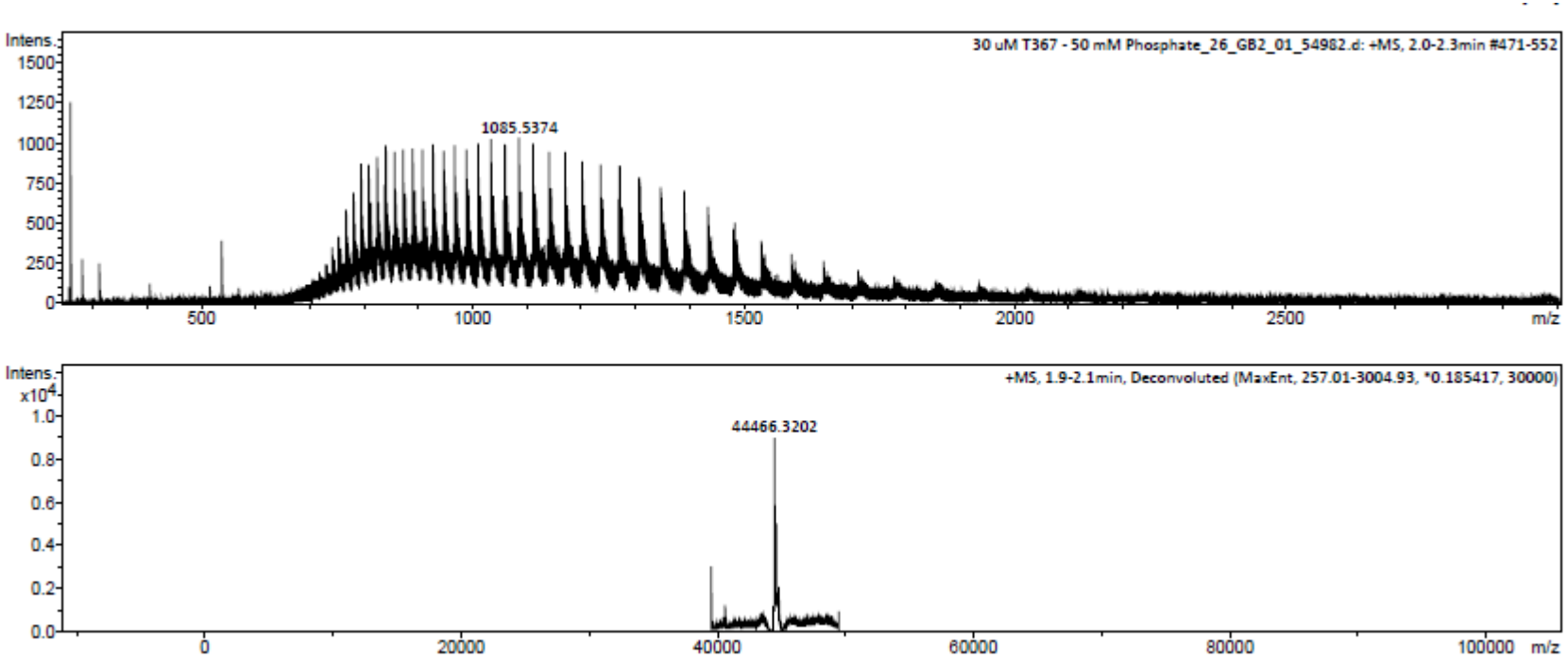
A4.10 Mass Spectrometry – His_6-MBP-N333

Expected protein mass = 44737.6 Da; observed protein mass = 44602.17 Da



A4.11 Mass Spectrometry – His_6-MBP-N333*

Expected protein mass = 44737.6 Da; observed protein mass = 44603.69 Da

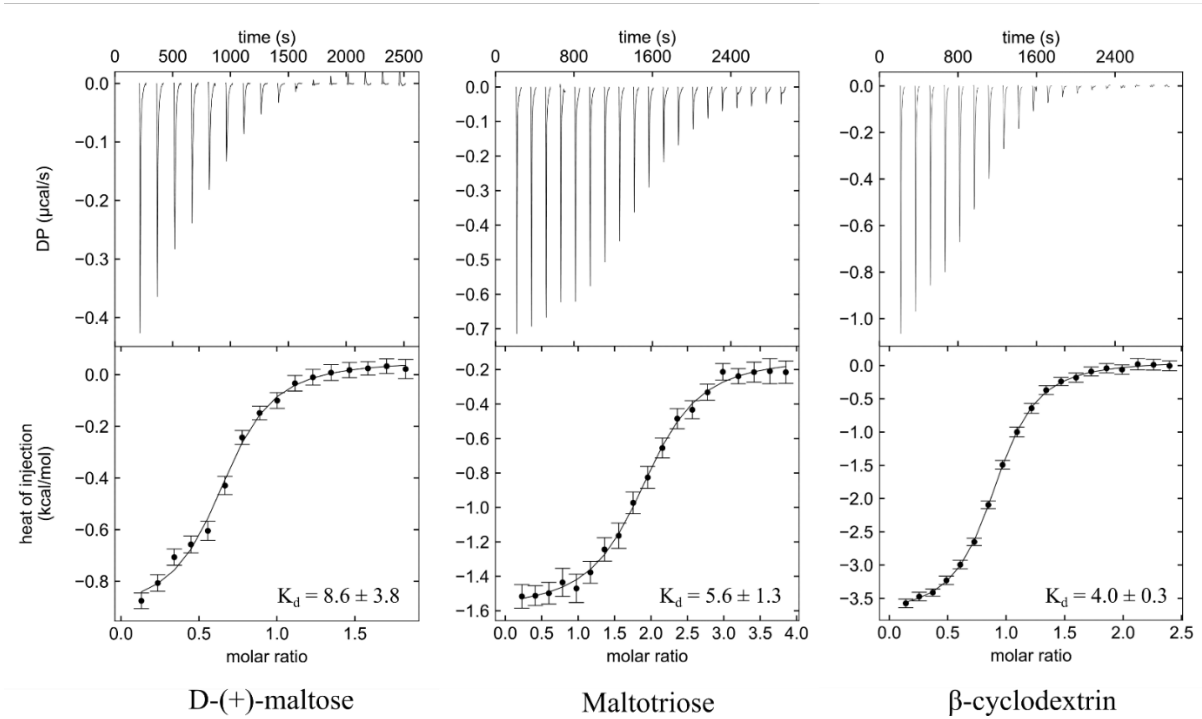


A4.12 Mass Spectrometry – His_6-MBP-T367

Expected protein mass = 44737.6 Da; observed protein mass = 44466.32 Da

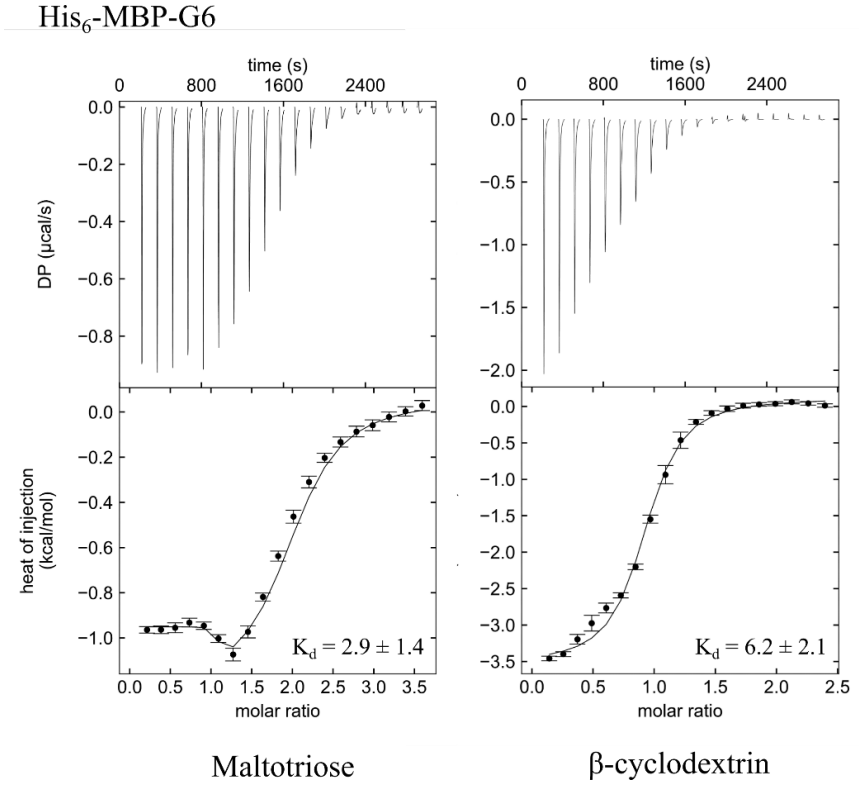
5 ITC Data

His_6-MBP (WT)



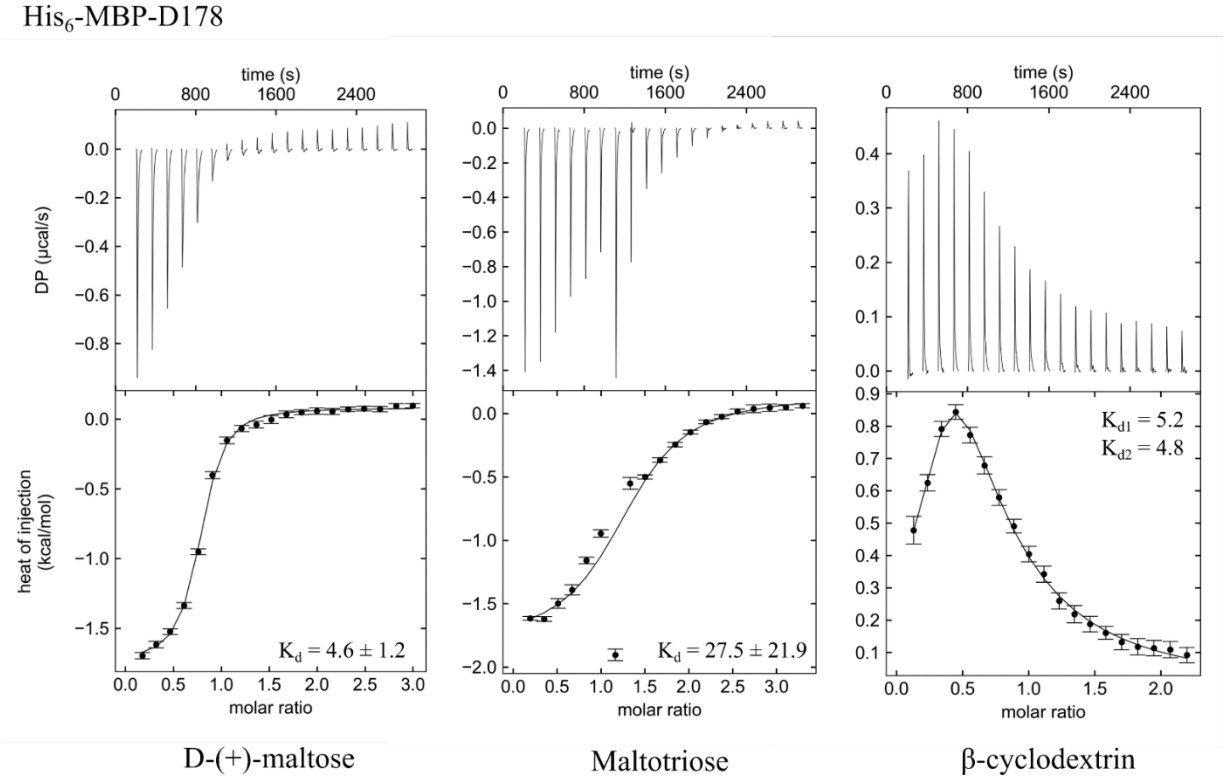
A5.1 ITC curves for His_6-MBP (WT) with sugar ligands

Concentrations: His_6-MBP (0.21 mM), D-(+)-maltose (3.0 mM), maltotriose (1.8 mM), β-cyclodextrin (1.2 mM)



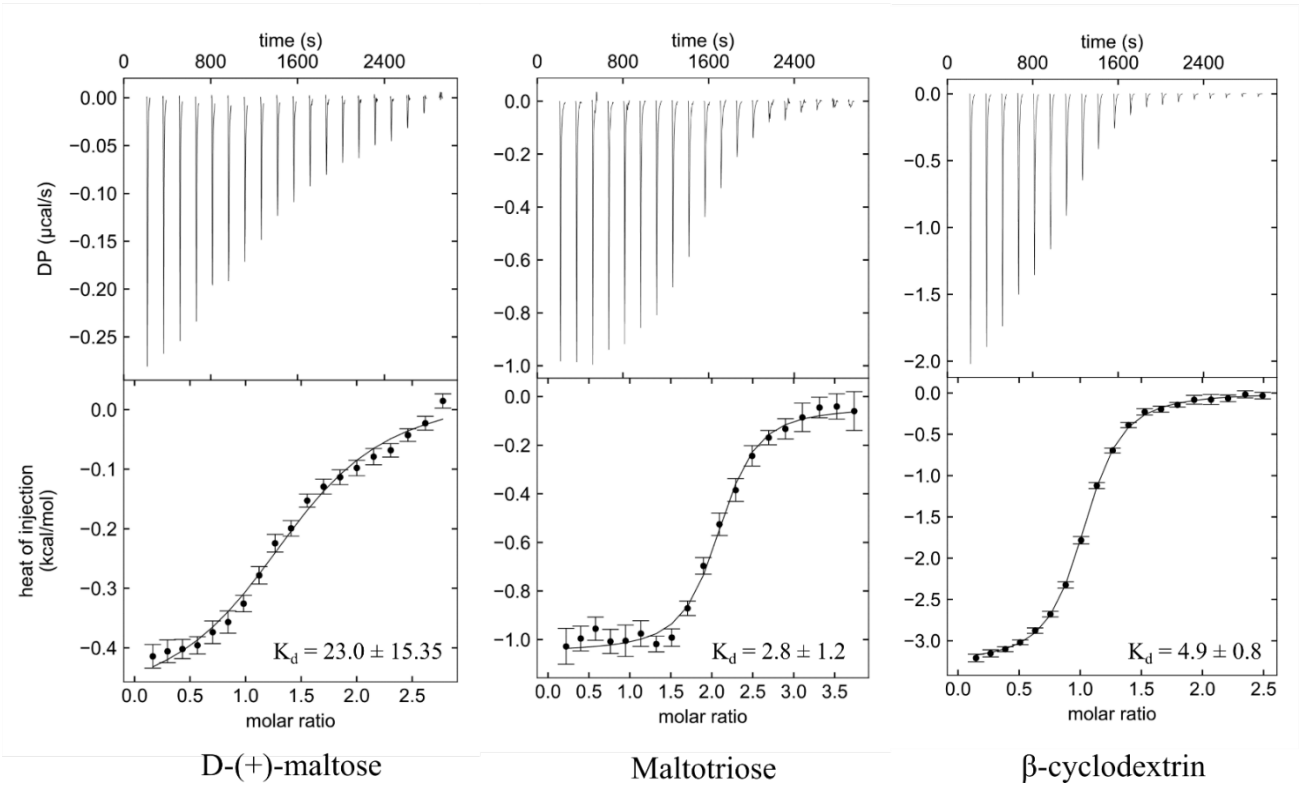
A5.1 ITC curves for His_6-MBP-G6 binding with sugar ligands

Concentrations: His_6-MBP-G6 (0.21 mM), maltotriose (3.6 mM), β-cyclodextrin (2.4 mM)



A5.3 ITC curves for His_6-MBP-D178 with sugar ligands

Concentrations: His_6-MBP-D178 (0.21 mM), D-(+)-maltose (3.0 mM), maltotriose (3.3 mM), β-cyclodextrin (2.2 mM)

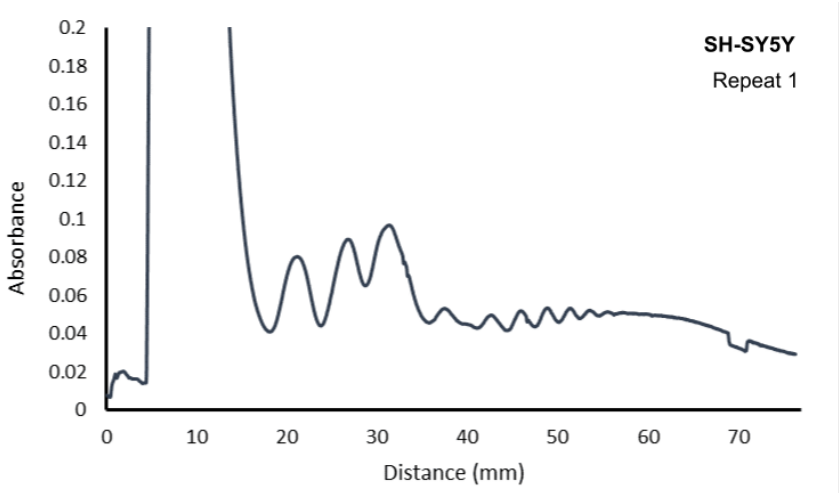


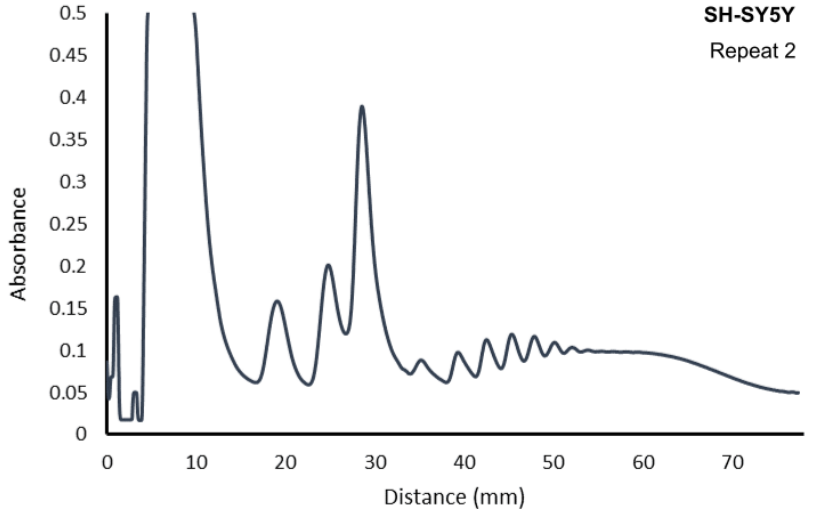
A5.4 ITC curves for His_6-MBP-T367 binding with sugar ligands

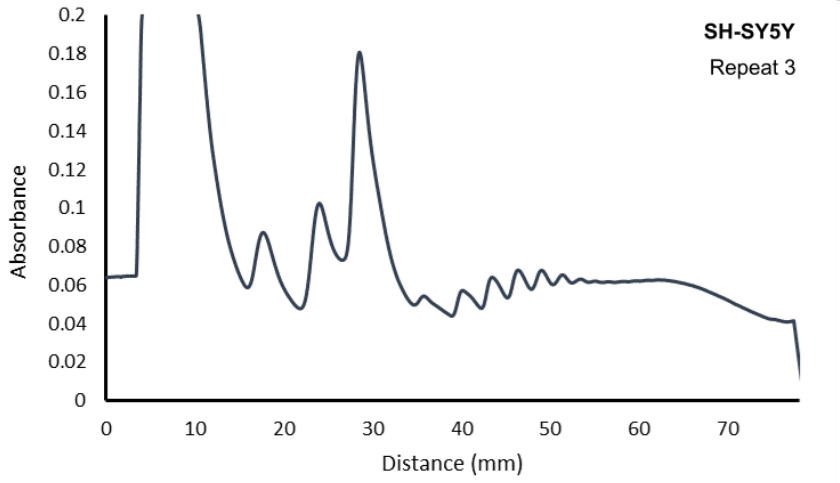
Concentrations: His_6-MBP-T367 (0.20 mM), D-(+)-maltose (2.3 mM), maltotriose (3.6 mM), β-cyclodextrin (2.4 mM)

6 Polysome profiles and Standard Curves

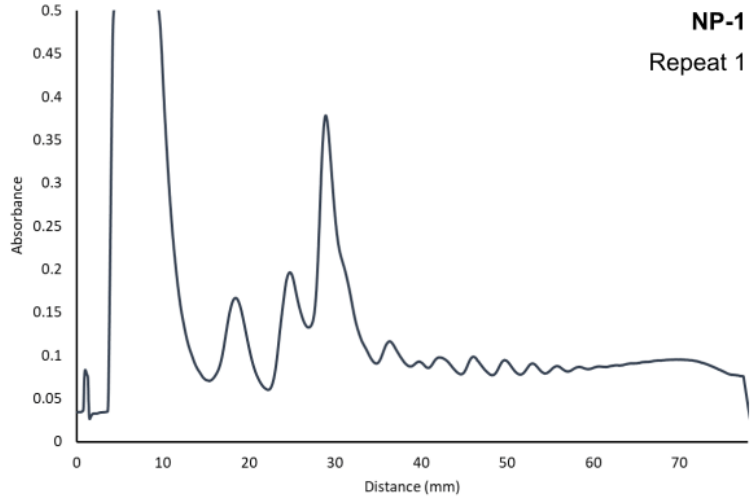
SH-SY5Y profiles

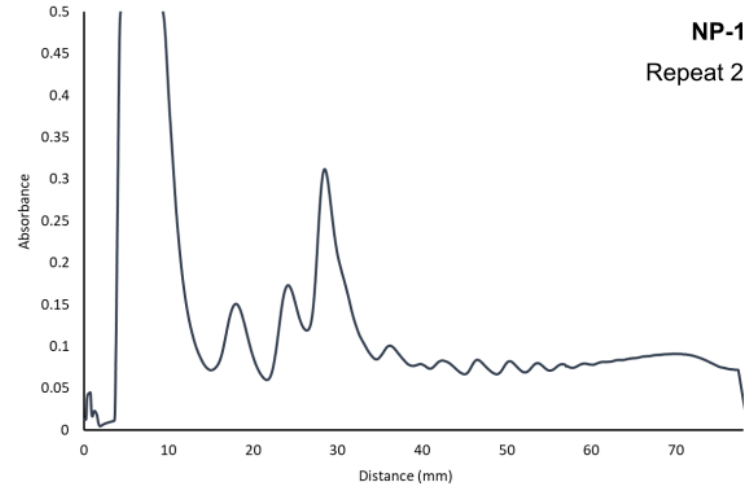


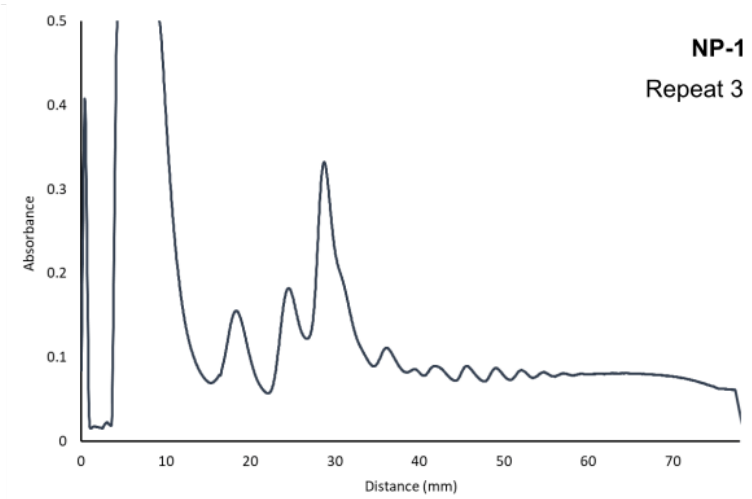




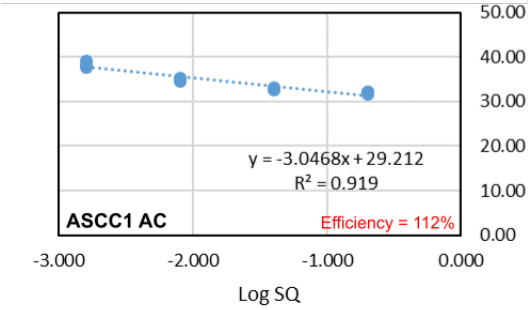
NP-1 profiles

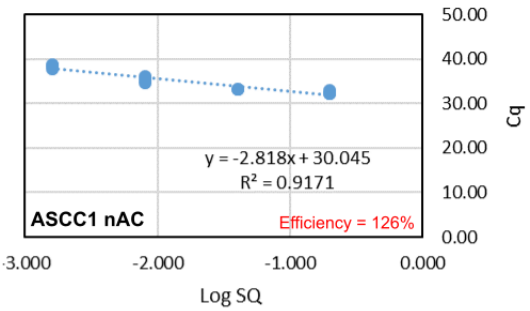


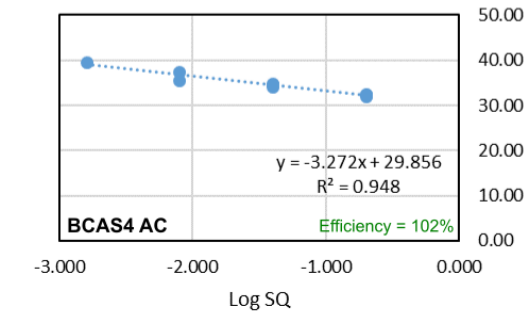


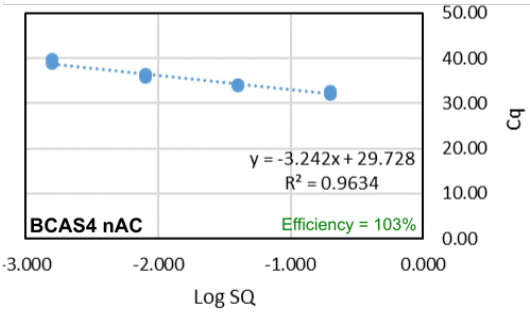


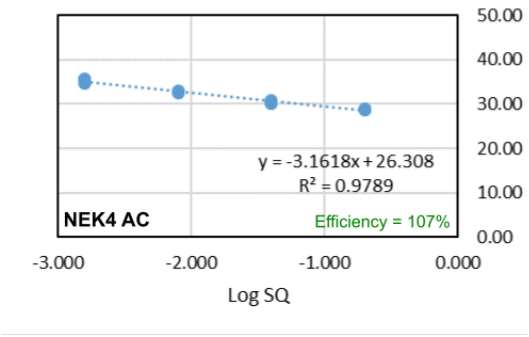
qPCR standard curves

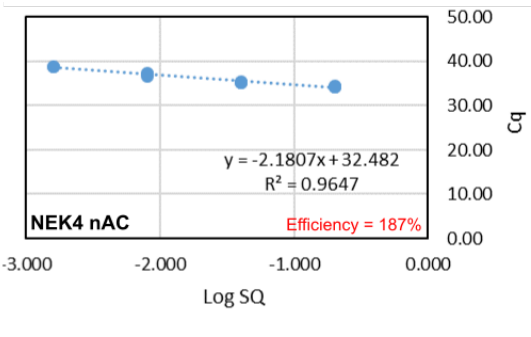












Bibliography

1. Smit, A. F. The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **6,** 743–748 (1996).
2. Smit, A. F. Interspersed repeats and other momentos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9,** 657–663 (1999).
3. Goodier, J. L. Restricting retrotransposons: a review. *Mob. DNA* **7,** 1–30 (2016).
4. Kazazian, H. H. *Mobile DNA: finding treasure in junk*. (FT Press, 2011).
5. Pace, J. K. & Feschotte, C. The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage. *Genome Res.* 422–432 (1995). doi:10.1101/gr.5826307
6. Muñoz-López, M. & García-Pérez, J. L. *DNA Transposons: Nature and Applications in Genomics*. *Current Genomics* **11,** (2010).
7. Biémont, C. A Brief History of the Status of Transposable Elements: From Junk DNA to Major Players in Evolution. *Genetics* **186,** 1085–1093 (2010).
8. Elbarbary, R. A., Lucas, B. A. & Maquat, L. E. Retrotransposons as regulators of gene expression. *Science (80-.).* **351,** 679–688 (2016).
9. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* **10,** 691–703 (2009).
10. Finnegan, D. J. Transposable elements: How non-LTR retrotransposons do it. *Curr. Biol.* **7,** 245–248 (1997).
11. Beauregard, A., Curcio, M. J. & Belfort, M. The take and give between retrotransposable elements and their hosts. *Annu Rev Genomics Hum Genet* **42,** 587–617 (2008).
12. Savage, A. L. *et al.* Retrotransposons in the development and progression of amyotrophic lateral sclerosis Neurodegeneration. *J Neurol Neurosurg Psychiatry* **90,** 284–293 (2019).
13. Sassaman, D. M. *et al.* Many human L1 elements are capable of retrotransposition. *Nat. Genet.* **16,** 37–43 (1997).
14. Flavell, A. J. Long terminal repeat retrotransposons jump between species. *PNAS* **96,** 12211–12212 (1999).
15. Havecker, E. R., Gao, X. & Voytas, D. F. The diversity of LTR

retrotransposons. *Genome Biol.* **5,** 1–6 (2004).

16. Han, J. S. Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. *Mob. DNA* **1,** 1–12 (2010).
17. Goodier, J. L. & Kazazian Jr, H. H. Retrotransposons Revisited: The Restraint and Rehabilitation of Parasites. *Cell* **135,** 23–35 (2008).
18. Feng, Y., Goubran, M. H., Follack, T. B. & Chelico, L. Deamination-independent restriction of LINE-1 retrotransposition by APOBEC3H. *Sci. Rep.* **7,** 1–11 (2017).
19. Lander, S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 861–921 (2001).
20. Khan, H., Smit, A. & Boissinot, S. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Res.* 78–87 (2006). doi:10.1101/gr.4001406
21. Deininger, P. Alu elements: know the SINEs. *Genome Biol.* **12,** 1–12 (2011).
22. Batzer, M. A. & Deininger, P. L. Alu Repeats and Human Genomic Diversity. *Nat. Rev. Genet.* **3,** 370–379 (2002).
23. Walter, P. & Blobel, G. Translocation of Proteins Across the Endoplasmic Reticulum III. Signal Recognition Protein (SRP) Causes Signal Sequence-dependent and Site-specific Arrest of Chain Elongation that is Released by Microsomal Membranes. *J. Cell Biol.* **91,** 557–561 (1981).
24. Ullu, E. & Tschudi, C. Alu sequences are processed 7SL RNA genes. *Nat. Lett.* **312,** 171–172 (1984).
25. Ullu, E., Murphy, S. & Melli, M. Human 7SL RNA Consists of a 140 Nucleotide Middle-Repetitive Sequence Inserted in an Ah Sequence. *Cell* **29,** 195–202 (1982).
26. Kramerov, D. A. & Vassetzky, N. S. Short retroposons in eukaryotic genomes. *Int. Rev. Cytol.* **247,** 165–221 (2005).
27. Quentin, Y. Origin of the Alu family: A family of Alu-like monomers gave birth to the left and the right arms of the Alu elements. *Nucleic Acids Res.* **20,** 3397–3401 (1992).
28. Quentin, Y. A master sequence related to a free left Alu monomer (FLAM) at the origin of the B1 family in rodent genomes. *Nucleic Acids Res.* **22,** 2222–2227 (1994).

29. Mighell, A. J., Markham, A. F. & Robinson, P. A. Alu sequences. *FEBS Lett.* **417,** 1–5 (1997).
30. Price, A. L., Eskin, E. & Pevzner, P. A. Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. *Genome Res.* **14,** 2245–2252 (2004).
31. Dagan, T., Sorek, R., Sharon, E., Ast, G. & Graur, D. AluGene: a database of Alu elements incorporated within protein-coding genes. *Nucleic Acids Res.* **32,** D489-92 (2004).
32. Wong, K. H. Y., Levy-Sakin, M. & Kwok, P.-Y. De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. *Nat. Commun.* 1–9 (2018). doi:10.1038/s41467-018-05513-w
33. Kim, D. D. Y. *et al.* Widespread RNA Editing of Embedded. *Genome Res.* 1719–1725 (2004). doi:10.1101/gr.2855504.has
34. Chen, L.-L. & Yang, L. ALUternative Regulation for Gene Expression. *Trends Cell Biol.* **27,** (2017).
35. Häsler, J. & Strub, K. Alu elements as regulators of gene expression. *Nucleic Acids Res.* **34,** 5491–5497 (2006).
36. Doucet, A. J., Wilusz, J. E., Miyoshi, T., Liu, Y. & Moran, J. V. A 3′ poly(A) tract is required for LINE-1 retrotransposition. *Mol. Cell* **60,** 728–741 (2015).
37. Roy-Engel, A. M. *et al.* Active Alu Element 'A-Tails': Size Does Matter. *Genome Res.* 1333–1344 (2002). doi:10.1101/gr.384802
38. Conti, A. *et al.* Identification of RNA polymerase III-transcribed Alu loci by computational screening of RNA-Seq data. *Nucleic Acids Res.* **43,** 817–835 (2014).
39. Ostertag, E. M., Goodier, J. L., Zhang, Y. & Kazazian, H. H. SVA Elements Are Nonautonomous Retrotransposons that Cause Disease in Humans. *Am. J. Hum. Genet* **73,** 1444–1451 (2003).
40. Boeke, J. D. LINEs and Alus - the poly A connection. *Nat. Genet.* **16,** 6–7 (1997).
41. Zhang, Z., Harrison, P. M., Liu, Y. & Gerstein, M. Millions of Years of Evolution Preserved: A Comprehensive Catalog of the Processed Pseudogenes in the Human Genome. *Genome Res.* 2541–2558 (2003). doi:10.1101/gr.1429003
42. Pray, L. A. Functions and Utility of Alu Jumping Genes. *Nat. Educ.* **1,** 93

(2008).

43. Ostertag, E. M. & Kazazian, H. H. Biology of Mammalian L1 Retrotransposons. *Annu. Rev. Genet.* **35,** 501–538 (2001).
44. Beck, C. R., Luis Garcia-Perez, J., Badge, R. M. & Moran, J. V. LINE-1 Elements in Structural Variation and Disease. *Annu Rev Genomics Hum Genet* **12,** 187–215 (2011).
45. Khazina, E. *et al.* Trimeric structure and flexibility of the L1ORF1 protein in human L1 retrotransposition. *Nat. Struct. Mol. Biol.* **18,** 1006–1015 (2011).
46. Feng, Q., Moran, J. V, Kazazian, H. H. & Boeke, J. D. Human L1 Retrotransposon Encodes a Conserved Endonuclease Required for Retrotransposition. *Cell* **87,** 905–916 (1996).
47. Mathias, S. L., Scott, A. F., Kazazian Jr, H. H., Boeke, J. D. & Gabriel, A. Reverse Transcriptase Encoded by a Human Transposable Element. *Science (80-.).* **254,** 1808–1810 (1991).
48. Moran, J. V *et al.* High Frequency Retrotransposition in Cultured Mammalian Cells. *Cell* **87,** 917–927 (1996).
49. Dewannieux, M., Esnault, C. & Heidmann, T. LINE-mediated retrotransposition of marked Alu sequences. *Nat. Genet.* **35,** 41–48 (2003).
50. Bovia, F., Wolff, N., Ryser, S. & Strub, K. The SRP9/14 subunit of the human signal recognition particle binds to a variety of Alu-like RNAs and with higher affinity than its mouse homolog. *Nucleic Acids Res.* **25,** 318–325 (1997).
51. Berger, A. *et al.* Direct binding of the Alu binding protein dimer SRP9/14 to 40S ribosomal subunits promotes stress granule formation and is regulated by Alu RNA. *Nucleic Acids Res.* **42,** 11203–11217 (2014).
52. Bennett, E. A. *et al.* Active Alu retrotransposons in the human genome. *Genome Res.* **18,** 1875–1883 (2008).
53. Christian, C. M., Deharo, D., Kines, K. J., Sokolowski, M. & Belancio, V. P. Identification of L1 ORF2p sequence important to retrotransposition using Bipartile Alu retrotransposition (BAR). *Nucleic Acids Res.* **44,** 4818–4834 (2016).
54. Jurka, J., Krnjajic, M., Kapitonov, V. V, Stenger, J. E. & Kokhanyy, O. Active Alu Elements Are Passed Primarily through Paternal Germlines. *Theor. Popul. Biol.* **61,** 519–530 (2002).

55. Jurka, J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 1872–1877 (1997).

56. Chen, W. & Zhang, L. The pattern of DNA cleavage intensity around indels. *Sci. Rep.* **5,** 8333 (2015).

57. Viollet, S., Monot, C. & Cristofari, G. L1 retrotransposition. *Mob. Genet. Elements* **4,** 1–6 (2014).

58. Cordaux, R., Hedges, D. J., Herke, S. W. & Batzer, M. A. Estimating the retrotransposition rate of human Alu elements. *Gene* **373,** 134–137 (2006).

59. Li, X. *et al.* Frequency of recent retrotransposition events in the human factor IX gene. *Hum. Mutat.* **17,** 511–519 (2001).

60. Brouha, B. *et al.* Hot L1s account for the bulk of retrotransposition in the human population. *PNAS* **100,** 5280–5285 (2003).

61. Kazazian Jr, H. H. An estimated frequency of endogeneous insertional mutations in humans. *Nat. Genet.* **22,** 130 (1999).

62. Morales, M. E. *et al.* The Contribution of Alu Elements to Mutagenic DNA Double-Strand Break Repair. *PLOS Genet.* **11,** 1–26 (2015).

63. Lieber, M. R. The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End Joining Pathway. *Annu Rev Biochem* **79,** 181–211 (2010).

64. Lieber, M. R., Ma, Y., Pannicke, U. & Schwarz, K. Mechanism and regulation of human non-homologous DNA end-joining. *Nat. Rev. Mol. Cell Biol.* **4,** 712–720 (2003).

65. Teixeira-Silva, A., Silva, R. M., Carneiro, J., Amorim, A. & Azevedo, L. The Role of Recombination in the Origin and Evolution of Alu Subfamilies. *PLoS One* **8,** (2013).

66. Han, K. *et al.* Alu recombination-mediated structural deletions in the chimpanzee genome. *PLoS Genet.* **3,** 1939–1949 (2007).

67. Takata, M. *et al.* Homologous recombination and non-homologous end-joining pathways of DNA double-strand break repair have overlapping roles in the maintenance of chromosomal integrity in vertebrate cells. *EMBO J.* **17,** 5497–5508 (1998).

68. Sen, S. K. *et al.* Human genomic deletions mediated by recombination between Alu elements. *Am. J. Hum. Genet.* **79,** 41–53 (2006).

69. Lengyel, P. & Söll, D. Mechanism of protein biosynthesis. *Bacteriol. Rev.* **33,** 264–301 (1969).
70. Hahn, S. Structure and mechanism of the RNA Polymerase II transcription machinery. *Nat Struct Mol Biol* **11,** 394–403 (2004).
71. Berg, J. M., Tymoczko, J. L. & Stryer, L. *Biochemistry*. (W H Freeman, 2002).
72. Alberts, B. *et al. Molecular Biology of the Cell*. (Garland Science, 2002).
73. Karp, G. *Cell and Molceular Biology*. (Wiley and Sons, 2008).
74. Revyakin, A. *et al.* Transcription initiation by human RNA polymerase II visualized at single-molecule resolution. *Genes Dev.* **26,** 1691–1702 (2012).
75. Carey, L. B. RNA polymerase errors cause splicing defects and can be regulated by differential expression of RNA polymerase subunits. *Elife* **4,** 1–10 (2015).
76. Rogers, J. & Wall, R. A mechanism for RNA splicing. *Proc. Natl. Acad. Sci. USA* **77,** 1877–1879 (1980).
77. Tatei, K., Takemura, K., Tanaka, H., Masaki, T. & Ohshima, Y. Recognition of 5' and 3' splice site sequences in pre-mRNA studied with a filter binding technique. *J. Biol. Chem.* **262,** 11667–11674 (1987).
78. Wu, Q. & Krainer, A. R. AT-AC Pre-mRNA Splicing Mechanisms and Conservation of Minor Introns in Voltage-Gated Ion Channel Genes. *Mol. Cell. Biol.* **19,** 3225–3236 (1999).
79. Burge, C. B., Padgett, R. A. & Sharp, P. A. Evolutionary fates and origins of U12-type introns. *Mol. Cell* **2,** 773–785 (1998).
80. Thanaraj, T. A. Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. *Nucleic Acids Res.* **29,** 2581–2593 (2001).
81. Zhang, X. *et al.* Branch point identification and sequence requirements for intron splicing in plasmodium falciparum. *Eukaryot. Cell* **10,** 1422–1428 (2011).
82. *Spinal Muscular Atrophy: Disease Mechanisms and Therapy*. (Academic Press, 2017).
83. Reed, R. & Maniatis, T. Intron sequences involved in lariat formation during pre-mRNA splicing. *Cell* **41,** 95–105 (1985).
84. Kiss, T. Biogenesis of small nuclear RNPs. *J. Cell Sci.* **117,** 5949–5951

(2004).

85. Wang, Z. & Burge, C. B. Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA* **14,** 802–813 (2008).
86. Hancks, D. C. & Kazazian, H. H. Active human retrotransposons: Variation and disease. *Curr. Opin. Genet. Dev.* **22,** 191–203 (2012).
87. Ule, J. Alu elements: at the crossroads between disease and evolution The role of transposable elements in gene regulation. *Biochem. Soc. Trans.* **41,** 1532–1535 (2013).
88. Larsen, P. A. *et al.* The Alu neurodegeneration hypothesis: A primate-specific mechanism for neuronal transcription noise, mitochondrial dysfunction, and manifestation of neurodegenerative disease. *Alzheimer's Dement.* **13,** 828–838 (2017).
89. Fitzpatrick, T. & Huang, S. 3'-UTR-located inverted Alu repeats facilitate mRNA translational repression and stress granule accumulation. *Nucleus* **3,** 359–369 (2012).
90. Hormozdiari, F. *et al.* Alu repeat discovery and characterization within human genomes. *Genome Res.* **21,** 840–849 (2011).
91. Zhang, X. O. *et al.* Complementary sequence-mediated exon circularization. *Cell* **159,** 134–147 (2014).
92. Nakama, M. *et al.* Intonic antisense Alu elements have a negative splicing effect on the inclusion of adjacent downstream exons. *Gene* **664,** 84–89 (2018).
93. Payer, L. M. *et al.* Alu insertion variants alter mRNA splicing. *Nucleic Acids Res.* **47,** 421–431 (2019).
94. Häsler, J., Samuelsson, T. & Strub, K. Useful 'junk': Alu RNAs in the human transcriptome. *Cell. Mol. Life Sci.* **64,** 1793–1800 (2007).
95. Häsler, J. & Strub, K. Alu RNP and Alu RNA regulate translation initiation in vitro. *Nucleic Acids Res.* **34,** 2374–2385 (2006).
96. Chu, W., Ballard, R., Carpick, B. W., Williams, B. R. G. & Schmid, C. W. Potential Alu Function : Regulation of the Activity of Double-Stranded RNA-Activated Kinase PKR. *Mol. Cell. Biol.* **18,** 58–68 (1998).
97. Rubin, C. M. Selective stimulation of translational expression by Alu RNA. *Nucleic Acids Res.* **30,** 3253–3261 (2002).
98. Salem, A. H., Kilroy, G. E., Watkins, W. S., Jorde, L. B. & Batzer, M. A.

Recently integrated Alu elements and human genomic diversity. *Mol. Biol. Evol.* **20,** 1349–1361 (2003).

99. Carroll, M. L. *et al.* Large-scale analysis of the Alu Ya5 and Yb8 subfamiles and their contribution to human genomic diversity. *J. Mol. Biol.* **311,** 17–40 (2001).
100. Batzer, M. A. & Deininger, P. L. A human-specific subfamily of Alu sequences. *Genomics* **9,** 481–487 (1991).
101. Batzer, M. A. *et al.* Amplification dynamics of human-specific (HS) alu family members. *Nucleic Acids Res.* **19,** 3619–3623 (1991).
102. Watkins, W. S. *et al.* Genetic variation among world populations: Inferences from 100 Alu insertion polymorphisms. *Genome Res.* **13,** 1607–1618 (2003).
103. Lin, L. *et al.* The contribution of Alu exons to the human proteome. *Genome Biol.* **17,** 1–14 (2016).
104. Sorek, R., Ast, G. & Graur, D. Alu-Containing Exons are Alternatively Spliced. *Genome Res.* **12,** 1060–1067 (2002).
105. Taft, R. J., Pang, K. C., Mercer, T. R., Dinger, M. & Mattick, J. S. Non-coding RNAs: regulators of disease. *J. Pathol.* **220,** 126–139 (2010).
106. Payer, L. M. *et al.* Structural variants caused by Alu insertions are associated with risks for many human diseases. *PNAS* 3984–3992 (2017). doi:10.1073/pnas.1704117114
107. Erez, A. *et al.* Alu-specific microhomology-mediated deletions in CDKL5 in females with early-onset seizure disorder. *Neurogenetics* **10,** 363–369 (2009).
108. Mustajoki, S., Ahola, H., Mustajoki, P. & Kauppinen, R. Insertion of Alu element responsible for acute intermittent porphyria. *Hum. Mutat.* **13,** 431–438 (1999).
109. Lei, H., Day, I. N. M., Vořechovsky, I. & Vořechovsky´*, V. Exonization of AluYa5 in the human ACE gene requires mutations in both 3' and 5' splice sites and is facilitated by a conserved splicing enhancer. *Nucleic Acids Res.* **33,** 3897–3906 (2005).
110. Kehoe, P. G. *et al.* Haplotypes extending across ACE are associated with Alzheimer's disease. *Hum. Mol. Genet.* **12,** 859–867 (2003).
111. Oldridge, M. *et al.* De Novo Alu-Element Insertions in FGFR2 Identify a Distinct Pathological Basis for Apert Syndrome. *Am. J. Hum. Genet.* **64,** 446–461 (2002).

112. Morrish, T. A. *et al.* Patterns of Transposable Element Expression and Insertion in. *Cancer. Front. Mol. Biosci* **3,** 76 (2016).
113. Lock, F. E. *et al.* A novel isoform of IL-33 revealed by screening for transposable element promoted genes in human colorectal cancer. *PLoS One* **12,** 1–30 (2017).
114. Makalowski, W. Not Junk After All. *Science (80-.).* **300,** 1246–1248 (2003).
115. Mitchell, G. A. *et al.* Splice-mediated insertion of an Alu sequence inactivates ornithine d-aminotransferase: A role for Alu elements in human mutation. *Proc. Natl. Acad. Sci. USA* **88,** 815–819 (1991).
116. Sorek, R., Ast, G. & Graur, D. Alu -Containing Exons are Alternatively Spliced. *Genome Res.* **12,** 1060–1067 (2002).
117. Lev-Maor, G., Sorek, R., Shomron, N. & Ast, G. The Birth of an Alternatively Spliced Exon: 3' Splice-Site Selection in Alu Exons. *Science (80-.).* **300,** 1288–1292 (2003).
118. Geer, L. Y. *et al.* The NCBI BioSystems database. *Nucleic Acids Res.* **38,** 492–496 (2009).
119. Gasteiger, E. *et al.* ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* **31,** 3784–1788 (2003).
120. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol* **215,** 403–410 (1990).
121. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).
122. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modelling, prediction and analysis. *Nat. Protoc.* **10,** 845 (2015).
123. Yang, J. & Zhang, Y. Protein Structure and Function Prediction Using I-TASSER. *Curr. Protoc. Bioinforma.* **52,** 1–24 (2016).
124. Hubley, R. *et al.* The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44,** 81–89 (2015).
125. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14,** 755–763 (1998).
126. Akerman, M. & Mandel-Gutfreund, Y. Alternative splicing regulation at tandem 3′ splice sites. *Nucleic Acids Res.* **34,** 23–31 (2006).
127. Uhlen, M. *et al.* A pathology atlas of the human cancer transcriptome. *Science*

(80-.). **357,** 1–11 (2017).

128. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science (80-.).* **356,** 1–12 (2017).
129. Uhlen, M. *et al.* Tissue-based map of the human proteome. *Science (80-.).* **347,** 1260419–1260419 (2015).
130. Kim, S., Kim, D., Cho, S., Kim, J. & Kim, J.-S. Highly Efficient RNA-guide genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res.* **128,** 1–32 (2014).
131. Ezkurdia, I. *et al.* Multiple evidence strands suggest that theremay be as few as 19 000 human protein-coding genes. *Hum. Mol. Genet.* **23,** 5866–5878 (2014).
132. Tiwawech, D. *et al.* Alu methylation in serum from patients with nasopharyngeal carcinoma. *Asian Pacific J. Cancer Prev.* **15,** 9797–9800 (2014).
133. Saeliw, T. *et al.* Integrated genome-wide Alu methylation and transcriptome profiling analyses reveal novel epigenetic regulatory networks associated with autism spectrum disorder. *Mol. Autism* **9,** 27 (2018).
134. Patchsung, M. *et al.* Alu siRNA to increase Alu element methylation and prevent DNA damage. *Epigenomics* **10,** 175–185 (2018).
135. Gerber, A., O'Connell, M. A. & Keller, W. Two forms of human double-stranded RNA-specific editase 1 (hRED1) generated by the insertion of an Alu cassette. *Rna* **3,** 453–463 (1997).
136. Gu, Y. *et al.* The first reported case of Menkes disease caused by an Alu insertion mutation. *Brain Dev.* **29,** 105–108 (2007).
137. Taylor, P. C., Clark, A. J., Marsh, A., Singer, D. R. J. & Dilly, S. J. A chemical genomics approach to identification of interactions between bioactive molecules and alternative reading frame proteins. *Chem. Commun. (Camb).* **49,** 9588–90 (2013).
138. Fukuyo, Y., Hunt, C. R. & Horikoshi, N. Geldanamycin and its anti-cancer activities. *Cancer Lett.* **290,** 24–35 (2010).
139. Supko, J. G., Hickman, R. L., Grever, M. R. & Malspeis, L. Preclinical pharmacologic evaluation of geldanamycin as an antitumor agent. *Cancer Chemother. Pharmacol.* **36,** 305–315 (1995).
140. Deboer, C., Meulman, P. A., Wnuk, R. J. & Peterson, D. H. Geldanamycin, a

new antibiotic. *J. Antibiot. (Tokyo)*. 442–447 (1970).

141. Rinehart Jr., K. L. & Shield, L. S. Chemistry of ansamycin antiobiotcs. *Prog. Chem. Org. Natur. Prod.* **33,** 231–307 (1976).
142. Rascher, A. *et al.* Cloning and characterization of a gene cluster for geldanamycin production in Streptomyces hygroscopicus NRRL 3602. *FEMS Microbiol. Lett.* **218,** 223–230 (2003).
143. He, W., Wu, L., Gao, Q., Du, Y. & Wang, Y. Identification of AHBA Biosynthetic Genes Related to Geldanamycin Biosynthesis in Streptomyces hygroscopicus 17997. *Curr. Microbiol.* **52,** 197–203 (2006).
144. Heisey, R. M. & Putnam, A. R. Herbicidal effects of geldanamycin and nigericin, antibiotics from Streptomyces hygroscopicus. *J. Nat. Prod.* **49,** 859–865 (1986).
145. Blagosklonny, M. V. Hsp-90-associated oncoproteins: multiple targets of geldanamycin and its analogs. *Leukemia* **16,** 455–462 (2002).
146. Franke, J., Eichner, S., Zeilinger, C. & Kirschning, A. Targeting heat-shock-protein 90 (Hsp90) by natural products: geldanamycin, a show case in cancer therapy. *Nat. Prod. Rep.* **30,** 1299 (2013).
147. Maloney, A. & Workman, P. HSP90 as a new therapeutic target for cancer therapy: the story unfolds. *Expert Opin. Biol. Ther.* **2,** 3–24 (2005).
148. Solit, D. B. *et al.* Phase I T rial of 17-Allylamino-17-Demethoxygeldanamycin in Patients with Advanced Cancer. *Cancer Ther. Clin.* **13,** 1775–1783 (2007).
149. Schulte, T. W. The benzoquinone ansamycin 17-allylamino-17-demethoxygeldanamycin binds to HSP90 and shares important biologic activities with geldanamycin. *Cancer Chemother. Pharmacol.* **42,** 273–279 (1998).
150. Hartmann, F. *et al.* Effects of the tyrosine-kinase inhibitor geldanamycin on ligand-induces Her-2/Neu activitation, receptor expression and proliferation of Her-2 positive malignant cell lines. *Int. J. Cancer* **229,** 221–229 (1997).
151. Whitesell, L., Sutphin, P. D., Pulcini, E. J., Martinez, J. D. & Cook, P. H. The Physical Association of Multiple Molecular Chaperone Proteins with Mutant p53 Is Altered by Geldanamycin , an hsp90-Binding Agent. *Mol. Cell. Biol.* **18,** 1517–1524 (1998).
152. Jogula, S. *et al.* European Journal of Medicinal Chemistry Geldanamycin-

inspired compounds induce direct trans- differentiation of human mesenchymal stem cells to neurons. *Eur. J. Med. Chem.* **135,** 110–116 (2017).

153. Neckers, L., Schulte, T. W. & Mimnaugh, E. Geldanamycin as a potential anti-cancer agent: Its molecular target and biochemical activity. *Invest. New Drugs* **17,** 361–373 (2000).
154. Riggs, D. L. *et al.* Functional Specificity of Co-Chaperone Interactions with Hsp90 Client Proteins. *Crit. Rev. Biochem. Mol. Biol.* **39,** 279–295 (2004).
155. Kimura, T., Uesugi, M., Takase, K., Miyamoto, N. & Sawada, K. Hsp90 inhibitor geldanamycin attenuates the cytotoxicity of sunitinib in cardiomyocytes via inhibition of the autophagy pathway. *Toxicol. Appl. Pharmacol.* **329,** 282–292 (2017).
156. Citri, A. *et al.* Hsp90 Recognizes a Common Surface on Client Kinases. *J. Biol. Chem.* **281,** 14361–14369 (2006).
157. Roe, S. M. *et al.* Structural Basis for Inhibition of the Hsp90 Molecular Chaperone by the Antitumor Antibiotics Radicicol and Geldanamycin. *J. Med. Chem.* **42,** 260–266 (1999).
158. Whitesell, L., Mimnaugh, E. G., Costat, B. D. E., Myers, C. E. & Neckers, L. M. Inhibition of heat shock protein HSP90-pp6Ov-src heteroprotein complex formation by benzoquinone ansamycins: Essential role for stress proteins in oncogenic transformation. *Proc. Natl. Acad. Sci. U. S. A.* **91,** 8324–8328 (1994).
159. Pick, E. *et al.* High HSP90 Expression Is Associated with Decreased Survival in Breast Cancer. *J. Cancer Res.* **67,** 2932–2938 (2007).
160. Chatterjee, S. & Burns, T. F. Targeting Heat Shock Proteins in Cancer : A Promising Therapeutic Approach. *Int. J. Mol. Sci.* **18,** 1–39 (2017).
161. Stebbins, C. E. *et al.* Crystal structure of an Hsp90-geldanamycin complex: Targeting of a protein chaperone by an antitumor agent. *Cell* **89,** 239–250 (1997).
162. Solit, D. B., Scher, H. I. & Rosen, N. Hsp90 as a therapeutic target in prostate cancer. *Semin. Oncol.* **30,** 709–716 (2003).
163. Clark, C. B. *et al.* Role of oxidative stress in Geldanaycin-induced cytotoxicity and disruption of Hsp90 signaling complex. *Free Radic Biol Med.* **47,** 1440–1449 (2009).
164. Samuni, Y. *et al.* Free Radical Biology & Medicine Reactive oxygen species

mediate hepatotoxicity induced by the Hsp90 inhibitor geldanamycin and its analogs. *Free Radic. Biol. Med.* **48,** 1559–1563 (2010).

165. Ladwa, S. R., Dilly, S. J., Clark, A. J., Marsh, A. & Taylor, P. C. Rapid Identification of a Putative Interaction between b 2 -Adrenoreceptor Agonists and ATF4 using a Chemical Genomics Approach. *Chem. Med. Chem.* **3,** 742–744 (2008).
166. Bazan, J., Całkosiński, I. & Gamian, A. Phage display—A powerful technique for immunotherapy. *Hum. Vaccin. Immunother.* **8,** 1817–1828 (2012).
167. Rami, A., Behdani, M., Yardehnavi, N., Habibi-Anbouhi, M. & Kazemi-Lomedasht, F. An overview on application of phage display technique in immunological studies. *Asian Pacific Journal of Tropical Biomedicine* **7,** 599–602 (2017).
168. Danner, S. & Belasco, J. G. T7 phage display: A novel genetic selection system for cloning RNA-binding proteins from cDNA libraries. *Proc. Natl. Acad. Sci.* **98,** 12954–12959 (2001).
169. Dilly, S. J. *et al.* A chemical genomics approach to drug reprofiling in oncology: Antipsychotic drug risperidone as a potential adenocarcinoma treatment. *Cancer Lett.* **393,** 16–21 (2017).
170. Smith, G. P. & Petrenko, V. A. Phage display. *Chem. Rev.* **97,** 391–410 (1997).
171. Scientific, T. F. *The Molecular Probes Handbook*. (2010).
172. Lea, W. A. & Simeonov, A. Fluorescence Polarization Assays in Small Molecule Screening. *Expert Opin. Drug Discov.* **6,** 17–32 (2011).
173. Berezin, M. Y. & Achilefu, S. Fluorescence Lifetime Measurments and Biological Imaging. *Chem. Rev.* **110,** 2641–2684 (2010).
174. Owicki, J. C. Fluorescence Polarization and Anisotropy in High-Throughput Screening: Perspectives and Primer. *J. Biomol. Screen.* **5,** 297–306 (2000).
175. Haas, J. A. & Fox, B. G. Fluorescence Anisotropy Studies of Enzyme-Substrate Complex Formation in Stearoyl-ACP Desaturase. *Biochemistry* **41,** 14472–14481 (2002).
176. Lakowicz, J. R. *Principles of Fluorescence Anisotropy*. (Springer, 2006).
177. Kinosita Jr, K., Kawato, S. & Ikegami, A. A theory of fluorescence polarization decay in membranes. *Biophys. J.* **20,** 289–305 (1977).

178. Jameson, D. M. & Ross, J. A. Fluorescence Polarisation/Anisotropy in Diagnostics and Imaging. *Chem. Rev.* **3,** 226–236 (2011).
179. Valeur, B. & Berberan-Santos, M. N. Principles and Applications. in *Molecular Fluorescence* 132 (2012).
180. Schröder, G. F., Alexiev, U. & Grubmüller, H. Simulation of fluorescence anisotropy experiments: Probing protein dynamics. *Biophys. J.* **89,** 3757–3770 (2005).
181. Jameson, D. M. & Sawyer, Wi. H. Fluorescence anisotropy applied to biomolecular interactions. *Methods Enzymol.* **246,** 283–300 (1995).
182. Burghardt, T. P. Fluorescence depolarization by anisotropic rotational diffusion of a luminophore and its carrier molecule. *J. Chem. Phys.* **78,** 5913–5919 (1983).
183. Heyduk, T., Ma, Y. & Tang, H. Fluorescence Anisotropy: Rapid, Quantitative Assay for Protein-DNA and Protein-Protein Interaction. *Methods Enzymol.* **274,** 492–503 (1996).
184. Weber, P. C., Ohlendorf D, H., Wendoloski, J. J. & Salemme, F. R. Structural Origins of High-Affinity Biotin Binding to Streptavidin. *Sci. Reports* **243,** 85–88 (1988).
185. Lundström, I. Real-time biospecific interaction analysis. *Biosens. Bioelectron.* **9,** 725–736 (1994).
186. Green, R. J. *et al.* Surface plasmon resonance analysis of dynamic biological interactions with biomaterials. *Biomaterials* **21,** 1823–1835 (2000).
187. Schasfoort, R. B. & Tudos, A. J. Introduction to Surface Plasmon Resonance. in *Handbook of Surface Plasmon Resonance* 1–14 (2008).
188. Merwe, P. A. V. Der. Surface plasmon resonance in Protein-Ligand Interactions: hydrodynamics and calorimetry. *Protein-Ligand Interact. ...* 137–170 (2001).
189. Pattnaik, P. Surface plasmon resonance: applications in understanding receptor-ligand interaction. *Appl. Biochem. Biotechnol.* **126,** 79–92 (2005).
190. De Mol, N. J. & Fischer, M. J. E. How SPR Developed into a Biomolecular Interaction Analysis Tool. in *Surface Plasmon Resonance: Methods and Protocols* 3–4 (2012). doi:10.1007/978-1-60761-670-2
191. MacKenzie, C. R. *et al.* Analysis by surface plasmon resonance of the influence of valence on the ligand binding affinity and kinetics of an anti-

carbohydrate antibody. *J. Biol. Chem.* **271,** 1527–1533 (1996).

192. Müller, K. M., Arndt, K. M. & Plückthun, A. Model and simulation of multivalent binding to fixed ligands. *Anal. Biochem.* **261,** 149–158 (1998).
193. Freire, E., Mayorga, O. L. & Straume, M. Isothermal Titration Calorimetry. *Anal. Chem.* **62,** 950–959 (1990).
194. Leavitt, S. & Freire, E. Direct measurement of protein binding energetics by isothermal titration calorimetry. *Curr. Opin. Struct. Biol.* **11,** 560–566 (2001).
195. Freyer, M. W. & Lewis, E. A. Isothermal Titration Calorimetry: Experimental Design, Data Analysis, and Probing Macromolecule/Ligand Binding and Kinetic Interactions. *Methods Cell Biol.* **84,** 79–113 (2008).
196. Velazquez-Campoy, A., Ohtaka, H., Nezami, A., Muzammil, S. & Freire, E. Isothermal Titration Calorimetry. *Curr. Protoc. Cell Biol.* **17,** 1–24 (2004).
197. Keller, Sandro; Vargas, Carolyn; Zhao, Huaying; Piszchek, Grzegorz; Brautigam, Chad; A. Schuck, P. High-Precision Isothermal Titration Calorimetry with Automated Peak Shape Analysis. *Anal. Chem.* **84,** 5066–5073 (2012).
198. Zhao, H., Piszczek, G. & Schuck, P. SEDPHAT - a platform for global ITC analysis and global multi-method analsysis of molecular interactions. *Methods* **76,** 137–148 (2015).
199. Betts, M. J. & Russell, R. B. Amino Acid Properties and Consequence of Substitutions. in *Bioinformatics for Geneticists* 289–316 (2003). doi:10.1103/PhysRevA.93.053607
200. Makalowski, W., Mitchell, G. A. & Labuda, D. Alu sequences in the coding regions of mRNA: a source of protein variability. *Trends Genet.* **10,** 188–193 (1994).
201. Keren, H., Lev-Maor, G. & Ast, G. Alternative splicing and evolution: diversification, exon definition and function. *Nat. Rev. Genet.* **11,** 345–355 (2010).
202. Georgiou, G. & Valax, P. Expresision of correctly folder protein in Escherichia coli. *Curr. Opin. Biotechnol.* **7,** 190–197 (1996).
203. Baolei, J. & Ok Jeon, C. High-throughput recombinant protein expression in Escherichia coli: current status and future perspectives. *Open Biol.* **6,** 1–17 (2016).
204. Steinmetz, E. J. & Auldridge, M. E. Screening Fusion Tags for Improved

Recombinant Protein Expression in E. coli with the Expresso® Solubility and Expression Screening System. *Curr. Protoc. protein Sci.* **90,** 5.27.1-5.27.20 (2017).

205. Chen, D. & Texada, D. E. Low-usage codons and rare codons of Escherichia coli. *Gene Ther. Mol. Biol.* **10,** 1–12 (2006).
206. Rosano, G. L. & Ceccarelli, E. A. Recombinant protein expression in Escherichia coli: Advances and challenges. *Front. Microbiol.* **5,** 1–17 (2014).
207. Hartl, F. U. & Hayer-Hartl, M. Protein folding. Molecular chaperones in the cytosol: From nascent chain to folded protein. *Science (80-.).* **295,** 1852–1858 (2002).
208. Brown, C. W. *et al.* Large-scale analysis of post-translational modifications in E.coli under glucose-limiting conditions. *BMC Genomics* **18,** 1–21 (2017).
209. Bärlund, M. *et al.* Cloning of BCAS3 (17q23) and BCAS4 (20q13) Genes That Undergo Amplification , Overexpression , and. *Genes Chromosom. Cancer* **35,** 311–317 (2002).
210. Nguyen, C. L. *et al.* Nek4 regulates entry into replicative senescence and the response to DNA damage in human fibroblasts. *Mol. Cell. Biol.* **32,** 3963–77 (2012).
211. Orloff, M. *et al.* Germline mutations in MSR1, ASCC1, and CTHRC1 in patients with Barrett esophagus and esophageal adenocarcinoma. *JAMA* **306,** 410–419 (2011).
212. Lai, Y. *et al.* Downregulation of long noncoding RNA ZMAT1 transcript variant 2 predicts a poor prognosis in patients with gastric cancer. *Int J Clin Exp Pathol* **8,** 5556–5562 (2015).
213. Basei, F. L. *et al.* New interaction partners for Nek4.1 and Nek4.2 isoforms: from the DNA damage response to RNA splicing. *Proteome Sci.* **13,** 1–13 (2015).
214. Tan, R. *et al.* Nek7 Protects Telomeres from Oxidative DNA Damage by Phosphorylation and Stabilization of TRF1. *Mol. Cell* **65,** 818–831 (2017).
215. Zhao, D. & Huang, Z. Effect of His-Tag on Expression, Purification, and Structure of Zinc Finger Protein, ZNF191 (243-368). *Bioinorg. Chem. Appl.* 1–6 (2016). doi:10.1155/2016/8206854
216. Vorráčková, I., Suchanová, Š., Ulbrich, P., Diehl, W. E. & Ruml, T. Purification of proteins containing zinc finger domains using Immobilized

Metal Ion Affinity Chromatography. *Protein Expr. Purif.* **79,** 88–95 (2011).

217. Ota, T. *et al.* Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat. Genet.* **36,** 40–45 (2004).
218. Zeng, F. *et al.* A restriction-free method for gene reconstitution using two single-primer PCRs in parallel to generate compatible cohesive ends. *BMC Biotechnol.* **17,** 1–7 (2017).
219. Van Den Ent, F. & Löwe, J. RF cloning: A restriction-free method for inserting target genes into plasmids. *J. Biochem. Biophys. Methods* **67,** 67–74 (2006).
220. Bornhorst, J. A. & Falke, J. J. Purificaiton of Proteins Using Polyhistidine Affinity. *Methods Enzymol.* **326,** 245–254 (2000).
221. Hilgarth, R. S. *et al.* Regulation and function of SUMO modification. *J. Biol. Chem.* **279,** 53899–53902 (2004).
222. Johnson, E. S. Protein modification by SUMO. *Annu. Rev. Biochem.* **73,** 355–382 (2004).
223. Peroutka III, R. J., Orcutt, S. J., Strickler, J. E. & Butt, T. R. SUMO Fusion Technology for Enhanced Protein Expression and Purification in Prokaryotes and Eukaryotes. in *Heterologous Gene Expression in E.coli: Methods and Protocols* (eds. Evans Jr., T. C. & Xu, M.-Q.) 15–30 (Humana Press, 2011). doi:10.1007/978-1-61737-967-3_2
224. Hickey, C. M., Wilson, N. R. & Hochstrasser, M. Function and Regulation of SUMO Proteases. *Nat. Rev. Mol. Cell Biol.* **13,** 755–766 (2012).
225. Harper, S. & Speicher, D. W. Purification of proteins fused to glutathione S-transferase. *Methods Mol. Biol.* **681,** 259–280 (2011).
226. O'Regan, L., Blot, J. & Fry, A. M. Mitotic regulation by NIMA-related kinases. *Cell Div.* **2,** 1–12 (2007).
227. O'Connell, M. J., Krien, M. J. E. & Hunter, T. Never say never. The NIMA-related protein kinases in mitotic control. *Trends Cell Biol.* **13,** 221–228 (2003).
228. Levedakou, E. N. *et al.* Two novel human serine/threonine kinases with homologies to the cell cycle regulating Xenopus MO15, and NIMA kinases: cloning and characterization of their expression pattern. *Oncogene* **9,** 1977–88 (1994).
229. Khow, O. & Suntrarachun, S. Strategies for production of active eukaryotic

proteins in bacterial expression system. *Asian Pac. J. Trop. Biomed.* **2,** 159–162 (2012).

230. Jensen, L. J. *et al.* STRING 8 - A global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* **37,** 412–416 (2009).
231. Laity, J. H., Lee, B. M. & Wright, P. E. Zinc finger proteins: New insights into structural and functional diversity. *Curr. Opin. Struct. Biol.* **11,** 39–46 (2001).
232. Li, T., Mo, X., Fu, L., Xiao, B. & Guo, J. Molecular mechanisms of long noncoding RNAs on gastric cancer. *Oncotarget* **7,** 8601–8612 (2016).
233. Rink, L. *et al.* Gene expression signatures and response to imatinib mesylate in gastrointestinal stromal tumor. *Mol. Cancer Ther.* **8,** 2172–2182 (2009).
234. Hahn, Y. *et al.* Finding fusion genes resulting from chromosome rearrangement by analyzing the expressed sequence databases. *Proc. Natl. Acad. Sci.* **101,** 13257–13261 (2004).
235. Edgren, H. *et al.* Identification of fusion genes in breast cancer by pair-end RNA-sequencing. *Genome Biol.* **12,** 1–13 (2011).
236. Torices, S. *et al.* A Truncated Variant of ASCC1, a Novel Inhibitor of NF- B, Is Associated with Disease Severity in Patients with Rheumatoid Arthritis. *J. Immunol.* **195,** 5415–5420 (2015).
237. Xia, Z. Bin *et al.* Inhibition of NF-κB signaling pathway induces apoptosis and suppresses proliferation and angiogenesis of human fibroblast-like synovial cells in rheumatoid arthritis. *Med. (United States)* **97,** 1–7 (2018).
238. Oliveira, J., Martins, M., Pinto Leite, R., Sousa, M. & Santos, R. The new neuromuscular disease related with defects in the ASC-1 complex: report of a second case confirms ASCC1 involvement. *Clin. Genet.* **92,** 434–439 (2017).
239. Kellerman, O. K. & Ferenci, T. Maltose-Binidng Protein from Escherichia coli. *Methods Enzymol.* **90,** 459–463 (1982).
240. Kapust, R. B. & Waugh, D. S. *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci.* **8,** 1668–1674 (1999).
241. Bertz, M. & Rief, M. Ligand Binding Mechanics of Maltose Binding Protein. *J. Mol. Biol.* **393,** 1097–1105 (2009).
242. Quiocho, F. A., Spurlino, J. C. & Rodseth, L. E. Extensive features of tight

oligosaccharide binding revealed in high-resolution structures of the maltodextrin transport/chemosensory receptor. *Structure* **5,** 997–1015 (1997).

243. Hemsley, A. *et al.* A simple method for site-directed mutagenesis using the polymerase chain reaction. *Nucleic Acids Res.* **17,** 6545–6551 (1989).
244. Edelheit, O., Hanukoglu, A. & Hanukoglu, I. Simple and efficient site-directed mutagenesis using two single-primer reactions in parallel to generate mutants for protein structure-function studies. *BMC Biotechnol.* **9,** 1–8 (2009).
245. Provencher, S. W. & Glöckner, J. Estimation of Globular Protein Secondary Structure from Circular Dichroism. *Biochemistry* **20,** 33–37 (1981).
246. Greenfield, N. J. Using circular dichroism spectra to estimate protein secondary structure. *Nat. Protoc.* **1,** 2876–2890 (2007).
247. Kelly, S. & Price, N. The Use of Circular Dichroism in the Investigation of Protein Structure and Function. *Curr. Protein Pept. Sci.* **1,** 349–384 (2000).
248. Chen, Y.-H., Yang, J. T. & Chau, K. H. Determination of the helix and β form of proteins in aqueous solution by circular dichroism. *Biochemistry* **13,** 3350–3359 (1974).
249. Brahms, S. & Brahms, J. Determination of protein secondary structure in solution by vacuum ultraviolet circular dichroism. *J. Mol. Biol.* **138,** 149–178 (1980).
250. Curtis Johnson Jr, W. Protein secondary structure and circular dichroism: A practical guide. *Proteins Struct. Funct. Bioinforma.* **7,** 205–214 (1990).
251. Barrow, C. J., Yasuda, A., Kenny, P. T. M. & Zagorski, M. G. Solution conformations and aggregational properties of synthetic amyloid β-peptides of Alzheimer's disease. *J. Mol. Biol.* **225,** 1075–1093 (2004).
252. Gahn, L. G. & Roskoski Jr, R. Thermal Stability and CD Analysis of Rat Tyrosine Hydroxylase. *Biochemistry* **34,** 252–256 (1995).
253. Greenfield, N. & Fasman, G. D. Computed Circular Dichroism Spectra for the Evaluation of Protein Conformation. *Biochemistry* **8,** 4108–4116 (1969).
254. Sturtevant, J. M. The thermodynamic effects of protein mutations. *Curr. Opin. Struct. Biol.* **4,** 69–78 (1994).
255. Weber, P. C. & Salemme, F. R. Applications of calorimetric methods to drug discovery and the study of protein interactions. *Curr. Opin. Struct. Biol.* **13,** 115–121 (2003).

256. Minetti, C. A. S. A. & Remeta, D. P. Energetics of membrane protein folding and stability. *Arch. Biochem. Biophys.* **453,** 32–53 (2006).

257. Lukas, K. & LeMaire, P. K. Differential scanning calorimetry: Fundamental overview. *Resonance* **14,** 807–817 (2009).

258. Aggarwal, V. *et al.* Supplementary Information Ligand modulated parallel mechanical unfolding pathways of Maltose Binding Proteins (MBPs). *J. Biol. Chem.* 1–20 (2011).

259. Beena, K., Udgaonkar, J. B. & Varadarajan, R. Effect of Signal Peptide on the Stability and Folding Kinetics of Maltose Binding Protein. *Biochemistry* **43,** 3608–3619 (2004).

260. Booth, W. T. *et al.* Impact of an N-terminal polyhistidine tag on protein thermal stability. *ACS Omega* **3,** 760–768 (2018).

261. Cattoli, F., Boi, C., Sorci, M. & Sarti, G. C. Adsorption of pure recombinant MBP-fusion proteins on amylose affinity membranes. *J. Memb. Sci.* **273,** 2–11 (2006).

262. Duan, X. & Quiocho, F. A. Structural evidence for a dominant role of nonpolar interactions in the binding of a transport/chemosensory receptor to its highly polar ligands. *Biochemistry* **41,** 706–712 (2002).

263. Sharff, A. J., Rodseth, L. E., Spurlino, J. C. & Quiocho, F. A. Crystallographic Evidence of a Large Ligand-Induced Hinge-Twist Motion between the Two Domains of the Maltodextrin Binding Protein Involved in Active Transport and Chemotaxis. *Biochemistry* **31,** 10657–10663 (1992).

264. Spurlino, J. C., Lu, G. Y. & Quiocho, F. A. The 2.3-A resolution structure of the maltose- or maltodextrin-binding protein, a primary receptor of bacterial active transport and chemotaxis. *J. Biol. Chem.* **266,** 5202–5219 (1991).

265. Walker, I. H., Hsieh, P. C. & Riggs, P. D. Mutations in maltose-binding protein that alter affinity and solubility properties. *Appl. Microbiol. Biotechnol.* **88,** 187–197 (2010).

266. Miller, D. M., Olson, J. S., PflugrathlT, J. W. & Quiochol, F. A. *Rates of Ligand Binding to Periplasmic Proteins Involved in Bacterial Transport and Chemotaxis*. *Journal of Biological Chemistry* **258,** (1983).

267. Hall, J. A., Gehring, K. & Nikaido, H. *Two Modes of Ligand Binding in Maltose-binding Protein of Escherichia coli*. *THE JOURNAL OF BIOLOGICAL CHEMISTRY* **272,** (1997).

268. Telmer, P. G. & Shilton, B. H. Insights into the Conformational Equilibria of Maltose-binding Protein by Analysis of High Affinity Mutants*. *J. Biol. Chem.* **278,** 34555–34567 (2003).

269. Liang, S. *et al.* Polysome-profiling in small tissue samples. *Nucleic Acids Res.* **46,** 2–13 (2017).

270. Roux, P. P. & Topisirovic, I. Regulation of mRNA translation by signaling pathways. *Cold Spring Harb. Perspect. Biol.* **4,** 1–23 (2012).

271. Bhat, M. *et al.* Targeting the translation machinery in cancer. *Nat. Rev. Drug Discov.* **14,** 261–278 (2015).

272. Moreno, J. A. *et al.* Sustained translational repression by eIF2α-P mediates prion neurodegeneration. *Nature* **485,** 507–511 (2012).

273. Piccirillo, C. A., Bjur, E., Topisirovic, I., Sonenberg, N. & Larsson, O. Translational control of immune responses: From transcripts to translatomes. *Nat. Immunol.* **15,** 503–511 (2014).

274. Gandin, V. *et al.* Polysome Fractionation and Analysis of Mammalian Translatomes on a Genome-wide Scale. *J. Vis. Exp.* 1–9 (2014). doi:10.3791/51455

275. Aspden, J. L. *et al.* Extensive translation of small open reading frames revealed by poly-ribo-seq. *Elife* **3,** 1–19 (2014).

276. Faye, M. D., Graber, T. E. & Holcik, M. Assessment of Selective mRNA Translation in Mammalian Cells by Polysome Profiling. *J. Vis. Exp.* 1–8 (2014). doi:10.3791/52295

277. Coudert, L., Adjibade, P. & Mazroui, R. Analysis of Translation Initiation During Stress Conditions by Polysome Profiling. *J. Vis. Exp.* 1–7 (2014). doi:10.3791/51164

278. Lacsina, J. R., Lamonte, G., Nicchitta, C. V. & Chi, J. T. Polysome profiling of the malaria parasite Plasmodium falciparum. *Mol. Biochem. Parasitol.* **179,** 42–46 (2011).

279. Krishnan, K. *et al.* Polysome profiling reveals broad translatome remodeling during endoplasmic reticulum (ER) stress in the pathogenic fungus Aspergillus fumigatus. *BMC Genomics* **15,** 159 (2014).

280. Biedler, J. L., Roffler-Tarlov, S., Schachner, M. & Freedman, L. S. Multiple neurotransmitter synthesis by human neuroblastoma cell lines and clones. *Cancer Res.* **38,** 3751–7 (1978).

281. Polson, E. S. *et al.* KHS101 disrupts energy metabolism in human glioblastoma cells and reduces tumor growth in mice. *Sci. Transl. Med.* **10,** (2018).

282. Wurdak, H. *et al.* An RNAi Screen Identifies TRRAP as a Regulator of Brain Tumor-Initiating Cell Differentiation. *Cell Stem Cell* **6,** 37–47 (2010).

283. Del Prete, M. J., Vernal, R., Dolznig, H., Müllner, E. W. & Garcia-Sanz, J. A. Isolation of polysome-bound mRNA from solid tissues amenable for RT-PCR and profiling experiments. *Rna* **13,** 414–421 (2007).