UNIVERSITY OF SOUTHAMPTON

FACULTY OF NATURAL AND ENVIRONMENTAL SCIENCES

Biological Sciences

Alternative Non-Canonical Translation Initiation Codons are Used To Synthesise Novel Isoforms of the Transcription Factor GATAD1

by

Helen Coral Knight

BSc (Hons)

Thesis for the degree of Doctor of Philosophy

April 2017

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF NATURAL AND ENVIRONMENTAL SCIENCES

BIOLOGICAL SCIENCES

Thesis for the degree of Doctor of Philosophy

ALTERNATIVE NON-CANONICAL TRANSLATION INITIATION CODONS ARE USED TO SYNTHESISE NOVEL ISOFORMS OF THE TRANSCRIPTION FACTOR GATAD1

Helen Coral Knight

Alternative translation initiation from upstream non-AUG codons contributes towards the diversity of the eukaryotic proteome; protein isoforms with varying N-terminal extensions can be generated from one mRNA transcript. Investigations are being carried out in order to elucidate how N-terminal extensions affect subcellular localisation, binding partners and thus the function of a protein as well as how the choice of alternative initiation codon (AIC) is regulated.

This thesis is focussed on GATAD1 (GATA Zinc Finger Domain-Containing 1), which is a ubiquitously expressed transcription factor, forming part of a transcriptionally repressive histone demethylase complex. GATAD1 regulates the expression of specific genes by forming an indirect interaction with the activating trimethyl marker of lysine four on histone three (H3K4me3); this interaction is made through a lysine demethylase chromatin 'reader', KDM5A (Jarid1A). GATAD1 is also involved in retinal development and heart disease, whereby a single mutation in the gene is the cause of dilated cardiomyopathy (DCM).

The GATAD1 mRNA transcript can be translated at alternative translation initiation codons resulting in the synthesis of three protein isoforms. The two isoforms with N-terminal extensions are initiated from a CUG and an unusual AUU codon. CRISPR genome editing has been used to tag genomic GATAD1, confirming endogenous expression of all three isoforms. Translation from the AICs is regulated by various factors, including eukaryotic initiation factors (eIFs) 1, 1A and 5, the context and position of the AICs, as well as secondary structure downstream of the AUU codon. Cell type and stresses such as hypoxia also influence the use of each GATAD1 AIC.

Although all three GATAD1 isoforms complex with KDM5A, it has been observed that the extended isoforms have a greater tendency to remain in the cytoplasm, potentially forming part of a cytoplasmic demethylase complex, whilst the annotated protein functions as a nuclear transcription factor.

Table of Contents

List of Tables

List of Figures

DECLARATION OF AUTHORSHIP

I, HELEN KNIGHT

declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

Alternative Non-Canonical Translation Initiation Codons are Used To Synthesise Novel Isoforms of the Transcription Factor GATAD1

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission

Signed: Helen Knight .....

Date: 18/04/2017.....

Acknowledgements

I would like to express the deepest appreciation for my supervisor, Dr Mark Coldwell for the patient guidance, unrelenting encouragement and advice he has provided me throughout my time as his student. I have been extremely lucky to have a supervisor who has gone beyond his duty to instil confidence in both myself and my work and whose door was always open.

I am grateful to all members of the Coldwell laboratory, past and present for making my PhD such a pleasure and ensuring a fun laboratory environment. Especially, I thank Dr Joanne Cowan and Dr Jim Schofield for their invaluable training and advice over the years, preparing me for years of science to come.

Finally, I acknowledge my husband and true supporter, Samuel, for his patience, advice and encouragement through both the good and challenging times of my PhD, thank you.

Abbreviations

AIC	Alternative Initiation Codon
APS	Ammonium persulfate
aTIS	alternative Translation Initiation Site
ATF4	Activating Transcription Factor 4
BSA	Bovine Serum Albumin
cDNA	complementary DNA
CDS	Coding sequence
CHX	Cycloheximide
dH_2O	distilled H_2O
DHFR	Dihydrofolate reductase
D-MEM	Dulbecco's Modified Eagle Medium
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
dNTP	deoxyribonucleotide
D-PBS	Dulbecco's Phosphate Buffered Saline
DSHP	Downstream Hairpin
eEF	eukaryotic Elongation Factor
eIF	eukaryotic Initiation Factor
ER	Endoplasmic Reticulum
eORF	extended ORF
FCS	Fetal Calf Serum
GATAD1	GATA Domain-containing protein 1
GDP	Guanosine-Di-Phosphate

GFP	Green Fluorescent Protein
GTP	Guanosine-Tri-Phosphate
GWIPS	Genome Wide Information on Protein Synthesis
HDAC	Histone Deacetylase
HeLa	Henrietta Lacks
HF	High Fidelity
HIF	Hypoxia Inducible Factor
IPTG	Isopropyl β-D-1-thiogalactopyranoside
IRES	Internal Ribosomal Entry Site
KDM5A	Lysine Demethylase 5A
LTM	Lactimidomycin
LDH	Lactate Dehydrogenase
M-PER	Mammalian Protein Extraction Reagent
mRNA	messenger RNA
mTORC1	Mammalian Target of Rapamycin Complex 1
MTP	Mitochondrial Targeting Peptide
NEB	New England BioLabs
NES	Nuclear Export Signal
NLS	Nuclear Localisation Signal
NPC	Nuclear Pore Complex
ORF	Open Reading Frame
ODD	Oxygen-Dependent Degradation
PCR	Polymerase Chain Reaction
PERK	Protein kinase RNA-like Endoplasmic Reticulum Kinase
PIC	Pre-Initiation Complex

pIC	plasmid for Initiation Codon
Post-TC	Post-Termination Complex
PPI	Protein-Protein Interaction
RNA	Ribonucleic acid
rRNA	ribosomal RNA
SD	Shine-Dalgarno
SERCA	Sarco/Endoplasmic Reticulum Ca^{2+} ATPase
SDM	Site Directed Mutagenesis
SILAC	Stable Isotope Labeling by Amino acids in Cell culture
siRNA	small interfering RNA
shRNA	small hairpin RNA
SNP	Single Nucleotide Polymorphism
SRP	Signal Recognition Particle
STED	Stimulated Emission Depletion microscopy
tRNA	transfer RNA
uORF	upstream ORF
UPP	Ubiquitin Proteasome Pathway
UPR	Unfolded Protein Response
USHP	Upstream Hairpin
USSDL	Upstream Shine Dalgarno-Like
UTR	Untranslated region
ZnF	Zinc Finger

Chapter One

Introduction

1. Introduction

1.1 Overview

Alternative translation initiation is a relatively recently discovered mechanism of increasing protein diversity. Translation from upstream alternative initiation codons (AICs) within the 5'UTR produces protein isoforms with N-terminally extended regions, as well as the annotated protein. Both the regulation of their translation as well as the function of these alternative protein isoforms is the focus of this thesis, using GATAD1 as a candidate gene in order to decipher whether the N-terminal extensions cause alternative subcellular localisation, alter the repertoire of binding partners or other functions within the cell, with respect to the annotated isoform. The introductory chapter of this thesis will cover the regulation of eukaryotic translation initiation as well as AICs and information on GATAD1.

1.2 Eukaryotic Gene Expression

The central dogma of molecular biology refers to the flow of genetic information from DNA encoding a gene, to a functional gene product which is usually in the form of a protein (Crick, 1970). Since the generation of macromolecules is central to cell survival, gene expression is utilised by all three domains of life. The greater complexity of eukaryotic genomes requires more elaborate mechanisms for gene regulation. These regulatory controls occur at every stage of the process, from transcription, through to RNA processing, to post-translational modifications of a protein. There are two main, independent events involved in eukaryotic gene expression; transcription, followed by translation. Transcription of double-stranded DNA into single-stranded RNA copies occurs in the nucleus. The pre-mRNA is co-transcriptionally processed in various ways which increases the stability of the RNA molecule (Laurencikiene et al., 2006); addition of a 5' 7-methylguanosine (m^7G) cap and a 3' polyA tail prevent ribonuclease degradation from either end of the transcript, whilst splicing removes non-coding introns from the RNA. Once processing is complete, mature mRNA leaves the nucleus via a nuclear pore and enters the cytoplasm, where translation takes place (Figure 1-1). The mRNA sequence is then used as a template by ribosomes to assemble specific amino acids into proteins during translation, which takes place in three main stages; initiation, elongation and finally, termination. Multiple auxiliary factors are required to regulate this whole process.

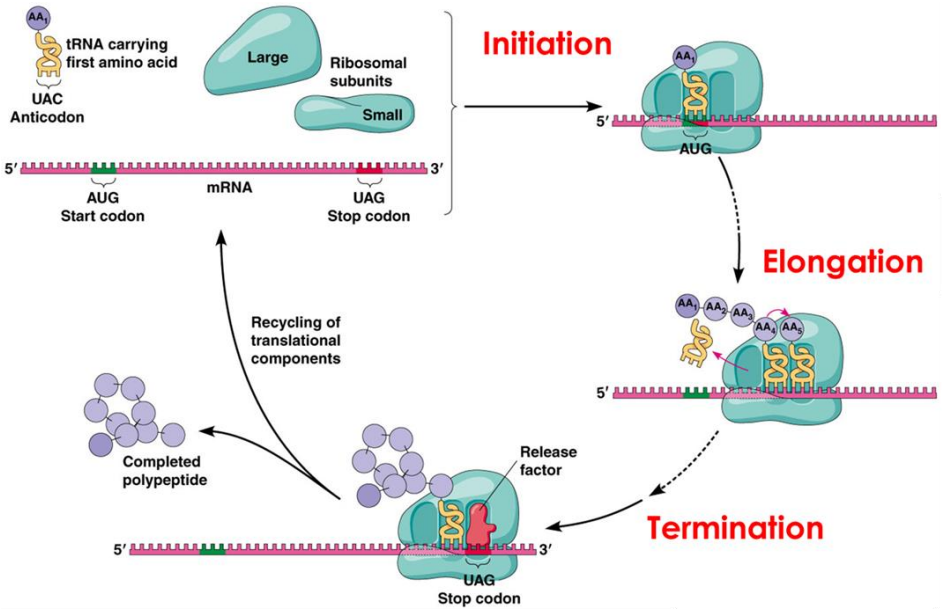


Figure 1-1: Eukaryotic Translation

mRNA is translated to protein in the cytoplasm in three main phases, initiation, elongation and termination, followed by recycling of the translational components. Image adapted from: http://www.mun.ca/biology/desmid/brian/BIOL2060.

1.3 Eukaryotic Translation

The mRNA molecule generated through transcription contains the protein coding sequence required to generate a full length protein. The translation of this nucleotide sequence into a linear amino acid sequence begins with translation initiation, which may proceed in one of two ways. Initiation via the cap-dependent pathway is the most prevalent; most cellular mRNAs utilise the m^7G cap to begin ribosomal scanning in order to identify the start codon, and ultimately assemble the initiation complex. A further method of translation initiation is the cap-independent pathway which relies on an internal ribosome entry site (IRES) in the 5'UTR. The IRES is an RNA structure which replaces the function of the 5' m^7G cap as well as many of the factors required for cap-dependent initiation, allowing translation to take place without the eukaryotic 40S ribosomal subunit scanning from the 5' end of the mRNA (Hellen, 2001). This method is less common in eukaryotes and is utilised mostly by viruses (Jackson et al., 2010), although in either case, the start codon is canonically AUG. A ternary complex consisting of GTP-bound eIF2 delivers the initiator tRNA (Met-$tRNA_i^{Met}$) to the ribosomal P-site, allowing complementary base pairing between the anticodon of the charged

initiator tRNA (Met-tRNA$_i^{Met}$) and the start codon. Translation initiation has also been observed from alternative initiation codons (AICs), although the efficiency of translation from AICs is affected by several different factors; the codon itself, the flanking sequences as well as the presence or absence of secondary structures downstream of the AIC (Kozak, 1989).

Following translation initiation, eukaryotic initiation factors (eIFs) dissociate from the 40S ribosome. This causes recruitment of the 60S ribosome to the mRNA to form the 80S ribosome, which is poised on the mRNA with the start codon and Met-tRNA$_i^{Met}$ placed in the P-site of the ribosome. Elongation of polypeptides can now begin, with two eukaryotic elongation factors (eEFs) required to control the process. eEF1A recruits the correct aminoacyl-tRNA to the A-site, dependent on the next codon in the ORF. A peptidyltransferase reaction then takes place between the amino-acyl tRNA in the A-site and the peptidyl-tRNA in the P-site, extending the polypeptide chain. Once the peptide bond formation is complete, eEF2 assists in translocation of the mRNA to the next codon, relative to the ribosome (Browne and Proud, 2002). The deacylated tRNA now shifts to the E-site, whilst the peptidyl-tRNA moves into the P-site, ensuring that the A-site of the ribosome is now free to accept the next amino-acyl tRNA and continue the polypeptide elongation.

Once an in-frame stop codon enters the A-site, it is recognised by eukaryotic release factor (eRF) eRF1. A further factor, eRF3, triggers the release of the polypeptide chain from the peptidyl tRNA (Baierlein and Krebber, 2010).

1.4 Eukaryotic Translation Initiation

Eukaryotic translation initiation is a vital control point in protein synthesis, and tight regulation is therefore required. As well as selecting the start codon and the translation reading frame, initiation of translation is also implemented in the cellular stress response, whereby translation initiation factors are post-translationally regulated (Sheikh and Fornace, 1999).

1.4.1 Formation of the 43S Preinitiation Complex (PIC)

The 43S PIC is a multifactor complex, formed by the association of the 40S ribosomal subunit, with numerous translation initiation factors as well as the ternary complex (Figure 1-2). Following translation termination, ribosomal recycling must take place in order to initiate a new round of translation. Post-termination complexes (post-TCs) must be dissociated in order to release the mRNA, P-site deacylated tRNA and eRF1/eRF3 which remain associated with the ribosomes (Figure 1-2-step 1). The ribosome anti-association factor eIF3 is the only initiation factor able to promote splitting of the ribosomal subunits independently, although this action is enhanced by its loosely

associated subunit eIF3j, as well as eIF1 and eIF1A (Pisarev et al., 2007). eIF3 binds to the solvent side of the 40S ribosomal subunit, both disrupting an inter-subunit bridge domain, and causing conformational changes in the 40S subunit which may contribute to post-TC dissociation (Hinnebusch, 2006). eIF6 binds to the 60S ribosomal subunit, further preventing re-association (Gandin et al., 2008). Meanwhile, eIF1 promotes deacylated tRNA release from the dissociated 40S subunit, after which eIF3j can mediate mRNA dissociation from the same 40S subunit, but only when it is not stabilised by P-site tRNA (Pisarev et al., 2007).

As well as promoting ribosomal recycling, the high affinity and cooperative binding of eIF1 and eIF1A on the interface side of the 40S subunit, together promote an open conformation of the mRNA channel. Cryo-EM studies show that the mRNA entry channel latch is open in the 40S-eIF1-eIF1A structure, whilst a new connection has made between the 18S rRNA helix 16 and the ribosomal protein rpS3 (h16-rpS3 connection) which prevents the latch from reforming on the solvent side (Passmore et al., 2007). These allosteric conformational changes both accelerate ternary complex binding, as well as ensuring that the complete 43S PIC is scanning-competent. eIF1 and eIF1A also interact cooperatively to maintain start codon fidelity, preventing initiation from occurring at non-AUG codons (Pestova and Kolupaeva, 2002). eIF1 monitors base pairing between the mRNA and the Met-tRNAi^{Met}, rejecting non-AUG codons by preventing the release of Pi from eIF2-GTP in the PIC (Sonenberg and Hinnebusch, 2009). On the other hand, the N and C-terminal regions of eIF1A have opposing functions; the N-terminal region acts to stall ribosomal scanning, promoting eIF1 release and subsequent eIF2-GTP hydrolysis at AUG codons. Whereas, the C-terminal region of eIF1A rejects non-AUG codons, promoting scanning of the PIC along the 5'UTR in search of a start codon (Fekete et al., 2007).

Once formed, binding of the eIF2 ternary complex (eIF2·GTP·Met-tRNAi^{Met}) to the 40S ribosomal subunit completes the formation of the 43S PIC; recruitment of Met-tRNAi^{Met} to the 40S ribosome is a key point of regulation in translation initiation. Formation of the ternary complex (Figure 1-2 – step 2) is dependent on the heterotrimeric GTP-binding protein eIF2, being GTP-bound. At physiological conditions, eIF2 has a high affinity for GDP and so a guanine nucleotide exchange factor (GEF) eIF2B is required to convert eIF2 to its active GTP-bound state (Jennings et al., 2013). eIF2B is regulated in many ways; one of which involves phosphorylation of its substrate eIF2 at a conserved serine residue in its α-subunit (Ser51). Phosphorylated eIF2α acts as a competitive inhibitor of eIF2B, decreasing the rate of formation of ternary complexes and therefore stalling 43S PIC formation, causing a decrease in overall translation rates (Williams et al., 2001). Binding of the completed ternary complex to the 40S ribosomal subunit (Figure 1-2 – step 3) is mediated by both eIF1 and eIF3 and involves the direct delivery of Met-tRNAi^{Met} to the P-site (Dever, 2002).

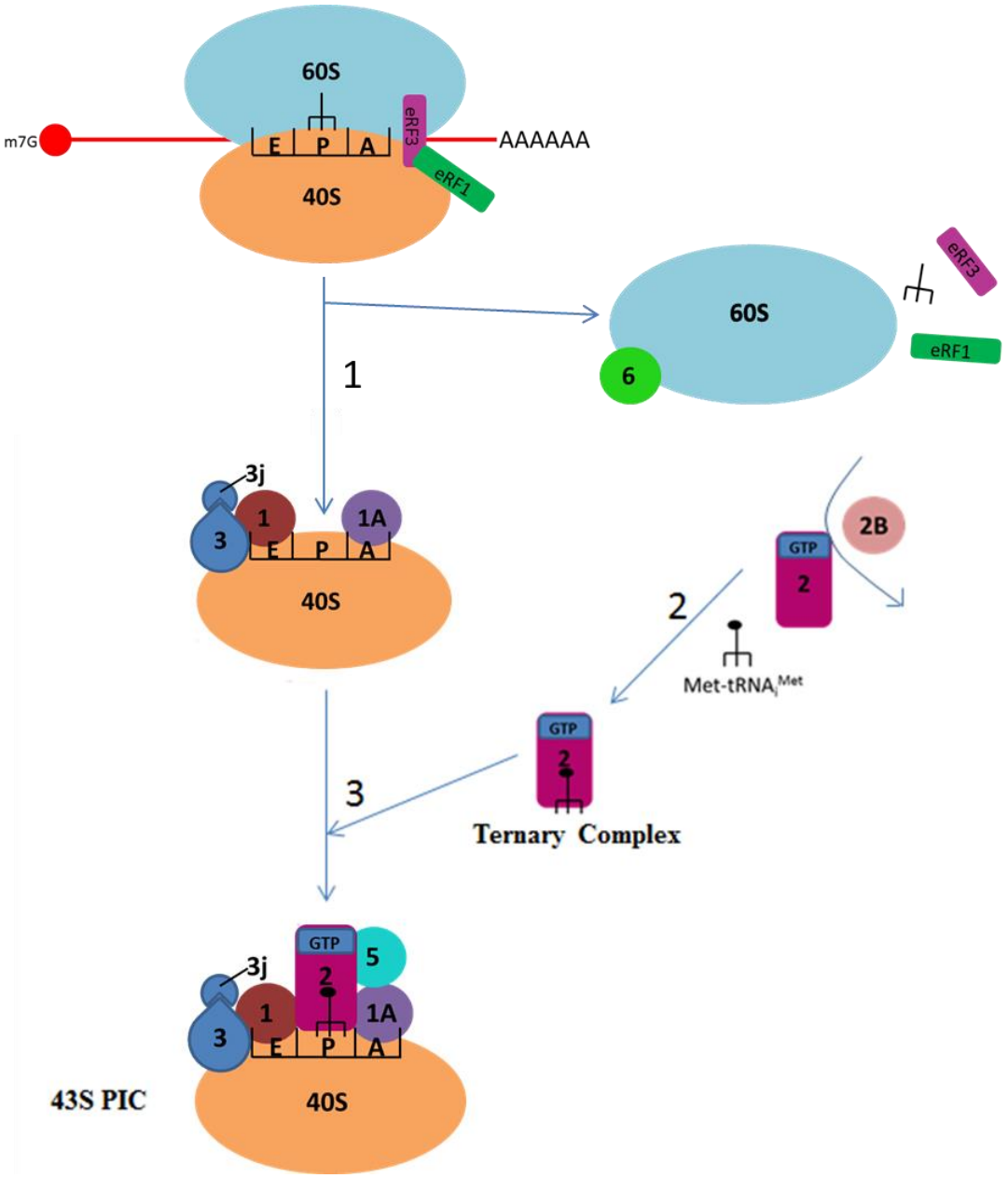


Figure 1-2: Formation of the 43S Preinitiation Complex

N.B. All objects labelled with a number are eukaryotic initiation factors and are prefixed with 'eIF'; i.e. 1 = eIF1.

(1) 80S post-termination complexes (post-TCs) require numerous eIFs to separate into their 60S and 40S ribosomal subunits. Ribosomal anti-association factor eIF3 binds to the 40S subunit and has the main role in separating the post-TC. eIF3j, eIF1 and eIF1A also bind to the 40S subunit and mediate the release of eRF1, eRF3, uncharged tRNA and the mRNA molecule from the ribosome. eIF6 binds to the 60S subunit and helps prevent re-association of the two subunits. (2) Formation of the eIF2·GTP·Met-tRNA$_i^{Met}$ ternary complex involves the exchange of GDP for GTP on eIF2, facilitated by the GEF eIF2B. eIF2 can then bind the initiator tRNA which completes the formation of the ternary complex. (3) Delivery of the ternary complex to the 40S ribosomal subunit completes the formation of the 43S PIC.

1.4.2 Formation of the 48S Initiation Complex

Following attachment of the 43S PIC to mRNA near the 5' cap and subsequent scanning of the 5'UTR, recognition of an initiation codon results in eIF5-mediated hydrolysis of GTP-bound eIF2 and the formation of a 48S initiation complex.

1.4.2.1 eIF4F Complex

The heterotrimeric eIF4F complex consists of eIF4E, A and G and binds the 5' cap of mRNAs (Figure 1-4 – step 1). eIF4E is the cap-binding protein, which has a high affinity for the m^7G cap of mRNA. Specific contacts are made between the cap and eIF4E upon binding; the cap is able to stack between two Trp residues on the concave surface of the initiation factor (Jackson et al., 2010). Regulation of eIF4E activity is important in determining translation rates. Phosphorylation of eIF4E occurs at Ser209 by protein kinase Mnk1 (mitogen-activated protein kinase (MAPK)- interacting kinase 1), which increases the affinity of eIF4E for the cap (Waskiewicz et al., 1999). There are two isoforms of Mnk; Mnk1 phosphorylates eIF4E upon MAPK activation, whereas Mnk2 is responsible for the basal, constitutive phosphorylation of eIF4E. Mnk1 is activated by both p38 MAP kinase as part of the cell stress response, as well as the ERK (Extracellular Signal-related Kinase) pathway. In contrast, eIF4E-binding proteins (4E-BP's) are regulatory proteins which compete for the same binding site on eIF4E as the scaffold protein eIF4G, inhibiting cap binding and thus blocking translation initiation (Wang et al., 1998). The initial binding of eIF4E at the m^7G cap positions the whole eIF4F complex at the 5' terminal end of the mRNA transcript, enabling eIF4A to exert its helicase activity (Merrick, 2003).

eIF4A is an ATP-dependent DEAD-box RNA helicase, with two binding domains in eIF4G which are connected by a polypeptide linker (Rogers et al., 2001). The helicase activity is low in free eIF4A and the accessory proteins eIF4B and eIF4H are required for successful unwinding of secondary structures within the 5' UTRs of mRNA transcripts, as well as removal of ribonucleoproteins (RNPs); therefore facilitating 40S ribosomal subunit binding and subsequent scanning (Rozovsky et al., 2008). The protein scaffold eIF4G contains three HEAT domains, two of which bind eIF4A. HEAT-1 stimulates eIF4A helicase activity whilst HEAT-2 regulates the interaction (Marintchev et al., 2009).

eIF4G acts as a protein scaffold which is involved in recruiting the 40S ribosomal subunit to the mRNA. There are two isoforms of eIF4G in mammalian systems; although both eIF4GI and eIF4GII can form functional eIF4F complexes, eIF4GI is the predominant isoform involved in translation initiation and forms 85% of total eIF4F complexes in HeLa cells (Gradi et al., 1998). The eIF4G isoforms contain homologous binding domains for numerous initiation factors, including the other members of the eIF4F complex (eIF4E and eIF4A), as well as PABP, eIF3 and a C-terminal

Mnk binding site (Figure 1-3), although the isoforms are only 46% conserved overall (Gradi et al., 1998).

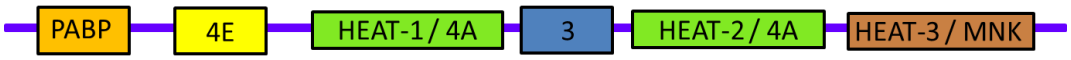


Figure 1-3: eIF4G Domain Structure

The domain organisation of human eIF4GI is shown. The N-terminal region of the protein contains binding sites for PABP and eIF4E. The middle region of the protein contains the first of three HEAT domains, which binds to eIF4A to stimulate its helicase activity, as well as an eIF3 binding site; this site forms the bridge between the 43S PIC and the activated mRNA. The C-terminal region of the protein contains a second HEAT domain which acts as a further binding site for eIF4A, as well as a third HEAT domain which interacts with protein kinase MNK.

1.4.2.2 Recruitment of the 43S PIC to mRNA

Whilst the eIF4F complex binds the 5' cap of the mRNA, poly(A)-binding protein (PABP) synergistically binds the poly(A) tail of the mRNA transcript. Binding of both structures together ensures translational integrity, confirming that the mRNA transcript has been fully processed before translation initiation can begin (Figure 1-4 – step1) (Eckmann et al., 2011). As well as binding the poly(A)-tail, PABP also binds the scaffold protein eIF4G, circularising the mRNA molecule, a process which is important for efficient translation (Preiss and Hentze, 1999). Regulation of PABP activity is mediated by PABP-interacting proteins (Paips), which can both stimulate as well as inhibit translation. Paip-1 acts as a translational enhancer, whilst Paip-2A and Paip-2B compete with Paip-1 for PABP binding, causing translational inhibition of capped-mRNA transcripts (Derry et al., 2006).

The 40S ribosomal subunit within the 43S PIC recruits the circularised, activated mRNA using numerous initiation factors. A molecular bridge is formed between eIF4G and eIF3, providing a direct interaction between the activated mRNA and the 40S ribosomal subunit, which is stabilised by the activation of the mammalian target of Rapamycin complex 1 (mTORC1) (Villa et al., 2013). eIF3 is a large multisubunit complex, but photoaffinity cross-linking studies as well as cryo-EM structures show that only eIF3e is involved in the interaction with eIF4G (LeFebvre et al., 2006). The attachment of an activated mRNA to the 43S PIC results in scanning of the 5' UTR in search of an initiation codon (Figure 1-4 – step 2).

1.4.2.3 48S Ribosomal Scanning

The 43S PIC consisting of the 40S ribosomal subunit and eIFs 1, 1A, 3 and 5 has formed an interaction with mRNA through eIF3 to eIF4G, which completes the 48S preinitiation complex. The scaffold protein eIF4G is the link to the activated mRNA transcript, along with eIFs 4A, 4E and PABP. The 48S PIC is then able to scan along the 5' UTR of the mRNA transcript in a 5' to 3' direction, in search of an initiation codon. eIF4A, aided by eIF4B and eIF4H unwind secondary structures that would otherwise inhibit scanning. The mRNA enters the 40S ribosomal subunit on the solvent side through a twelve nucleotide entry channel. As the mRNA is threaded through the channel it is analysed in the P-site and checked for complementarity to the anticodon of the initiator tRNA. If no match is made, the mRNA exits through the 12 nucleotide exit channel and the ribosome continues scanning along the transcript (Hinnebusch, 2011). Upon identification of an initiation codon (Figure 1-4 – step 3), a GTPase-activator protein (GAP), eIF5 acts as a molecular switch, causing hydrolysis of eIF2-bound GTP on the β-subunit, leaving eIF2-GDP and releasing P_i as well as other factors which were associated with the ribosome (Paulin et al., 2001). A further factor, eIF5B uses eIF1A to mediate the joining of the 60S subunit to the 48S initiation complex to form an 80S active ribosome (Pestova et al., 2000) (Figure 1-4 – step 4). The correct aminoacyl-tRNA is then recruited to the A-site, dependent on the next codon in the transcript. A peptidyl transferase reaction takes place between the aminoacyl-tRNA in the A-site and the methionyl-tRNA in the P-site, extending the polypeptide chain; this causes the deacylated tRNA to shift to the E-site ready to exit from the ribosome. Polypeptide elongation then continues until a stop codon is recognised (Browne and Proud, 2002). Certain mRNAs, including two-thirds of oncogenes contain an AUG in their 5'UTRs, which is associated with the start of an upstream open reading frame (uORF). Once the ribosome has translated the uORF, it is possible for scanning of the transcript to continue, causing re-initiation of translation at the canonical start codon (Morris and Geballe, 2000).

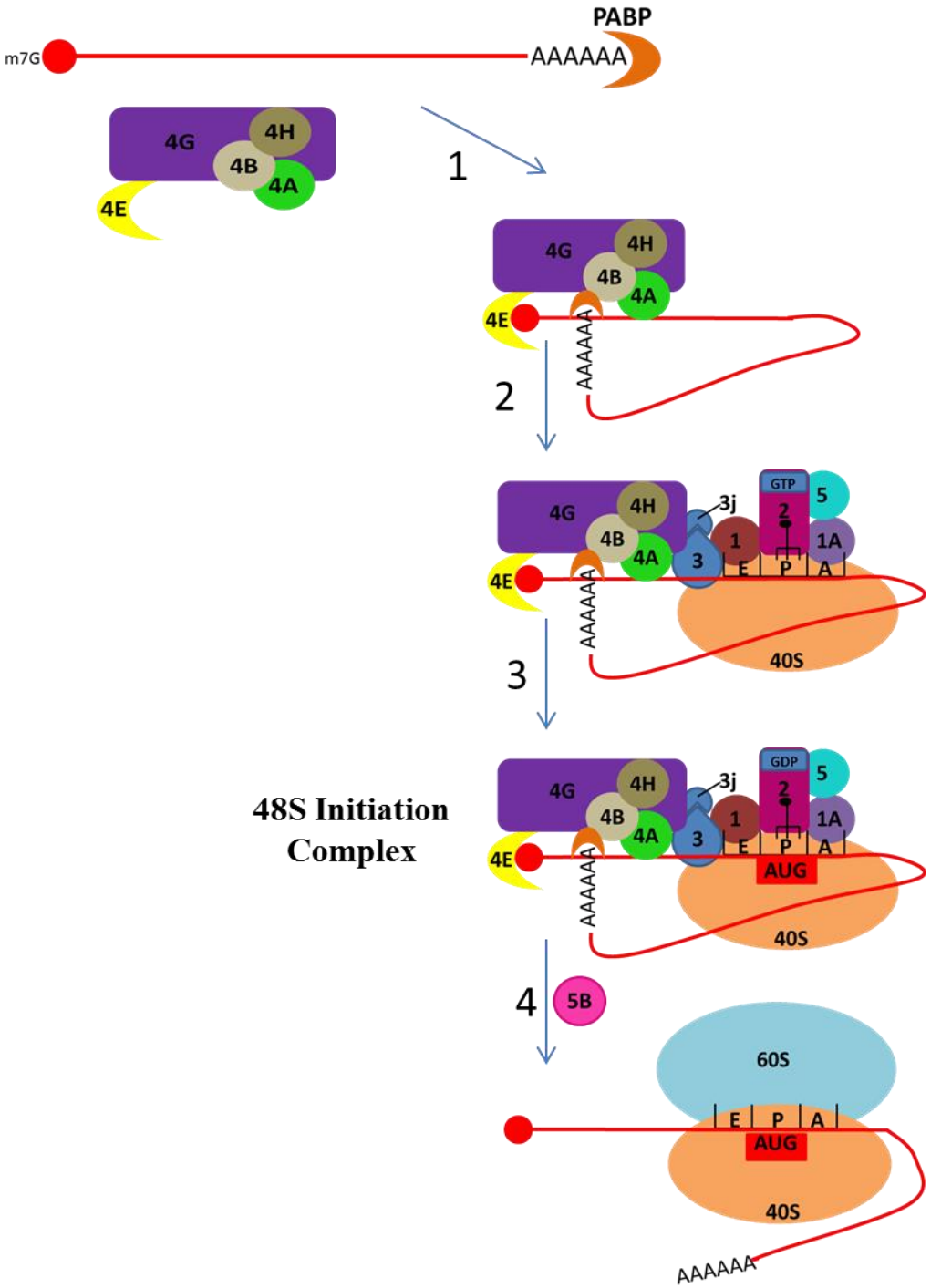


Figure 1-4: Formation of the 48S Initiation Complex

(1) The eIF4F complex forms first, consisting of eIF4G, eIF4E and eIF4A. Accessory proteins eIF4H and eIF4B work with eIF4A to increase its helicase activity. eIF4E binds to the m^7G cap on the mRNA, whilst PABP binds the poly(A) tail, causing circularisation and activation of the mRNA. (2) Activated mRNA recruits the 43S PIC via an interaction between eIF4G and eIF3. Scanning of the mRNA is then able to take place, in a 5' to 3' direction. (3) Recognition of an initiation codon, which is usually but not always an AUG, halts ribosomal scanning. eIF5 causes hydrolysis of eIF2-bound GTP, releasing P_i. (4) eIF5B then uses eIF1A to mediate the joining of the 60S ribosomal subunit, forming the final 80S ribosome which can carry out peptide elongation.

1.5 Start Codon Selection

1.5.1 Kozak Consensus

Although alternative initiation codons (AICs) have now been identified, the scanning model of translation initiation postulates that the canonical start codon for every protein is the first AUG codon found within a good context. In 1987, Kozak proposed a consensus sequence required for successful recognition of the AUG start codon by the scanning 43S ribosome, based on a study of 699 vertebrate mRNAs. Analysis of the 5' UTRs of these mRNAs showed that the sequence flanking the start codon is highly conserved (Kozak, 1987a). 97% of the mRNAs had a purine at position -3; whilst a guanine nucleotide was conserved at position +4. Site directed mutagenesis studies were carried out at all positions flanking the initiator AUG and it was shown that mutations at positions -3 and +4 have the largest influence on translational efficiency (Figure 1-5).

More recently however, it has been shown that the +4G consensus may not actually be necessary for efficient translation. Instead, amino acid constraints at the second codon cause the overuse of alanine (GCN) and glycine (GGN) as the second amino acid in the polypeptide, hence requiring a +4G (Xia, 2007). N-terminal methionine excision (NME) is a co-translational process carried out by methionine aminopeptidase (MAP); this enzyme cleaves all residues with small side chains on the second residue (P1'). The MAP binding pocket cannot tolerate the larger side chains of residues such as asparagine and leucine; the optimal residues for cleavage at P1' are alanine and glycine (Frottin et al., 2006). Both alanine and glycine are therefore over-represented as the second amino acid in the polypeptide in order for efficient MAP cleavage to take place; hence requiring a G at the +4 position in the mRNA.

Similarly, the N-end rule states that the N-terminal amino acid of a protein determines how rapidly it is degraded. Therefore, proteins with a stabilizing N-terminal amino acid such as valine/glycine (GNN) are over-represented. Similarly, rather than increasing the efficiency of translation, having a G at the +4 position relative to the start codon is over-represented due to the long half-life of proteins with these GNN amino acids at their N-terminal end (Varshavsky, 2011). This may account for the high occurrence of a G at the +4 position as shown by Kozak. The nature of the nucleotide at the +5 position also has an effect on translational efficiency; an adenine increases start codon recognition most efficiently, followed by a cytosine. Uracil at the +6 position was also shown to increase recognition (Grunert and Jackson, 1994). Translation was initiated from the first AUG in 90-95% of the mRNAs, although if the local sequence context of the codon was weak, the ribosome was able to bypass the first AUG and continue scanning to the second AUG which may have a stronger context. This phenomenon is known as 'leaky scanning' (Kozak, 1987a). Leaky scanning may also take place if the initiation codon is too close to the 5' cap-proximal AUG (Kozak, 1991).

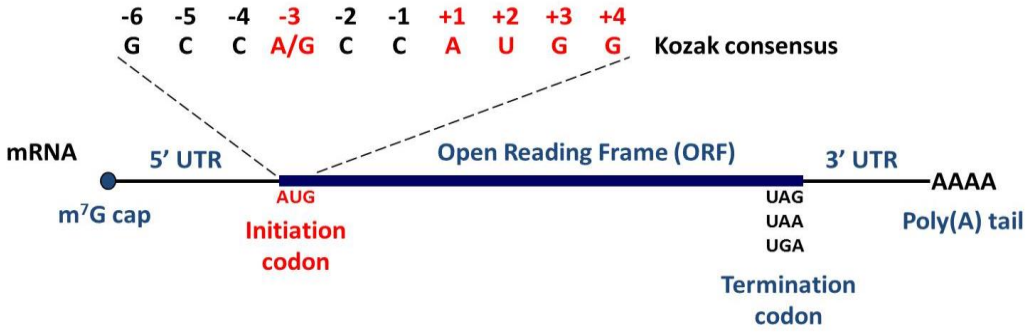


Figure 1-5: Kozak Consensus

The Kozak consensus is shown surrounding the canonical AUG at position +1. In order to ensure that efficient translation takes place, the start codon must be in a strong context. This requires a purine (A/G) at position -3 and a guanine at position +4.

1.5.2 Internal Ribosome Entry Site (IRES)

In addition to leaky scanning, the selection of the initiation codon can also be made by internal entry of the ribosome (Touriol et al., 2003). Internal ribosomal entry sites (IRESs) were first identified in viral mRNAs and allow cap-independent translation of the ORF to take place using tertiary RNA structures. The eIF4E subunit of the eIF4F complex, which would usually recruit the 43S PIC to the 5' capped end of the mRNA is not utilised when translation occurs from an IRES. Viral IRES elements vary greatly in structure and also in the way that they are able to recruit the 40S ribosome to the start site. For example, the EMCV (Encephalomyocarditis virus)-IRES requires most of the initiation factors used in cap-dependent translation initiation in order to recruit the 40S ribosome (except the cap binding protein eIF4E), whereas CrPV (cricket paralysis virus)-IRES elements require no eIFs whatsoever (Hellen, 2001). Binding of the CrPV-IRES to the ribosome places a pseudoknot in the ribosomal P-site which mimics tRNA in the absence of eIF2 and Met-tRNA$_i^{Met}$, resulting in the direct joining of 60S and formation of an 80S ribosome. Elongation factor 1 (EF1) delivers the aminoacyl-tRNA to the empty A-site which is followed by pseudo-translocation of the ribosome (no peptide bond formation) and standard elongation of the polypeptide (Fernandez et al., 2014).

1.5.3 The Wegrzyn Consensus

Canonical AUG start codons conform to the Kozak consensus relatively efficiently; however, an extended consensus is thought to be required for alternative initiation codons. In order for an aTIS to be recognised efficiently by the scanning ribosome, Wegrzyn postulated a consensus (CG**CG**UCGCGxxxG) which incorporates more than the -3 and +4 position said to be important in AUG start codon recognition via the Kozak consensus. The Wegrzyn consensus states that it is the nucleotide at positions -6 (G/C) and -7 (C) which are responsible for translational efficiency (Wegrzyn et al., 2008), rather than the positions -3 and +4 in the Kozak consensus. However, the Wegrzyn consensus is based on a study using only 45 mammalian mRNAs whereas the Kozak consensus is based on a study of 699 vertebrate mRNAs, rendering it more widely accepted and reliable.

1.6 Generating Protein Diversity

1.6.1 Production of alternative mRNA species

The human genome is composed of a relatively low number of genes, which cannot account for our complexity. Increasing protein heterogeneity accounts for the complexity of the biological systems and signalling pathways which we rely on. There are numerous ways of generating protein diversity by the production of different mRNA species within eukaryotes, resulting in the translation of several protein products from each individual gene. These mechanisms include alternative promoter usage, which results in the transcription of mRNA molecules with varying 5' ends, thus the potential to encode protein isoforms with different N-termini (Ayoubi and Van De Ven, 1996).

The regulation of 3' processing by alternative polyadenylation also results in the generation of several mRNAs with differing 3' sequences, depending on poly(A) site selection and subsequent position of pre-mRNA cleavage. Alternative poly(A) sites present within coding regions of a pre-mRNA will result in the subsequent translation of alternative protein isoforms (Colgan and Manley, 1997). On the other hand, alternative poly(A) sites within the 3'UTR will not increase protein diversity, instead generating mRNAs with differing 3'UTRs. The 3'UTR is capable of regulating gene expression qualitatively by containing regulatory sequences such as microRNA (miRNA) binding sites or AU-rich elements (AREs) which negatively regulate translation (Fabian et al., 2010).

Alternative splicing is a further mechanism of producing numerous forms of mature mRNA from a single gene when ligating exons to form mature mRNA. Several alternative splicing events may take place in eukaryotes, including selection of alternative 5' or 3' splice sites within an exon sequence, or exon skipping (Kim et al., 2008).

RNA editing involves post-transcriptional insertion, deletion or base modification within an mRNA molecule and has the potential to result in the translation of alternative protein isoforms. The most prevalent form of RNA editing is the adenosine to inosine (A-to-I) base modification whereby adenosine is deaminated to inosine which can subsequently pair with guanosine and cytosine as well as uracil (Wulff and Nishikura, 2010).

1.6.2 Production of multiple protein isoforms from a single mRNA

As well as producing several mRNAs from a single gene, protein diversity may also be increased by producing several protein isoforms from a single eukaryotic mRNA. Examples of this include translation of upstream open reading frames (uORFs), which often negatively regulate the subsequent translation of the annotated ORF (Cao and Geballe, 1995). The structure of the uORF peptide can result in poor translation termination, leading to ribosomal stalling and a blockade of leaky scanning downstream to the annotated initiation codon (Lovett and Rogers, 1996). After translation of certain uORFs however, the ribosome remains associated with the mRNA and reinitiation occurs at a downstream initiation codon, resulting in the translation of several protein isoforms from a single mRNA transcript (Kozak, 1987b). General control protein 4 (GCN4) encodes a transcription factor in yeast, which is responsible for the activation of numerous genes required for amino acid/purine biosynthesis in response to starvation of these factors. There are four inhibitory, small uORFs within the GCN4 mRNA 5'UTR, which negatively regulate the translation of the GCN4 ORF (Mueller and Hinnebusch, 1986). During starvation conditions however, whilst phosphorylation of eIF2α results in global translation inhibition in order for the cell to divert resources into amino acid biosynthesis (see section 1.4.1), GCN4 translation is derepressed in a dual-regulatory response. This is as a result of less ternary complex being available due to inhibition of eIF2B by phosphorylated eIF2α during amino acid starvation. Therefore the usual reinitiation at inhibitory uORF 4 cannot take place when the scanning 40S does not bind ternary complex (Dever et al., 1992). Bypassing uORF 4 results in leaky scanning and reinitiation at the GCN4 annotated start site, increasing translation of the activator protein GCN4.

Insertion of a selenocysteine (structurally similar to cysteine but with a selenium in place of sulphur) amino acid at a UGA (stop) codon, results in C-terminal extension of the protein rather than termination of translation (Allmang and Krol, 2006), which is a further mechanism of increasing protein diversity from a single mRNA.

Repeat-associated non-AUG (RAN) translation is a more recently discovered mechanism resulting in the translation of several protein isoforms from a single mRNA by alternative translation initiation without the requirement for an AUG codon (Zu et al., 2011). Instead, nucleotide repeat expansions can take place at CAG, CUG, GGGGCC, GGCCCC and CGG in all three reading frames,

resulting in the translation of toxic repeat proteins which are causative for various neurodegenerative and neuromuscular disorders (Labbadia and Morimoto, 2013).

Alternative translation of an mRNA may also occur through ribosomal frameshifting, which is mostly found in viral mRNAs and results in alteration of the reading frame and the subsequent translation of alternative protein isoforms (Kwun et al., 2014).

The recent identification of small ORF (smORF)-containing long noncoding RNAs (lncRNAs) which translate small conserved peptides with functional relevance, further increases protein diversity (Anderson et al., 2015). Many RNAs currently annotated as 'non-coding' may in fact translate functional peptides (Aspden et al., 2014) .

1.7 Alternative Initiation Codons

Alternative translation initiation is a further mechanism used to increase protein diversity, involving the initiation of translation from several start sites within an mRNA transcript, resulting in the translation of several protein isoforms.

Alternative initiation events were first identified in 1985, when capsid protein B of the adeno-associated virus type 2 (AAV2) was shown to initiate translation from an in-frame ACG codon (Becerra et al., 1985). This discovery fuelled mutagenesis studies to elucidate what codons a mammalian ribosome is capable of initiating translation from. Peabody (Peabody, 1989) mutated the initiator AUG of dihydrofolate reductase (DHFR) to every possible triplet which differs from AUG by one nucleotide (near cognate start codons). The ability of each mutated codon to initiate translation and produce a fully functional DHFR protein was analysed *in vitro*. A hierarchy regarding the efficiency of recognition of each AIC was put together as follows: ACG, CUG, AUC, AUU, AUA. No full length DHFR was produced when the start codon was mutated to GUG, UUG, AAG or AGG, suggesting no recognition takes places between the scanning 40S ribosome and these particular codons. However, more recent studies have shown that a GUG codon is recognised efficiently; mutagenesis studies identified a conserved GUG codon as the sole initiation codon in DAP5 (Death-associated protein), also known as NAT1/p97 (Takahashi et al., 2005), which utilises a downstream hairpin as well as an optimum Kozak consensus.

Since numerous AICs have been shown to initiate translation, questions have been posed regarding the identity of the initiator tRNA. There is evidence to suggest that the initiator tRNA would remain as ^{Met}tRNA, base pairing to anticodons which were not exactly complementary. Greater flexibility for mismatch is sustained in the P-site between the codon:anticodon duplex than

in the A site when elongator ^{Met}tRNA is being incorporated, because the decoding centre in the A site strictly prevents mismatches from being incorporated (Ivanov et al., 2011). This 'wobble' effect in the P-site would enable the same cellular machinery to carry out translation initiation at AICs as at the canonical AUG. The near-cognate mutants used in the studies on DHFR by Peabody (1989) were all able to incorporate methionine as the first amino acid, supporting the ^{Met}tRNA view. The DHFR mRNA initiator triplet was mutated to numerous AICs and exposed to an *in vitro* translation system (wheat germ and reticulocyte lysate) in the presence of [^{35}S]Met-tRNAi^{Met}. The radiolabelled methionyl-tRNA can only be incorporated at the N-terminus of a protein and is therefore often used to investigate translation initiation events. ^{35}S was incorporated into each of the DHFR isoforms, suggesting that the initiator tRNA does indeed remain as ^{Met}tRNA regardless of the initiation codon used (Peabody, 1989). However, this does not rule out incorporation of a non-methionine initiation as the first amino acid of the polypeptide in some cases.

Experiments in mammalian COS1 cells using anticodon sequence mutants have provided evidence that AICs can initiate translation with amino acids other than methionine (Drabkin and Rajbhandary, 1998). An initiator tRNA with mutated anticodon sequence CAU to GAC (G34C36 mutant) is aminoacylated with valine. The ability of the G34C36 mutant tRNA to initiate protein synthesis was analysed using a reporter CAT gene (GUC1) with a mutated initiation codon (AUG to GUC); the activity of the CAT gene was then analysed. The G34C36 mutant initiator tRNA is indeed able to initiate protein synthesis from the GUC initiation codon and produces 24-27% of the wild-type initiator tRNA activity, suggesting that GUC is likely to initiate translation with a valine rather than a methionine amino acid. In addition, overexpression of the G34C36 initiator tRNA mutant leads to a four-fold increase in valine acceptance of total tRNA; whilst the deacylation rate of the aminoacylated G34C36 tRNA is comparable to that of valyl-tRNA and very different to that of methionyl-tRNA which has a less stable linkage between the amino acid and tRNA.

Translation from AICs is generally less efficient than from the canonical AUG start codon. This is the basis of several diseases, including β-thalassemia; mutation of the canonical AUG to GUG, AUU, AUA, ACG or AAG in β-globin results in disease, even though the AIC is in a perfect Kozak consensus (Takahashi et al., 2005). This suggests that other elements are required to facilitate translation from an AIC.

The recently-developed ribosome profiling technique allows accurate identification of active initiation sites, using translation inhibitors to stall ribosomes at the initiation phase. Deep sequencing of ribosome-protected fragments is then carried out allowing ribosome positions to be identified. The emergence of ribosome profiling has provided new insights into the extensive use of upstream non-AUGs, with one study suggesting over one third of transcripts initiate from an upstream AIC (Lee et al., 2012).

1.7.1 Factors Promoting Translation from AICs

The first factor thought to impact on AIC recognition by the 40S ribosome is the sequence flanking the codon. Kozak proposed that non-AUG initiation codons have the same requirements for start codon flanking sequences as AUG start sites; claiming that a +4G increases ribosomal recognition, whereas the nucleotides at +5 and +6 do not (Kozak, 1987a). On the other hand, other studies (Boeck and Kolakofsky, 1994) have shown that positions +5 and +6 do have important effects on recognition of AICs. In three mRNAs with three different non-AUG initiation codons, it was shown that when the start codon was followed by GUA, recognition of the AIC did not take place and translation therefore did not occur. On the other hand, when the start codon was followed by GAU or GCU, initiation took place successfully (Boeck and Kolakofsky, 1994). Further bioinformatics studies have shown that it is most likely to be a G/C nucleotide at position -6 and a C nucleotide at position -7 which has the greatest impact on AIC recognition (Wegrzyn et al., 2008).

A secondary structure present downstream of an initiation codon can increase the recognition of a suboptimal start codon by the scanning ribosome and provide an important mechanism for regulating the expression of certain protein isoforms. GC-rich, stable hairpins found at a critical distance of 13-17 base pairs downstream of the AIC will cause the ribosome to pause until the secondary structure has been unwound by RNA helicases (Kozak, 1990). This temporary delay provides an optimal interaction between the start codon and the initiator tRNA, facilitating recognition and thus translation from the AIC (Kochetov et al., 2007). If the secondary structure is too far from, or too close to the AIC, no effect will be seen on translational efficiency; 13-17 base pairs corresponds to the distance between the 3' leading edge of the scanning 40S ribosomal subunit and its codon recognition centre and will therefore provide the optimal interaction between AIC and codon recognition.

Cellular factors are capable of affecting the stringency of start codon selection. eIF1 is involved in high-stringency start site selection, maintaining the open, scanning-competent conformation of the 43S PIC until an AUG in a strong Kozak consensus is identified (Passmore et al., 2007). Upon start site selection, eIF1 must be dissociated before the ribosome can enter the closed conformation, locking onto the mRNA. When eIF1 is knocked down in cells, translation from AIC's increases (Jackson et al., 2010). On the other hand, the N and C-termini of eIF1A have opposing effects; the N-terminal tail promotes the closed conformation of the ribosome on the mRNA, whilst the C-terminal tail promotes the open conformation of the scanning ribosome, increasing the stringency of start site selection (Fekete et al., 2007). It has also been shown that over-expression of eIF5 relaxes the stringency of start site selection, increasing initiation from AICs (Loughran et al., 2012). Increasing levels of eIF5 promotes hydrolysis of eIF2·GTP leading to release of factors from the 43S ribosome and 60S subunit joining. A final factor thought to be involved in the stringency of start site selection is eIF2. If the interaction between eIF2β and eIF2γ

is weak, the GTP-Met-tRNA$_i^{Met}$ interaction in the ternary complex is destabilised, causing less stringent start site selection and more translation occurring from weak AUGs or AICs (Hashimoto et al., 2002).

1.7.2 Identifying Novel AICs

Numerous genes containing upstream non-AUG initiation codons have now been identified. Both eIF4GI and eIF4GII which form part of the eIF4F complex and regulate both cap-dependent and cap-independent translation initiation, use AICs to translate multiple isoforms. As well as containing an IRES, eIF4GI utilises five upstream in-frame AUG codons to initiate translation (Byrd et al., 2002) and the resultant isoforms differ in their ability to rescue translation rates upon endogenous eIF4G small interfering RNA (siRNA)-mediated knockdown (Coldwell and Morley, 2006). eIF4GII uses multiple promoters and alternative splicing events as well as a non-canonical CUG AIC to translate multiple isoforms with different N-termini (Coldwell et al., 2012). In addition, BAG-1 is a molecular chaperone regulator which interacts with Bcl-2 to suppress apoptosis and utilises an upstream CUG and downstream AUG initiation codon to translate isoforms p50 and p36 respectively. These isoforms localise differently within the cell, suggesting an alternative function (Packham et al., 1997). BAG-1 has also utilises an IRES element following heat shock to maintain expression of the p36 isoform (Coldwell et al., 2001).

Large scale prediction of novel AICs has been widely carried out via numerous *in silico* bioinformatic databases which use the context of the AIC to predict likelihood of translation taking place. On the other hand, advances in next generation sequencing has enabled identification of *in vivo* translation initiation sites by ribosome profiling. This investigation however used an experimentally-informed pipeline to identify novel AICs (Cowan et al, manuscript in preparation). Initially, a pICtest reporter system was utilised to analyse the efficiency of near-cognate non-AUG translation, to increase the stringency of bioinformatic prediction, Novel AUG, CUG, GUG and ACG AICs were subsequently identified via a bioinformatics pipeline utilising the Ensemble genome database to identify conserved translated N-terminal eORFs of more than 40 amino acids. The bioinformatic pathway used all available transcriptome data, unlike ribosome profiling which is limited by tissue/cell type. A similar approach was taken by Ivanov, who also analysed conserved translated 5'UTRs (Ivanov et al., 2011). The predicted AICs were subsequently screened in mammalian cell culture.

The bioinformatic pipeline identified GATAD1 as a candidate gene, which was investigated throughout this work. Initiation from an upstream, in-frame AIC results in the translation of an N-terminally extended protein (Figure 1-6). The N-terminal extension has the potential to direct the protein to a different subcellular location, provide a novel protein binding site and alter its function. Most genes which use AICs encode regulatory proteins, such as transcription factors, kinases, growth factors and proto-oncogenes (Touriol et al., 2003). This demonstrates that alternative initiation has a cellular impact, expanding the proteome and translating novel protein isoforms which may localise and function differently within the cell.

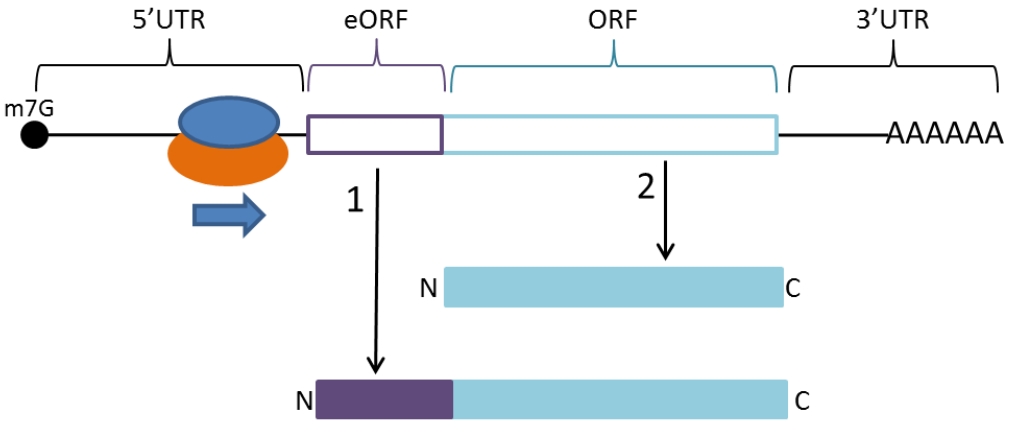


Figure 1-6: Alternative Translation Initiation

The 48S ribosome scans along the mRNA in search of a start codon. When the upstream AIC is recognised, the ribosome translates the eORF as well as the standard ORF, generating an N-terminally extended protein (1). When the ribosome recognises the canonical AUG start site, the wild-type protein is translated (2).

1.8 GATA Zinc Finger Domain-Containing Protein 1 (GATAD1)

GATA Zinc Finger Domain-Containing 1 (GATAD1) is also known as Ocular Development-Associated Gene (ODAG) and is located on chromosome seven. The protein encoded by the GATAD1 gene is ubiquitously expressed and contains a GATA zinc finger, as found in the GATA family of transcription factors. The GATA factors (GATA1-6) bind to DNA sequences with the consensus sequence (A/T)GATA(A/G), through their zinc finger domains (Znf). The zinc finger domains within this family of transcription factors have a highly conserved sequence identity (Ko and Engel, 1993); with the C-terminal zinc finger responsible for DNA binding and the N-terminal zinc finger involved in stabilising the protein-DNA complex formation (Ko and Engel, 1993). The NMR structure of GATA-1 shows that the central DNA-binding domain is composed of a zinc molecule coordinated by four cysteine residues and a carboxy-terminal tail. There are two binding regions associated with the zinc ion, following the form Cys-X-X-Cys-$(X)_{17}$-Cys-X-X-Cys which separate the binding regions by 29 residues (Omichinski et al., 1993). However, a single N-terminal zinc finger domain is present in GATAD1, located between amino acid 9 and 23 in the protein, which is composed of 269 amino acids when translated from the canonical start codon. Since GATAD1 lacks the C-terminal zinc finger present in the other GATA factors which is thought to be required for DNA binding, it is thought that GATAD1 regulates transcription indirectly through complex formation, rather than directly by DNA-binding through the GATA binding motif.

In addition to GATAD1, the paralogues GATAD2A (P66A) and GATAD2B (P66B) also contain a single GATA zinc finger; however, these are found at the C-terminal end of the proteins, not the N-terminal as is the case for GATAD1. These proteins are transcriptional repressors which act together, to enhance methyl-CpG-binding domain protein 2 (MBD2) mediated repression (Brackertz et al., 2002).

Mutations in GATAD1 are implicated in autosomal recessive dilated cardiomyopathy (DCM). Patients with DCM suffer from cardiac enlargement and hypertrophy of the left ventricle, causing systolic pump impairment. This can occasionally result in cardiac failure, although patients are usually able to live relatively normal lives. Theis et al (2011) carried out genome-wide linkage analysis on a family suffering from DCM to identify region 7q21 as the locus for this disease. Genome-wide homozygosity mapping then enabled fine mapping of the region containing the DCM gene in these patients; this was followed by exome sequencing of the genomic DNA of the siblings along with a filtering process to identify a mutation in GATAD1 as the solitary cause of DCM (Theis et al., 2011). A missense mutation resulting from a thymine to cytosine substitution at c.304 of exon 2 in 7q21, resulted in a S102P mutation in the GATAD1 protein of DCM sufferers. As well as this mutation, it was also shown that knock-down of the HDAC1/2 binding partners of GATAD1 also resulted in DCM in murine hearts (Theis et al., 2011). DCM sufferers have a different epigenetic

profile in their left ventricular myocardium when compared to healthy cardiac tissue, suggesting that GATAD1 complex interacts with/causes these histone modifications which lead to the disease state. Immunohistochemistry was carried out on left ventricular tissue of healthy patients versus DCM sufferers, using a monoclonal antibody against GATAD1. In healthy tissue GATAD1 was shown to localise to the nuclei/cytoplasm and exhibited a homogenous, striated pattern; this pattern was disturbed in tissue holding the S102P mutation. Mislocalisation of the mutant form of GATAD1 could contribute towards its disease causing capabilities in the left ventricular myocardium (Theis et al., 2011).

Experiments on mice have demonstrated that ODAG (GATAD1) is also implicated in development of the eye; mouse ODAG has a 92% similarity to human ODAG at the amino acid level. *In situ* hybridisation showed that expression of ODAG in many tissues in the eye was high at P2 (postnatal day 2). By P10 however, expression of ODAG was down-regulated dramatically and by P14 no expression could be detected whatsoever (Tsuruga et al., 2002). Work by Sasaki (2009) involving overexpressing ODAG in the eyes of transgenic mice demonstrated impaired ocular development; as well as optic nerve atrophy and elevated intraocular pressure, which is the main risk involved in developing glaucoma. Pull-down studies identified Rab6 and Rab6-GAP as binding partners of ODAG, suggesting ODAG is implicated in interfering with the Rab6/Rab6-GAP mediated signalling pathway. Along with other functions, Rab6 has been shown to assist in transporting newly synthesised rhodopsin from the trans-Golgi network to the outer rod segment in the eye (Sasaki et al., 2009). A further implication of GATAD1 in development has been identified in primates whereby GATAD1 represses the transcription of two key puberty-related genes, KISS1 and TAC3 (Lomniczi et al., 2015).

Recent work has shown that the GATAD1 gene is itself subject to epigenetic regulation in placental syncytiotrophoblasts – specialised transporting multinucleate epithelial cells, (Xiaoling Ma, 2014). GATAD1 contains CpG islands within both the 5 prime and 3 prime regions of the gene, which are known to subject the area to epigenetic modifications. The DNA methylation state was assessed using COBRA (Combined Bisulphite Restriction Analysis); the 5 prime region is largely unmethylated in all of the placental samples, whereas the 3 prime region is heavily methylated. A positive correlation was shown to exist between the 3 prime methylation levels and the GATAD1 RNA/protein expression levels. A 5-fold increase in GATAD1 expression was observed between healthy first-trimester compared to healthy third-trimester placentas; which was accompanied by a significant increase in 3' methylation of the GATAD1 gene. Interestingly, when considering preeclamptic placentas (the major cause of maternal and fetal mortality in pregnancy, associated with high blood pressure and proteinurea), a 2-fold decrease in GATAD1 expression was observed between healthy third-trimester compared to preeclampic third-trimester placentas; which was accompanied by significantly lower 3' methylation of the GATAD1 gene. The dramatic increase in GATAD1 expression in maturing placentas as well as the significant decrease in expression between

healthy and preeclamptic placentas suggests that GATAD1 is likely to play an important role in the physiology of the placenta as well as the pathophysiology of preeclampsia. DNA methylation of promoter sequences is typically a repressive mark associated with gene inactivation, yet DNA methylation of the gene body is more often associated with gene activation. As seen in GATAD1, gene-body methylation can be positively correlated with gene expression (Jjingo et al., 2012).

1.8.1 GATAD1 Chromatin Complex

It is thought that GATAD1 has a role in transcriptional repression (Islam et al., 2011); GATAD1 is able to regulate the expression of specific genes by forming an interaction with the stable trimethyl marker of lysine four on histone three (H3K4me3). The interaction of GATAD1 with the H3K4me3 modification is indirect and the contact is made through a chromatin 'reader'. GATAD1 binds to the lysine demethylase Jarid1A (Jumonji, AT-rich interactive domain 1), also known as KDM5A (lysine demethylase 5A) RBBP2 (retinoblastoma binding protein 2), which contacts the H3K4me3 histone modification site through a PHD (plant homeo domain) finger (Levy and Gozani, 2010). Histone modifications are a key method of chromatin regulation, determining how actively genes are transcribed in the nucleus; therefore any modification of this event is likely to affect the expression levels of many genes under the control of a single promoter.

A green-fluorescent protein (GFP) pull-down approach was used by Vermeulen et al in order to identify potential GATAD1 interactors; HeLa Kyoto cells expressing GFP-GATAD1 were stable isotope labelled by amino acids in cell culture (SILAC) and affinity purified using GFP-nanotrap beads before the proteins interacting with the bait GATAD1 were identified (Vermeulen et al., 2010). Mass spec analysis of the interacting proteins identified that the breast cancer associated protein EMSY as well as the Sin3/HDAC deacetylase complex also interact with GATAD1 when it is bound to Jarid1A. These proteins were present in similar ratios, suggesting the formation of a chromatin reading complex which may therefore repress transcription through consecutive demethylation of H3K4me3 as well as deacetylation of specific histone 3 acetylation sites (Figure 1-7) (Lee et al., 2006).

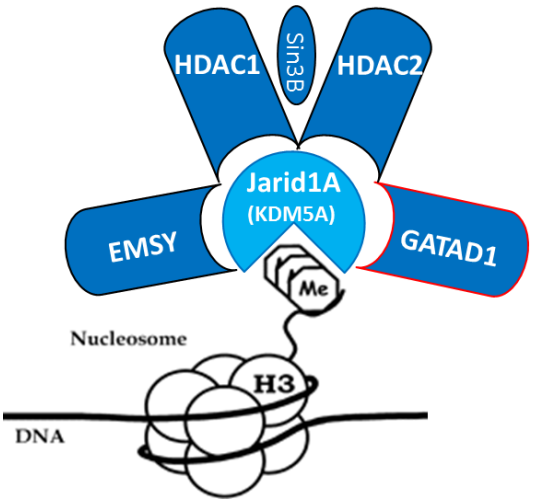


Figure 1-7: GATAD1 Interactors

Pull down studies by Vermeulen et al. (2010) showed that GATAD1 forms a complex with the Sin3B complex (including HDAC1 and HDAC2) as well as EMSY and Jarid1A, all of which have been implicated in transcriptional repression.

1.8.1.1 Jarid1A/KDM5A/RBP2

Jarid1A (Jumonji/AT-rich interactive domain-containing protein, also known as KDM5A (lysine demethylase 5A) is part of the Jarid family of four members, Jarid1A-D. These proteins contain a Jumonji C (JmjC) domain which catalyses the demethylation of mono-, di-, and tri-methylated lysine residues via an oxidative reaction (Accari and Fisher, 2015). JmjC-containing proteins represent the largest class of histone demethylases, with at least 30 proteins across 7 families. Jarid1A forms the direct link to the H3K4me3 histone mark within the chromatin reader complex, with a dissociation constant (K_d) of 0.75 μM (Wang et al., 2009). This histone contact is made through the last of 3 PHD (plant homeodomain) fingers (PHD1-3). The carboxyl-terminal PHD finger specifically recognises the methylation status of the lysine residue, while the amino-terminal PHD finger preferentially recognises the unmodified lysine on histone 3 (H3K4me0) (Torres et al., 2015). Recognition of H3K4me0 by PHD1 initiates a positive feedback loop whereby the enzymic activity of Jarid1A is increased. Jarid1A contains both a reader domain and a catalytic domain which enables increased demethylation to occur through the JmjC catalytic domain (Torres et al., 2015). The first histone demethylase to be identified in 2004 was LSD1 (KDM1A) , which is able to demethylate di-methylated H3K4 (Shi and Shi, 2004). Before the discovery of LSD1, methylated histones were thought to be enzymatically irreversible (Shi and Whetstine, 2007).

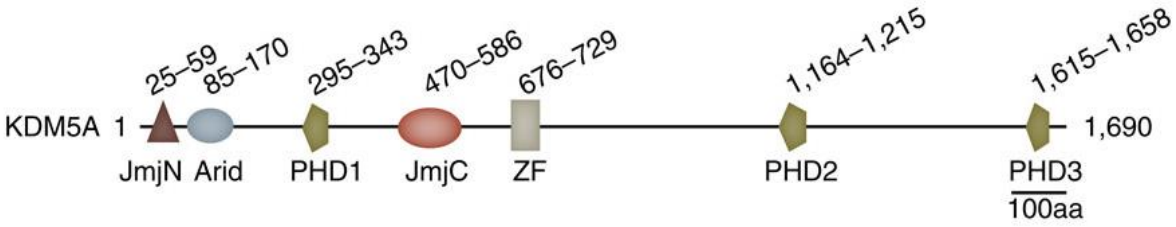


Figure 1-8: Jarid1A Domain Structure

The JmjN/C domains are common to the jumonji family of transcription factors; with the JmjC domain being the catalytic domain of the protein, catalysing the oxidative demethylation reaction. Jarid1A contains 3 PHD fingers; PHD 3 recognises the methylation status of H3K4. The single zinc finger may have a role in DNA binding. Image from: (Torres et al., 2015)

There are 5 main families of histone proteins; H2A, H2B, H3 and H4 are known as the core histones, whilst H1/H5 are known as linker histones. The octameric histone core is formed of two H2A-H2B dimers and a H3-H4 tetramer, around which 147 base pairs of DNA can wind to form the nucleosome (Marino-Ramirez et al., 2005). Each nucleosome is linked by either a H1 or H5 linker histone which locks the DNA in place. The N-terminal tail of the core histone proteins is the predominant site for histone post-translational modifications which can include methylation, acetylation, phosphorylation and ubiquitination, amongst others. Each of these modifications can affect the interaction between the protein and DNA in a different way depending on the type and position of the modification, which can alter processes such as gene regulation.

Jarid1A is a histone demethylase specific for the H3K4me3 site. Unlike acetylation of histones which promotes transcription by removing the positive charge on the histone and therefore decreasing the histone-DNA interaction, histone methylation has no effect on charge and acts indirectly to either promote or repress transcription from a certain site. Lysine residues can be mono-, di- and tri-methylated. Tri-methylation of lysine 4 on histone 3 is tightly associated with the promoters of active genes and so is an active mark for transcription (Lachner and Jenuwein, 2002). Therefore, when Jarid1A is able to remove this mark, the active signal is lost and transcription from this promoter is repressed.

Demethylation of histones by Jarid1A potentiates nuclear hormone receptor-mediated transcription (Chan and Hong, 2001); it also causes retinoblastoma-mediated gene silencing through direct binding to retinoblastoma protein during cellular senescence (Chicas et al., 2012) and is

implicated in regulating expression of the progesterone receptor (Stratmann and Haendler, 2011). As well as having a major role in the maintenance of the circadian clock (DiTacchio et al., 2011), Jarid1A has been shown to bind numerous proteins as well as the retinoblastoma protein, including rhombotin-2 which regulates red blood cell development and is implicated in T-cell mediated leukaemia (Hayakawa et al., 2007).

Certain histone demethylase enzymes have been found to function in the cytoplasm, away from their standard chromatin target. KDM4A, like KDM5A (Jarid1A) is also a JmjC domain-containing protein, which de-methylates two sites on histone 3, H3K9me3 and H3K36me3 when in the nucleus. KDM4A has been found to unexpectedly function in the cytoplasm, where it interacts with the translational machinery and affects initiation factor distribution (Van Rechem et al., 2015). Since KDM4A was found to associate with the initiation complex, polysome fractions were analysed for KDM5A (Jarid1A), which was found to be present in both the 40S and the 60S fractions, suggesting a potential similar cytoplasmic function.

1.8.1.2 Sin3B/HDAC and EMSY

Work by Vermeulen et al, 2010 found that subunits of the Sin3/HDAC deacetylase complex were also complexed with GATAD1 (Vermeulen et al., 2010), suggesting a possible link between the demethylation and deacetylation of histone tail residues in order to repress transcription. Amine groups normally present on lysine and arginine residues give histone tails a positive charge, allowing tight association of negatively charged phosphate groups of the DNA with histone proteins, repressing transcription. Histone acetylation neutralises this charge interaction by forming an amide in place of an amine group, allowing the chromatin structure to relax and so promoting transcription. HDAC enzymes specifically remove acetyl groups from lysine ε-amino residues therefore repressing transcription (de Ruijter et al., 2003).

Jarid1A forms the direct link to the H3K4me3 mark, and has been shown through immunoprecipitation experiments to complex with both the Sin3B/HDAC complex (Hayakawa et al., 2007) as well as GATAD1 (Vermeulen et al., 2010). It has been suggested that histones are cooperatively deacetylated by the HDACs and demethylated by Jarid1A; it is also likely that this complex is able to reposition the nucleosomes for more efficient transcription repression (Hayakawa and Nakayama, 2011). H3K4 is indeed an acetylated residue and this modification is enriched at the promoters of actively transcribed genes (Guillemette et al., 2011). Therefore, it is likely that the GATAD1 complex is involved in the mutual demethylation as well as deacetylation of the H3K4 residue, leading to repression of transcription.

Several HDAC complexes including Sin3 and NuRD, contain the same core proteins; including HDAC1, HDAC2, RbAp46 (RBBP7) and RbAp48 (RBBP4). The Sin3B complex additionally contains Sin3B (Swi-*in*dependent 3B), SAP18 and SAP30 (Ahringer, 2000)

Figure 1-9). Although it is unknown which of these proteins, if any, are directly bound to GATAD1, it has been shown that there is an association between the Sin3B complex, KDM5A and GATAD1 (Hayakawa et al., 2007). Both RBBP4 and RBBP7 are histone binding proteins often found in HDAC complexes; Sin3B is a HDAC which also acts as a large protein scaffold, with a central HID (HDAC Interaction Domain) which functions as a binding site for many transcriptionally repressive HDAC proteins (Grzenda et al., 2009). SAP18 (Sin3A-Associated-Protein 18kDa) and SAP30 (Sin3A-Associated-Protein 30kDa) enhance Sin3-mediated transcriptional repression.

EMSY has previously been pulled down with GATAD1 in the repressive chromatin remodelling complex (Vermeulen et al., 2010). As well as repressing transcription through complexes at gene promoters, EMSY also binds to and inactivates BRCA2, a tumour suppressor gene responsible for repairing DNA. This results in an increased risk of breast cancer (Hou et al., 2014).

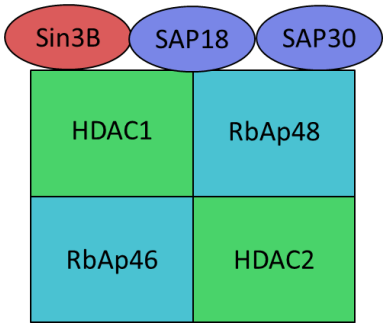


Figure 1-9: Sin3B/HDAC Complex

The histone deacetylase Sin3B complex is composed of HDAC1 and HDAC2, as well as Sin3B, SAP18 and SAP30.

1.9 Project Aims

Initial project aims are to confirm whether alternative translation initiation is taking place within the GATAD1 mRNA transcript and to determine which AICs are responsible for this translation. Sequences or factors influencing AIC choice will also be analysed in order to understand the translational regulation of each isoform. The function of the alternative isoforms will then be determined; the N-terminal extension of the alternative isoforms may result in relocalisation of GATAD1 within the cell, potentially resulting in the formation of different protein interactions compared to the annotated GATAD1 isoform.

Chapter Two

Methods

2. Methods

2.1 Bioinformatics

The candidate gene for this investigation, GATAD1, was identified using a macro designed by Dr. Richard Edwards, which used the Ensemble genome database to search for eukaryotic genes with N-terminal extensions of >40 amino acids. Possible AIC's were then identified by screening the translated 5' UTRs of >9000 shortlisted human genes. The 5'UTRs of the further shortlisted genes were aligned to murine 5'UTRs, and then other species; the 142 genes which contained conserved 5'UTRs were shortlisted as candidates. These were analysed gene-by-gene; ATGs in any context were identified, whilst AICs in the form of ACG, CTG and GTG were only identified when in a strong context (Figure 2-1). Further potential AICs were then identified by looking through the sequence manually in order to flag potential AICs which were not in a strong Kozak consensus and therefore not identified by the macro.

Annotated	Alternative	Description
Strong	Strong	ATG codon with purine at -3 and guanine at +4.
MidR	MidR	ATG codon with purine at -3 only.
MidG	MidG	ATG codon with guanine at +4 only.
Weak	Weak	ATG codon with neither purine at -3 nor guanine at +4.
ACG	ACG	ACG codon with purine at -3 and guanine at +4.
CTG	CTG	CTG codon with purine at -3 and guanine at +4.
GTG	GTG	GTG codon with purine at -3 and guanine at +4.

Figure 2-1: Identification of Potential Initiation Codons using a Macro

ATG codons were identified by the macro within a strong (coloured yellow), midR (green), midG (green) or weak context (orange). The most efficient AICs; ACG, CTG and GTG were also identified when in a strong Kozak consensus (blue).

Molecular Biology Techniques

2.2 RNA Extraction (Nucleospin RNA-Machery-Nagel)

In an RNase-free environment, cells were washed once with ice-cold PBS and then scraped in PBS from 6 cm plates. Cells were pelleted by centrifugation at 500x *g* for 5 minutes at 4°C and all supernatant was removed. The cells were then lysed with 350 µL buffer RA1 and 3.5 µL β-ME and vortexed vigorously. The lysate was passed through a NucleoSpin filter into a collection tube by centrifugation at 11,000x *g* for 1 minute and 350 µL 70% EtOH was added to the collection tube to adjust the RNA binding conditions. Once mixed thoroughly, the lysate was loaded on to an RNA binding column and centrifuged at 11,000x *g* for 30 seconds. 350 µL MDB (membrane desalting buffer) was added to the membrane which was centrifuged again at 11,000x *g* for 1 minute. DNA was removed from the membrane by incubation with 80 µL of reconstituted DNase at room temperature for 15 minutes. After incubation, 200 µL RAW2 buffer was added to the column to inactivate the DNase before being centrifuged at 11,000x *g* for 30 seconds. The membranes were then washed twice with 600 µL followed by 250 µL RA3 buffer, centrifuged at 11,000x *g* for 1 minute each time. The membrane was then dried by centrifugation at 11,000x *g* for 2 minutes, before the RNA was eluted in 60 µL RNase-free H_2O into a nuclease-free collection tube. Yield and concentration was analysed by Nanodrop spectrophotometer.

2.3 Reverse Transcription (ImProm-II RT System, Promega)

Complementary DNA (cDNA) was synthesised from HEK293 RNA, for PCR use. 750 ng RNA was combined with 1 µL supplied oligo(dT)15 primer (0.5 µg) on ice for each experimental reaction, and made up to 5 µL with nuclease-free H_2O. The positive control reaction combined 2 µL (10 ng) of 1.2 kb kanamycin positive control RNA with 0.5 µg oligo(dT)15 primer and 2 µL nuclease-free H_2O; whilst the negative (no template) control combined 0.5 µg oligo(dT)15 primer and 4 µL nuclease-free H_2O only. The reactions were incubated at 70˚C for 5 minutes to denature RNA secondary structure, before being placed immediately on ice for a further 5 minutes, allowing the primers to anneal to the RNA. The experimental reactions (Table 2-1), positive control and negative (no reverse transcriptase) control were then prepared, added to the 5 µL of annealed RNA/primer mix, which was incubated at 25˚C for 5 minutes, extended at 42˚C for 1 hour and the reverse transcriptase was inactivated by heating at 70˚C for 15 minutes.

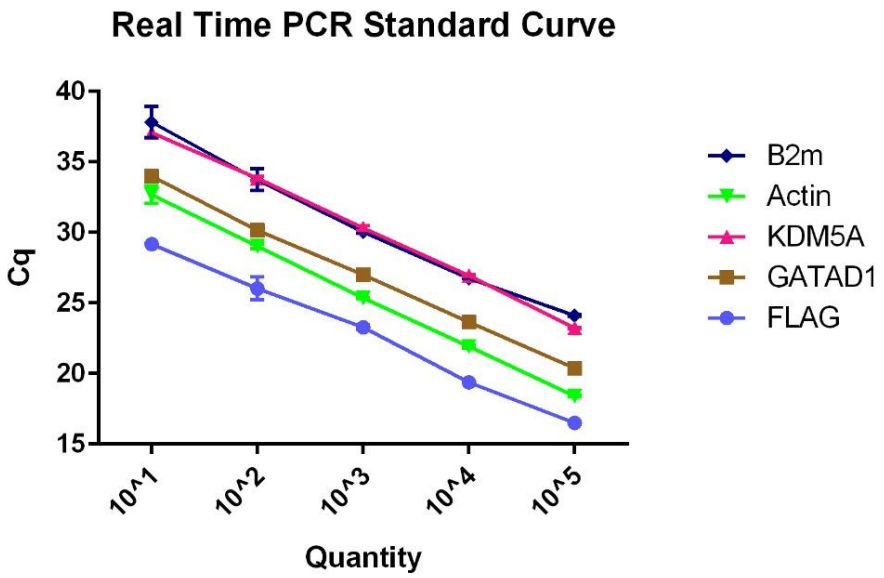
Table 2-1: Components of the Reverse Transcription Experimental Reaction

Final amount/concentration when in a final 20 µL volume. All components of the reaction were supplied by Promega with the ImProm-II RT System)

Experimental Reaction	Volume of Reactant	Final amount/ Concentration
Nuclease-free water	4.5	-
ImProm-II 5x reaction buffer	4.0	1x
$MgCl_2$	4.0	5 mM
dNTP mix	1.0	0.5 mM each dNTP
Rnasin ribonuclease inhibitor	0.5	20 U
ImProm-II reverse transcriptase	1.0	
Final volume	**15.0**	

2.4 Quantitative PCR (qPCR) – SYBR Green

Initially, the PCR amplification efficiency was calculated for each set of primers (Figure 2-2). qPCR enabled gene expression quantification (investigating mRNA levels). 20 µL qPCR reactions were made up of 10 µL SYBR green master mix solution (Thermo Fisher), 1 µL 10 µM forward primer, 1 µL 10 µM reverse primer, 1 µL cDNA (1 in 10 dilution) and 7 µL dH_2O. The reactions were made in triplicate for each sample and primer set (Table 2-2), along with a no template control reaction. The 20 µL reactions were then loaded into a 48-well qPCR plate which was sealed and spun down at 600x *g* for 1 minute. The plate was analysed using the Illumina Eco qPCR thermal cycler and Eco software.



Assay	Efficiency	Equation	R^2
Actin	90.79 ± 0%	y=-3.56x+36.18	0.998
GATAD1	97.98 ± 0%	y=-3.37x+37.15	0.998
KDM5A	94.03 ± 0%	y=-3.47x+40.73	0.999
FLAG	103.95 ± 0%	y=-3.23x+32.74	0.996
B2M DNA	97.31 ± 0%	y=-3.39x+40.61	0.990

Figure 2-2: Real Time PCR Standard Curve

A 100% efficient reaction will have a 10-fold increase in PCR amplicon every 3.32 cycles during the exponential phase of amplification. Therefore, the reactions with a gradient more negative than -3.32 (Actin, GATAD1, KDM5A) are less than 100% efficient, whilst the FLAG reaction has an efficiency of over 100%. Error bars indicate the standard deviation (n=2).

Table 2-2: qPCR Primer Sequences

Both GATAD1 and KDM5A qPCR primers (Eurofins MWG) targeted a cDNA sequence corresponding to the 3'UTR of the transcripts. The FLAG qPCR primers amplified a region of the 3xFLAG-tag as well as part of the 3'UTR transcribed from the vector before the polyadenylation signal.

qPCR Target	NCBI Ref Sequence	qPCR Primer Sequence (5'-3')	
		Forward	*Reverse*
FLAG	-	CAAGGATGACGATGACAAGTAG	AGGGGCAAACAACAGATGG
GATAD1	NM_021167.4	CCCCGTCGCTACTAAAAATAC	CTCACTACAACCTCCACCTC
KDM5A	NM_001042603.2	AGCAAGACCAGCAAATAACAG	ACTACAACCCCACATCCTC
β-actin	NM_001101.3	CTCGCCTTTGCCGATCC	CATCATCCATGGTGAGCTGG
Beta-2-microglobulin	NM_004048.2	CCCCCACTGAAAAAGATGAG	ATCCAATCCAAATGCGGC

2.5 Polymerase Chain Reaction (PCR)

A full primer list can be found in Supplementary Data, chapter 10. Primers were made by Sigma-Aldrich (Gillingham, UK) or Eurofins MWG (qPCR primers) (Frome, UK). Primers were supplied lyophilised, before being resuspended and diluted before use. PCR enabled the amplification of certain genes using sequence-specific primers, which could then be cloned into expression vectors. Similarly, site-directed mutagenesis enabled specific changes to the DNA sequence of the gene using mutagenic primers.

2.5.1 Phusion High-Fidelity PCR (NEB)

Phusion DNA Polymerase (NEB, London, UK) is a high-fidelity enzyme which is more than 50x more accurate than Taq DNA Polymerase. Phusion PCR reactions were made as follows: cDNA was amplified in 25 µL reactions containing 16.25 µL nuclease free H_2O, 5 µL 5x Phusion HF buffer, 0.5 µL 10 mM dNTPs, 1.25 µL of 10 µM forward and reverse primers (section 10.1), 0.5 µL template DNA (10-100 ng) and 0.25 µL Phusion DNA Polymerase. 1 M Betaine (Sigma) was also added to reactions containing GC-rich transcripts.

An initial 98˚C hold for 30 seconds heat-activated the Phusion DNA Polymerase before the PCR reactions were thermocycled (Table 2-3). The DNA was denatured at 98˚C, before being cooled to a primer-specific temperature for 25 seconds, allowing primers to anneal to the template DNA. Extension of the DNA strands by Phusion DNA Polymerase then took place at 72˚C (Figure

2-3). A final extension for 10 minutes at 72˚C ensured all single stranded DNA has been elongated and was followed by a 4˚C hold for storage purposes.

Table 2-3: Phusion PCR Thermocycling Conditions

Step	Temperature (˚C)	Duration
Initial denaturation	98	30 s
30 cycles	98	10 s
	45-72	25 s
	72	20 s/kb of amplicon
Final extension	72	10 mins
Hold	4	-

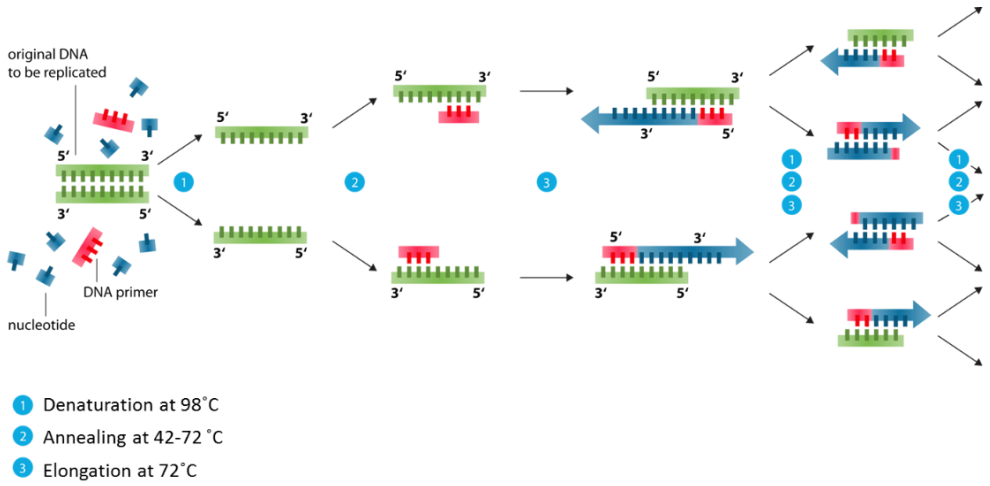


Figure 2-3: Polymerase Chain Reaction Amplification

Illustration of the main steps in PCR which enables amplification of the target sequence from template DNA; DNA is initially denatured at 98ºC, primers are then annealed at a primer-dependent temperature, before the polymerase incorporates complementary dNTPs in a 5'-3' direction at 72ºC.

2.5.1.1 QuikChange Lightning (Agilent) Site-Directed Mutagenesis

50 µL sample reactions were prepared, consisting of 5 µL 10x reaction buffer, 10-100 ng DNA template, 125 ng forward and reverse complementary mutagenic primer, 1 µL dNTP mix and 1.5 µL QuikSolution Reagent, made up to final volume with dH_2O. 1 µL QuikChange Lightning PfuUltra polymerase was then added. Tubes were briefly spun before being thermocycled (Table 2-4).

The DNA was initially denatured, allowing mutagenic primers to anneal at 68˚C. The mutant strand was then synthesised by a PfuUltra high-fidelity DNA polymerase, generating a mutant plasmid containing staggered nicks, as well as the parental plasmid. The endonuclease DpnI digested the methylated/hemimethylated parental plasmid and the nicked plasmid product was transformed directly into 30 µL XL10- Gold ultracompetent cells (Agilent, South Queensferry, UK) with 2 µL β-ME, where the nicks were repaired (Figure 2-4).

Table 2-4: QuikChange PCR Thermocycling Conditions

Step	Temperature (˚C)	Duration
Initial denaturation	95	2 mins
30 cycles	95	20 s
	68	10 s
	68	30 s/kb of plasmid
Final extension	72	5 mins
Hold	4	-
DpnI Treatment	37	60 mins

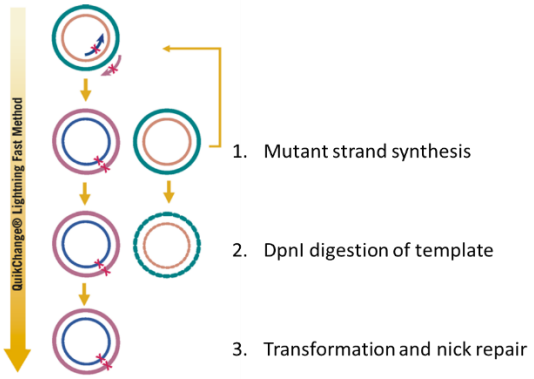


Figure 2-4: QuikChange Lightning PCR

Image from Agilent QuikChange Lightning Site-Directed Mutagenesis Kit instruction manual.

2.6 PCR Purification (Machery-Nagel Nucleospin Extract II Kit)

The PCR purification kit was used to remove unused primers, dNTPs, enzyme and other impurities from the DNA sample. 200 µL Buffer NT was added to 100 µL of PCR reaction, mixed and loaded into a spin column and centrifuged at 11,000x *g* for one minute, leaving the DNA bound to the membrane. The membrane was washed twice with 700 µL Buffer NT3 and centrifuged at 11,000x *g* for one minute each time. The column was dried by centrifugation at 11,000x *g* for 2 minutes before the DNA was eluted in 30 µL Buffer NE.

2.7 Gel Electrophoresis

TAE Buffer (pH8)*: 1x solution containing 40 mM Tris base (Fisher Scientific), 20 mM NaOAc (Sigma), 26.9 mM acetic acid (Sigma) and 2mM EDTA (Sigma)*

6x DNA Loading Dye*: 30% (v/v) glycerol (Fisher BioReagents), 0.25% (w/v) bromophenol blue (Sigma), 0.25% (w/v) xylene cyanol FF (Sigma)*

A 0.8% agarose gel required 0.8 g of agarose (Fisher Scientific) dissolved in 100 mL 1x TAE buffer. Once cool, 3 µL/100 mL of GelRed stain (Biotium, Cambridge, UK) was added before the gel was poured into a casting tray with 2 dams and a comb. Once set, the dams and comb were removed and the gel was immersed in 1x TAE buffer. 6x DNA loading dye was added to each sample to make a final 1x concentration (e.g. 2 µl to a 10 µl reaction), which were loaded alongside 10 µL 1:10 GeneRuler DNA ladder mix (Thermo Fisher). The gel was run at 120 V for 40 minutes, before the bands were visualised with a UV transilluminator.

2.7.1 Gel Purification (NucleoSpin Gel and PCR Clean-Up, Macherey-Nagel)

To allow separation of one fragment from another, the DNA was run on an agarose gel before the required band was purified. 200 µL Buffer NT was added to every 100 mg of agarose gel slice and incubated at 50˚C until the gel was completely dissolved. The sample was loaded into a spin column before the PCR clean up protocol as seen in 2.6 was followed in order to purify the DNA.

2.8 Cloning Digestion

The pc_DNA_3F vector (Coldwell et al, 2012) was used as vector backbone throughout this thesis, unless otherwise stated (Figure 2-5). Both the vector and the PCR products to clone were digested in 50 µL reactions containing DNA (1 µg vector/5 µL PCR product), enzyme (NEB or Promega, Southampton, UK) and the recommended buffer for a double digestion. The samples were centrifuged briefly and incubated at 37˚C for 90 minutes.

To prevent re-ligation of the vector, 2 µL shrimp alkaline phosphatase (T-SAP) (Promega) and 3 µL Multicore 10x buffer (Promega) was subsequently added to the digest and incubated at 37˚C for a further 30 minutes. The products of the vector digest were resolved on a 0.8% agarose gel and the upper band excised using scalpel. DNA from the gel slice was then purified following the protocol in 2.7.1.

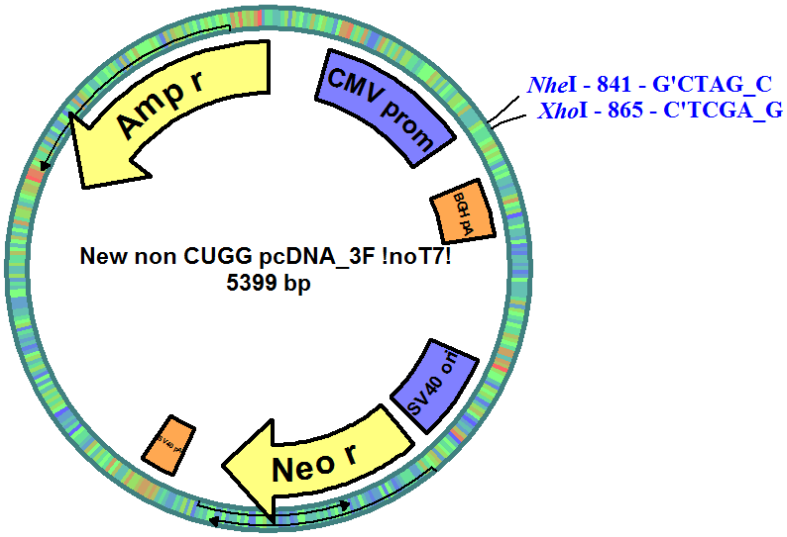


Figure 2-5: pcDNA_3F Vector

A virtual pcDNA_3F plasmid created using pDraw DNA analysis software. PCR products were cloned in using NheI/XhoI restriction sites. The plasmid has both ampicillin and neomycin resistance genes, as well as a CMV promoter to drive expression of the inserted gene. The plasmid also contains a 3xFLAG-tag which is utilised during an immunoblot to identify the expression of the insert, and transcripts include a polyadenylation signal from Bovine Growth Hormone to ensure efficient translation.

2.9 T4 Polynucleotide Kinase (PNK) Treatment of DNA Ends

Complementary oligonucleotides which do not require a restriction digest prior to cloning should be annealed and PNK-treated. PNK adds a phosphate group to the 5'-end of the DNA, ensuring efficient cloning. A 10 µL reaction mix was made as follows: 1 µL of both top and bottom oligo (100 µM), 1 µL T4 ligation buffer (NEB), 1 µL T4 PNK (NEB) and 6 µL dH_2O. The reaction mix was placed in a thermocycler set to the following parameters: 37°C for 30 minutes, 95°C for 5 minutes and then ramped down to 25°C at 5°C min^{-1}. The phosphorylated and annealed oligonucleotides were diluted 1:200 before subsequent ligation.

2.10 Ligation

10 µL reactions were prepared to ligate the digested DNA with the digested pcDNA_3F vector. The reactions included a no insert control, a no ligase control and both 1:1 as well as 3:1 insert:vector ratios and were left at room temperature for 60 minutes, or overnight at 4˚C (Table 2-5).

Table 2-5: Ligation Conditions

	No insert Control (µL)	No Ligase Control (µL)	1:1 Insert to Vector Ratio (µL)	3:1 Insert to Vector Ratio (µL)
Digested DNA insert	0	0	1	3
Digested pcDNA_3F	1	1	1	1
10x T4 Ligase Buffer	1	1	1	1
T4 Ligase	0.5	0	0.5	0.5
H_2O	7.5	8	6.5	4.5

2.10.1 Ligation into pGEM-Easy Vector (Promega)

This system was used in some cases to insert a blunt ended PCR product into a vector for amplification in order to improve the efficiency of subsequent cloning. A single 3'-terminal thymidine at both ends of the linearised pGEM-T Easy vector improves the efficiency of ligation of PCR products, by providing a compatible overhang generated by certain polymerases. A 10 µL reaction mixture was made to A-tail the PCR product, which consisted of 7 µL PCR product, 1 µL PeqLab Buffer S, 1 µL 2 mM dATPs and 1 µL Taq DNA Polymerase. The reaction was incubated at 70˚C for 30 minutes.

A 5 µL ligation was then set up, consisting of 2.5 µL 2x Rapid Ligation Buffer-T4 DNA ligase, 0.5 µL pGEM-T Easy vector, 0.5 µL A-tailed PCR product, 0.5 µL T4 DNA Ligase and 1 µL nuclease-free dH_2O. The ligation reactions were incubated at room temperature for 60 minutes, or overnight at 4˚C.

2.11 Transformation and Culture of DH5α Competent E. *coli*

***SOC Medium**: (Thermo Fisher) 0.5% Yeast Extract, 2% Tryptone, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$, 20 mM Glucose*

***Lysogeny Broth (LB)**: 10 g Tryptone (Sigma), 5 g Yeast Extract (Sigma), 10 g NaCl (Fisher Scientific), up to 1 L with dH_2O*

***LB Agar**: Dissolve 15 g Agar (Sigma) in 1 L LB*

2.11.1 Standard Transformation

5 µL of ligated plasmid DNA was added to 50 µL DH5α competent cells (Thermo Fisher) of subcloning efficiency, which were kept on ice for 30 minutes. The cells were then heat shocked for 30 seconds before being placed immediately back on ice. 150 µL of SOC media was then added to each transformation which was incubated with agitation at 37˚C for 60 minutes. The transformation was added to a pre-warmed agar plate supplemented with appropriate antibiotic and incubated at 37˚C overnight.

2.11.2 Transformation of pGEM-T Easy Vector

2 µL of ligated plasmid DNA was added to 50 µL DH5α cells. From here, the transformation protocol for 2.11.1 was followed until the transformation was spread onto the agar plate.

The MCS of a pGEM-T Easy Vector is present within the LacZ gene and so it is possible to screen for recombinants using blue-white selection. 100 µL of 100 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) and 20 µL of 50 mg/mL X-Gal was spread over the surface of a pre-warmed LB plate supplemented with the appropriate antibiotic and allowed to soak in for 30 minutes prior to use. The transformation was then added to the pre-warmed plate and incubated at 37˚C overnight. The white colonies produced when using blue-white selection contained the recombinant plasmid (Figure 2-6).

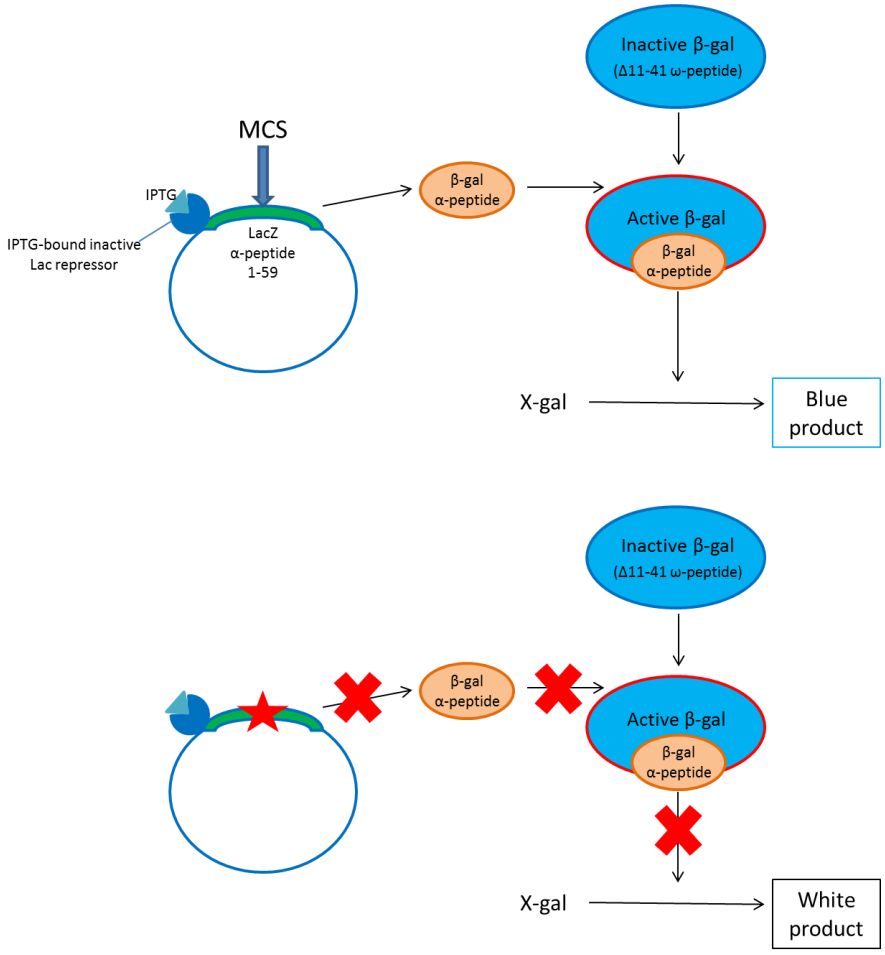


Figure 2-6: Blue-White Selection

Successful ligation into the MCS of the pGEM-T Easy vector prevents the production of the β-galactosidase α-peptide from the LacZ operon. α-complementation of the α-peptide and ω-peptide cannot then happen, and β-galactosidase remains inactive, unable to hydrolyse X-galactose into a blue product. The colonies from a successful ligation were therefore white.

2.12 Inoculation of E.*coli* in LB media

Single colonies of E. *coli* were selected from the agar plates and grown in LB media containing the appropriate antibiotic. If going on to purify the plasmid by midi-prep, day cultures were prepared in 5 mL LB for approximately 8 hours (logarithmic phase), before 500 µL of the culture was subsequently incubated overnight in 35 mL LB at 37˚C with constant agitation. For mini-prep purification, a single colony was grown overnight in 3 mL LB at 37˚C with constant agitation.

2.13 Plasmid Purification

2.13.1 Mini-Prep (NucleoSpin Kit, Machery-Nagel)

For small-scale isolation of plasmid DNA from the 3 mL E. *coli* culture, cells were pelleted by centrifugation for 1 minute at 11,000x *g*. The supernatant was discarded and the pellet was resuspended in 250 µL Buffer A1. Cells were lysed in 250 µL SDS/alkaline Buffer A2 which was mixed by gentle inversion 6-8 times. The reaction was then neutralised with 300 µL Buffer A3 which was mixed by gentle inversion. The lysate was clarified by centrifugation at 11,000x *g* for 5 minutes at room temperature and the supernatant was then loaded into a spin column and centrifuged for 1 minute at 11,000x *g* to bind the DNA. The silica membrane was washed with 500 µL Buffer AW, followed by 600 µL Buffer A4, centrifuging for 1 minute at 11,000x *g* each time. The membrane was dried with a 2 minute spin at 11,000x *g*, before the plasmid DNA was eluted in 50 µL Buffer AE after a final spin at 11,000x *g* for 1 minute. The yield, concentration and purity of the DNA were analysed with Nanodrop.

2.13.2 Midi-Prep (HiSpeed Plasmid Midi Kit, QIAGEN)

For larger-scale isolation of plasmid DNA from high copy number plasmids (up to 200 µg), the 35 mL culture was centrifuged at 6,000x *g* for 15 minutes at 4ºC to pellet the bacterial cells. The cells were resuspended in 6 mL of Buffer P1 and then lysed with 6 mL Buffer P2, which was mixed by vigorous inversion 4-6 times and incubated at room temperature for 5 minutes. Precipitation of genomic DNA, proteins and cell debris was enhanced with 6 mL of chilled Buffer P3, followed by inversion 4-6 times. The lysate was immediately loaded into the barrel of a QIAfilter Cartridge and incubated at room temperature for 10 minutes without inserting the plunger. A vacuum manifold was used for all subsequent supernatant and wash steps. 4 mL Buffer QBT was added to a HiSpeed Midi Tip to equilibrate the column, before the plunger was inserted into the QIAfilter Cartridge and

cell lysate was filtered into the HiSpeed Tip. The cleared resin entered the resin by gravity flow, before the HiSpeed Midi Tip was washed with 20 mL Buffer QC. The DNA was then eluted in 5 mL Buffer QF. DNA was precipitated by adding 3.5 mL isopropanol to the DNA which was centrifuged at 15,000x *g* for 30 minutes at 4ºC. The DNA pellet was then washed with 2 mL room temperature 70% ethanol, followed by a further 15,000x *g* spin for 10 minutes. The DNA was air-dried for 5 minutes before being resuspended in TE buffer. The yield, concentration and purity of the DNA were analysed with Nanodrop.

2.14 Diagnostic Digest

The recombinant plasmid DNA was digested with appropriate restriction enzymes to ensure that the correct insert was present. Digestions were set up consisting of 1 µL DNA (approximately 1-200 ng), 1 µL enzyme-specific buffer, and 0.25 µL restriction enzyme, made up to 10 µL with nuclease-free H_2O. The restriction digests were incubated at 37˚C for 90 minutes and analysed by agarose gel electrophoresis (Section 2.7). To confirm the presence of the insert, a sample was sent off to Eurofins MWG for sequencing.

Cell Culture Techniques

2.15 Cell Line Maintenance

***DMEM**: Dulbecco's Modified Eagle Medium (Thermo Fisher)*

***FBS**: Heat-inactivated foetal bovine serum (HyClone, Cramlington UK)*

***DPBS**: Dulbecco's Phosphate-Buffered Saline (1x), no Calcium, no Magnesium (Thermo Fisher)*

***TrypLE™ Express**: 1x solution in PBS containing 1mM EDTA (Thermo Fisher)*

Table 2-6: Cell Culture Media

Media	Formulation	Cell Line
DMEM (- Pyruvate)	High glucose, GlutaMAX™ Supplement, HEPES	HeLa, HEK293,
DMEM (+ Pyruvate)	High glucose, GlutaMAX™ Supplement, pyruvate	MDA-MB-231, MCF7, HaCaT
McCoy's 5A (Thermo Fisher)	High glucose, L-Glutamine, Bacto-peptone	HCT116
RPMI 1640 (Thermo Fisher)	L-Glutamine, Vitamins	H1299, PC-3

Sterile conditions were maintained in a laminar flow cabinet at all times when handling cells. Cell line stocks were maintained in sterile 75 cm^2 flasks containing appropriate media (Table 2-6) supplemented with 10% FBS. Passage of semi-confluent cells was carried out frequently to ensure 80% maximum confluency. This started with the removal of medium from the flask, before the cells were washed using 7 mL of Mg^{2+} and Ca^{2+}-free PBS. The PBS was then removed before 1.5 mL TrypLE (or Trypsin in the case of HaCaT and PC-3 cells) was added and the flask was incubated at 37˚C for enough time to encourage cell detachment (typically 5 minutes). Following incubation, fresh DMEM (+ FBS) was added to the appropriate volume of cells in a new flask for a 1:4 dilution, or 1:8 dilution over the weekend. The cells were incubated at 37˚C in 5% CO_2.

2.16 Transfection

Following dissociation by TrypLE or Trypsin (Section 2.15), cells were counted and plated out using a haemocytometer the day before transfection took place (Figure 2-7). 10 µL resuspended cell culture was pipetted under a haemocytometer coverslip and viewed under a microscope. Cells in all 4 corners of the haemocytometer were counted, and an average was found for each corner (0.1 µL). This figure was then multiplied by 10,000 to give an average value of cells per mL, which was used to calculate the volume of cell culture required to generate a stock solution of specific number and concentration of cells.

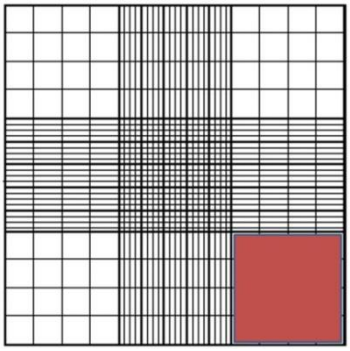


Figure 2-7: Haemocytometer

Under the microscope, cells were counted in all four 4x4 squares in each corner (one shown shaded red). An average number of cells was then calculated per corner, which is used to calculate the average number of cells per mL of media in the flask. The correct number of cells was then put into each plate.

Typically, a 6 cm plate was seeded with 100,000 cells for a transfection 24 hours later with 1 µg DNA, and this would be harvested after 72 hours, with experiments scaled up or down in different sized plates as required. The GeneJuice (Novagen, Watford, UK) transfection reagent was used, and the ratio of GeneJuice to DNA varies for different cell lines, but we typically used it at a ratio of 3 µL for every 1 µg of DNA. For example, to transfect HeLa cells with 1 µg DNA, 3 µL GeneJuice transfection reagent was added to 100 µL serum-free DMEM. This mixture was vortexed and incubated at room temperature for 5 minutes. 1 µg of DNA to be transfected was made up to 10 µL with dH_2O, added to the GeneJuice/DMEM mixture and incubated for 5 minutes at room temperature. 100 µL of this mixture was then added drop-wise to the plate which was incubated for 24-72 hours at 37˚C, 5% CO_2.

For cell lines that were difficult to transfect, an alternative transfection reagent – Lipofectamine LTX (Life Technologies) was used. DNA was diluted using both Opti-MEM and PLUS reagent, and added to Lipofectamine LTX diluted in Opti-MEM (5:1 DNA:LTX ratio). The DNA-lipid complexes were then added to the cells which were washed 6 hours later. Cells were then incubated as before.

2.17 CRISPR-Cas9

2.17.1 Co-transfection of CRISPR-Cas9 and HDR template

Cells were seeded at $1x10^5$ cells per 6 cm plate in a volume of 5 mL, 24 hours prior to transfection. 10 µg HDR targeting plasmid was linearised with PvuI (NEB), purified and eluted in 35 µL EB buffer (as in section 2.6). 1 µg pspCas9BB-gRNA plasmid was then mixed with 1 µg linearised pcDNA3-gBlock targeting plasmid in a total volume of 20 µL before the HEK293 cells were transfected as in section 2.16. After a 6 hour incubation, the cells were washed with DPBS and the complete medium was replaced. Puromycin selection was applied at a concentration of 2 µg/mL 24 hours post-transfection, for a total of 72 hours.

2.17.2 Isolation of clonal cell lines by dilution

Once puromycin selection had taken place, cells were allowed to recover for 48 hours. The transfected cells were then dissociated from the plate by gently washing two times with DPBS before adding 1 mL TrypLE Express. 3 mL complete media was then added before the cells were pelleted at 500x *g* for 5 minutes and resuspended in 10 mL DMEM. The cells were passed through a cell strainer two times before being counted (section 2.16) and subsequently serially diluted to a concentration of 0.5 cells per well (200 µL volume) of a 96-well plate. Five 96-well plates were seeded at this concentration.

One week post-plating, colonies were inspected for a clonal appearance and wells which were seeded with multiple cells were marked off. The cells were allowed to expand for a further 5 days before the clonal colonies were replica-plated in 96-well plates and grown to near confluence. One plate was frozen down; cells were washed once with DPBS and detached in 50 µL TrypLE followed by a 37°C incubation. The plate was then tapped gently to encourage dissociation and the cells were resuspended in 50 µL of cryo solution (80% FBS: 20% DMSO). A sealing film was put on the plate, which was also wrapped in bubble wrap and placed in a polystyrene box at -80°C until PCR analysis determined which colonies had incorporated the 3xFLAG tag and should be resurrected.

DNA was extracted from the second 96-well replica plate, allowing HDR to be detected by PCR screening of the modified region. The media was removed and cells were washed two times with DPBS, before 50 µL 1x OneTaq PCR buffer (NEB) containing 0.3 µg/µL proteinase K (Thermo) was added to each well. The plate was sealed and placed in a Tupperware box containing damp blue roll, before being incubated overnight at 55°C. The lysates were transferred to a tube and

incubated at 95°C for 10 minutes to heat-inactivate the proteinase K, before 1 µL was used in the Taq polymerase PCR reactions.

2.18 Cell Culture Treatments

Cells were treated with various compounds (Table 2-7), all of which were diluted to working concentrations in DMEM + 10% FBS.

Table 2-7: Cell Culture Compounds and Treatment Concentrations and Durations

Compound	Solubility	Supplier	Function	Final Conc	Treatment Duration
MG132	DMSO	Merck Millipore	26S proteasome inhibitor	10 µM	18 hr
Tunicamycin	DMSO	Simon Morley	induces eIF2α-P (induces UPR)	2.5 µg/mL	24 hr
Thapsigargin	DMSO	Simon Morley	induces eIF2α-P (inhibits SERCA)	0.1 µM	24 hr
Rapamycin	DMSO	Simon Morley	inhibits mTORC1	100 nM	24 hr
Cobalt Chloride	H_2O	Graham Packham	mimics hypoxia	0.2 mM	24 hr
Anisomycin	DMSO	Simon Morley	inhibits protein synthesis (blocks peptidyl transferase)	50 nM	24 hr

2.19 Small Interfering RNA (siRNA)

A pre-designed synthetic siRNA duplex to human GATAD1 was used to transiently silence GATAD1 gene expression (siRNA ID 29258). The siRNA was designed by Ambion (Life Technologies) and targeted exon 5 in the 3'UTR to ensure that the mRNA of all isoforms was targeted for degradation (Table 2-8). GATAD1 siRNA as well as a non-specific negative control siRNA (Ambion AM4611) were transfected using RiboJuice (Novagen) as in Table 2-9 and section 2.16. Once inside the cells, the duplex siRNA was able to activate the RNAi (RNA interference) pathway. Initially, the siRNA bound to RISC (RNA-induced silencing complex) and was unwound, followed by cleavage of the sense strand by endonucleases. The remaining antisense siRNA-RISC complex was used as a template to identify complimentary GATAD1 mRNA sequences which were then degraded by the endonuclease Argonaute.

Table 2-8: GATAD1 siRNA Sequences

	Sense	Antisense
Target Sequence (5'-3')	GGUAGUAUCUAUUUUUCUC	GAGAAAAAUAGAUACUACC
Length	19	19

Table 2-9: siRNA Transfection Conditions

Transfection of siRNA	6 cm Dish
Number of adherent cells ($x10^5$)	10
Volume of complete growth medium in dish (ml)	5
Volume of serum-free medium in the transfection mixture (µL)	976
Volume of RiboJuice siRNA Transfection Reagent (µL)	24
Volume of 1 µM siRNA stock (5 nM final concentration; µL)	30

2.20 Lactate Dehydrogenase (LDH) Cytotoxicity Assay (Pierce)

Chemical-mediated cytotoxicity was measured by quantifying the amount of extracellular LDH, a cytoplasmic enzyme. LDH in the cell culture media was therefore indicative of plasma membrane damage and can be measured using a colorimetric assay; a coupled enzymatic reaction results in a red formazan product which was directly proportional to the amount of LDH and can be measured at 490 nm (Figure 2-8).

The optimum cell number to use for the assay was first calculated by plating out two sets of triplicates for each serial dilution of cells (between 2000-20,000 cells per well of a 96 well plate), as the LDH signal must be in the linear range at the time of the experiment. Cells were seeded and incubated for 72 hours at 37˚C, 5% CO_2. 10 µL sterile H_2O was then added to one triplicate of cells (spontaneous LDH activity controls), whilst 10 µL lysis buffer was added to the other triplicate (maximum LDH activity controls) before the plate was returned to 37˚C, 5% CO_2 for 45 minutes. 50 µL of media was then transferred from each well to a 96 well flat-bottom plate and 50 µL of reaction mixture was added to each well and left at room temperature, protected from light. To end the reaction, 50 µL Stop solution was added to each well before the absorbance was measured at 490 and 690 nm. LDH activity was determined by subtracting the 690 nm value from the 490 nm value enabling the optimum cell number to be determined.

For the experimental reaction, the appropriate number of cells were seeded in each well of a 96-well plate, in triplicate for each treatment type. A scaled-down transfection was carried out 24 hours later as in section 2.16, followed by a second KDM5A transfection and 5 nM MG132 treatment 24 hours later. Both the spontaneous and maximum LDH activity controls were then treated and the rest of the experiment was carried out using the same method as when determining the optimum cell number.

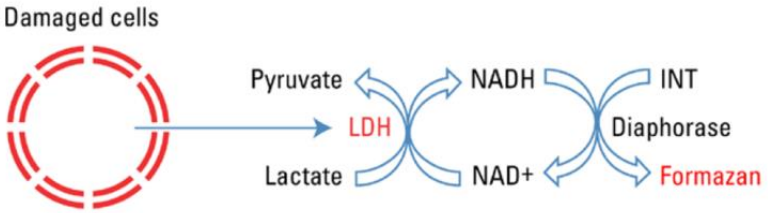


Figure 2-8: LDH Cytotoxicity Assay Mechanism

The amount of red Formazan produced in the coupled enzymatic reaction is directly proportional to the amount of LDH present in the cell culture media. Image from Pierce LDH Cytotoxicity Assay manual.

2.21 NanoBiT Protein:Protein Interaction (PPI) System (Promega)

NanoBiT – ***Nano****Luc* ***Bi****nary* ***T****echnology*

To detect a protein-protein interaction (PPI), the two proteins of interest were fused to LgBiT and SmBiT Nanoluciferase enzyme subunits and expressed in cells. Successful PPI would result in the structural complementation of the NanoBiT subunits (LgBiT:SmBiT), resulting in a functional NanoLuc enzyme with a luminescent signal. HeLa cells were seeded on day 0 at 5000 cells/well in a white 96-well tissue culture plate. 24 hours later, NanoBiT plasmids including positive and negative control constructs were transiently transfected at 50 ng/well.

Nano-Glo Live Cell Reagent was prepared by diluting Nano-Glo Live Cell Substrate 20-fold with Nano-Glo LCS Dilution Buffer. The cell medium was aspirated and replaced with 50 µL room temperature medium, to which 12.5 µL Nano-Glo Reagent was then added. The plate was mixed before the luminescence was monitored using GloMax luminometer, with an integration time of 1 second.

2.21.1 Dual Luciferase Reporter Assay (Promega)

Cells were seeded in a 96 well plate at 5000 cells per well, in a volume of 200 µL, 24 hours prior to transfection of required plasmid. 48 hours post-transfection, cells were washed once with ice-cold PBS, before cells were lysed in 20 µL 1x passive lysis buffer (PLB) at room temperature for 15 minutes. 5 µL of cell lysate was then transferred to a white reading plate, before being read by the luminometer.

The luminometer protocol required both injectors 1 and 2 to dispense 25 µL, with a 2 second delay, 10 second read and a speed of 200 µL/second. 25 µL for each reaction plus a residual 1 mL was required of both reconstituted LarII firefly reagent and 1x Stop & Glo renilla reagent diluted in Stop & Glo buffer. Once the luminometer had been primed and the photomultiplier tube (PMT) activated, the plate was read. Once the readings had been taken, the injectors were purged and flushed.

Protein Techniques

2.22 Cell Harvest

***M-PER**: Mammalian Protein Extraction Reagent (Pierce) used as lysis buffer, supplemented with 1x Halt protease and phosphatase inhibitor cocktail (Thermo Fisher)*

***RIPA**:150mM NaCl (Fisher Scientific), 1% IGEPAL (Sigma), 1% DOC (Sigma), 0.1% SDS (Bio-Rad), 50mM Tris HCl pH 7.6 (Fisher Scientific), 1 mM EDTA (Sigma), 1 mM EGTA (Sigma) and H_2O; supplemented with 1x Halt protease and phosphatase inhibitor cocktail*

On ice, the medium was aspirated from the cell culture plates before they were washed with ice-cold PBS. Cells were scraped into PBS and pelleted by centrifugation for 10 minutes at 500x *g* at 4°C. The PBS was aspirated and the pellet re-suspended in appropriate lysis buffer and incubated at 4°C for 30 minutes. A final chilled spin at 14,000x *g* for 15 minutes was required to remove the cell debris, whilst retaining the supernatant.

Alternatively, cells could also be lysed directly on the plate. Media was removed from the plates and cells washed once with cold PBS. Lysis buffer was then added directly to the plate (200-500 µL for 6cm plate) and cells were left to lyse for 30 minutes on ice. The cells were then scraped and transferred to an Eppendorf before being centrifuged at 14,000x *g* for 10 minutes. Cell lysate was removed and snap-frozen for future use.

2.23 Bradford Assay

1 µL of each lysate sample and 0, 1, 2, 4, 6 and 8 µL 0.5 mg/mL BSA was added in triplicate to 200 µL Protein Assay Dye Reagent Concentrate (Bio-Rad). The samples were vortexed immediately before 150 µL was loaded into a 96 well plate, which was read by the Biotek plate reader using 630 nm and 405 nm filters.

2.24 SDS-PAGE Gel (Sodium Dodecyl Sulphate-Polyacrylamide Gel Electrophoresis)

***Ammonium Persulphate (APS)**: (Sigma) 10% solution in water*

***SDS-PAGE Running Buffer (1x)**: 200 mL Laemmli buffer 10x diluted in 1800 mL dH_2O giving a final concentration of 25 µM Tris (HCl), 0.1% SDS, 192 mM Glycine, pH 8.5*

***SDS-PAGE Loading Buffer (3x)**: 187.5 mM Tris (HCl) (Fisher Scientific), pH 6.8, 30% (v/v) glycerol (Fisher BioReagents), 6% SDS (Bio-Rad), 0.03% (w/v) phenol red (Sigma)*

The composition of an SDS-PAGE resolving gel is shown in Table 2-10. The gel mix was pipetted between 2 glass plates in a cassette, which was overlaid with dH_2O before being left to set. When set, the dH_2O was removed before the stacking gel and comb were added (Table 2-11).

Protein lysates were prepared as determined by Bradford assay, before 3x sample buffer containing 10% v/v β-mercaptoethanol was added to each sample. The samples were then heated at 70˚C for 10 minutes. When set, the comb was removed from the gel and the cassette was clamped into the gel electrophoresis tank, which was filled with 1x Laemmli running buffer. Samples were then loaded into the gel, alongside 3 µL PageRuler Protein Ladder (Thermo) and resolved at 120V for 60-90 minutes.

Table 2-10: Resolving Gel Composition

Volumes of components required to make 2x 1 mm resolving gels. The gel percentage refers to the varying acrylamide concentration, which dictates the length of the polymer chains once polymerisation takes place, catalysed by TEMED. Proteins migrate through the denaturing gel according to their size, allowing estimations of their molecular weight.

Gel Percentage	6%	7.5%	10%	12%
Lowest Mw gel can resolve	50 kDa	40 kDa	35 kDa	25 kDa
dH_2O (mL)	5.4	4.9	4	3.4
1.5 M Tris, pH 8.8 (mL)	2.5	2.5	2.5	2.5
10% SDS (µL)	100	100	100	100
Protogel (mL)	2	2.5	3.4	4
10% APS (µL)	50	50	50	50
TEMED (µL)	10	10	10	10

Table 2-11: Stacking Gel Composition

Volumes of reaction components required to make 2x 1 mm stacking gels. The stacking gel provides a discontinuous buffer system; this increases the resolution of protein separation by concentrating the proteins at the interface of the two gels, before the proteins are resolved by size.

	Stacking Gel
dH_2O (mL)	2.4
0.5 M Tris, pH 6.8 (mL)	1
10% SDS (µL)	80
Protogel (µL)	520
10% APS (µL)	40
TEMED (µL)	10

2.25 Immunoblotting

Tris-Glycine (TG) 1x Transfer Buffer*: 100 mL 10x Tris-Glycine stock (Bio-Rad) diluted in 700 mL dH_2O and 200 mL methanol (Acros Organics), giving a final concentration of 25 mM Tris, 192 mM glycine, pH 8.3 and 20% v/v methanol*

TBS-Tween*: 50 mM Tris (HCl) (Fisher Scientific), 150 mM NaCl (Fisher Scientific), 0.1% (v/v) Tween-20, pH 7.5-8.0 (Acros Organics)*

SDS-resolved proteins were transferred to 0.45 µM nitrocellulose membrane (Whatman) using 9x6 cm stacks soaked in 1x TG buffer, consisting of 3x 3 mm filter papers, nitrocellulose membrane, SDS-PAGE gel and a final 3x 3 mm stack of filter paper. The semi-dry transfer method used the Bio-Rad Trans-bot Turbo system to transfer the proteins at 0.1 A, 25 V for 30 minutes.

The wet transfer method is more efficient for higher molecular weight proteins and used the Bio-Rad Mini Trans-blot Electrophoretic Transfer Cell. Once the components of the stacks had been equilibrated in transfer buffer for 20 minutes, they were assembled in the same way as the semi-dry transfer protocol, but with additional fibre pads on either side of the stacks. The cassette was then assembled in the tank before transfer buffer was poured to the blotting line. Proteins were then transferred to the membrane at 100 V, constant 350 mA for 1 hour, on ice. Membranes were blocked using 3% BSA (Sigma) in TBS-Tween for 30 minutes and incubated in primary antibody overnight at 4˚C (Table 2-12). Membranes were then washed three times for 10 minutes in TBS-Tween, whilst secondary antibodies were prepared in 3% BSA in TBS-Tween. Membranes were incubated on a rocker with Pierce DyLight secondary antibodies in a dark box for 60 minutes. A further three 10 minute TBS-Tween washes were required before visualisation.

Table 2-12: Antibody Dilutions

Company	Antibody	Raised In	Dilution WB	Dilution IF
Sigma	Anti-Flag M2 Affinity Purified	Mouse	1:5000	1:500
Sigma	Anti-α actin	Rabbit	1:5000	-
Abcam	Anti-β actin	Mouse	1:5000	-
Promega	Anti-Halo Tag	Rabbit	1:1000	1:100
Sigma	Anti-GATAD1	Rabbit	1:1000	1:100
Cell Signalling	Anti-Jarid1A	Rabbit	1:500	1:100
Cell Signalling	Anti-Lamin A/C Monoclonal	Mouse	1:1000	1:50
Home-made	Anti-eIF1	Rabbit	1:1000	-
Novus	Anti-eIF1AY	Mouse	1:5000	-
Cell Signalling	Anti-eIF1A	Rabbit	1:5000	-
Cell Signalling	Anti-eIF5	Rabbit	1:1000	-
Cell Signalling	Anti-HIF1α	Mouse	1:500	-
Cell Signalling	Anti-4E-BP1 (9644)	Rabbit	1:1000	-
Sigma	Anti-GATAD1 (SAB2100901)	Rabbit	1:1000	-
Sigma	Anti-GATAD1 (Prestige)	Rabbit	1:1000	-
Bioss	Anti-GATAD1 (bs-11921R)	Rabbit	1:1000	1:100
Abcam	DyLight 680 Anti-Mouse IgG	Goat	1:20,000	-
Abcam	DyLight 800 Anti-Rabbit IgG	Goat	1:20,000	-
Cell Signalling	Anti-Cox IV Monoclonal	Rabbit	-	1:125
Cell Signalling	Anti-mouse IgG (H+L) F(ab')2 conjugated to Alexa Fluor 488	Goat	-	1:2000
Cell Signalling	Anti-rabbit IgG (H+L) F(ab')2 conjugated to Alexa Fluor 555	Goat	-	1:2000

2.25.1 Visualisation of Nitrocellulose Membrane using LI-COR

The LI-COR Odyssey Imaging System was used to scan nitrocellulose membranes using both 700 nm and 800 nm channels, before viewing on Image Studio Lite, which was also used for relative quantification of bands to a housekeeping gene using densitometry.

2.26 Immunofluorescence

Cells were grown on cover slips in 6-well plates until they were ready to harvest. Cells were then washed with PBS and fixed with 1 mL PBS containing 4% (v/v) paraformaldehyde (Thermo Fisher) per well for 15 minutes. Three PBS washes were then carried out before the cells were permeablised with PBS containing 0.1% (v/v) Triton-X (Fisher Scientific) for 5 minutes. After three further PBS washes, cells were blocked with 200 µL 5% goat serum (Sigma) in PBS for 20 minutes. Cells were then incubated with 100 µL diluted primary antibody for 45 minutes at room temperature, followed by five PBS washes. This was then repeated with secondary antibodies before a final wash with dH_2O. 10 µL ProLong antifade reagent (Thermo) with DAPI nuclear stain was used to mount the coverslips. Microscope slides were kept in the dark at room temperature for 12 hours before being stored at 4˚C and visualised by fluorescence microscopy using a Zeiss microscope and MetaMorph software (Table 2-13).

When preparing slides for STED microscopy, DAPI-free mounting medium was used since the dye may be excited by the 592 nm depletion beam, resulting in increased background. Slides were otherwise prepared in the same manner as previously described. The images obtained were processed by deconvolution, whereby additional information obtained during acquisition as a result of light diffraction through the microscope is removed using a Fourier Transform, increasing the signal-to-noise ratio of the image.

Table 2-13: Fluorescence Microscopy Filters

Excitation and emission spectrum of the fluorescent dyes used to stain HeLa cells. Filtersets were manufactured by Chroma. The DAPI emission filter is long-pass, transmitting all light above the stated wavelength.

Fluorescent Dye	Excitation (nm)	Emission (nm)
Alexa Fluor 488	465-495	495-575
Alexa Fluor 555	515-565	550-660
DAPI	347-403	435

2.27 Co-Immunoprecipitation (Pierce Co-IP Kit)

All Co-IP centrifugation steps were carried out at 1000x g for 1 minute, unless otherwise stated.

Antibody Immobilisation

Antibody was covalently crosslinked onto an amine-reactive resin, preventing antibody fragments from eluting and interfering with pull-down detection. 50 µL of AminoLink Plus Coupling Resin was added to a Pierce Spin Column for each reaction, which was centrifuged to remove storage buffer. The resin was then washed twice using 1x Coupling Buffer, before 10 µg of antibody made in 200 µL (10 µL of 20x Coupling Buffer, 180 µL H_2O and 10 µL of antibody at 1 µg/µL) was added to the column containing the resin. 3 µL of Sodium Cyanoborohydride Solution was then added before the column was sealed and incubated on a rotator for 2 hours at room temperature. After incubation, the column was centrifuged, before two washes were carried out using 200 µL of 1x Coupling Buffer. 200 µL of quenching buffer was then added to the column to block the remaining N-Hydroxysuccinimide (NHS)-ester sites. The column was spun and the flow-through is discarded, before a further 200 µL of quenching buffer was added to the resin with 3 µL Sodium Cyanoborohydride Solution. The column was then incubated for 15 minutes with end-over-end mixing. After incubation, the column was spun and the flow-through discarded. The resin was then washed twice with 200 µL of 1x Coupling Buffer, followed by six washes with 150 µL Wash Solution, centrifuging between washes.

Pre-Clear Lysate

Cells were lysed (section 2.22) using IP/Lysis Wash Buffer. The lysate was then pre-cleared using 80 µL of the Control Agarose Resin per 1 mg of lysate. Storage buffer was first removed from the control resin by centrifuging the spin column, before the lysate was added to the column containing the resin and incubated at 4˚C for 1 hour. After incubation, the column was centrifuged at 4°C and the flow-through was retained to be used in the co-IP.

Co-IP

The antibody-coupled resin was washed twice using 200 µL of IP Lysis/Wash Buffer, centrifuging and discarding the flow-through between washes. The pre-cleared lysate was added to the resin and incubated end-on-end for 2 hours at room temperature. Following incubation, the columns were spun and the flow-through saved for analysis. The columns were then washed three times using 200 µL IP Lysis/Wash Buffer and once with 100 µL 1x Conditioning Buffer, with a spin between each wash.

Elution of Co-IP

Antibody-resin complexes were retrieved by boiling in 50 µL reducing sample buffer at 90ºC for 10 minutes. The supernatant was spun through to retrieve both the proteins within the complex as well as the contaminating antibody fragments. The proteins eluted with low efficiency when using the provided Pierce elution buffer.

Chapter Three

Identification of Alternative GATAD1 Isoforms

3. Identification of Alternative GATAD1 Isoforms

3.1 Introduction

GATAD1 was initially shortlisted following an initial bioinformatics pipeline which identified conservation between human and mouse translated 5'UTRs and was then also listed as a likely alternative translation initiation candidate using a bioinformatic macro designed by Dr. Richard Edwards (ExTATIC). If the predicted in-frame AICs within the 5'UTR are recognised by the ribosome, protein isoforms with N-terminal extensions would be translated. The N-terminal extension may cause the protein to localise differently within the cell, have different binding partners or may even alter the function of the protein, when compared to the annotated GATAD1 isoform. Initial experiments must identify whether alternative translation initiation is taking place on the GATAD1 mRNA, and if so, which AICs are used. This will enable identification of the N-terminal extension present on the alternative isoforms and allow further experiments to take place.

Initial AIC identification will be carried out by overexpression of plasmids containing the GATAD1 cDNA, including ones altered to strengthen or weaken predicted initiation codons, and analysis of the translated protein isoforms. Following identification of AIC usage in GATAD1, we will confirm that multiple forms of GATAD1 also exist in the endogenous protein.

3.2 Hypothesis and Aims

3.2.1 Hypothesis

Alternative translation initiation takes place from upstream CUGs within the 5'UTR, at position -207 and -144 relative to the annotated AUG, as predicted using the bioinformatic macro.

3.2.2 Aims

The aim of this chapter was to determine whether alternative translation initiation was taking place within the GATAD1 mRNA transcript. This was investigated by:

- Mutagenesis of the potential AICs identified by the macro, to increase or decrease the translational efficiency of the codon.
- Western blotting to identify which AICs are utilised, when compared to the wild-type GATAD1.
- qPCR of GATAD1 isoforms to ensure that the changes in protein expression observed when mutagenesis is carried out are not due to changes in the level of mRNA, but solely due to a translational effect.
- Confirming the presence of -207 CUG and -45 AUU AIC-synthesised isoforms of GATAD1 in the endogenous protein.

3.3 Results

3.3.1 Bioinformatic Prediction of GATAD1 AICs

Accurate identification of translational start sites is essential in understanding the function of the translated proteins. A bioinformatic macro designed by Dr. Richard Edwards was used to predict potential alternative initiation codons, following an initial bioinformatics pipeline which shortlisted translated 5'UTRs where there was conservation between species. The macro, which was limited to predicting in-frame initiation from CUG, GUG and ACG codons in a good Kozak consensus, highlighted two potential alternative initiation codons, a CUG at position -207 and another at -144 (Figure 3-1).

The GATAD1 annotated AUG is not within a strong Kozak consensus (ACCAUGC), lacking the guanine at position +4 required for efficient translation initiation to take place (Kozak, 1987a). The ribosome therefore has the potential to scan through this codon and leak downstream to an AUG (or other initiation codon) within a stronger context. The downstream sequence was therefore analysed and a further potential AUG start codon at position +55 was identified. Use of the +55 AUG to initiate translation would generate a truncated GATAD1 isoform with only a partial zinc finger. A further three potential upstream, in-frame AICs were identified by visual analysis of the GATAD1 5'UTR; a GUG at position -39 and a CUG at position -33, which both lacked a +4 guanine as well as a -3 purine, as well as an AUU at position -45 which lacked a +4 guanine, Figure 3-5. These codons were not identified by the macro as they are within a weak sequence context, however they should be considered when identifying the AICs utilised within GATAD1 (Table 3-1).

Annotated	Alternative	Description
Strong	Strong	ATG codon with purine at -3 and guanine at +4.
MidR	MidR	ATG codon with purine at -3 only.
MidG	MidG	ATG codon with guanine at +4 only.
Weak	Weak	ATG codon with neither purine at -3 nor guanine at +4.
ACG	ACG	ACG codon with purine at -3 and guanine at +4.
CTG	CTG	CTG codon with purine at -3 and guanine at +4.
GTG	GTG	GTG codon with purine at -3 and guanine at +4.

#	Pos	Codon	Context	Strength	eLen/tLen
1	-207	CTG	ATCCTGG	CTG	69
2	-144	CTG	GGCCTGG	CTG	48
3	1	ATG	ACCATGC	MidR	0

Figure 3-1: GATAD1 Ensemble Macro AIC Predictions

ATG codons were identified by the macro within a strong, midR, midG or weak context. The most efficient AICs (ACG, CTG and GTG) were also identified when in a strong Kozak consensus. The context of the codon as well as the length of the N-terminal extension provided by the upstream AIC is also given.

Table 3-1: Bioinformatic Analysis of GATAD1

An overview of predicted initiation codons within the GATAD1 mRNA transcript.

Position	Initiation Codon	Context	Length of potential N-terminal extension/truncation
-207	CUG	AUCCUGG (strong)	69 Amino Acids
-144	CUG	GGCCUGG (strong)	48 Amino Acids
-45	AUU	GCCAUUC (relatively weak)	15 Amino Acids
-39	GUG	CCCGUGU (weak)	13 Amino Acids
-33	CUG	UCUCUGC (weak)	11 Amino Acids
+1	AUG	ACCAUGC (relatively weak)	---
+55	AUG	UCCAUGU (weak)	18 Amino Acids

A Clustal Omega multiple protein sequence alignment was carried out to determine how conserved the potential alternative protein isoforms are between species (Figure 3-2). If an alternative protein isoform is very well conserved, it is likely that it may have a specific and necessary function that the cell cannot afford to lose. The isoforms translated from both -207 CUG and -144 CUG are well conserved between *Homo sapiens* (human) and *Macaca* (macaque) throughout the sequence, whereas *Bos taurus* (cow) and *Mus musculus* (mouse) only align for most of the final 16 amino acids of the extension nearest the annotated AUG. The extension translated from -45 AUU onwards is perfectly conserved between all four species, suggesting that this part of the sequence of the GATAD1 extension may be functionally important, increasing the likelihood of this initiation codon being used.

```
mus       SRAGSRAAGARFPPPGVSGGRWVPARAAQHLRPCSSS-RRGGRPTPGRPVQTAAIPVSLRPRGPPEPASM
bos       ---RPGLGS-------------------SALRPRRSFVSGSGGPARGPPRSAVAIPVSLRPRGPPEPATM
homo      LASACGAGGTRFPPPRGSASGLVLSPAAPCRRSHRSSYRREWRADQGAAGLPSAIPVSLRPRGPPEPATM
macaca    GRRAGGTGGTRFPPPRGSASGLVPSRAARCCRPHRSSYRREWWADQGAAGLASAIPVSLRPRGPPEPATM
                 ..                        *   *          *      ***************:*
```

Initiation codon	Translated protein sequence
-207 CUG	LASACGAGGTRFPPPRGSASGLVLSPAAPCRRSHRSSYRREWRADQGAAGLPSAIPVSLRPRGPPEPATM
-144 CUG	----------LVLSPAAPCRRSHRSSYRREWRADQGAAGLPSAIPVSLRPRGPPEPATM
-45 AUU	----------IPVSLRPRGPPEPATM
-39 GUG	----------VSLRPRGPPEPATM
-33 CUG	----------LRPRGPPEPATM

Figure 3-2: GATAD1 5'UTR Translated Protein Alignment

Multiple sequence alignment of the GATAD1 5'UTR of 4 species; *Mus musculus* (mouse), *Bos taurus* (cow), *Homo sapiens* (human) and *Macaca* (macaque). The green font indicates species alignment, whilst the red font indicates a methionine residue, indicating the start of the annotated protein. The lower table shows the translated protein sequence of human GATAD1 from each potential initiation codon.

3.3.1.1 GATAD1 Ribosome Profiling Data

Ribosome profiling is an experimental approach used to identify translation initiation sites by using translation inhibitors to stall ribosomes at the initiation phase. Deep sequencing of ribosome-protected fragments is then carried out allowing ribosome positions to be identified. A ribosome profiling study carried out on HEK293 cells (Lee et al., 2012) identified a CUG initiation codon at position -144 relative to the annotated AUG, within the GATAD1 mRNA transcript (Figure 3-3). This CUG was also identified in the bioinformatic macro used by our laboratory, and so was a promising alternative initiation site.

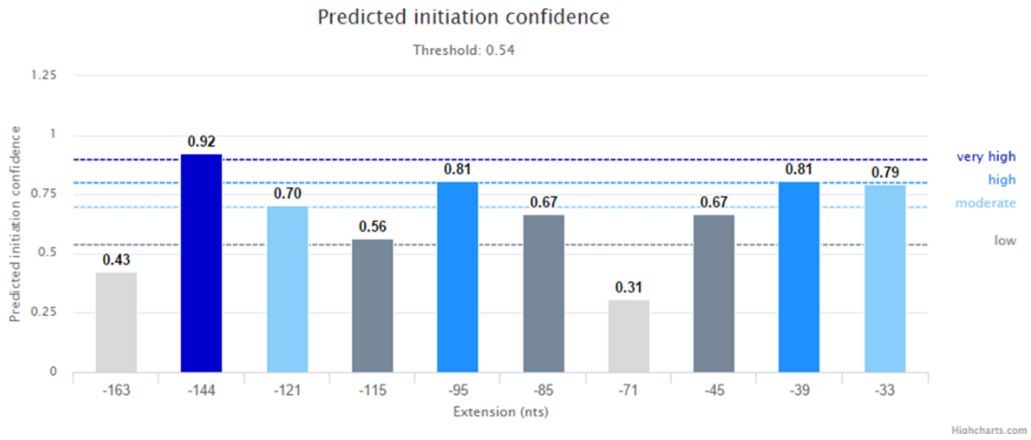
RefSeq accesion	GeneID	Gene Symbol	Position	Annotation	R_{LTM}-R_{CHX}	LTM reads	CHX reads	ORF length	Codon	CHX density (reads/nt)
NM_021167	57798	GATAD1	-144	5'UTR	0.86720918	18	1	954	CTG	0.201257862

Figure 3-3: CTG Identified by Ribosome Profiling of GATAD1

The ribosome profiling dataset by Lee et al. identified the use of -144 CUG as an *in vivo* AIC within the GATAD1 transcript. Both cycloheximide and lactimidomycin were used as inhibitors in this study, which both bind to the ribosomal E-site and allow differentiation between initiating and elongating ribosomes.

3.3.1.2 Pre-TIS Prediction of Translation Initiation Events

Numerous studies have produced tools used to predict alternative initiation sites within mRNA transcripts. The Pre-TIS prediction tool (Reuter et al., 2016) was the first to combine ribosome profiling datasets with mRNA sequence information to predict both AUG and near-cognate initiation codons and their ability to initiate translation. Two prediction models were made, the first was based on the HEK293 ribosome profiling dataset by Lee (Lee et al., 2012) and the second was based on the mouse ES dataset by Ingolia (Ingolia et al., 2011). Numerous features based on mRNA sequence information were also considered when calculating the initiation confidence, including start site conservation – pairs of human and mouse sequences (5'UTR and CDS) were aligned using *blastn* using MUSCLE sequence alignment. The flanking sequence context of the predicted initiation site was also considered as well as mRNA secondary structure. The PreTIS search of the GATAD1 sequence returned numerous predicted AICs (both in and out of frame with the annotated AUG) (Figure 3-4). In-frame AIC predictions were -144 CUG, -45 AUU, -39 GUG and -33 CUG.



Extension (nts)	Codon	Frame	Position stop codon	ORF length	Predicted initiation confidence
-163	ACG	-1	251	417	0.43
-144	**CUG**	**0**	**807**	**954**	**0.92**
-121	**GUG**	**-1**	**251**	**375**	**0.7**
-115	**ACG**	**-1**	**251**	**369**	**0.56**
-95	**AUC**	**-2**	**13**	**111**	**0.81**
-85	**GUG**	**-1**	**251**	**339**	**0.67**
-71	AGG	-2	13	87	0.31
-45	**AUU**	**0**	**807**	**855**	**0.67**
-39	**GUG**	**0**	**807**	**849**	**0.81**
-33	**CUG**	**0**	**807**	**843**	**0.79**

Figure 3-4: PreTIS GATAD1 Results

Predicted AICs in-frame with the annotated AUG included -144 AUG (high confidence) as well as -45 AUU (low confidence), -39 GUG and -33 CUG (moderate confidence).

3.3.2 QuikChange Site-Directed Mutagenesis (SDM) to Confirm AICs

In order to determine whether the AICs predicted by the macro are indeed utilised by the ribosomal machinery, individual point mutations were introduced at each possible initiation codon to strengthen and weaken the codon, encouraging and discouraging translation from each position. Overlapping mutagenic primers were used to generate each mutant using the QuikChange Lightning Kit. The -207 AUG and UAC mutants were obtained from previous work carried out on GATAD1 in the MJC lab; Alice Fletcher observed definite extra bands on a GATAD1 Western blot, however these were not conclusively identified. The CUG at position -144, -45 AUU, -39 GUG and -33 CUG were each mutated to both an AUG and a UAC. The +1 AUG as well as the +55 AUG were mutated to CUU codons, to prevent initiation from taking place from these positions (Figure 3-5). A CUU (Leu) codon was used here instead of the previously used UAC (Tyr) when making the non-initiating AUG mutants because Leu is the closest residue in structure to Met, therefore minimising interference with the function of the Ext/Mid forms.

NcoI diagnostic digests showed that each SDM plasmid contained GATAD1 of the correct size (Figure 3-6B), which was confirmed by sequencing (Figure 3-6C).

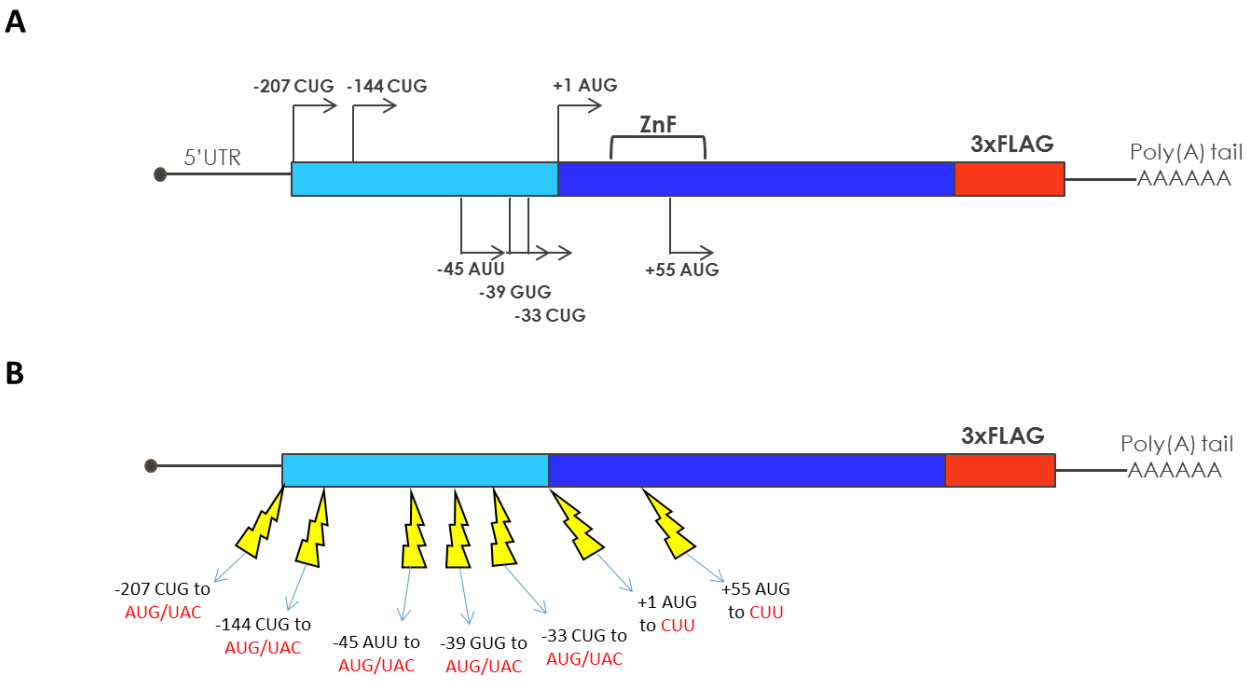


Figure 3-5: GATAD1 QuikChange Mutagenesis

(A) GATAD1 construct indicating the translation initiation sites predicted by the bioinformatic macro (above the construct), as well as other potential AICs (below the construct). Use of the +55 AUG would result in translation of a truncated GATAD1 isoform with a partial zinc finger. The N-terminal zinc finger is 25 residues in length (position 9-33 with respect to the annotated isoform). (B) The point mutations made at each position are indicated in yellow. Mutating an AIC to an AUG will increase expression from that position, whilst mutating to a UAC/CUU should completely prevent expression from that codon. The mutagenesis template was C-terminally truncated GATAD1, containing only part of the CDS (202 of 269 amino acids); this is in order to resolve the protein isoforms more clearly on a Western blot. The mutations made were at a single codon in each construct, with the rest of the sequence remaining the same as wild-type GATAD1.

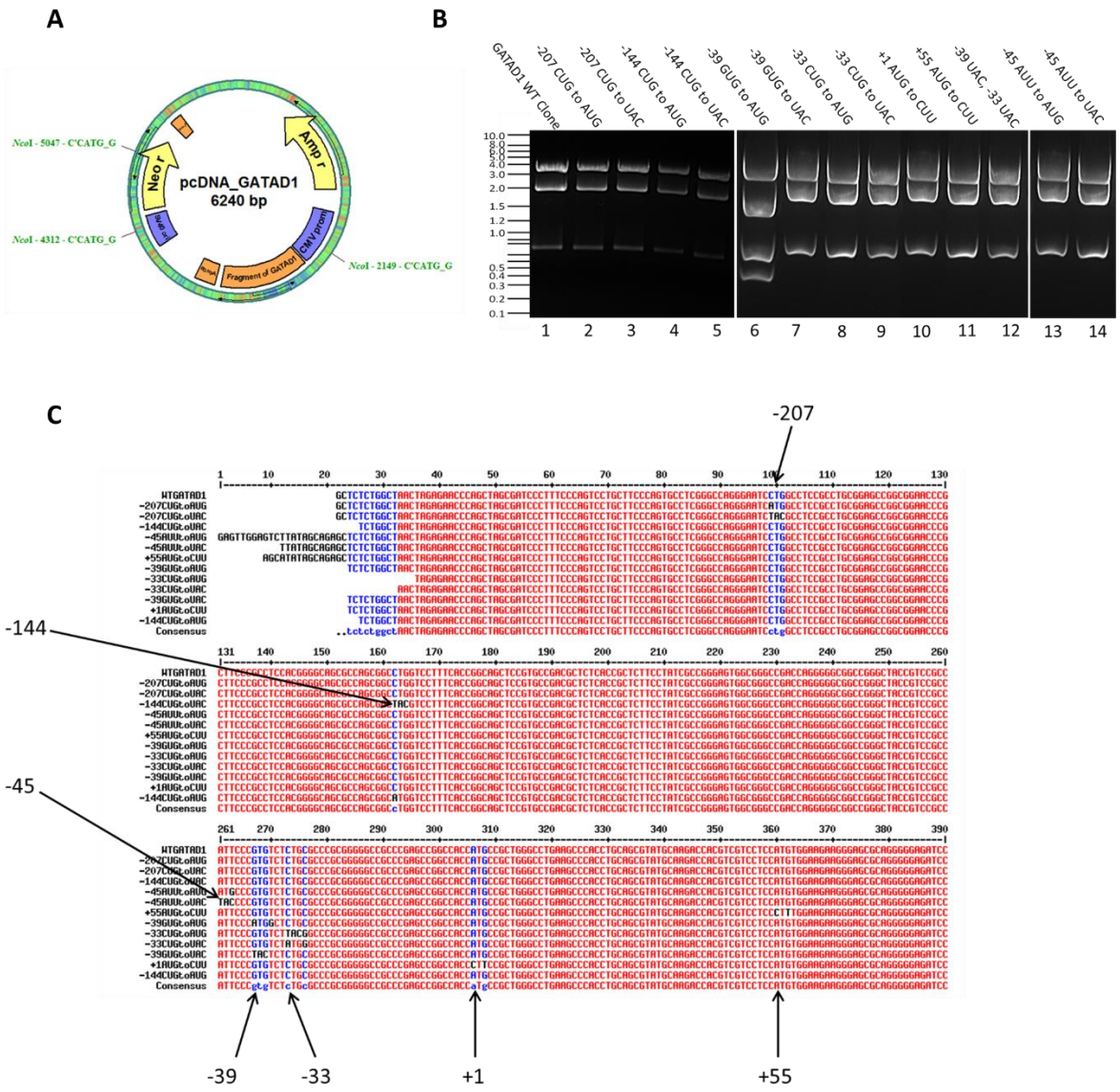


Figure 3-6: Restriction Digest and Sequencing of GATAD1 SDM Clones

(A) Virtual GATAD1 SDM plasmid showing NcoI restriction sites. (B) NcoI restriction digests of the SDM plasmids run on a 0.8% agarose gel, which should result in three fragments of 3342, 2613 and 735 bp. The -39 GUG to AUG mutation introduces an extra NcoI restriction site (CCATGG), resulting in an additional band on the gel. The GATAD1 insert is 883 bp. (C) Sequence alignment of the GATAD1 mutants to wild-type GATAD1 were carried out using MultAlin, which confirmed that the correct mutations had been made at each potential initiation site.

The FLAG-tagged GATAD1 mutants were expressed in HeLa cells, before the whole cell lysate was used to perform a Western blot (Figure 3-7). Each mutant was designed to either encourage or prevent expression from a predicted initiation codon. The initiation codons being utilised by the ribosome to initiate translation were identified on the Western blot by comparing each expression pattern to wild-type GATAD1, as well as the expected molecular weights for each isoform (Figure 3-7A). The GATAD1 isoforms run slightly higher on a Western blot than predicted, which may be due to a post-translational modification increasing the molecular weight of the protein.

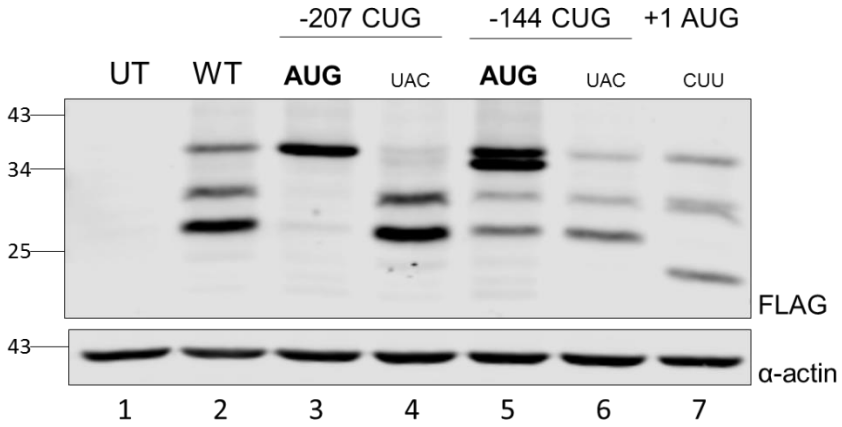
The initial Western blot (Figure 3-7B) shows mutants of the initiation codons identified by the bioinformatic macro, which are within a strong Kozak consensus and are therefore the most likely codons to initiate translation. The untransfected negative control (lane 1) showed no FLAG signal. Wild-type GATAD1 (lane 2) expressed 3 isoforms ofFigure 3-7 approximately 35 (upper), 30 (middle) and 27 (lower) kDa in a 1:1.3:3 ratio respectively, using α-actin as a loading control. The -207 CUG to AUG mutant (lane 3) expressed a single upper band, whilst the -207 CUG to UAC mutant (lane 4) only expressed the middle and lower bands. This confirms that the upper band represents GATAD1 expressed from -207 CUG, since strengthening the codon to an AUG prevented leaky scanning downstream, whereas the UAC mutant caused the ribosome to scan through the -207 position and leak down to other initiation codons. The -144 CUG to AUG mutant (lane 5) expressed the upper band as well as an additional smaller band not seen in the wild-type lane. The -144 CUG to UAC mutant (lane 6) expressed the same three protein isoforms as wild-type GATAD1, indicating that the -144 CUG codon was not used to initiate translation, as was predicted by the AIC macro and apparently confirmed in the Lee ribosome profiling dataset. The +1 AUG to CUU mutant (lane 7) expressed both the upper and middle bands. This confirms that the lower band represents the annotated isoform translated from an AUG, since the +1 CUU mutant allowed the ribosome to leaky scan down to the internal +55 AUG which is normally not used as the 40S ribosomal subunit would ordinarily be used for translation from the +1 AUG.

The second Western blot (Figure 3-7C) was used to identify the initiation codon responsible for translation of the middle isoform. The -45 AUU to AUG mutant (lane 3) expressed the upper and middle band, whilst the -45 AUU to UAC mutant (lane 4) expressed the upper and lower bands. This confirms that the middle band represents GATAD1 expressed from -45 AUU, since strengthening the codon to an AUG prevented leaky scanning downstream to the annotated AUG start codon, whereas the UAC mutant prevented ribosome recognition at the -45 position, causing leaky scanning through to the annotated AUG. Both the -39 GUG to UAC (lane 6) and the -33 CUG to UAC mutants (lane 8) expressed the same three protein isoforms as wild-type GATAD1, confirming that the -39 GUG and the -33 CUG were not used to initiate translation. Finally, the +55 AUG to CUU mutant (lane 9) expressed the same three protein isoforms seen in wild-type GATAD1 and is therefore not used efficiently to initiate translation. These mutants have been run on a single Western (Figure 3-7D) to show the -207 CUG, -45 AUU and annotated AUG together.

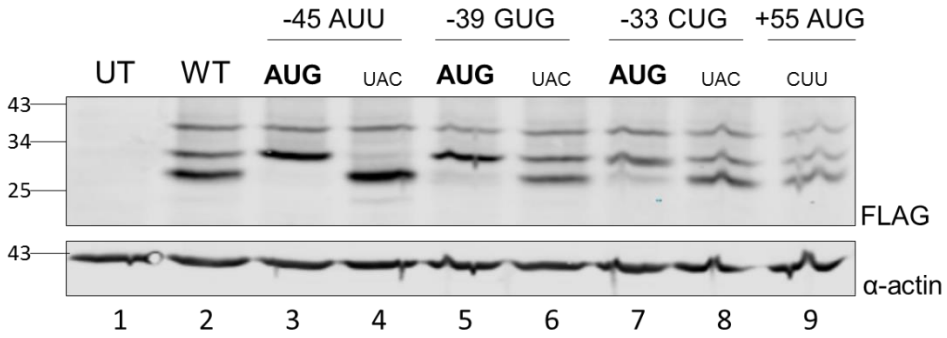
A

Initiation Codon Position	Expected Size (kDa)
-207	31.2
-144	29.3
-45	25.7
-39	25.5
-33	25.3
+1	24.1
+55	22.3

B



C



D

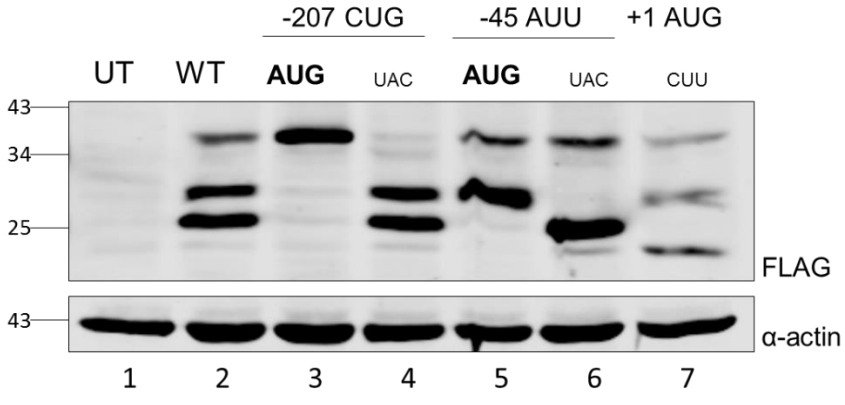


Figure 3-7: CUG and AUU AIC Confirmed in GATAD1 5'UTR

(A) Expected sizes of protein isoforms produced when translated from each potential initiation codon. The SDM isoforms have part GATAD1 CDS (202 of 269 aa), the whole 5'UTR and a 3x-FLAG tag. (B) FLAG-probed Western blot showing expression of GATAD1 protein isoforms when translation was encouraged and prevented at each AIC predicted in the bioinformatic macro. 10 µg of transfected HeLa whole cell lysate was loaded on to a 10% SDS-PAGE gel, where UT = untransfected and WT = wild-type. (C) FLAG-probed Western blot showing expression of GATAD1 protein isoforms when translation was encouraged and prevented at each AIC identified using the GATAD1 5'UTR sequence. These predicted AICs are in a weak Kozak consensus. 10 µg of transfected HeLa whole cell lysate was loaded on to a 10% SDS-PAGE gel. (D) FLAG-probed western blot showing GATAD1 mutants of the confirmed initiation codons only. 10 µg of transfected HeLa whole cell lysate was loaded on to a 10% SDS-PAGE gel.

In order to confirm that the upper bands on the Western blot were not due to post-translational modifications of the annotated isoform, the 5'UTR upstream of each initiation codon was deleted. A Western blot confirms that the bands disappear when the 5'UTR is truncated (Figure 3-8).

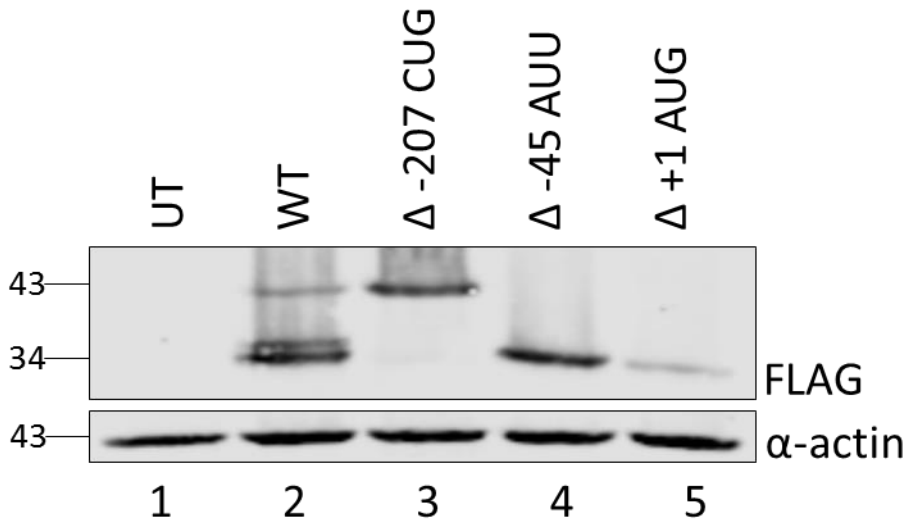


Figure 3-8: GATAD1 Isoforms Truncated Upstream of AICs

FLAG-probed western blot showing FCS GATAD1 truncated upstream of each initiation codon. 10 µg of transfected HeLa whole cell lysate was loaded on to a 10% SDS-PAGE gel, where UT = untransfected, WT = wild-type and Δ represents 5'UTR truncations at each initiation codon.

3.3.3 qPCR Confirms Translational Effect

When mutagenesis was carried out at each initiation codon, changes in protein expression were observed (Figure 3-7)Figure 3-7. qPCR was carried out to ensure that the mutagenesis was not affecting the GATAD1 mRNA levels and that the changes in isoform expression were solely due to a translational effect. Initially, the efficiencies of the qPCR primers were calculated, (Figure 2-2) which were within a range of 90-104%. Generally, efficiencies between 90-110% are considered acceptable, meaning all five PCR reactions analysed are within the recommended efficiency range. The Eco qPCR software also takes in to account the amplification efficiency of each qPCR reaction. Relative quantification of FLAG vs β2M mRNA shows approximately equal levels of GATAD1 mRNA in each sample (Figure 3-9), irrespective of the changes in protein expression observed.

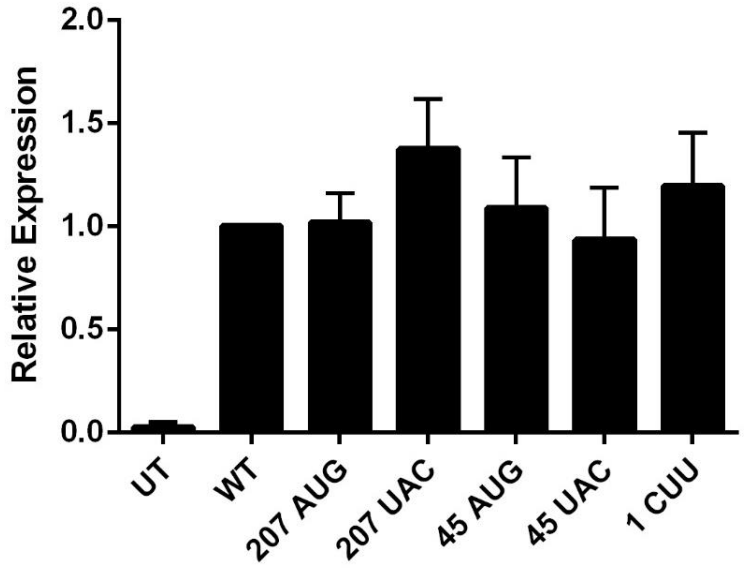


Figure 3-9: Relative Quantification of FLAG vs β2M mRNA levels

mRNA from transfected HeLa cells was converted to cDNA and assayed using qPCR for FLAG. β2M was used as a housekeeping gene internal control. UT = untransfected and WT = wild-type. Data representative of 3 independent experimental repeats. Error bars indicate the standard deviation (n=3).

3.3.4 HEK293 CRISPR/Cas9 Confirms Endogenous Alternative GATAD1 Isoforms

Initially, detection of endogenous GATAD1 was attempted by Western blot using commercially available antibodies. However, the antibodies either gave very little signal or were non-specific, making it difficult to identify endogenous GATAD1 isoforms. Therefore, CRISPR genome editing was carried out to 3xFLAG-tag the C-terminus of the GATAD1 gene, allowing identification of endogenous GATAD1 protein isoforms by carrying out a FLAG Western blot.

CRISPR is a modified bacterial immune system that is utilised for genome editing and has two components, a guide RNA (gRNA) and a Cas9 endonuclease. The short, synthetic gRNA is composed of a scaffold region for Cas9 binding and a user-defined 20 nucleotide targeting sequence which modifies the target of the Cas9 enzyme. The target sequence must be immediately upstream of a PAM (Protospacer Adjacent Motif), which is necessary for target binding and varies in sequence depending on the species of Cas9 (5'-NGG-3' for *Streptococcus pyogenes*). Co-transfection and expression of the gRNA and Cas9 results in the formation of a riboprotein complex capable of DNA binding. The Cas9-gRNA complex binds the target sequence, where Cas9-mediated DNA cleavage results in a double stranded break (DSB). In order for a specific modification (eg, 3xFLAG-tag insertion) to be made, the DSB must be repaired via homology-directed repair (HDR), rather than the more common, but error-prone non-homologous end joining (NHEJ) pathway. In order to promote HDR, a DNA repair template containing the sequence modification as well as homology domains, is also transfected with the gRNA and Cas9. The modified sequence is incorporated in the repair, however NHEJ has low efficiency and the proportion of repair via this pathway resulting in the desired DNA modification is <10%, (Cong et al., 2013).

3.3.4.1 Design and Cloning of the CRISPR sgRNA expression construct

The Zhang Lab CRISPR design tool (http://crispr.mit.edu/) (Cong and Zhang, 2015) enabled identification of suitable single GATAD1 guide RNA (sgRNA) target sites surrounding the stop codon, within the human genomic sequence (NCBI ref seq NG_032807.1) (Figure 3-10). Each sgRNA target was located directly upstream of a protospace adjacent motif (PAM) site (5'-NGG), which would target the *S. pyogenes* Cas9 nuclease to cleave the DNA 3 bp upstream of the PAM (Figure 3-11). GATAD1 sgRNA guide #3 was chosen to target the Cas9 nuclease to cleave 26 bp downstream of the GATAD1 stop codon.

The U6 RNA polymerase III promoter which controls expression of the sgRNA prefers a guanine as the first base of its transcript; a G was therefore appended to the 5' of the sgRNA. 5' overhangs were also required for ligation into the pair of BbsI sites within the pSpCas9(BB)-2A-Puro (PX459) vector (Figure 3-12). The GATAD1 sgRNA oligonucleotides were synthesised by Sigma Aldrich. Before cloning of the sgRNA into the pspCas9(BB) vector for co-expression with Cas9, the forward and reverse oligo inserts were first annealed and phosphorylated using T4 PNK and the pspCas9(BB) vector was linearised with BbsI, prior to dephosphorylation with rSAP. An EcoRI diagnostic digest of the recombinant clones confirmed approximately the correct size inserts, which was confirmed by sequencing with U6 sequencing primers (Figure 3-13).

	score	sequence	
Guide #1	63	TACTCCTGCAATAACAATTA	AGG
Guide #2	61	GATTCCTTAATTGTTATTGC	AGG
Guide #3	61	TAAAACTGGGTTTCCAGGCC	TGG
Guide #4	57	TTAATTGTTATTGCAGGAGT	AGG
Guide #5	57	CAAATTAAAACTGGGTTTCC	AGG
Guide #6	55	ATTTGTGAACTACAAATGGT	TGG
Guide #7	46	GGTTTCCAGGCCTGGTGTGG	TGG
Guide #8	44	CTGGGTTTCCAGGCCTGGTG	TGG
Guide #9	43	TGTAGTTCACAAATTAAAAC	TGG
Guide #10	42	TTTAATTTGTGAACTACAAA	TGG
Guide #11	38	GTGAGCCACCACACCAGGCC	TGG
Guide #12	36	GTAGTTCACAAATTAAAACT	GGG
Guide #13	1	CAGGCGTGAGCCACCACACC	AGG

Figure 3-10: GATAD1 sgRNA target selection

Guide sequences were ranked by inverse likelihood of off-target binding. A high quality guide which would target the Cas9 nuclease to cut as close as possible (within 50 bp) to the GATAD1 stop codon target site was required. Guide #3 provided the maximum on-target activity (cleaving closest to the stop codon target), whilst being categorised as a high quality guide meant that computationally predicted off-target activity was minimised.

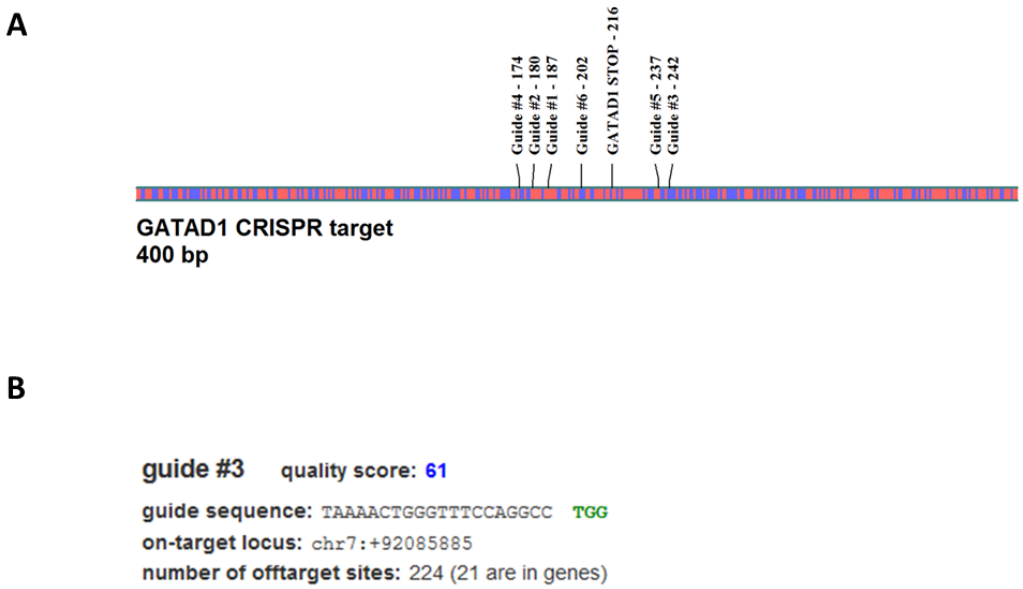


Figure 3-11: Position of high quality sgRNA targets

(A) High quality guide sequences #1-6 are shown along a portion of genomic GATAD1, surrounding the target stop codon. The cas9 cleavage sites are indicated. (B) Guide number 3 was used. The likelihood of a sgRNA having off-target sites is computationally assessed upon sequence similarity, whereby more than three mismatching bases is usually not tolerated, preventing off-target binding of the sgRNA.

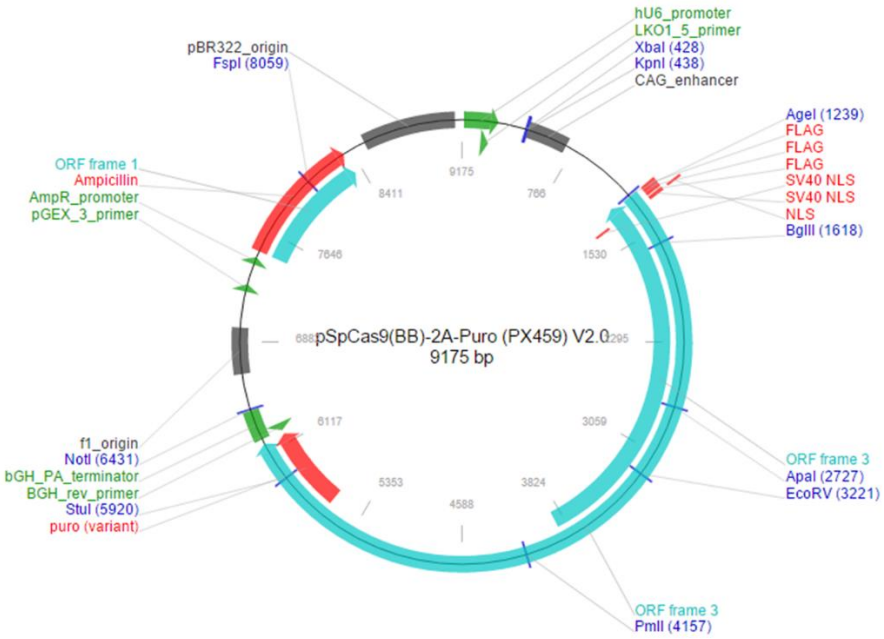


Figure 3-12: pSpCas9(BB)-2A-Puro Vector

Full pSpCas9(BB)-2A-Puro sequence map generated by Addgene. The sgRNA guide sequence oligo pair was cloned into the expression vector cloning site immediately downstream of the U6 promoter. The invariant scaffold section of the sgRNA immediately followed the cloning site. The vector also contained Cas9 with a fused 2A-puro which was later used for screening transfected cells.

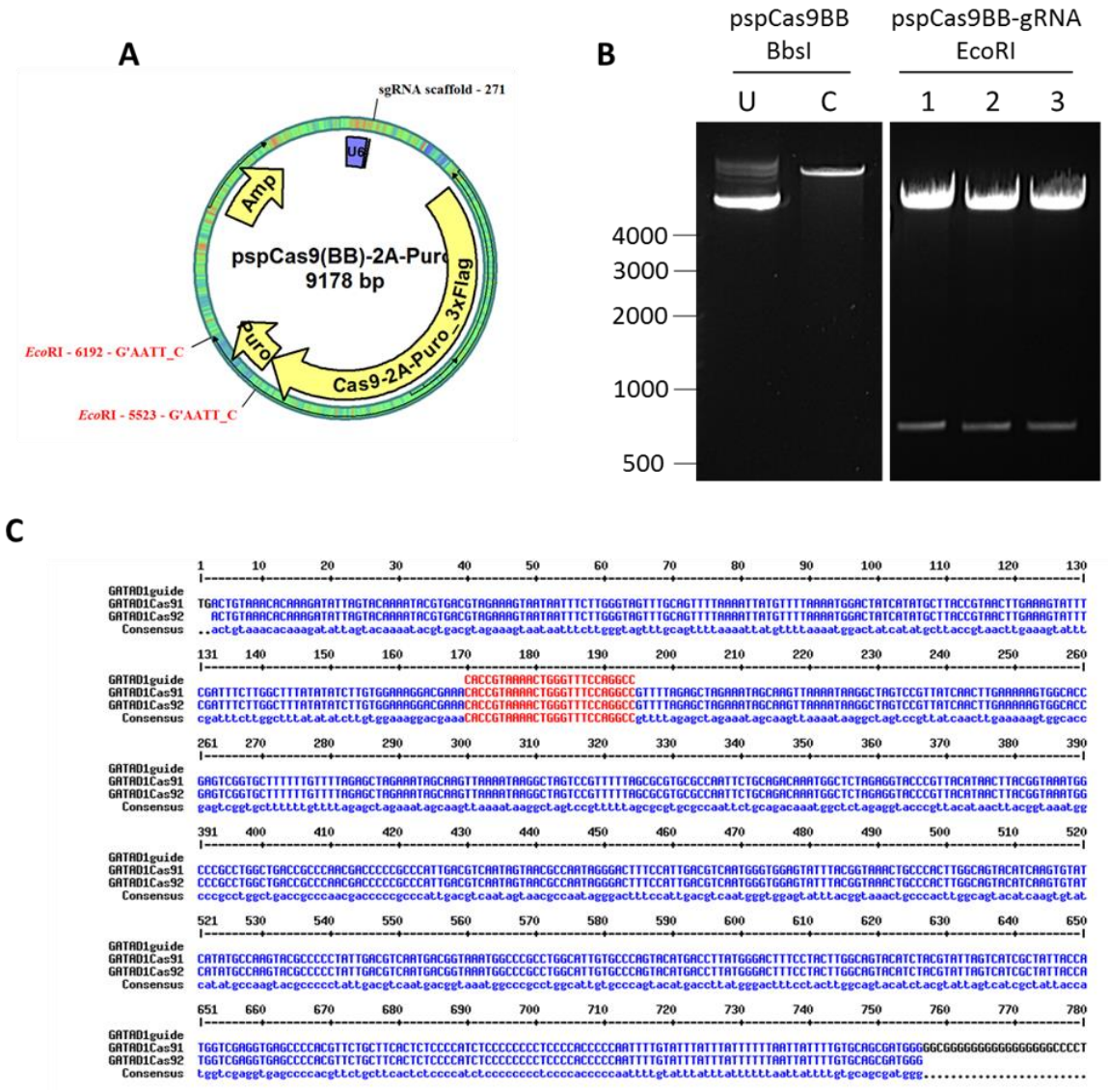


Figure 3-13: CRISPR sgRNA Cloning

(A) Virtual GATAD1 pspCas9-gRNA plasmid, indicating the position of the EcoRI restriction sites which were used for subsequent diagnostic digest. (B) A 0.8% agarose gel showing successful linearization of the vector prior to cloning. The supercoiled to linearised plasmid transition is evident by a decrease in mobility through the gel. An EcoRI diagnostic digest of recombinant pspCas9BB-gRNA produced a 669 bp fragment as well as the 8509 bp backbone. (C) The 20 base pair guide sequence sgRNA was successfully cloned into the Cas9 vector, downstream of the U6 RNA polymerase III promoter (U6 sequencing primer was used).

3.3.4.2 Design and Cloning of the CRISPR repair template expression construct

The GATAD1 CRISPR HDR template (Figure 3-14) was supplied by IDT and cloned into pcDNA3, which was first checked for complete linearization on an agarose gel (Figure 3-15A). An EcoRI/XhoI diagnostic digest of the recombinant clone dropped out the insert of approximately the correct size (Figure 3-15A). The pcDNA-gBlock clones were then sequenced with CMV forward to confirm the successful insert of the gBlock sequence (Figure 3-15B).

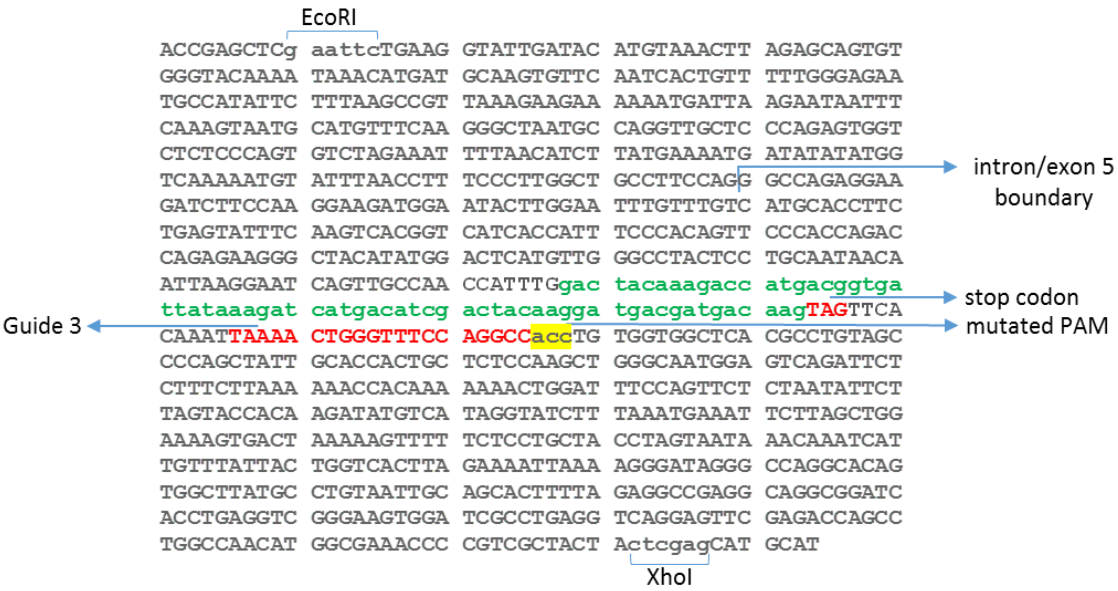


Figure 3-14: GATAD1 Repair Template Design

GATAD1 genomic sequence was used to design the repair template with homology arms, containing the 3xFLAG insert (green text), followed by the stop codon and a mutated PAM site (TGG to ACC) to prevent repeated rounds of HDR.

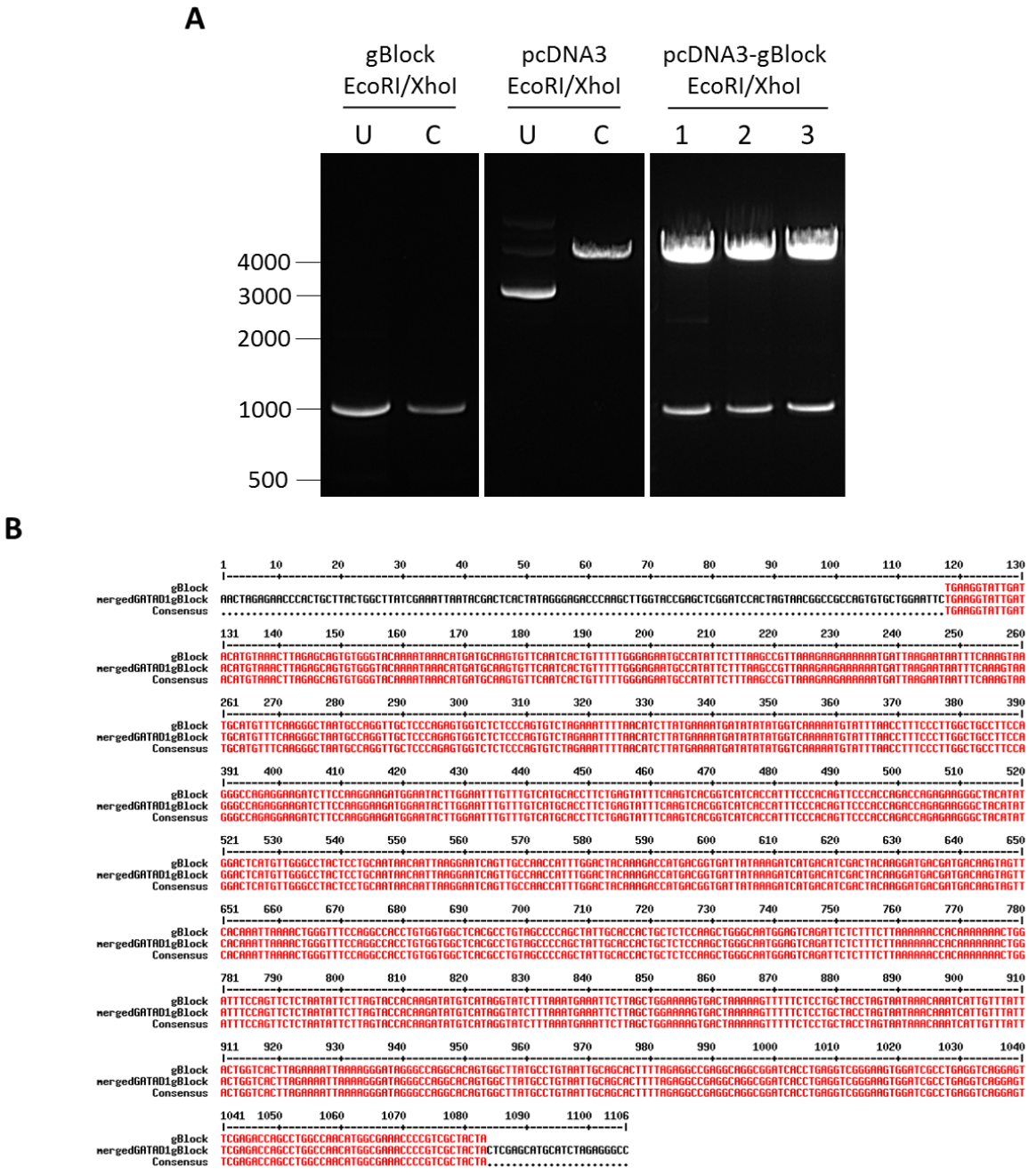


Figure 3-15: CRISPR Repair Template Cloning

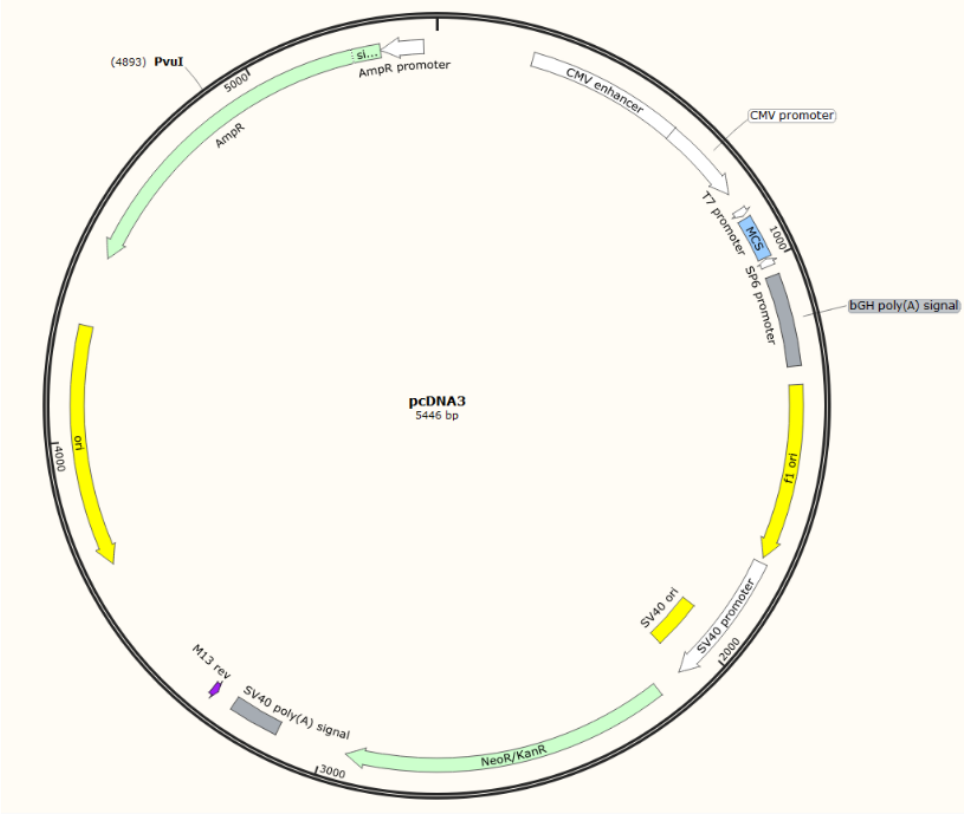
(A) A 0.8% agarose gel showing EcoRI/XhoI digest of the gBlock template as well as EcoRI/XhoI linearization of the pcDNA3 plasmid prior to cloning. A diagnostic digest was then carried out on the recombinant gBlock clone using EcoRI/XhoI to successfully drop out the gBlock insert. (B) Sequencing files (CMV forward and BGH reverse) were merged using the Emboss merger tool (http://emboss.bioinformatics.nl/cgi-bin/emboss/merger). The full GATAD1 gBlock repair template sequence was successfully cloned into the pcDNA3 vector.

3.3.4.3 HEK293 CRISPR Transfection and Screening

HEK293 cells were initially chosen to carry out the CRISPR genome editing over the standard HeLa cell line used in our laboratory. As well as being easy to transfect and grow, HEK293 cells are mostly diploid, except for being triploid for the X chromosome and tetraploid for chromosomes 17 and 22. HEK293 cells carry only 2 copies of GATAD1 (chromosome 7), increasing the efficiency of homozygous 3xFLAG insertion, compared to HeLa cells, (Lin et al., 2014). On the other hand, HeLa cells have a high level of aneuploidy with at least 3 copies of chromosome 7 which would make obtaining a clone homozygous for 3xFLAG-tagged GATAD1 more challenging, (Landry et al., 2013). Confirmation of AIC selection taking place within HEK293 cells as well as other cell lines was made prior to the CRISPR work taking place (Figure 5-3A, panel 3).

Prior to transfection of HEK293 cells, the HDR targeting plasmid (pcDNA3_gBlock) was linearised using PvuI in order to increase transfection efficiency (Figure 3-16). The single PvuI site was away from the homology arms and promoter sequences. Once clonal cell expansions had been made from successfully transfected, single-seeded cells (15% transfection efficiency), validation of the 3xFLAG-tag insertion was made using PCR screening. A panel of 47 clonally derived HEK293 cell lines were screened using primers flanking the 3xFLAG-tag insertion, ensuring that the correct size insertion had taken place at the correct position in the gene (Figure 3-17). 14 homozygous clones were then screened using a forward primer from within the 3xFLAG-tag and the reverse flanking primer, confirming the presence of a 3xFLAG-tag at the correct position in the gene in all 14 clones (30% complete HDR efficiency) (Figure 3-18). Five clones were sequenced by Eurofins using the initial genomic flanking forward primer, (Figure 3-19A). All five clones had inserted the 3xFLAG sequence upstream of the GATAD1 stop codon; this was also confirmed by a BLAST search (Figure 3-19B) and shown to be in-frame using the pDraw translation tool (Figure 3-19C).

CRISPR clones A8, A9, B7 and B10 were cultured (clone C11 did not resurrect), harvested and whole cell lysates were run on a Western blot alongside over-expressed full length 3xFLAG-tagged GATAD1. The three isoforms observed in over-expressed GATAD1 were also seen in the CRISPR clones, confirming the endogenous use of -207 CUG and -45 AUU as well as the annotated +1 AUG codon (Figure 3-20A). CRISPR 3xFLAG-GATAD1 was required in order to detect endogenous GATAD1 protein, since commercially available antibodies were poor or produced non-specific bands on a Western blot (Figure 3-20C).



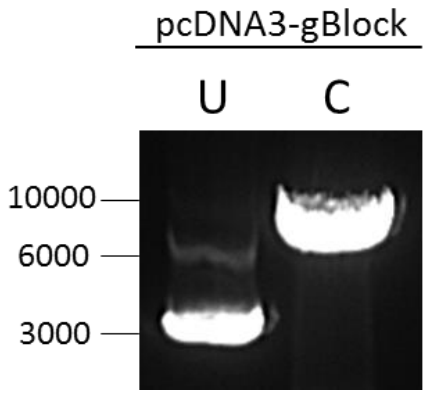


Figure 3-16: PvuI Linearisation of pcDNA3_gBlock Prior to Transfection

Complete linearization of the repair plasmid was made using the PvuI site (where U=uncut, C=cut), indicated in the virtual plasmid.

A

Primers Used for Screen	Primer sequence (5'-3')
GATAD1 CRISPR F	AATGCCAGGTTGCTCCCAGAGTGGT
GATAD1 CRISPR R	GCCCTATCCCTTTTAATTTTCTAAG

B

```
ACCGAGCTCg aattcTGAAG GTATTGATAC ATGTAAACTT AGAGCAGTGT
GGGTACAAAA TAAACATGAT GCAAGTGTTC AATCACTGTT TTTGGGAGAA
TGCCATATTC TTTAAGCCGT TAAAGAAGAA AAAATGATTA AGAATAATTT
CAAAGTAATG CATGTTTCAA GGGCTAATGC CAGGTTGCTC CCAGAGTGGT
CTCTCCCAGT GTCTAGAAAT TTTAACATCT TATGAAAATG ATATATATGG
TCAAAAATGT ATTTAACCTT TCCCTTGGCT GCCTTCCAGG GCCAGAGGAA
GATCTTCCAA GGAAGATGGA ATACTTGGAA TTTGTTTGTC ATGCACCTTC
TGAGTATTTC AAGTCACGGT CATCACCATT TCCCACAGTT CCCACCAGAC
CAGAGAAGGG CTACATATGG ACTCATGTTG GGCCTACTCC TGCAATAACA
ATTAAGGAAT CAGTTGCCAA CCATTTGgac tacaaagacc atgacggtga
ttataaagat catgacatcg actacaagga tgacgatgac aagTAGTTCA
CAAATTAAAA CTGGGTTTCC AGGCCacctG TGGTGGCTCA CGCCTGTAGC
CCCAGCTATT GCACCACTGC TCTCCAAGCT GGGCAATGGA GTCAGATTCT
CTTTCTTAAA AAACCACAAA AAAACTGGAT TTCCAGTTCT CTAATATTCT
TAGTACCACA AGATATGTCA TAGGTATCTT TAAATGAAAT TCTTAGCTGG
AAAAGTGACT AAAAAGTTTT TCTCCTGCTA CCTAGTAATA AACAAATCAT
TGTTTATTAC TGGTCACTTA GAAAATTAAA AGGGATAGGG CCAGGCACAG
TGGCTTATGC CTGTAATTGC AGCACTTTTA GAGGCCGAGG CAGGCGGATC
ACCTGAGGTC GGGAAGTGGA TCGCCTGAGG TCAGGAGTTC GAGACCAGCC
TGGCCAACAT GGCGAAACCC CGTCGCTACT ActcgagCAT GCAT
```

C

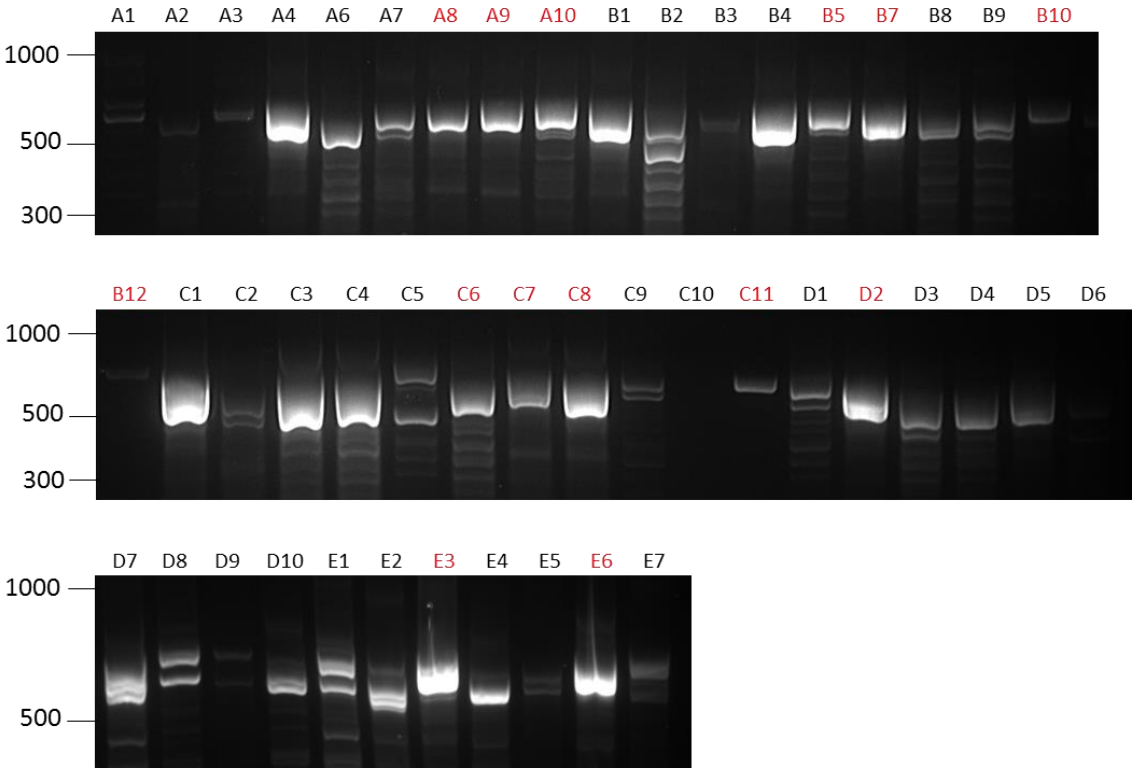


Figure 3-17: HEK293 GATAD1 CRISPR PCR Screening

(A) Sequencing primers were designed to amplify the region surrounding the 3xFLAG insert, as well as the 3xFLAG insert itself, ensuring the 66 bp homozygous modification had been made in the correct position within the gene. (B) Initial screening used primers (red arrows) flanking the 3xFLAG-tag insertion (green text), which produced a 600/666 bp amplicon dependant on successful insertion of the 3xFLAG-tag. The GATAD1 guide target is shown in red text, with the PAM sequence highlighted. (C) 2% agarose gel showing initial CRISPR screens, using GATAD1 3xFLAG flanking primers. The 14 clones which are labelled with red text appear to be homozygous for the 3xFLAG-tag insertion and are taken through to the next screening step. Clones A7, B9, C2, D1, D8, E1 and E5 are heterozygous for the insertion.

A

Primers Used for Screen	Primer sequence (5'-3')
FLAG CRISPR F	gacagccGACTACAAAGACCA
GATAD1 CRISPR R	GCCCTATCCCTTTTAATTTTCTAAG

B

```
ACCGAGCTCg aattcTGAAG GTATTGATAC ATGTAAACTT AGAGCAGTGT
GGGTACAAAA TAAACATGAT GCAAGTGTTC AATCACTGTT TTTGGGAGAA
TGCCATATTC TTTAAGCCGT TAAAGAAGAA AAAATGATTA AGAATAATTT
CAAAGTAATG CATGTTTCAA GGGCTAATGC CAGGTTGCTC CCAGAGTGGT
CTCTCCCAGT GTCTAGAAAT TTTAACATCT TATGAAAATG ATATATATGG
TCAAAAATGT ATTTAACCTT TCCCTTGGCT GCCTTCCAGG GCCAGAGGAA
GATCTTCCAA GGAAGATGGA ATACTTGGAA TTTGTTTGTC ATGCACCTTC
TGAGTATTTC AAGTCACGGT CATCACCATT TCCCACAGTT CCCACCAGAC
CAGAGAAGGG CTACATATGG ACTCATGTTG GGCCTACTCC TGCAATAACA
ATTAAGGAAT CAGTTGCCAA CCATTTGgac tacaaagacc atgacggtga
ttataaagat catgacatcg actacaagga tgacgatgac aagTAGTTCA
CAAATTAAAA CTGGGTTTCC AGGCCacctG TGGTGGCTCA CGCCTGTAGC
CCCAGCTATT GCACCACTGC TCTCCAAGCT GGGCAATGGA GTCAGATTCT
CTTTCTTAAA AAACCACAAA AAAACTGGAT TTCCAGTTCT CTAATATTCT
TAGTACCACA AGATATGTCA TAGGTATCTT TAAATGAAAT TCTTAGCTGG
AAAAGTGACT AAAAAGTTTT TCTCCTGCTA CCTAGTAATA AACAAATCAT
TGTTTATTAC TGGTCACTTA GAAAATTAAA AGGGATAGGG CCAGGCACAG
TGGCTTATGC CTGTAATTGC AGCACTTTTA GAGGCCGAGG CAGGCGGATC
ACCTGAGGTC GGGAAGTGGA TCGCCTGAGG TCAGGAGTTC GAGACCAGCC
TGGCCAACAT GGCGAAACCC CGTCGCTACT ActcgagCAT GCAT
```

C

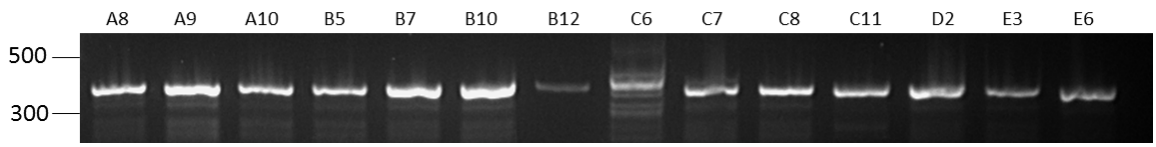
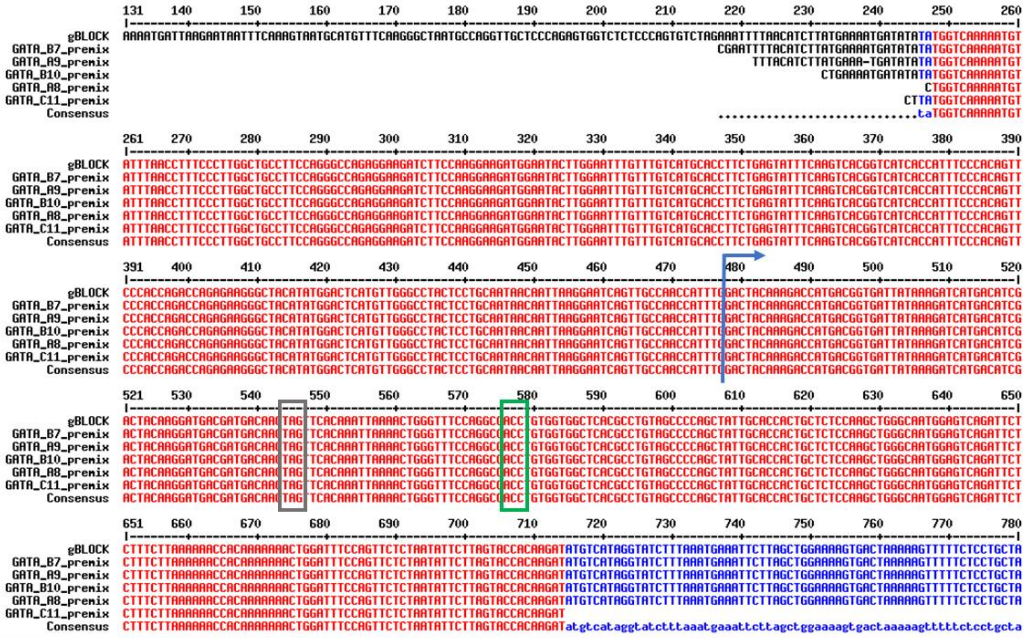


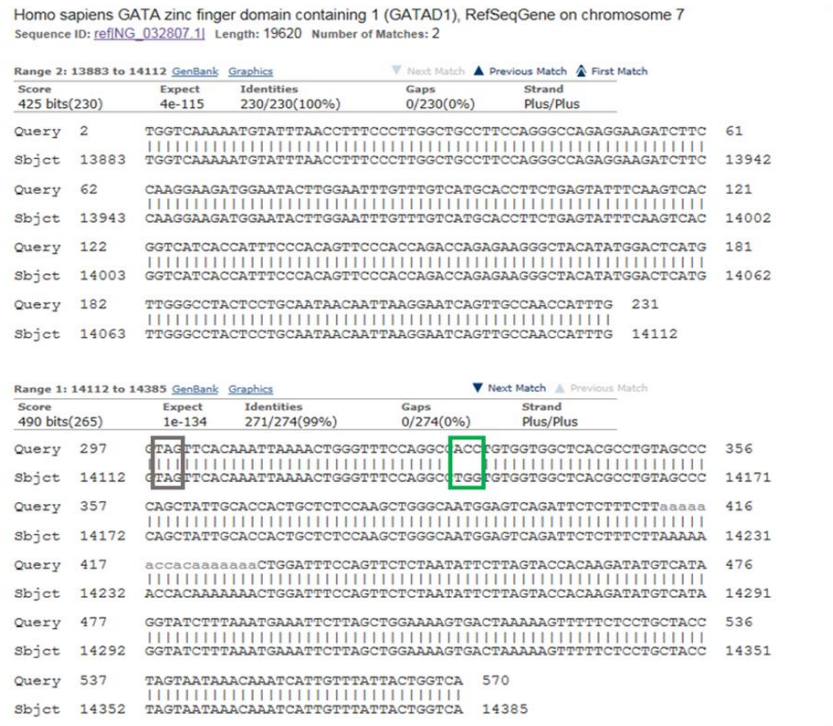
Figure 3-18: HEK293 FLAG CRISPR PCR Screen

(A) The second round of screening used the FLAG forward primer with the original flanking reverse primer. (B) This resulted in amplification of a region of 364 bp upon successful insertion of 3xFLAG-tag. (C) FLAG screens of the homozygous HEK293 clones produced a 364 bp amplicon, indicating that a 3xFLAG-tag had been introduced in all 14 clones.

A


gBLOCK
GATA_B7_prenix
GATA_A9_prenix
GATA_B10_prenix
GATA_A8_prenix
GATA_C11_prenix
Consensus

B


Homo sapiens GATA zinc finger domain containing 1 (GATAD1), RefSeqGene on chromosome 7
Sequence ID: refING_032807.1| Length: 19620 Number of Matches: 2
Range 2: 13883 to 14112 GenBank Graphics
Next Match Previous Match First Match
Score 425 bits(230)
Expect 4e-115
Identities 230/230(100%)
Gaps 0/230(0%)
Strand Plus/Plus
Range 1: 14112 to 14385 GenBank Graphics
Next Match Previous Match
Score 490 bits(265)
Expect 1e-134
Identities 271/274(99%)
Gaps 0/274(0%)
Strand Plus/Plus

C

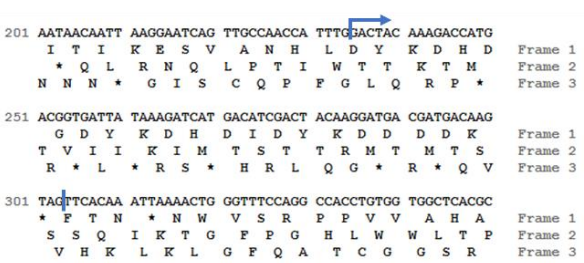

Frame 1
Frame 2
Frame 3

Figure 3-19: HEK293 GATAD1_3xFLAG-CRISPR Clone Sequencing

(A) All 5 sequenced clones (A8, A9, B7, B10 and C11) are homozygous for the 3xFLAG-tag insertion at the C-terminus of the GATAD1 gene. The blue arrow indicates the start of the 3xFLAG sequence, the grey box shows the stop codon and the green box indicates the position of the mutated PAM sequence (TGG to ACC). (B) BLAST returned the GATAD1 sequence in two sections, with query sequence 231-297 absent where the 66 bp FLAG insert is present in the CRISPR clone. The grey box indicated the stop codon and the green box indicated the mutated PAM sequence within the clone. (C) The pDraw translation tool showed that the 3xFLAG-tag was in-frame (Frame 1) with the GATAD1 sequence ending in VANHL.

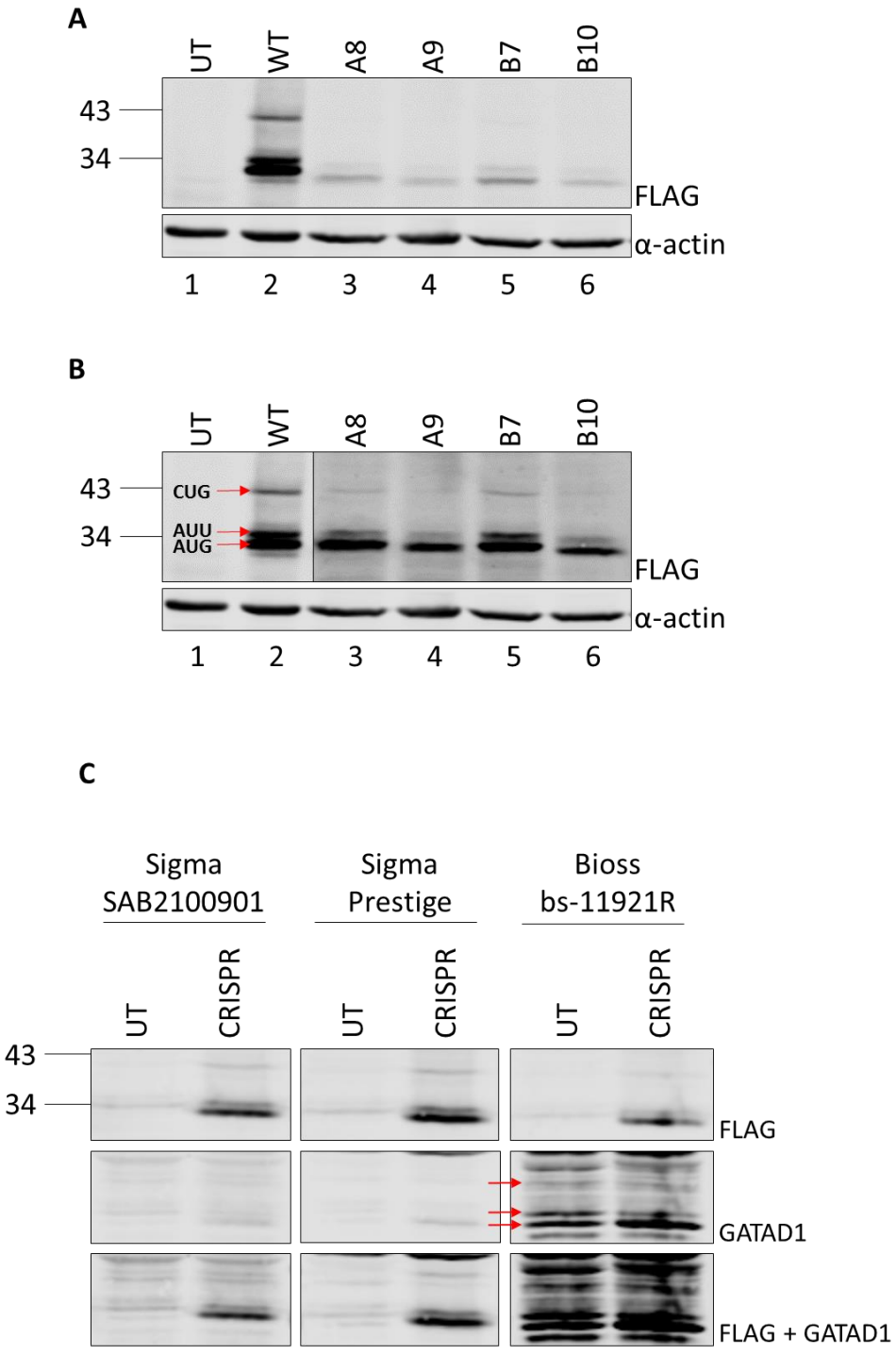


Figure 3-20: HEK293 CRISPR-GATAD1_3xFLAG Clones Express Three Isoforms

(A) FLAG-probed western blot showing HEK293 3xFLAG-tagged GATAD1 homozygous clones, alongside overexpressed WT GATAD1 and untransfected lysate (UT). 20 µg of whole cell lysate was loaded on to a 10% SDS-PAGE gel. (B) Increasing the contrast of the CRISPR clones enables clearer identification of the three endogenous GATAD1 isoforms. (C) Both Sigma antibodies (SAB2100901/Prestige) were unable to detect endogenous GATAD1. The Bioss bs-11921R antibody detected various non-specific bands, however comparison to the CRISPR 3xFLAG lane enabled identification of the three GATAD1 isoforms.

3.3.5 HeLa CRISPR/Cas9 Attempt to Confirm Endogenous Alternative GATAD1 Isoforms

Since HeLa cells were the standard cell line used in the Coldwell laboratory and were previously used to identify GATAD1 AICs (section 3.3.2) as well as being used in subsequent localisation and functional experiments, it would be succinct to confirm endogenous GATAD1 expression in this cell line as well as in the HEK293 cell line shown in the previous section (section 3.3.4). Once clonal cell expansions had been made from successfully transfected, single-seeded HeLa cells (15% transfection efficiency), validation of the 3xFLAG-tag insertion was made using PCR screening. A panel of 48 cell lines were screened using primers flanking the 3xFLAG-tag insertion, ensuring that the correct size insertion had taken place at the correct position in the gene. Only 16 of the original 48 screens produced a PCR product, since many of the clones were not confluent enough when lysed. Unsuccessful PCR clones were disregarded as they would be unlikely to resurrect if required. All 16 potentially successful clones were also screened using a forward primer from within the 3xFLAG-tag and the reverse flanking primer (Figure 3-21A). Clone D3 appeared to be the only clone (2% HDR efficiency) containing the homozygous insert and was therefore sequenced using the initial genomic flanking forward primer (Figure 3-21B). However, the D3 positive cell line would not resurrect from -80°C and so a further round of HeLa CRISPR was attempted.

The second round of HeLa CRISPR resulted in screening 11 clonal cell lines (4% transfection efficiency) using both flanking primers and a FLAG primer, as before (Figure 3-22). Clones A4 and B2 appeared to have the homozygous 3xFLAG insertion and were sent to be sequenced, however did not contain the 3xFLAG insert (Figure 3-22B).

The third round of HeLa CRISPR screening resulted in screening 35 clonal cell lines (7% transfection efficiency) using the same PCR screening methods as previously described (Figure 3-23). None of the clones successfully incorporated the 3xFLAG tag. Clone A8 was heterozygous for the 3xFLAG insertion which would be sufficient to run a Western blot and observe the expression pattern, however A8 did not resurrect from -80°C storage. As mentioned previously, the difficulties in obtaining a homozygous HeLa 3xFLAG-GATAD1 CRISPR clone is likely due to the aneuploidy of HeLa cells, which contain at least three copies of the GATAD1 gene (Landry et al., 2013).

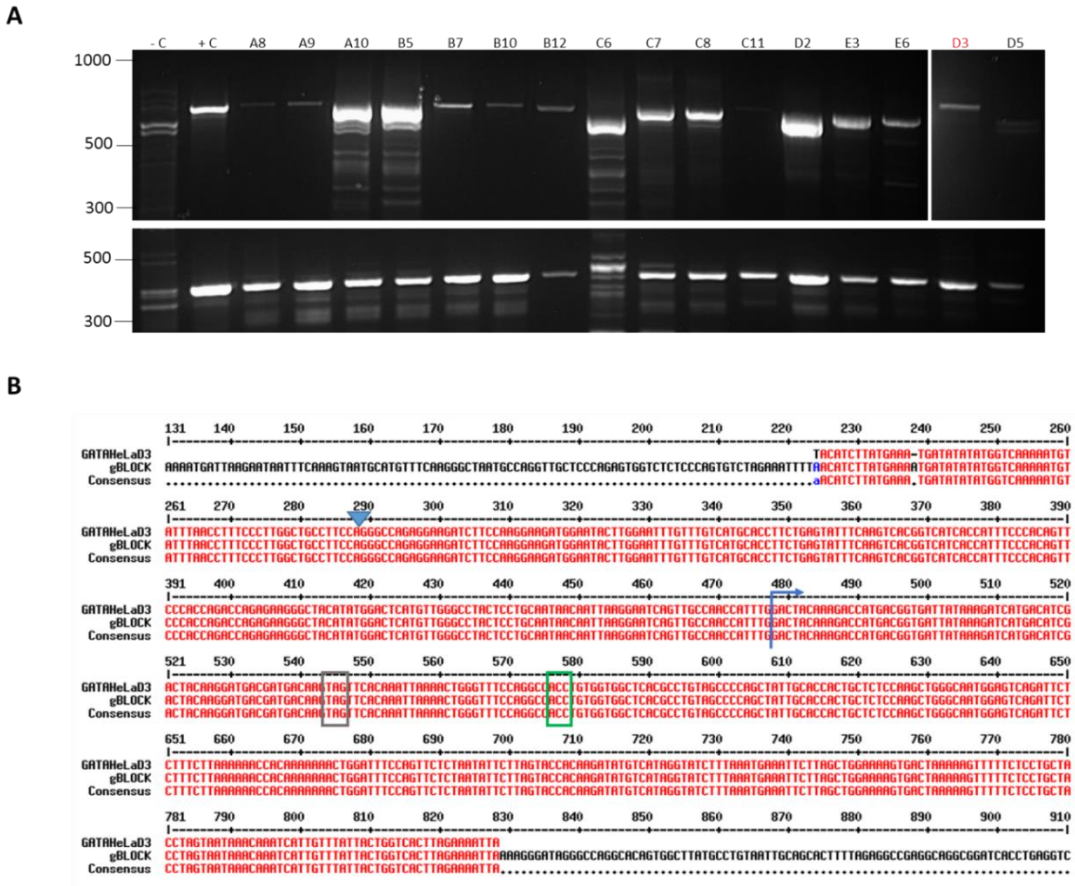


Figure 3-21: Initial HeLa GATAD1 CRISPR PCR Screens

(A) The upper 2% agarose gel shows PCR screening using both flanking GATAD1 primers, which would amplify either 600/666 bp dependent on the FLAG insertion. The negative control used untransfected genomic GATAD1, whilst the gBlock template plasmid was used for the positive control reactions. The lower 2% agarose gel shows PCR screens using the FLAG forward primer, amplifying a 364 bp product. (B) The sequencing results confirmed that clone D3 was homozygous for the 3xFLAG-tag insertion at the C-terminus of the GATAD1 gene. The blue arrow indicates the start of the 3xFLAG sequence, the grey box shows the stop codon and the green box indicates the position of the mutated PAM sequence (TGG to ACC). There was a deletion at position 238, however this was within the intronic sequence (intron-exon boundary indicated by the blue arrow at position 288) and would have been spliced before translation was to take place, therefore not affecting the ORF.

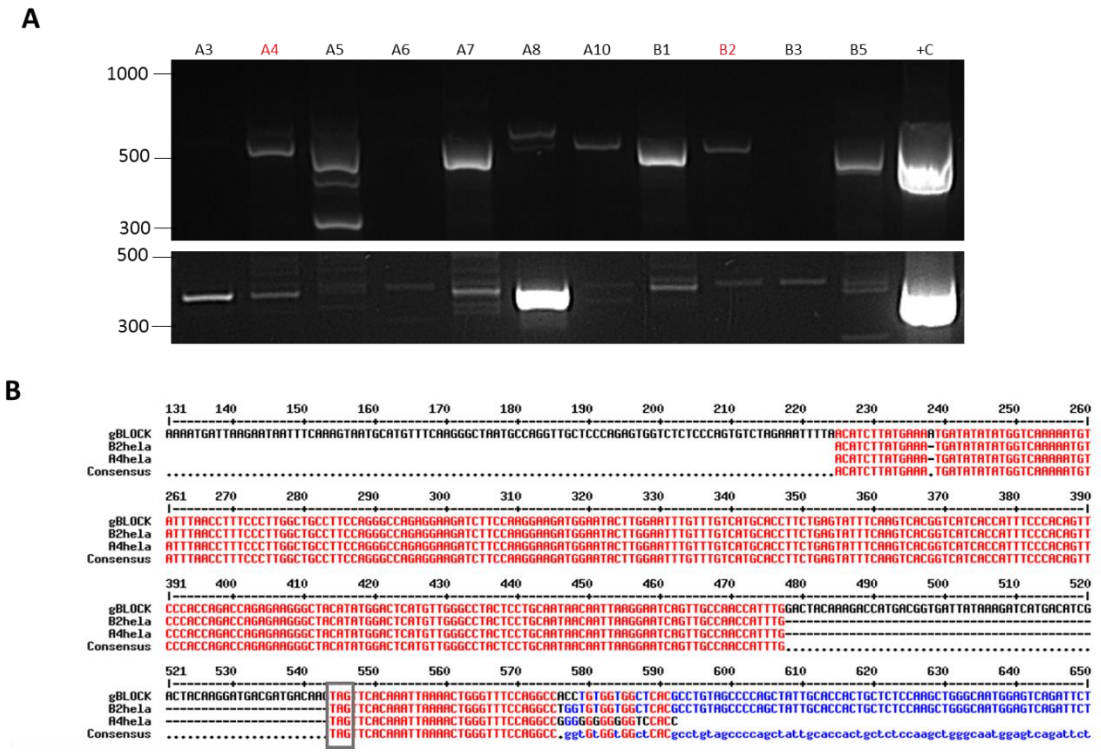


Figure 3-22: Second Round of HeLa GATAD1 CRISPR PCR Screens

(A) The upper 2% agarose gel shows PCR screening using both flanking GATAD1 primers, which would amplify either 600/666 bp dependent on the FLAG insertion. The lower 2% agarose gel shows PCR screens using the FLAG forward primer, amplifying a 364 bp product. Clones A4 and B2 appeared to have the homozygous 3xFLAG insertion. (B) Neither A4 nor B2 HeLa CRISPR clone contained the 3xFLAG insert.

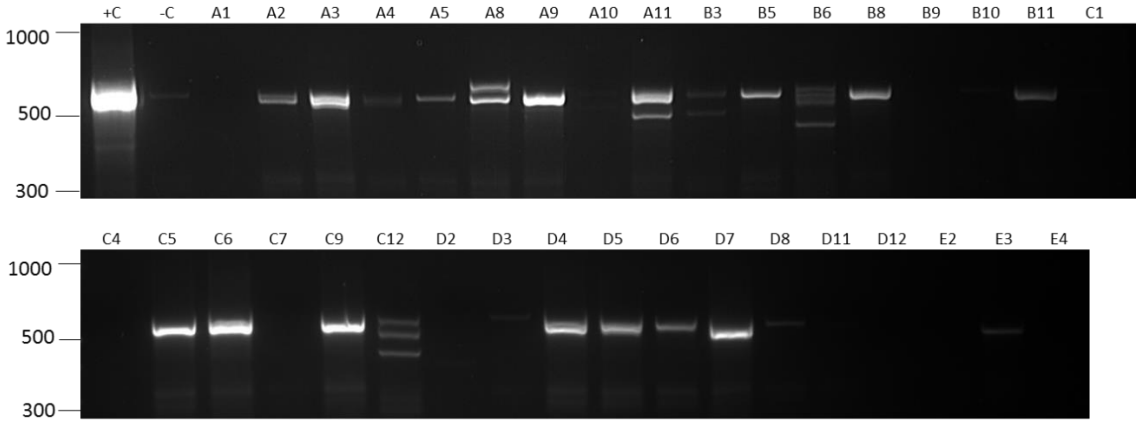


Figure 3-23: Third Round of HeLa GATAD1-CRISPR PCR Screens

PCR screening using flanking GATAD1 primers, which amplify 600/666 bp dependent on the presence of the 3xFLAG insertion. None of the clones were homozygous for the insertion. Clone A8 was heterozygous, but did not successfully resurrect from -80°C.

3.4 Summary of Main Findings

- Alternative initiation takes place within the GATAD1 transcript to generate N-terminally extended isoforms.
- A CUG at position -207 as well as an AUU at position -45 are utilised by the ribosome within the 5'UTR to initiate translation.
- The levels of GATAD1 mRNA within the cell remains unchanged when mutagenesis takes place, indicating that the changes in protein expression observed on the Western blot are due to translational control (alternative initiation).

3.5 Discussion

3.5.1 CUG and AUU AICs Are Utilised by GATAD1

The initial aim of this project was to confirm that AICs are recognised within the 5'UTR of the GATAD1 transcript and are subsequently utilised by the translational machinery, which would result in the translation of N-terminally extended proteins. Both potential AIC's identified by the macro, -207 CUG and -144 CUG were within a strong Kozak consensus; AUCCUGG and GGCCUGG respectively. Both the -45 AUU and the canonical +1 AUG have an intermediate Kozak consensus, lacking a G at position +4 relative to the first base of the start codon; GCCAUUC and ACCAUGC, respectively. All other potential AICs (-39 GUG, -33 CUG and +55 AUG) are within a weak Kozak consensus, lacking both a G at position +4 and a purine at position -3. It would be reasonable to predict that the ribosome would utilise the initiation codons within a strong Kozak consensus to initiate translation. The Wegrzyn consensus is thought to encompass the context of non-AUG initiation codons more accurately, which require CG at position -7 and -6 respectively for efficient translation to take place. All of the predicted AICs are within an intermediate Wegrzyn context, except -33 CUG which has a full Wegrzyn context.

Wild-type GATAD1 produced three protein isoforms when run on a Western blot (Figure 3-7B, lane 2). Mutagenesis identified two AICs within the GATAD1 transcript, -207 CUG and -45 AUU as well as the annotated AUG. The UAC mutants of each initiation codon confirmed that the isoforms were individually translated and are not breakdown products of a larger isoform, or modifications of the annotated isoform. The ratio of expression of each GATAD1 isoform is 1:1.3:3 (-207, -45, +1 isoform respectively), suggesting that the -45 AUU is translated slightly more

efficiently than the -207 CUG. Although CUG codons are well established AICs, AUU codons are far less common. Numerous experiments have calculated relative efficiencies of translation initiation from a CUG codon versus an AUU codon. Firstly, Peabody (Peabody, 1989) showed that a CUG codon initiates translation of full length DHFR in rabbit reticulocyte lysates with 82% efficiency, as opposed to 36% in wheat germ extracts; whereas an AUU had efficiencies of 67% and 14% respectively. Ivanov used a firefly reporter in HEK293 cells to show that a CUG codon has a 19.5% efficiency relative to an AUG, whereas an AUU has just 3.2% efficiency (Ivanov et al., 2011). Similarly, a firefly reporter with a C-terminal PEST domain has been used to show that when the CUG codon is in a strong Kozak consensus, it has an 18% relative efficiency when in HeLa cells and 16% efficiency in HEK293 cells; whereas an AUU codon has just a 1.1% efficiency in HeLa cells and 0.4% efficiency in HEK293 cells, (Stewart et al., 2015). Stewart et al also showed that when in a Wegrzyn non-AUG context, CUG has a 16% efficiency in HeLa cells and a 12% efficiency in 293 cells; as opposed to the AUU codon which has just 0.7% efficiency in HeLa cells and a 0.4% efficiency in HEK293 cells. In all cases, the AUU is far less efficient than the CUG at initiating translation. An interesting point to consider is why the -45 AUU within a weak Kozak consensus was used to initiate translation of GATAD1, whereas the -144 CUG within a strong Kozak consensus was not used, even though it had been picked up during ribosome profiling by Lee et al, (Lee et al., 2012). This may be due to the structure or sequence of the RNA itself, or indeed other factors acting on the mRNA transcript, which will be discussed in the following chapters.

AUU codons are rarely used to initiate translation and a single eukaryotic example has currently been experimentally identified in addition to GATAD1. An AUU-initiated PTEN (phosphatase and tensin homologue on chromosome ten) isoform is more abundant than multiple CUG-initiated isoforms (Tzani et al., 2016). PTEN tumour suppressor gene is often mutated in human cancers and autism spectrum disorders and all alternative isoforms are still able to negatively regulate cell survival by dephosphorylating PIP3 and therefore inhibiting PI3K signalling. The HEK293 ribosome profiling dataset (Lee et al., 2012) identified AUU translation initiation sites within many genes, including ZNF737 (DNA-binding-zinc finger protein), TIMM23 (translocase of inner mitochondrial membrane 23), RBM6 (ribosome binding motif 6), DDX5, DDX6, DDX17, DDX39B and DDX46 (DEAD-box helicases). AUU as an initiation codon has been more extensively studied in prokaryotes; several prokaryotic genes translate from an AUU and use this as a mechanism to regulate gene expression. For example, the *Enterococcus faecalis* EbpA (endocarditis and biofilm-associated pili A) uses -120 AUU to decrease EbpA surface display in order to attenuate biofilm and reduce adherence to fibrinogen, (Montealegre et al., 2015).

3.5.2 Conflicts Between Predicted and Experimentally Proven AICs

There are conflicts between the sites predicted to initiate translation and those experimentally proven to initiate translation of GATAD1 isoforms. All prediction models as well as ribosome profiling predicted translation to initiate from the -144 CUG. The macro used sequence context to predict potential in-frame initiation sites, and the -144 CUG (GGCCUGG) is within a strong Kozak consensus. However, other factors also play a role in start site selection, including secondary structure of the mRNA downstream of the initiation codon as well as various trans-acting factors, which will be investigated in following chapters. On the other hand, the ribosome profiling study (Lee et al., 2012) uses GTI-seq (global translation initiation-sequencing) to precisely identify the position of initiating ribosomes. The most common protein synthesis inhibitor used in ribosome profiling studies is cycloheximide (CHX), which stabilises ribosomes on mRNA and binds to the E-site of the 60S ribosomal subunit of an assembled 80S ribosome, near to where the deacylated tRNA normally binds. Lee used lactimidomycin (LTM), which is similar in structure to CHX, but larger in size. This causes preferential inhibition of initiating ribosomes since LTM can only bind to the empty E-site. Side-by-side comparisons of CHX and LTM-inhibited ribosomes allowed reduction of background noise and precise identification of initiating ribosome positions by deep-sequencing of ribosome-protected fragments (RPFs), including position -144 CUG of GATAD1. A possible explanation is that GATAD1 may be subject to tissue-specific regulation of translation initiation sites. On the other hand, GTI-seq has since been shown to generate a high number of artefacts, due to the 37°C incubation period required for run-off of elongating ribosomes after LTM-treatment (Gao et al., 2015). A final discrepancy between predictions and experimental data was that the -207 CUG was not predicted in the Pre-TIS search or ribosome profiling data. The Pre-TIS tool used start site conservation between human and mouse as a classifier of translation initiation sites. Both codon and amino acid conservation were considered – although the codon at -207 CUG is conserved between human and mouse, the protein sequence is not (Figure 3-2) which may explain the absence of this CUG in the prediction results.

3.5.3 CRISPR Confirms Endogenous CUG and AUU

CRISPR genome editing was used to insert a 3xFLAG-tag at the C-terminus of genomic GATAD1 in HEK293 cells, which confirmed the endogenous use of the -207 CUG and -45 AUU identified when over-expressing GATAD1. Over-expressing a gene creates an artificial enviroment and may cause the proteins to interact or localise differently in the cell.

Although the CRISPR 3xFLAG-tag insertion was successful in HEK293 cells, it was not successful following several rounds in HeLa cells. This may have been due to HeLa cells containing high levels of aneuploidy and chomosomal instability (Landry et al., 2013).

3.5.4 qPCR Confirms Translational Effect of SDM

GATAD1 isoform expression changed when SDM was carried out in order to identify AICs within the transcript. qPCR confirmed that all of the mutants retained approximately the same GATAD1 mRNA levels, when compared to the wild-type. SDM was therefore not affecting the GATAD1 mRNA levels and all changes seen in protein expression were due to alternative translation initiation taking place.

Chapter Four

Regulation of GATAD1 Translation via mRNA Signals

4. Regulation of GATAD1 Translation by mRNA Signals

4.1 Introduction

As discussed in the previous chapter, the sequence context of an AIC plays an important role in the efficiency of translation initiation. Alternative translation initiation most commonly occurs when the annotated start codon has a weak Kozak consensus. Similarly, an AIC is most likely to be recognised by the ribosome if it is within a strong Kozak/Wegrzyn context. In the case of GATAD1, the annotated AUG is indeed within a weak Kozak consensus, whereas the -207 CUG is within a strong Kozak, but intermediate Wegrzyn consensus (GGGAAUCCUGG) and the -45 AUU is within a weak Kozak but intermediate Wegrzyn consensus (GUCCGCCAUUC). In addition to the interesting context of the AUU initiation codon in particular, AUU codons have been shown to be very inefficient at initiating translation under physiological conditions, with a 0.7% efficiency compared to an AUG codon in HeLa cells, (Stewart et al., 2015). The -144 CUG codon which was predicted but not utilised by the ribosome, is within a strong Kozak consensus, yet the ribosome leaks through this codon to the AUU which is within a weak context. This suggests that factors other than sequence context must influence AIC selection by the ribosome.

A downstream secondary structure within the 5'UTR is able to increase the efficiency of AIC recognition and compensate for a suboptimal start codon, (Kozak, 1990). The structure is most effective if positioned between 13-17 nucleotides downstream of the AIC, since this corresponds to the approximate distance between the leading edge of the ribosome and the P-site where the Met-tRNAi anticodon is positioned. Scanning of the 40S ribosomal subunit is stalled whilst helicases melt the mRNA structure, resulting in increased recognition of an AIC within a sub-optimal context.

A further signal within the mRNA that may also cause increased recognition of sub-optimal AICs via stalling of the ribosome, are sequences which encode polyproline sequences. Proline has a slow rate of amino acid polymerisation due to the pyrrolidine ring spanning the α-carbon and nitrogen of the backbone, which causes proline to be a poor A-site acceptor of amino-acyl tRNA during peptide-bond formation, as well as a poor donor when in the P-site (Pavlov et al., 2008). Consequently, polyproline sequences cause ribosomal stalling. The sequence context of the proline motif influences the strength of ribosomal stalling. The presence of Arg/His upstream of a PPP motif increased stalling, whilst Asp/Ala preceding or Trp/Asp/Asn/Gly following a PP motif was necessary for ribosomal stalling to take place (Starosta et al., 2014).

4.2 Hypothesis and Aims

4.2.1 Hypothesis

mRNA sequences and structures influence start codon selection within the GATAD1 transcript

4.2.2 Aims

The overall aim of this chapter was to determine whether signals within the mRNA were influencing the ribosome to select alternative initiation codons within the GATAD1 transcript. This was carried out by:

- Using prediction models to determine whether secondary structure is present within the GATAD1 5'UTR.
- Mutagenesis of secondary structures to identify whether these could be affecting start codon selection.
- Mutagenesis of polyproline motifs present within the transcript to identify whether these too could be affecting start codon selection.

4.3 Results

4.3.1 Signals within GATAD1 5'UTR Which May Regulate Translation

Secondary structure within the 5'UTR may promote translation from an AIC which may otherwise be inefficient or not recognised. The CENTROIDFOLD web server (http://www.ncrna.org/centroidfold/) accurately predicts RNA secondary structures. Since the GATAD1 5'UTR (upstream of the annotated AUG initiation codon) had a high GC content of 75%, the sequence (Figure 4-1) was run through the prediction software to see whether secondary structure could be affecting AIC usage (Figure 4-2A). For enhanced translation initiation, the optimal distance between the AIC and a 3' hairpin is 13-17 nucleotides (Kozak, 1990). There are no obvious hairpins downstream of either the CUG or AUU AIC in the predicted structure, however there is a strong hairpin directly upstream of the -45 AUU which may regulate translation in an alternative way.

Within this upstream hairpin structure, there is a sequence resembling one which has been shown to pair with the 18S rRNA, in a manner analogous to the Shine-Dalgarno (SD) sequence. The SD sequence is commonly found 7 nucleotides upstream of the initiation codon in prokaryotic mRNA (Betts and Spremulli, 1994), and is used to align the ribosome with the initiation codon through direct 16S rRNA base pairing. Variations on this 9 nucleotide sequence (CCGGCGGGU) have shown direct 40S ribosomal subunit binding within eukaryotic messages, (Chappell et al., 2004). In addition, there is a GC rich sequence downstream of the -45 AUU, although it does not appear to form a hairpin structure. Since there are several potential points of translation regulation around the -45 AUU, mutagenesis was carried out to alter each signal in turn. The upstream hairpin (USHP) was mutated so that the structure would no longer form (Figure 4-2B), the GC-rich sequence downstream (DSHP) was mutated (Figure 4-2C), as well as the potential upstream SD-like sequence (USSDL) (Figure 4-2D). All mutagenesis used silent mutations to retain the amino acid sequence.

The effect of the GATAD1 mutants (USHP, DSHP and USSDL) on -45 AUU usage and subsequent -45 isoform expression was analysed by Western blotting (Figure 4-3A). Significantly less ($p=0.011$) -45 isoform was expressed by the DSHP mutant compared to WT GATAD1 (Figure 4-3B). The GC-rich sequence downstream of the -45 AUU may be regulating translation from the AUU codon, which would not usually be efficiently recognised. Both mutations within the hairpin upstream of -45 AUU (USHP and USSDL) had no effect on the expression of GATAD1 isoforms.

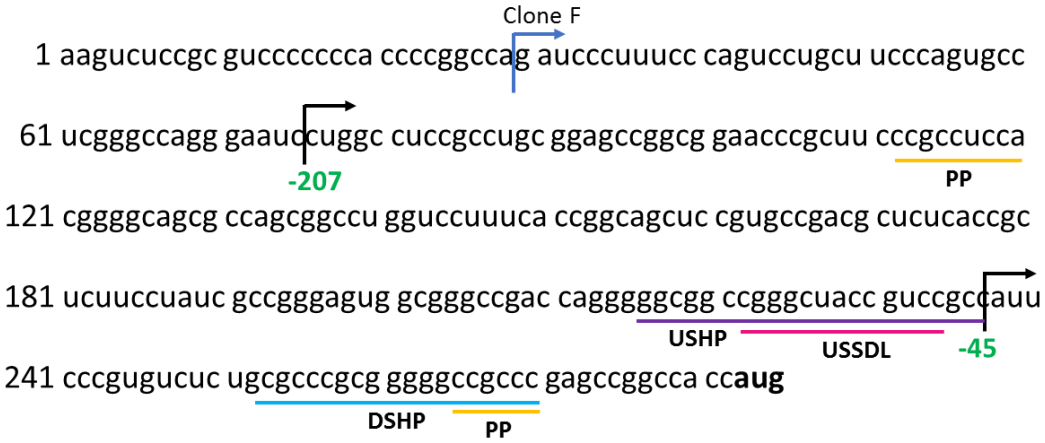
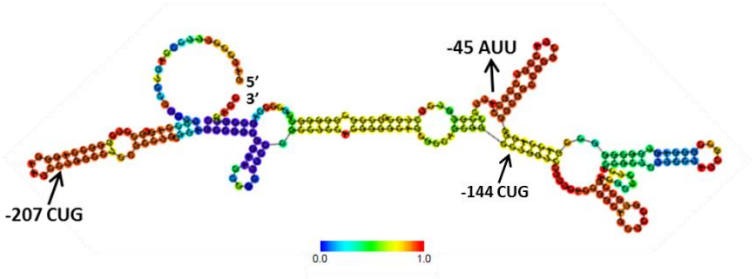


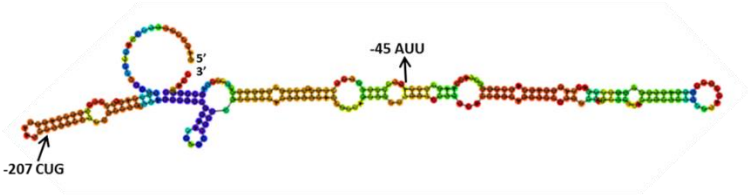
Figure 4-1: Human GATAD1 5'UTR Sequence Used to Predict RNA Secondary Structure

GATAD1 5'UTR (NM_021167.4) from Clone F (the sequence cloned into the 3xFLAG reporter) was analysed for secondary RNA structure using the CENTROIDFOLD server. The AICs are indicated with arrows and other sequences of interest are also shown, including polyproline sequences (PP), upstream hairpin (USHP), upstream Shine-Dalgarno-like sequence (USSDL) and downstream hairpin (DSHP).

A **WT**



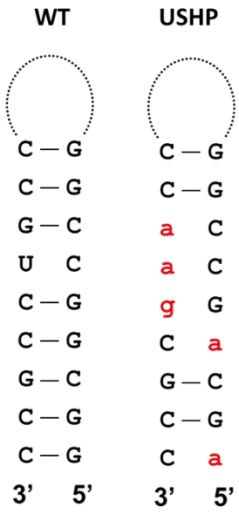
B **USHP**



CAG GGG GCG GCC GGG CUA CCG UCC GCC AUU

CAG GGa GCa GCC GGG CUA CCa agC GCC AUU

Q G A A G L P S A I



C **DSHP**

AUU CCC GUG UCU CUG CGC CCG CGG GGG CCG CCC GAG CCG GCC

AUU CCC GUG UCU CUG CGa CCa Cga GGa CCa CCa GAa CCG GCC

I P V S L R P R G P P E P A

D **USSDL**

CAG GGG GCG GCC GGG CUA CCG UCC GCC AUU

CAG GGG GCG GCa GGa CUA CCa UCa GCC AUU

Q G A A G L P S A I

Figure 4-2: GATAD1 5'UTR Mutants

(A) CentroidFold predicts structures that the RNA can adopt based on the minimum free energy taken to form these structures. The WT GATAD1 RNA folding result used the human 5'UTR sequence, which indicated a hairpin structure directly upstream of -45 AUU. (B) Silent mutations were made to the predicted hairpin upstream of -45 AUU (USHP) in order to prevent it from forming. (C) Silent mutations were made to the GC-rich sequence downstream of -45 AUU (DSHP), preventing any structures from forming which may not have been identified by the prediction software. (D) Silent mutations were also made to the GC-rich sequence upstream of -45 AUU which may be acting as a eukaryotic alternative to the Shine-Dalgarno sequence (USSDL), allowing direct ribosomal binding to initiate translation at the -45 AUU AIC.

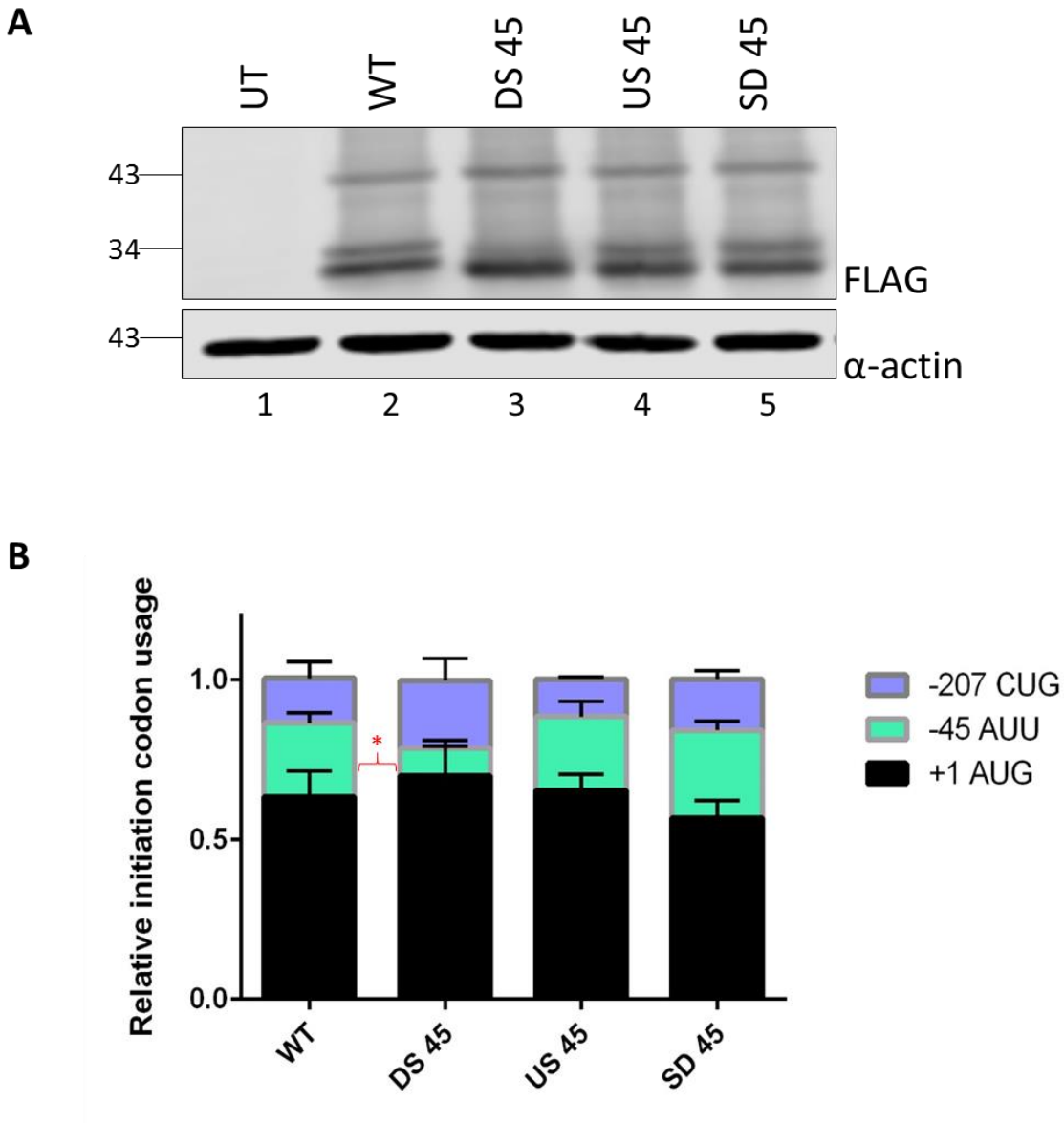


Figure 4-3: DSHP Sequence Regulates -45 AUU Translation

(A) FLAG-probed Western blot showing expression of GATAD1 protein isoforms when mutations were made to the mRNA sequence within the 5'UTR. 10 µg of transfected HeLa whole cell lysate was loaded on to a 10% SDS-PAGE gel, where UT = untransfected and WT = wild-type. Representative blot from three independent experiments. (B) Quantification made using GraphPad Prism (p=0.011, n=3, Tukey's multiple comparisons test). Error bars indicate the standard deviation (n=3).

4.3.2 DHX Helicases

As well as potentially forming hairpin structures as seen in Figure 4-2, the GATAD1 5'UTR sequence was analysed by QGRS Mapper (Quadruplex forming G-Rich Sequences Mapper), (http://bioinformatics.ramapo.edu/QGRS/index.php) (Kikin et al., 2006). Several QGRS were predicted to form within the 5'UTR of GATAD1, one of which is 13 nucleotides upstream of the -45 AUU AIC (Figure 4-4). QGRS within the 5'UTR can regulate the efficiency of translation initiation in two ways; generally, GQRS are translational repressors as they inhibit ribosomal progression, causing the 40S to drop off of the mRNA transcript; on the other hand, GQRS can positively regulate translation by stabilising IRES structures required for cap-independent translation as in the case of both VEGF and FGF-2 (Bugaut and Balasubramanian, 2012). In order to determine whether the structured 5'UTR of GATAD1 enables the AICs to be recognised and translated by the ribosome more efficiently, the GATAD1 reporter was cotransfected with vectors overexpressing several DEAH (Asp-Glu-Ala-His)-box RNA helicases. The DHX helicases differentially regulate translation initiation by unwinding RNA structures. DHX29 preferentially unwinds strong RNA hairpins and DHX36 specifically unwinds RNA G-quadruplex structures, whilst the specific function of DHX30 is unknown (Linder and Jankowsky, 2011). Therefore co-transfection of GATAD1 with DHX29, DHX30 or DHX36 may modify any translation initiation which is particularly dependent on RNA secondary structures. The vectors expressing FLAG-tagged helicases were created by Jim Schofield as part of his PhD work (Schofield, 2016). As shown by immunoblotting, the three DHX helicases did co-express with the GATAD1 reporter (Figure 4-5A), although it would have been advantageous to have included a reporter known to be helicase sensitive as a positive control, to confirm that each helicase was functional. None of the DHX helicases used have an effect on expression of GATAD1 isoforms, suggesting that there may be no secondary structure influencing translation initiation and the GQRS may not form as predicted *in vivo* (Figure 4-5B). Co-transfection of GATAD1 with DHX30 results in an additional band on the Western blot (Lane 4), above the -45 AUU isoform. The extra band is likely to be a breakdown product of the FLAG-tagged DHX30 helicase, since there are no near-cognate AICs with the potential to translate a GATAD1 isoform of this size.

Gene Information	
Gene ID:	Number of Products: 1
Gene Symbol:	Number of poly A Signals:
Gene Size: 285 nt.	QGRS found: 3
	QGRS found (including overlaps): 192

```
000001 AAGTCTCCGC GTCCCCCCCA CCCCGGCCAG ATCCCTTTCC CAGTCCTGCT TCCCAGTGCC TCGGGCCAGG GAATCCTGGC CTCCGCCTGC GGAGCCGGCG
000101 GAACCCGCTT CCCGCCTCCA CGGGGCAGCG CCAGCGGCCT GGTCCTTTCA CCGGCAGCTC CGTGCCGACG CTCTCACCGC TCTTCCTATC GCCGGGAGTG
000201 GCGGGCCGAC CAGGGGGCGG CCGGGCTACC GTCCGCCATT CCCGTGTCTC TGCGCCCGCG GGGGCCGCCC GAGCCGGCCA CCATG
```

Figure 4-4: QGRS Predictions within GATAD1 5'UTR

The GATAD1 5'UTR has the potential to form several G-quadruplex structures (highlighted), which may influence translation initiation from non-optimal AICs, indicated by red arrows.

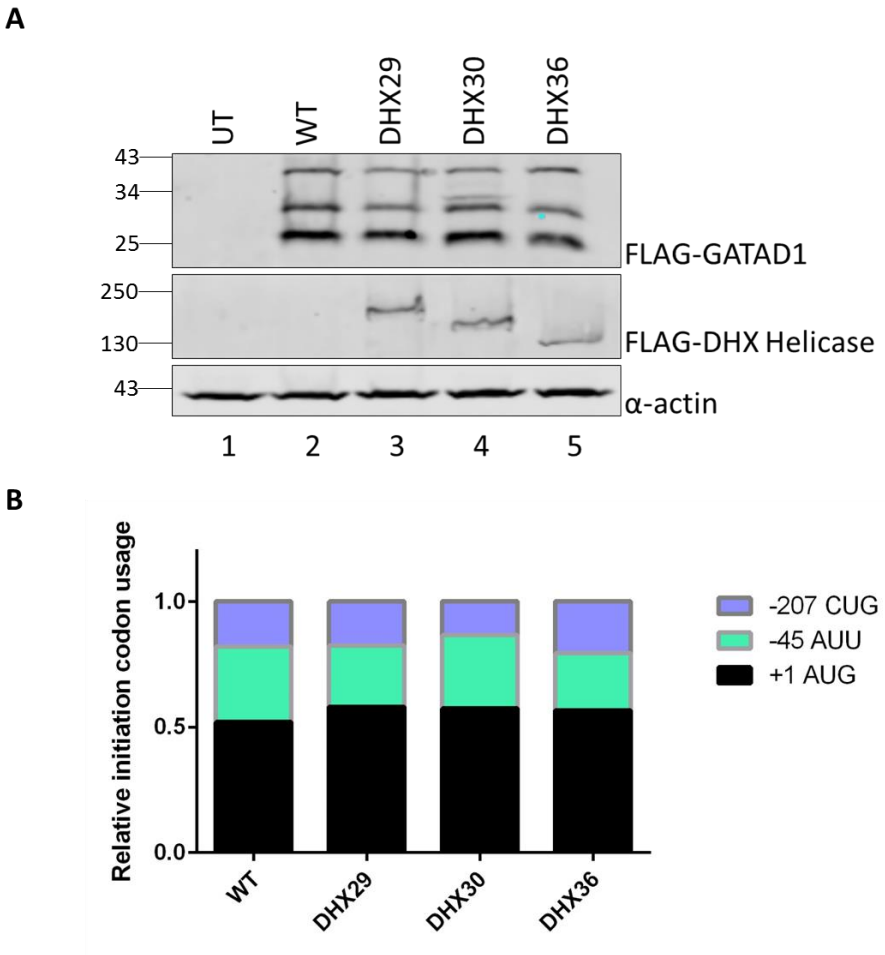


Figure 4-5: DHX Helicases have no Effect on GATAD1 Isoform Expression

(A) FLAG-probed Western blot showing expression of GATAD1 protein isoforms when co-transfected with DHX29, 30 or 36 alongside untransfected and wild-type lysate. 10 µg of transfected HeLa whole cell lysate was loaded on to a 10% SDS-PAGE gel. (B) Quantification of the FLAG expression relative to α-actin (n=1).

4.3.3 Proline – Alanine mutants

Several polyproline sequences are present within the extended GATAD1 N-terminus, translated from the AICs, including a triple-proline 12 residues downstream of the -207 CUG AIC and a double-proline 9 residues downstream of the -45 AUU AIC. Increased recognition of a suboptimal AIC may be possible via ribosomal stalling at polyproline sequences (Pavlov et al., 2008). Mutagenesis of the sequences from polyproline to polyalanine (Figure 4-6) was carried out prior to GATAD1 expression pattern analysis by Western blotting (Figure 4-7A). Although not statistically significant, the -207 polyproline to polyalanine mutation appeared to result in decreased translation of the extended isoform when compared to WT GATAD1. Similarly, the -45 polyproline to polyalanine mutation appeared to result in decreased translation of the mid isoform when compared to WT GATAD1 (Figure 4-7B). To check whether the non-significant results were due to a lack of statistical power as a result of small sample size, power analyses were carried out using GraphPad StatMate, where α=0.05, two tailed. The statistical power of analysing -207 CUG with respect to WT and -207 PPP-AAA, as well as -45 AUU with respect to WT and -45 PP-AA were both above the universally accepted power value of 0.8. This means that the non-significant results are not attributed to a limited sample size.

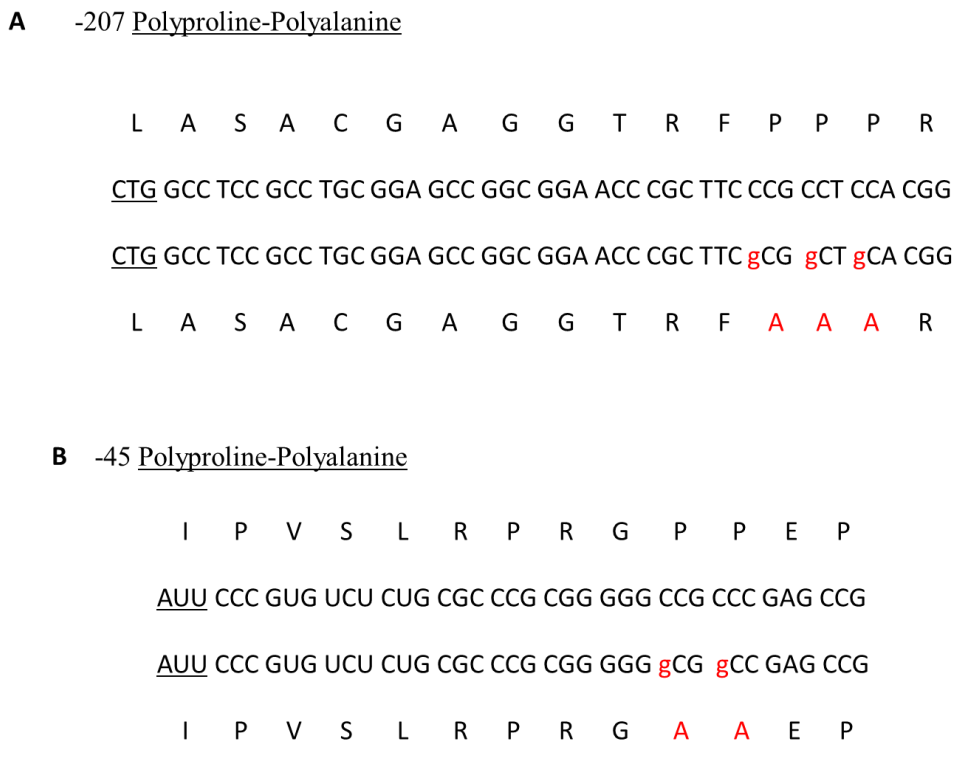


Figure 4-6: Proline-Alanine Mutants

(A) Proline-Alanine mutations were made at the triple-proline sequence downstream of -207 CUG. (B) Proline-Alanine mutants were also made at the double-proline sequence downstream of -45 AUU.

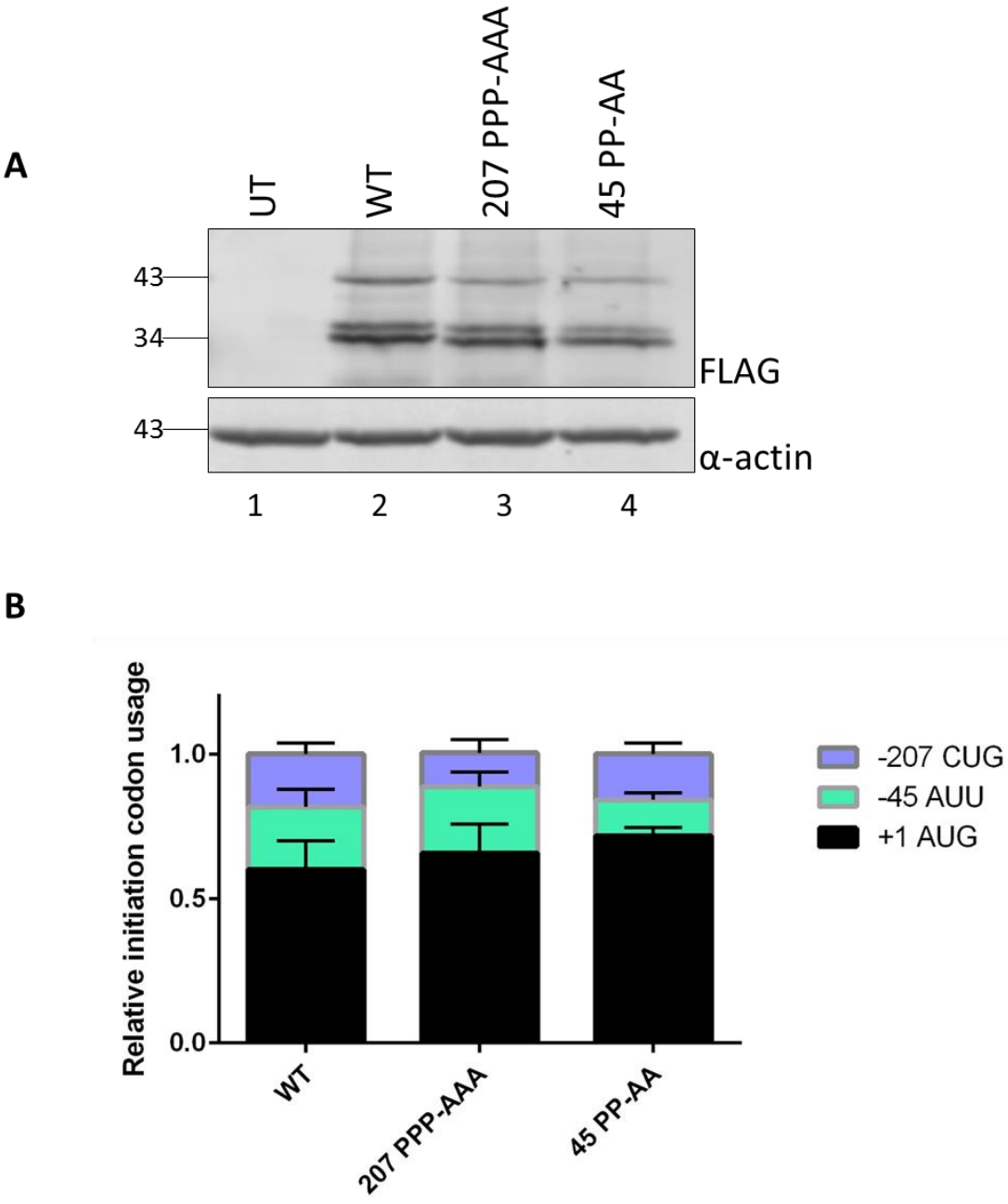


Figure 4-7: GATAD1 Isoform Expression Following Mutation of Polyproline Sequences

(A) FLAG-probed Western blot showing expression of GATAD1 protein isoforms when mutations were made to polyalanine sequences downstream of AICs within the 5'UTR, alongside untransfected and wild-type lysates. 10 µg of transfected HeLa whole cell lysate was loaded on to a 10% SDS-PAGE gel. Representative blot from three independent experiments. (B) Quantification was made using GraphPad Prism. Error bars indicate the standard deviation (n=3) Results were not significant, ($p = 0.38$ and $p = 0.1$ for WT versus -207 PPP-AAA and -45 PP-AA respectively).

4.3.4 Single Nucleotide Polymorphisms (SNPs)

A search on dbSNP (NCBI) indicated 15 SNPs within the 5'UTR of the GATAD1 transcript. Two of these were of interest with regard to alternatively translated isoforms as they were around the -207 CUG. The first (rs112545210) was a cytosine to guanine mutation at position -208 (ATC to ATG), for which there was no information regarding the minor allele frequency (MAF). The second SNP (rs192745223) was a thymine to cytosine mutation at position -206 (CTG to CCG), C is observed 11 times in a sample population of 2500 people (Table 4-1).

Mutagenesis was carried out to mimic the SNPs in the reporter plasmid and Western blot analysis was used to determine the effect on GATAD1 isoform expression. The -206 T to C SNP (rs192745223) prevented expression of the extended GATAD1 isoform, which required the -207 CUG to initiate translation. The-208 C to G SNP (rs112545210) caused 48% of translation to initiate from a novel -210 AUG, no translation from the -207 CUG and decreased levels of translation from both the -45 AUU and annotated AUG, when compared to wild-type GATAD1 (Figure 4-8).

Table 4-1: GATAD1 SNP Data

rs112545210	rs192745223
-208 C→G	-206 T→C
GGAATCCTG → GGAATGCTG	ATCCTGGCC → ATCCCGGCC
No MAF data	MAF: C=0.0022/11
	
	

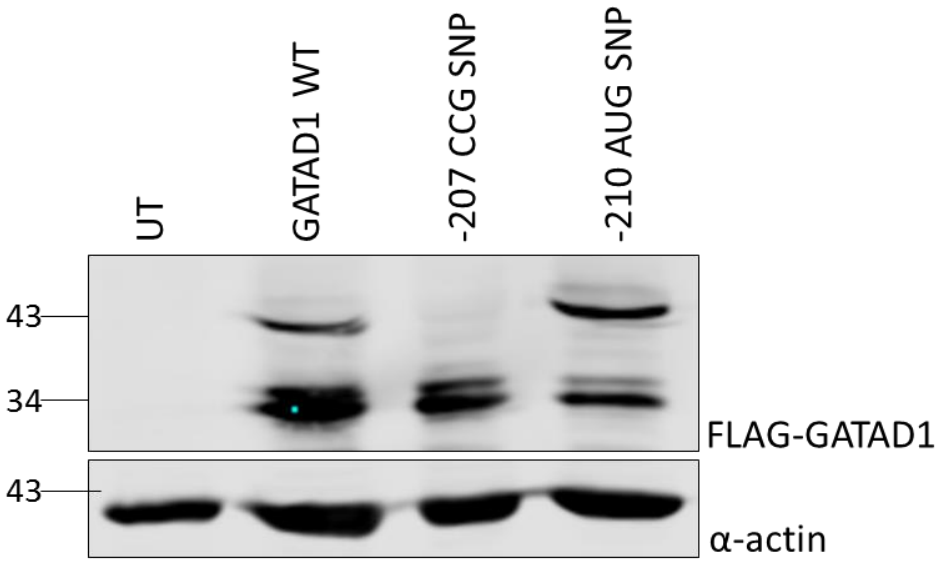


Figure 4-8: SNPs Change GATAD1 Isoform Expression

FLAG-probed Western blot showing expression of GATAD1 protein isoforms when physiologically relevant SNP mutations were made within the 5'UTR, alongside untransfected and wild-type lysate. 10 µg of transfected HeLa whole cell lysate was loaded on to a 10% SDS-PAGE gel.

4.4 Summary of Main Findings

- There is likely to be a secondary structure forming downstream of the -45 AUU, which is encouraging translation to take place from a sub-optimal AIC.
- Polyproline sequences downstream of both -207 CUG and -45 AUU AICs appear to be encouraging translation initiation.
- Two SNPs surrounding the -207 CUG AIC affect GATAD1 isoform expression.

4.5 Discussion

4.5.1 Sequence Downstream of -45 AUU Is Encouraging Translation

The primary sequence of an mRNA is able to regulate translation initiation in numerous ways. As well as determining the initiation codon itself and context surrounding it, the mRNA sequence may also form structures able to influence translation initiation from sub-optimal initiation codons, which would otherwise be scanned through. Hairpin structures 13-17 nucleotides downstream of an AIC may increase recognition and prevent leaky scanning from taking place, (Kozak, 1990). No hairpins were predicted to form downstream of -207 CUG or -45 AUU within the GATAD1 mRNA using the CENTROIDFOLD software. However, the sequence 14 nucleotides downstream of -45 AUU is GC-rich and silent mutations made within this stretch of sequence significantly decreased translation from this AIC. The sequence downstream of the AUU AIC is therefore regulating expression of the mid-GATAD1 isoform, strengthening the AUU which is an unusual initiation codon within a weak Kozak context. Since CENTROIDFOLD was limited to analysing 400 nucleotides for secondary structure, only the 5'UTR GATAD1 sequence was analysed. Submitting the whole 5'UTR and CDS sequence into an alternative RNA fold programme may have revealed a secondary structure within the GC-rich sequence downstream of -45 AUU, since long-range RNA-RNA interactions may have been taking place within the transcript which would not have been considered by only entering the 5'UTR sequence. On the other hand, rather than secondary structure impeding ribosomal scanning, the sequence downstream of -45 AUU may instead harbour a site for an RNA binding protein (RBP) which could also retard ribosomal scanning and influence translation from the AUU. This could be investigated by carrying out a RNA pull-down assay, which involves the isolation of a protein-RNA complex using a biotinylated RNA probe, followed by pull-down with strepdavidin beads and elution of protein-RNA complexes which can subsequently be analysed by Western blot.

4.5.2 Upstream Hairpin Does Not Regulate Translation From -45 AUU

Several mutations were also made to the sequence surrounding the -45 AUU, in order to further investigate regulation of translation by the GATAD1 mRNA sequence. The first set of silent mutations unwound the hairpin which was predicted to form directly upstream of the AIC. It was predicted that the ribosome may have scanned through the hairpin sequence more slowly, as RNA helicases melted the secondary structure. The ribosome would therefore have more time to recognise the AIC, increasing translation from the AUU. However, attempts to unwind a potential structure by overexpressing different helicases had no effect on translation of the mid-GATAD1 isoform, suggesting this sequence may not influence translation initiation from the -45 AUU AIC. Similarly, the hairpin structure formed by the prediction software may not form in physiological conditions. Selective 2'-hydroxyl acylation analysed by primer extension (SHAPE) quantification uses local backbone flexibility to determine RNA structures to single-nucleotide resolution in live cells. The RNA 2'-hydroxyl group is highly reactive when single stranded, but less reactive when the RNA is engaged in base pairing through secondary structure formation. Acylated sites prevent the primer extension from taking place and the resultant quantitative SHAPE data can be used to determine physiological mRNA secondary structure (Wilkinson et al., 2006).

A further method of determining secondary structures within mRNAs which also relies on chemical modification of secondary structures, is dimethyl sulphate (DMS) mutational profiling with sequencing (DMS-MaPseq). DMS modifies unpaired adenines and cytosines which results in mutations rather than cDNA truncations as in SHAPE analysis. Mutational profiling has advantages over truncation methods as more than one modification per molecule may be analysed on both high and low-abundance RNAs (Zubradt et al., 2016). DMS-MAPseq identified a secondary structure involving a GUG AIC within the FXR2 mRNA, which was also predicted by the CentroidFold software used here for initial GATAD1 secondary structure predictions. The GUG was positioned at the top of a hairpin, rather than the more commonly positioned secondary structure downstream of a sub-optimal AIC. The GATAD1 -207 CUG is positioned near the top of a hairpin (Figure 4-2), which was previously discounted as a regulatory structure, however further research into this structure may reveal a regulatory feature of the structure, which could encourage translation from the CUG AIC.

4.5.3 SD-Like Sequence Does Not Regulate Translation From -45 AUU

Mutations were made to a GC-rich sequence upstream of the -45 AUU within the predicted hairpin, which resembled a Shine Dalgarno (SD) sequence. Often found in prokaryotic mRNAs, a SD sequence approximately 7 nucleotides upstream of an initiation codon enables direct eukaryotic 16S rRNA base pairing, promoting the ribosome to initiate translation following scanning of the 5'UTR.

The 9 nucleotide Shine-Dalgarno-like sequence (CCGGCGGGU) has been shown to act within eukaryotic mRNAs and is 100% complementary to a segment of 18S rRNA within the 40S ribosomal subunit. Variations on this sequence can also recruit the ribosome to an initiation site, with the level of complementarity reflected in the binding affinities of the mRNA to 18S rRNA. Omission of the final GU has no effect on the efficiency of 18S binding and therefore the consensus sequence for efficient ribosomal recruitment is CCGGCGG (Chappell et al., 2004). Adenine mutations were made to the GC-rich sequence upstream of AUU as no adenine bases are present within the consensus sequence. However, mutating these nucleotides had no effect on the expression of the GATAD1 isoform expressed from -45 AUU. This may be because the GATAD1 mRNA sequence was not complementary enough to bind the 18S rRNA with a high affinity (CCGGGCU), with only four nucleotides of the minimum required five binding to 18S rRNA.

4.5.4 Polyproline Sequences Downstream of AICs Encourage Alternative Translation

The polyproline sequence present downstream of the GATAD1 alternative translation initiation sites appear to show a trend towards encouraging translation from each AIC, although these results were not significant (p= 0.38 and p=0.1 for -207 PPP-AAA and -45 PP-AA respectively). Ribosomal stalling occurs at these sequences due to the slow incorporation of proline residues during translation. Proline peptide-bond formation is slower due to the N-alkyl group on the nitrogen nucleophile used to form the peptide-bond (Pavlov et al., 2008). Proline is therefore both a poor donor and acceptor of peptide-bond formation and polyproline sequences can result in ribosomal stalling when the peptidyl-Pro-Pro-tRNA is located in the ribosomal P site. The strength of ribosomal stalling is influenced by the residues flanking the polyproline sequences. Triprolyl motifs preceded with Arg and His residues have a strong pausing effect, whereas Thr and Cys preceding a triproline reduce ribosomal stalling. On the other hand, diprolyl motifs may also cause ribosomal pausing when either Asp or Ala precede the sequence, or Trp, Asp, Asn or Gly follow the diproline (Starosta et al., 2014). 12 residues (36 nt) downstream of the -207 CUG is a triproline motif (FPPPR) and 9 residues (27 nt) downstream of the -45 AUU is a diproline (GPPE), neither of which fit the consensus thought to increase the strength of the stalling. Although the approximate distance between the leading edge of the ribosome and the P-site where the Met-tRNAi anticodon is positioned is 15 nt, polyproline sequences cause ribosomal pausing which can lead to queueing of ribosomes along the transcript (Woolstenhulme et al., 2015), which may in turn lead to increased recognition of the sub-optimal -207 CUG and -45 AUU AICs. As well as ribosomal pausing due to slow peptide-bond formation, N-alkyl groups within polyproline sequences can form kinks within the polypeptide chain. This can cause a block in the ribosome channel, slowing translation and causing queuing of ribosomes along the transcript, potentially also increasing translation from sub-optimal initiation codons.

4.5.5 SNPs around -207 CUG Alter Expression of GATAD1

The SNPs flanking the -207 CUG within GATAD1 are of particular interest because they significantly alter the expression pattern of GATAD1. The first (rs112545210) SNP at position -208 (ATC to ATG), resulted in the translation of a novel GATAD1 isoform and prevented all translation from the extended isoform usually translated from -207 CUG. There is no information regarding the minor allele frequency (MAF) of this SNP. The second SNP (rs192745223) at position -206 (CTG to CCG) also prevents all expression from -207 CUG which is normally responsible for the translation of the extended GATAD1 isoform. This SNP was observed 11 times (0.22%) in the 1000 Genome project phase 3 sample population of 2500 (5000 alleles). This T to C SNP is population-specific and all occurrences submitted were within the African population, with a MAF of 0.83% in a sample population of 661 (1322 alleles). If the extended GATAD1 isoform has a function unique to the annotated isoform which forms part of a transcription factor complex, then the SNPs may have implications in the transcriptome landscape and potentially in health and disease.

Chapter Five

Factors Regulating GATAD1 Translation

5. Factors Regulating GATAD1 Translation

5.1 Introduction

Regulation of translation provides rapid control over cellular protein levels. As noted in the previous chapter, the primary mRNA sequence can regulate translation initiation causing preferential selection of sub-optimal alternative translation initiation codons. In addition to this, environmental factors causing cell stress and cellular machinery including initiation factors are also able to regulate translation initiation (Pestova and Kolupaeva, 2002). These regulatory factors may cause preferential translation of alternatively translated GATAD1 isoforms.

Stress signalling results in activation of pro-survival pathways and rapid reprogramming of gene expression within mammalian cells. This often results in inhibition of cap-dependent translation resulting from changes in the phosphorylation status of specific components of the translational machinery, including p-eIF2α (Patel et al., 2002). Although global translation is therefore down-regulated, certain mRNAs must still be translated during the integrated stress response. In stress conditions with low levels of available eIF2α, eIF2A-dependent translation of uORFs from these mRNAs is increased (Starck et al., 2016). Similarly, eIF2A-dependent non-canonical translation takes place during tumour formation, under conditions of hypoxia (Sendoel et al., 2017).

During cap-dependent translation, initiation factors including eIF1, eIF1A and eIF5 are known to influence the fidelity of start codon selection (Pestova and Kolupaeva, 2002). Changing the availability of these initiation factors within the cell influences translation from AICs, acting as a method of gene regulation which can greatly impact the proteome. In particular, mutations within eIF1 or depletion of eIF1 results in a lack of discrimination between cognate and non-cognate initiation codons, as mismatches between initiation codons and the anticodon of initiator tRNA are disregarded (Pestova and Kolupaeva, 2002). eIF1 plays the largest role in initiation codon selection and is also able to dissociate incorrectly assembled ribosomal complexes, increasing start codon stringency. Efficient mRNA scanning requires the cooperative binding of both eIF1 and eIF1A near the ribosomal P-site and A-site respectively, of the 40S ribosomal subunit. Together, eIF1 and eIF1A stabilise the 43S PIC in an open, scanning-competent confirmation until an AUG initiation codon in a good context is recognised, when a conformational change causes release of eIF1 and other subsequent factors also dissociate (Passmore et al., 2007). The exact mechanism of action of eIF1 in start codon selection is not known, however a model suggests that the eIF1-containing scanning-competent 43S ribosomal subunit ensures that the initiator tRNA anticodon loop is only able to form

stable interactions with cognate AUG codons within a strong context. On the other hand, in the absence of eIF1, the 40S subunit is in a closed conformation which enables partial base-pairing to take place with non-canonical initiation codons (Hussain et al., 2014).

eIF1A works synergistically with eIF1 to enhance recruitment of eIF2-GTP-Met-tRNAi^{Met} ternary complex to the 40S ribosomal subunit and to promote an open, scanning-competent PIC. The effect of eIF1A on the stringency of start codon selection is modulated in opposing ways by its N and C-terminal tail (NTT/CTT). Mutations within the NTT results in leaky scanning past the GCN4 uORF1 AUG and an increase in the accuracy of start codon selection, whilst mutations in the CTT increase translation initiation from non-AUG codons (Fekete et al., 2007). The conserved 25 amino acid NTT stabilises the closed 43S PIC conformation, decreasing start codon stringency and allowing more translation to take place from non-canonical initiation codons, whilst the 20-35 amino acid CTT of eIF1A stabilises the open, scanning-competent conformation, increasing the stringency of start codon selection. This is possible because unlike the CTT, the NTT does not interact with the mRNA and Met-tRNAi^{Met} (Yu et al., 2009). Together, the dual-modulation of eIF1A is able to ensure accurate selection of an AUG initiation codon.

Whilst both eIF1 and eIF1A increase the accuracy of start codon selection, eIF5 itself has the opposite effect. eIF5 is a GTPase-activating protein (GAP), which promotes hydrolysis of eIF2-bound GTP and subsequent P_i release from the 43S PIC (Das and Maitra, 2001). This results in the release of bound initiation factors and eIF5B-mediated joining of the 60S ribosomal subunit. Although eIF5 plays an important role in start codon selection, its GAP function must be highly regulated in order to prevent aberrant eIF2-GTP-Met-tRNAi^{Met} hydrolysis and initiation of translation at a non-AUG codon. A block in GTP hydrolysis is mediated by initiation factors bound to the C-terminus of eIF5 and the 43S PIC, namely eIF1, 1A, 3 and 4F (Majumdar and Maitra, 2005). These factors may sterically prevent interaction of eIF5 with the GTPase eIF2 until an AUG codon is recognised.

5.2 Hypothesis and Aims

5.2.1 Hypothesis

Stress signalling as well as levels of initiation factors eIF1, eIF1A and eIF5 influence translation from non-AUG initiation codons within the GATAD1 transcript.

5.2.2 Aims

The aim of this chapter was to determine the factors influencing ribosomal AIC selection within the GATAD1 transcript. This was carried out by investigating:

- Whether inducing cell stress pathways has an effect on start codon preference.
- Whether changing levels of initiation factors 1 and 1A causes a change in ribosomal start codon preference.
- Whether cell type has an effect on alternative translation initiation, and if so, whether this is a result of the presence/absence of particular *trans*-acting factors.

5.3 Results

5.3.1 Cell Stress Influencing AIC Selection

Stress signalling results in increased non-canonical translation, through a variety of mechanisms. HeLa cells over-expressing the GATAD1-3xFLAG reporter were treated with multiple stress-inducing agents, including anisomycin, cobalt (II) chloride ($CoCl_2$), rapamycin, thapsigargin and tunicamycin.

Anisomycin is a peptidyl-transferase A-site inhibitor, which inhibits elongation of protein synthesis, also resulting in activation of the p38(MAPK) pathway (Grollman, 1967). On the other hand, $CoCl_2$ induces an oxidative stress response by mimicking hypoxia. Under normoxia, the α-subunits of HIF (hypoxia-inducible factor) transcription factors are bound by the von Hippel-Lindau protein (pVHL) which mediates their ubiquitination and subsequent degradation. The HIF-α/pVHL interaction is dependent on the hydroxylation of a proline within the oxygen-dependent degradation (ODD) domain of the HIF-α proteins. Cobalt inhibits the HIF-α/pVHL interaction under normoxic conditions, stabilising HIF-α and inducing a stress response within the cell (Yuan et al., 2003).

Rapamycin is an allosteric inhibitor of mTORC1, which results in inhibition of cap-dependent translation through several pathways (Sehgal, 1995). Inhibition of mTORC1 prevents phosphorylation of its downstream substrates S6Ks and 4E-BPs. The S6K signalling pathway regulates ribosome biogenesis, whilst phosphorylation of 4E-BPs allows eIF4E to bind to eIF4G, forming part of the scanning-competent eIF4F complex (section 1.4.2.1).

Both thapsigargin and tunicamycin induce phosphorylation of eIF2α, repressing translation. Thapsigargin inhibits the sarco/endoplasmic reticulum Ca^{2+} ATPase (SERCA), resulting in ER stress and a resultant unfolded protein response (UPR) (Lytton et al., 1991). The UPR results in both transcriptional and translational changes in gene expression through ER stress sensors, including protein kinase RNA-like endoplasmic reticulum kinase (PERK). Activation of PERK results in phosphorylation of eIF2α and subsequent translation inhibition, however translation of activating transcription factor 4 (ATF4) is up-regulated, activating transcription of genes involved in the integrated stress response and is used here as a marker of cell stress. Tunicamycin also induces UPR through eIF2α-P, by inhibiting glycoprotein synthesis (Heifetz et al., 1979). Tunicamycin inhibits the UDP-HexNAc:polyprenol-P HexNAc-1-P family of enzymes, blocking N-linked glycosylation which also results in a cell stress response similar to thapsigargin.

Treatment of GATAD1-expressing HeLa cells with each of the previously mentioned cell stress-inducing agents was carried out for 24 hours, before the resultant cell lysates were run on a Western blot. The oxidative stress response induced by $CoCl_2$ appeared to result in increased

translation of the alternatively translated GATAD1 isoforms (Figure 5-1) and increased HIF-1α levels upon $CoCl_2$ treatment confirmed stabilisation of the protein due to inhibition of the HIF-1α/pVHL interaction, which is observed during the oxidative stress response. However, the apparent decrease in annotated GATAD1 translated compared to the extended GATAD1 isoforms was not significant ($p= 0.28$ for +1 AUG with respect to WT and $CoCl_2$-treated). To check whether the non-significant result was due to a lack of statistical power as a result of small sample size, power analyses were carried out using GraphPad StatMate, where $\alpha=0.05$, two tailed. The statistical power of analysing +1 AUG with respect to WT and $CoCl_2$-treated cells was between 0.5-0.6, which is below the universally accepted power value of 0.8. This suggests that statistical power is limited by the sample size. Increasing in the n number from n=3 to n=7 would obtain statistical power at the recommended 0.8 level and may result in a significant decrease in annotated GATAD1 translation in $CoCl_2$-treated cells.

A decrease in p-4E-BP is observed in rapamycin-treated cells, concurrent with mTORC1 inhibition. Although a p-eIF2α/ATF4 blot was not possible with the antibodies available, 24 hour treatment with Thapsigargin (0.1 µM) and Tunicamycin (2.5 µg/mL) should induce the UPR stress response.

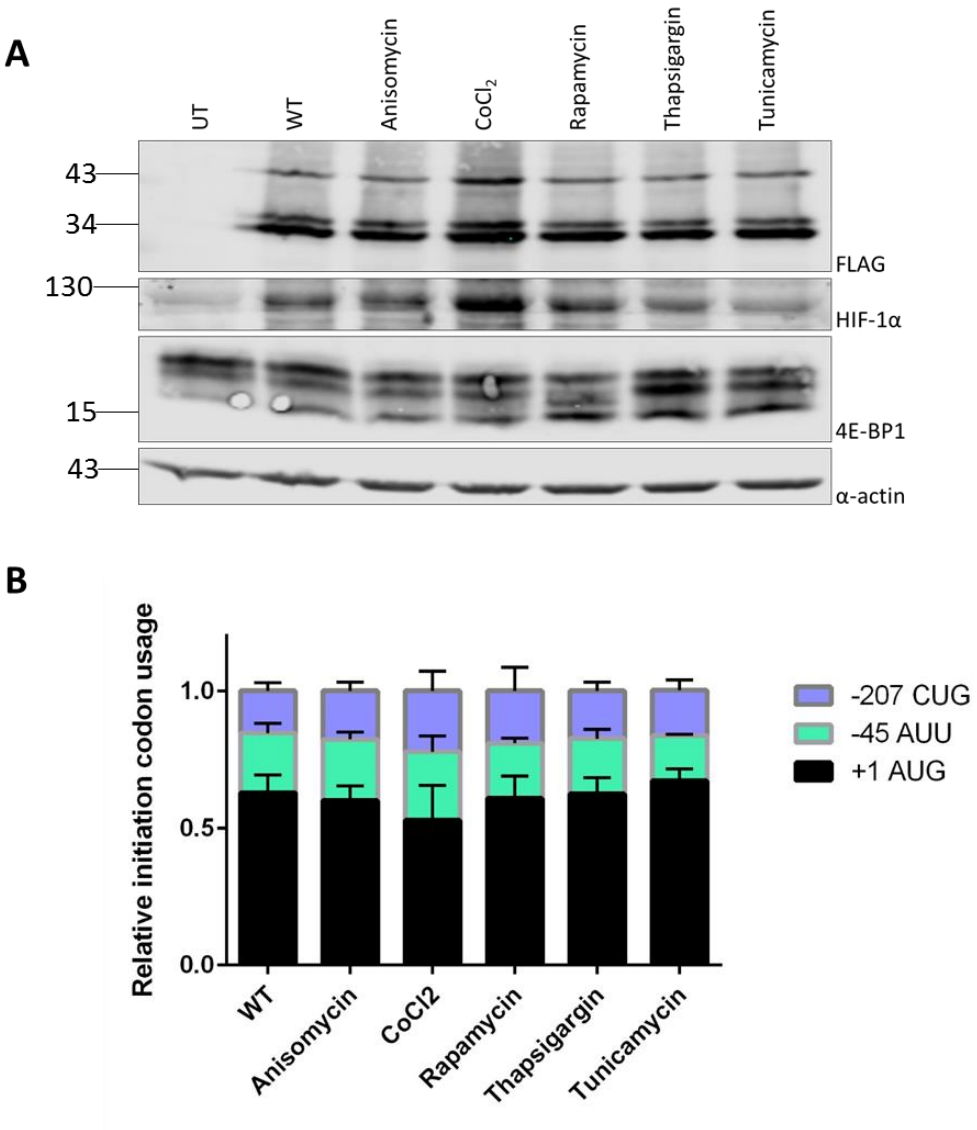


Figure 5-1: $COCl_2$ Increases Translation of GATAD1 Isoforms

(A) FLAG-probed Western blot showing expression of GATAD1 protein isoforms when treated with cell stress-inducing agents for 24 hours; anisomycin (50 nM), $CoCl_2$ (0.1 mM), rapamycin (100 nM), thapsigargin (0.1 µM), tunicamycin (2.5 µg/mL) as well as untransfected and wild-type lysate. 10 µg of transfected HeLa whole cell lysate was loaded on to a 10% SDS-PAGE gel. Representative blot from three independent experiments. (B) Quantification of relative expression of the different GATAD1 isoforms, using the FLAG signals from three independent experiments. Error bars indicate the standard deviation (n=3), results were not significant.

5.3.2 Initiation Factors Influencing AIC Selection

Both eIF1 and eIF1A regulate cap-dependent translation by maintaining the stringency of start codon selection (Passmore et al., 2007). Changing the availability of these factors within the cell influences translation from AICs and therefore regulates the protein isoforms which are translated. Plasmids were used to either overexpress myc-tagged eIF1 or eIF1A (from Samantha Hodges and Kevin Jones), or knockdown eIF1 or eIF1A (small hairpin RNA (shRNA) from Lucinda Eaton) using specific short hairpins to knockdown expression via RNA interference. When the GATAD1-3xFLAG reporter was overexpressed with eIF1, the stringency of start codon selection was increased and less translation initiated from both the CUG and AUU AICs, with significantly more translation taking place from the annotated AUG ($p=0.0006$, $n=3$, Tukey's multiple comparisons test) (Figure 5-2A). The myc-tagged proteins can be observed as their migration is retarded compared to the endogenous protein (lanes 4 and 6). Only a partial shRNA-mediated knockdown of eIF1 or eIF1A was observed (lanes 5 and 7), and this was not sufficient to result in a decrease in the stringency of start codon selection, which would have resulted in more translation from the upstream non-AUG codons within the GATAD1 mRNA transcript. There was however, a significant difference in the expression of the mid-GATAD1 isoform expressed from the -45 AUU when eIF1 is knocked-down compared to when eIF1 is over-expressed ($p=0.0043$, $n=3$, Tukey's multiple comparisons test) (Figure 5-2B).

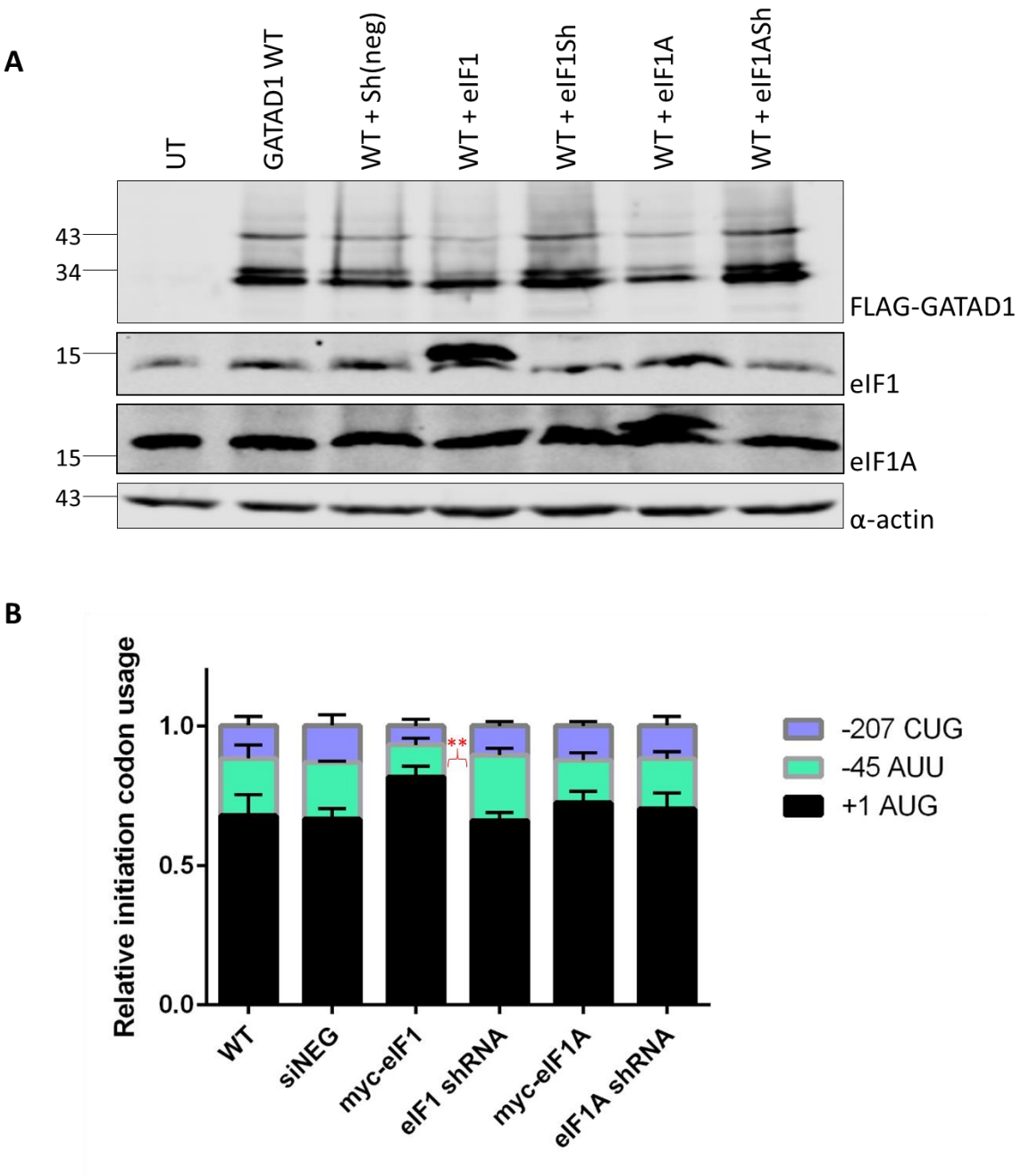


Figure 5-2: Overexpression of eIF1 Encourages Alternative Translation Initiation in GATAD1

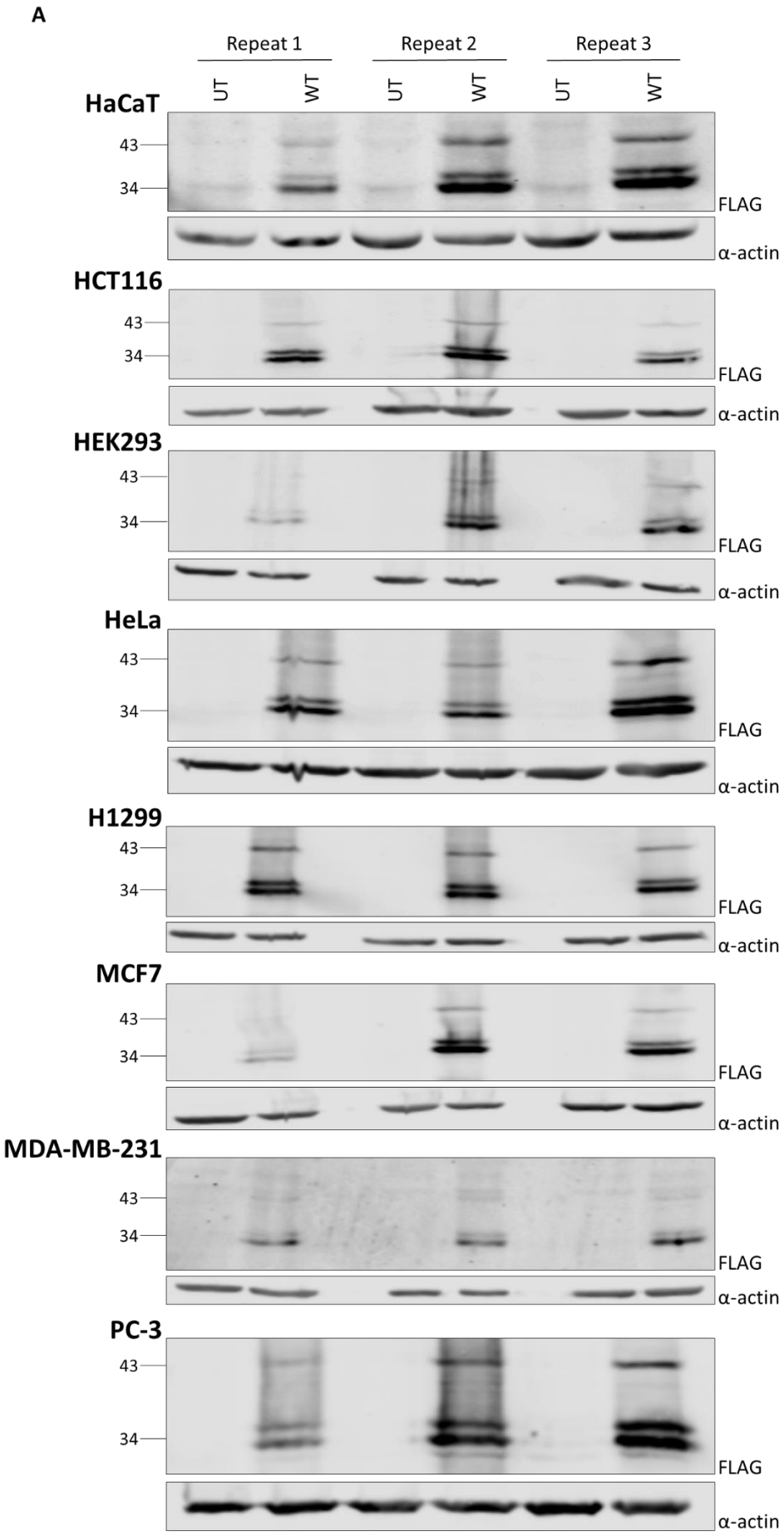
(A) FLAG-probed Western blot showing expression of GATAD1 protein isoforms when eIF1 and eIF1A are both over-expressed and knocked-down using shRNA for 48 hours, alongside untransfected and wild-type lysate. 10 µg of transfected HeLa whole cell lysate was loaded on to a 10% SDS-PAGE gel. Representative blot from three independent experiments. (B) Quantification made using GraphPad Prism ($p=0.0043$, $n=3$, Tukey's multiple comparisons test). Error bars indicate the standard deviation ($n=3$).

5.3.3 Alternative Translation Initiation in Various Cell Lines

GATAD1 expression was analysed by Western blot in a variety of different cell lines (cancer-derived and normal cell lines from a variety of tissue types) available to the laboratory, to observe any effect on alternative translation initiation. GATAD1 was over-expressed in a panel of cell lines; HaCaT keratinocyte cells, HCT116 colon cancer cells, HeLa cervical cancer cells, HEK293 embryonic kidney cells, H1299 non-small cell lung carcinoma cells, MCF-7 (oestrogen +, progesterone +, HER2-) and MDA-MB-231 breast cancer cells (triple negative) as well as PC-3 prostate cancer cells. Both H1299 and PC-3 cell lines translated more of the extended GATAD1 isoforms (Figure 5-3).

The levels of eIF1, eIF1A and eIF5 in each of the cell lines were quantified by Western blot, to determine if changes in their expression correlated with the differences between non-AUG translation and therefore stringency of start codon selection (Figure 5-4A). The eIF1AY antibody had a higher signal to noise ratio than the antibody to the X-linked homologue eIF1AX, which made the blots easier to quantify (data not shown). eIF1AX and eIF1AY only differ by a single amino acid and the antibody epitope is present in both forms.

The level of regulation over start codon selection was analysed using the ratio of eIF5:eIF1 (Figure 5-4B) or eIF5:eIF1A (Figure 5-4C) in the cell. Higher ratios indicate less control over stringency of start codon selection by initiation factors within the cell line. HaCaT cells showed higher rations for both eIF5:eIF1 and eIF5:eIF1A, suggesting less stringent start codon selection. All other cell lines, including H1299 and PC-3 showed similar levels of start codon stringency.


A
Repeat 1
Repeat 2
Repeat 3
UT
WT
UT
WT
UT
WT
HaCaT
43
34
FLAG
α-actin
HCT116
43
34
FLAG
α-actin
HEK293
43
34
FLAG
α-actin
HeLa
43
34
FLAG
α-actin
H1299
43
34
FLAG
α-actin
MCF7
43
34
FLAG
α-actin
MDA-MB-231
43
34
FLAG
α-actin
PC-3
43
34
FLAG
α-actin

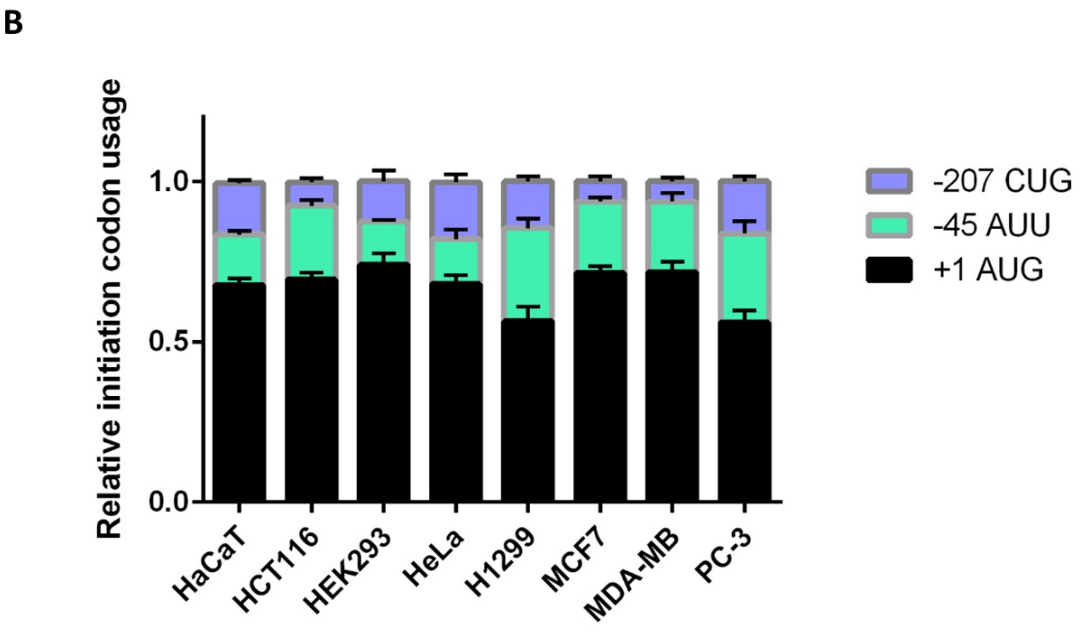


Figure 5-3: H1299 and PC-3 Cell Lines Express More Non-AUG GATAD1 Isoforms

(A) FLAG-probed Western blot showing expression of GATAD1 protein isoforms in different cell lines, where UT = untransfected and WT = wild-type GATAD1. 20 µg of transfected whole cell lysate was loaded on to a 10% SDS-PAGE gel. (B) Error bars indicate the standard deviation. Quantification made using GraphPad Prism, (n=3, Tukey's multiple comparisons test). For +1 AUG in both H1299 and PC-3 vs all other cell lines, p=<0.0001. For -45 AUU, H1299 vs HaCaT, HEK293, HeLa, MCF7 and MDA-MB, p= <0.05. For -45 AUU, PC-3 vs HaCaT, HEK293 and HeLa, p= <0.0001. For -207 CUG, H1299 vs HCT116, MCF7 and MDA-MB, p= <0.05. For -207 CUG, PC-3 vs HCT116, MCF7, MDA-MB, p= <0.005.

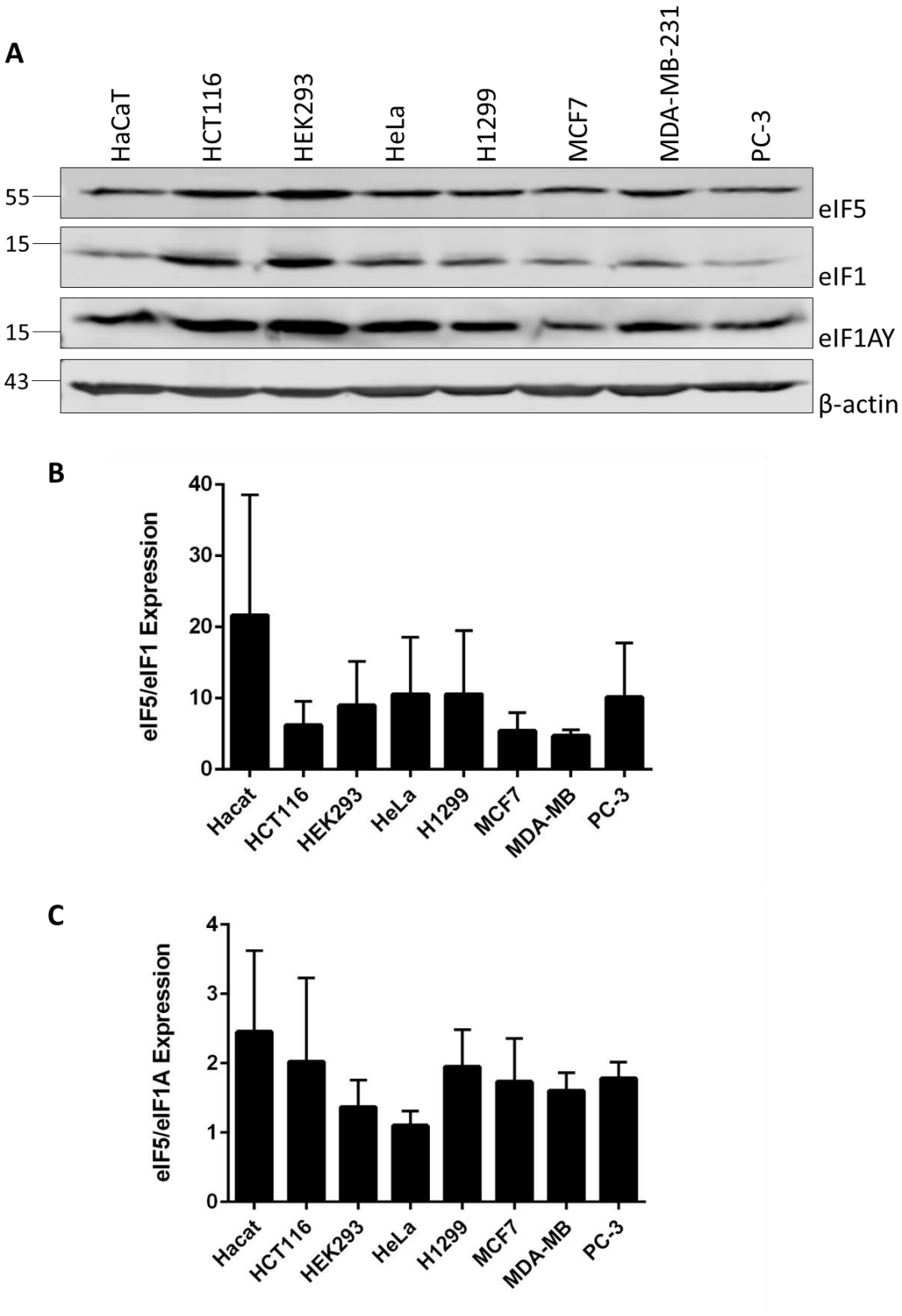


Figure 5-4: Levels of eIF1, eIF1A and eIF5 Between Cell Lines

(A) FLAG-probed Western blot showing levels of eIF1, eIF1A and eIF5 between cell lines. 20 µg of untransfected whole cell lysate was loaded on to a 10% SDS-PAGE gel. (B) The ratio of eIF5 to eIF1 or (C) eIF5 to eIF1A was calculated, indicating the level of regulation surrounding start codon stringency between cell lines. Quantification was made from three independent experiments. Error bars indicate the standard deviation (n=3), results not significant.

5.3.4 pIC (plasmid for Initiation Codon) Test Assay

A further method for assaying how much non-AUG translation takes place between cell lines is the pICtest. The pICtest plasmids are dual-luciferase vectors which enable the analysis of initiation codon efficiency, (Stewart et al., 2015). Renilla luciferase (Rluc) is used as an internal transfection control and is always translated from an AUG in an optimal Kozak consensus, GCCACCAUGG. Firefly luciferase is also expressed within the same plasmid backbone and it is the initiation codon of this ORF which is altered from an AUG, to a CUG, GUG, UUG, AAG, ACG, AGG, AUA, AUC, AUU and finally to a stop codon as a negative control (Figure 5-5). Otherwise the firefly luciferase reporter is very simple, with only a short unstructured 5' UTR which has a GC-content of 43.3%, allowing efficient ribosomal scanning.

The pICtest assay was carried out on HeLa cells as these are the standard cell line used in most of the experiments throughout this work (Figure 5-6A). H1299 and PC-3 cell lines were also investigated using pIC test as they allowed more translation of GATAD1 from AICs (Figure 5-6B, Figure 5-6C). When compared to HeLa cells, translation from non-AUGs is far less efficient in H1299 cells in the pIC test, which contradicts our findings with GATAD1 which may have been a result of GATAD1 transcript-specific regulation rather than the H1299 cell line carrying out more non-AUG translation on the whole. On the other hand, the H1299 cells used in the GATAD1 Western blots may have been more or less confluent during the experiments which could cause certain cell stress responses to affect the level of non-AUG translation taking place. Similarly, the discrepancy could be a passage-dependent effect, whereby cells at a high passage number can exhibit changes in gene expression (Briske-Anderson et al., 1997). However, the pIC test results do confirm that PC-3 cells translate from non-AUGs and in particular AUU codons with higher efficiency than HeLa cells.

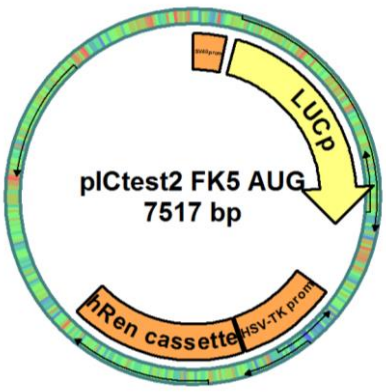


Figure 5-5: pICtest Plasmid

The pICtest plasmid shown is the positive control plasmid, which uses an AUG initiation codon within a full Kozak consensus to initiate translation of Firefly luciferase. Transcription of the Renilla reporter mRNA is under the control of the herpes simplex virus (HSV) thymidine kinase promoter, whereas the Firefly luciferase is expressed from a Simian virus 40 promoter.

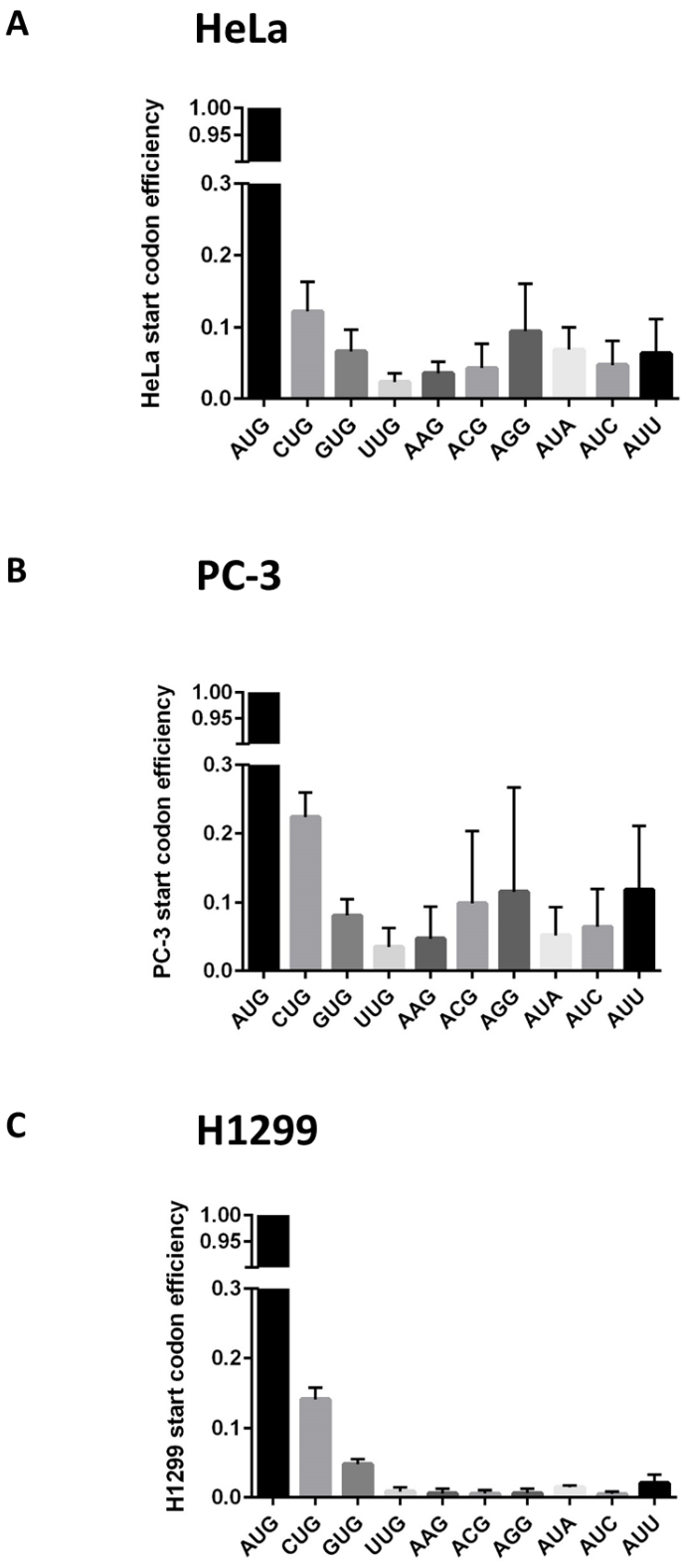


Figure 5-6: pICtest Analysis of Non-AUG Translation Efficiencies

Using the pICtest plasmids, initiation codon efficiency between cell lines (A) HeLa (B) PC-3 (C) H1299, were calculated from the Renilla/Firefly ratio, normalising to the AUG codon. Quantification was made from five independent experiments. Error bars indicate the standard deviation (n=5).

5.4 Summary of Main Findings

- Activation of oxidative stress pathways increases translation from alternative initiation codons within GATAD1.
- The GATAD1 alternative initiation codons are regulated by eIF1 and eIF1A levels.
- Levels of initiation factors regulating start codon stringency varies between cell types, which has an effect on non-AUG translation.

5.5 Discussion

5.5.1 Specific Cell Stress Pathways Regulate non-AUG Translation of GATAD1

Stress signalling causes activation of pro-survival pathways, which results in rapid reprogramming of gene expression. Alternative translation initiation provides a further layer of regulation to gene expression and is thought to be promoted in periods of cell stress. However, when HeLa cells were treated with the translation inhibitors anisomycin, rapamycin, thapsigargin and tunicamycin, there was no effect on alternative translation initiation of GATAD1.

On the other hand, mimicking hypoxia using $CoCl_2$ resulted in increased alternative translation initiation of GATAD1. It would have been advantageous to consider the GATAD1 mRNA levels by qPCR in order to ensure that the effect seen upon $CoCl_2$ treatment was a result of translational regulation and not increased transcription of GATAD1. HIF-1α is stabilised upon $CoCl_2$ treatment (Figure 5-1A, lane 4), confirming that the oxidative stress response pathways have been activated. Translation from the alternative initiation codons CUG and AUU are selectively regulated by certain stresses only. $CoCl_2$ mimics the effects of hypoxia under normoxic conditions; confirming these results using a hypoxia incubator chamber would provide more accurate data. Cells become hypoxic when oxygen cannot be efficiently delivered to cells, as in the case of tumour cells which rapidly outgrow their blood supply. Further research may be able to generalise that hypoxia promotes alternative translation initiation; tumour cells may therefore be under this regulation and express higher levels of N-terminally extended protein isoforms.

GATAD1 forms part of a histone demethylase complex with KDM5A (lysine demethylase 5A), which demethylates the H3K4 marker linked to transcriptional activation (section 1.8.1).

Hypoxia inactivates the activity of KDM5A, increasing H3K4me3 levels in both normal epithelial Beas-2B cells as well as lung carcinoma A549 cells, (Zhou et al., 2010). Since the GATAD1 complex is inactive during hypoxia, increased translation of extended non-AUG GATAD1 isoforms may be a regulatory mechanism preventing formation of functional histone demethylase complex.

5.5.2 eIF1 Regulates non-AUG Translation of GATAD1

eIF1 regulates cap-dependent translation by maintaining the stringency of start codon selection. Changing the availability of this factor within the cell influences translation from AICs and therefore regulates the protein isoforms which are translated. Overexpressing eIF1 resulted in increased start codon stringency and a decreased non-AUG translation, whilst knock-down of eIF1 relaxed start codon stringency, increasing alternative translation initiation. However, this effect was not significant when considering eIF1A (Figure 5-2B). This may be because eIF1 plays the main role in initiation codon selection, acting as the switch between the open and closed conformation of the 43S PIC. On the other hand, the eIF1A shRNA knock-down did not appear to be working efficiently, resulting in increased start codon stringency compared to the eIF1 knock-down.

Non-canonical translation initiation is common in oncogenes as a result of eIF2A translational reprograming during tumorigenesis. eIF2A delivers tRNAs to the P-site of the ribosome and is implicated in uORF translation initiation. eIF2A also competes with eIF2α in ternary complex formation. Stress-induced p-eIF2α (as seen in tumour cells) results in a global inhibition of protein synthesis, however ribosome profiling has shown that certain cancer-associated mRNAs are able to remain efficiently translated. The eIF2A-dependent translation of these oncogenic messages is up-regulated when eIF2α is inhibited, (Sendoel et al., 2017). Regulation of non-canonical translation by initiation factors therefore influences tumour development and malignancy and has an important therapeutic relevance. It would be interesting to investigate the influence of eIF2A on GATAD1 AIC selection in future experiments.

5.5.3 Regulatory Initiation Factors and non-AUG Translation Differs between Cell Type

It was observed that the ratio of GATAD1 isoform expression varied between cell lines, with more alternative translation initiation taking place in both H1299 and PC-3 cells. It would have been advantageous to confirm that this effect seen was not due to verying levels of GATAD1 transcription between cell lines, which could be confirmed by qPCR. In order to determine whether the apparent increase in non-AUG translation was due to levels of specific initiation factors within the cell lines, ratios of eIF5:eIF1/eIF1A were calculated. The higher this ratio, the less stringent start codon selection is likely to be, allowing more alternative translation to take place. Most of the cell lines had similar ratios of these initiation factors and H1299 and PC-3 cells did not show increased levels of eIF5 versus eIF1 or eIF1A.

The pICtest assays calculated initiation codon efficiency between cell lines, relative to the AUG codon. Unexpectedly, H1299 cells showed very low translation efficiencies from all AICs, although previously translating more GATAD1 from CUG and AUU than other cell lines. Interestingly, translation from both CUG and AUU codons as used in GATAD1, is almost twice as efficient in PC-3 cells as in HeLa cells. This may have interesting implications in prostate cancer cells expressing higher levels of N-terminally extended protein isoforms translated from upstream non-AUG codons. This may be as a result of the phosphorylation and inactivation of eIF2α during cancer progression, directing translation to take place using the alternative initiation factor eIF2A which has a preference for non-AUG translation initiation. As a result of initiation codon selection becoming less regulated, increased translation of uORFs takes place. The alternative protein isoforms expressed in the prostate cancer cells may have alternative localisations and functions to the annotated proteins, (Sendoel et al., 2017).

Chapter Six

Subcellular Localisation of GATAD1 Isoforms

6. Subcellular Localisation of GATAD1 Isoforms

6.1 Introduction

The N-terminal extension of the alternatively translated GATAD1 isoforms may contain a protein targeting signal not present within the annotated protein. This could therefore cause the alternative protein isoforms to localise differently within the cell, which may affect the availability of binding partners, potentially resulting in a novel function. On the other hand, localising an alternative protein isoform away from its normal binding partners and function could be a form of regulation, sequestering excess translated protein as a way of controlling certain pathways within the cell.

The N-terminus of a protein is important in mediating protein targeting. Membrane-bound and secretory proteins are synthesised on ribosomes associated with the endoplasmic reticulum (ER) through co-translational translocation. This class of proteins are directed to the ER by a signal recognition particle (SRP), which recognises an N-terminal signal sequence between five and thirty hydrophobic amino acids long (Blobel, 1980).

On the other hand, proteins destined for the mitochondria are translated in the cytoplasm on free ribosomes, before being post-translationally targeted initially to the mitochondrial TOM20 receptor by cytosolic chaperones recognising an N-terminal presequence (Goping et al., 1995). The N-terminal, cleavable mitochondrial targeting peptide (mTP) contains interspersed positively charged residues, allowing an amphiphilic α-helix to form which allows translocation across the mitochondrial membranes (Gavel and von Heijne, 1990).

Proteins are targeted to the nucleus by a nuclear localisation signal (NLS). Most NLSs consist of clusters of basic residues, either monopartite or bipartite (Boulikas, 1993). Importin-α recognises and binds to the positively charged residues within the NLS and transports the protein through the nuclear pore complex (NPC) into the nucleus (Gorlich and Kutay, 1999). The NLS can be located either at the N-terminus or within the main core of the protein and is not cleaved; this allows for nuclear shuttling to take place as well as ensuring proteins can re-enter the nucleus once mitosis has taken place.

The alternatively translated proteins have different N-terminal sequences, which have the potential to re-localise the isoform within the cell, with respect to the annotated isoform.

6.2 Hypothesis and Aims

6.2.1 Hypothesis

The N-terminal extension present on the two extended GATAD1 isoforms results in re-localisation of the protein within the cell.

6.2.2 Aims

The aim of this chapter was to determine whether the alternatively translated GATAD1 isoforms localised differently in the cell with respect to the annotated isoform, due to the presence of the N-terminal extension.

This was investigated by carrying out:

- Immunofluorescence experiments using full length GATAD1 isoforms.
- Nuclear export signal (NES) studies to determine the route of each GATAD1 isoform in/out of the nucleus.

6.3 Results

6.3.1 Full Length GATAD1 Cloning

In order to carry out cellular or functional studies on GATAD1 isoforms, constructs were made containing the full CDS in order to be physiologically relevant, since the previous reporters were truncated at residue 202 of 269 to allow simpler differentiation of isoforms. Reporters were truncated at each native start site and also mutated to an AUG codon, to prevent leaky scanning and ensure individual expression of -207, -45 and annotated isoforms. Phusion PCR was carried out using different forward primers to amplify each GATAD1 isoform (Figure 6-1A).

Each PCR product was cloned into the pcDNA3F backbone using NheI/XhoI restriction sites. Insertion of the GATAD1 PCR products into the pcDNA3F plasmid vector ensured that all transcripts contained a C-terminal 3xFLAG-tag, which was utilised in expression studies. NcoI diagnostic digests confirmed that each clone contained GATAD1 constructs of the correct size (Figure 6-1B), before the plasmids were sequenced.

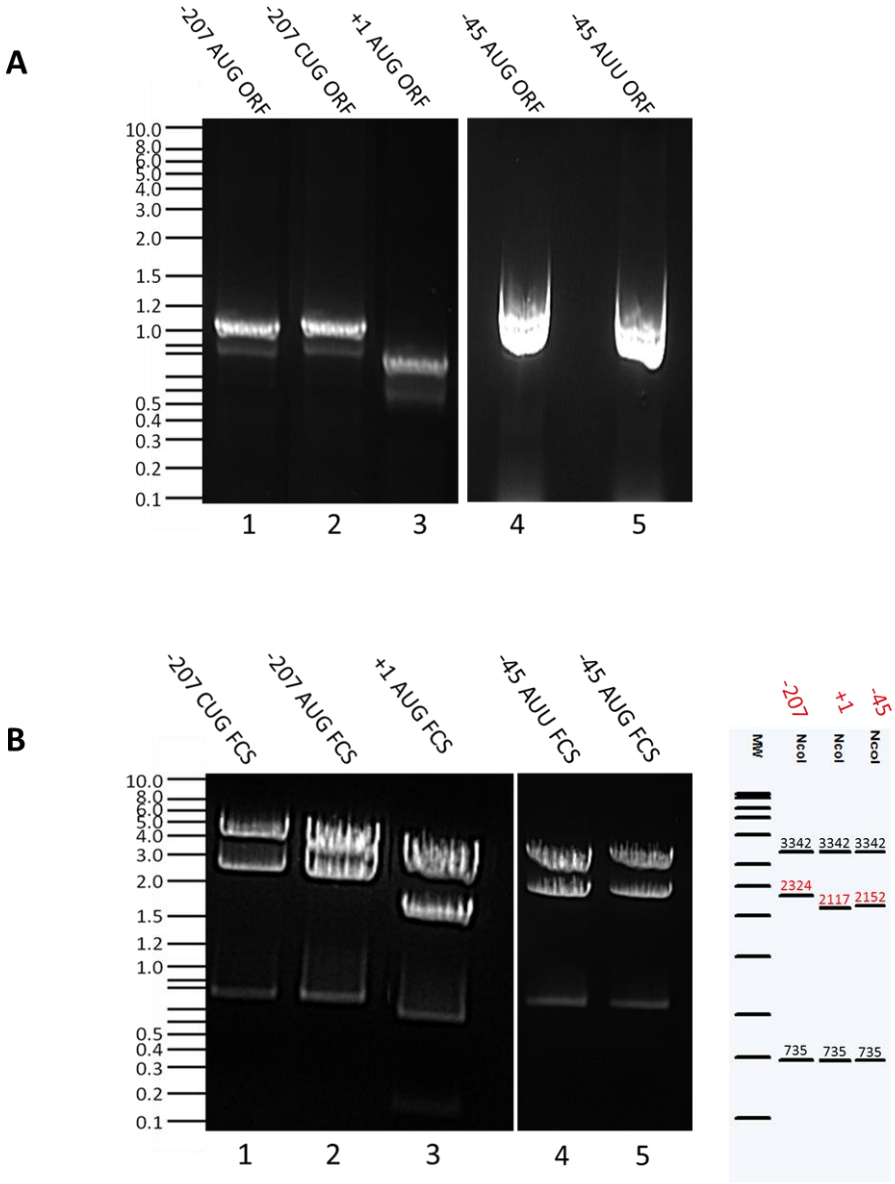


Figure 6-1: GATAD1 FCS Cloning

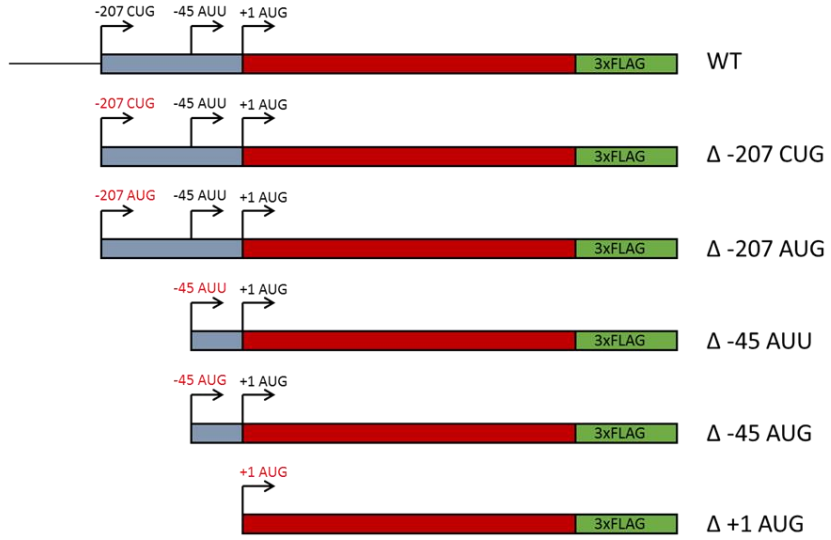
(A) 0.8% agarose gel resolving the GATAD1 PCR fragments truncated at each initiation codon and containing the whole CDS. The amplified -207, -45 and annotated fragments are the correct size, 1040 bp, 872 bp and 833 bp respectively. (B) NcoI diagnostic digest of the recombinant GATAD1 FCS plasmids. The middle band generated by the NcoI digest will change in size depending on the size of the insert, since the insert sits between two NcoI restriction sites.

6.3.2 Expression of GATAD1 Full Coding Sequence Constructs

Expression of the full length GATAD1 constructs were analysed by Western blot (Figure 6-2C), to ensure individual expression of isoforms and minimal leaky scanning was taking place. The wild type FCS GATAD1 in lane 2 of the FLAG-probed immunoblot produced 3 bands representing the 3 GATAD1 isoforms (-207, -45, +1), running slightly higher than expected (Figure 6-2B). Expression from the truncated -207 CUG and AUG results in most of the expression occurring from the -207 AIC as intended, however a small amount of leaky scanning does occur, resulting in minimal expression from the -45 AUU and +1 AUG. Expression from the truncated -45 AUU prevents initiation from the AUU start site, causing total initiation from the annotated AUG. However, when the AUU is mutated to an AUG, the majority of translation takes place from the -45 position with a small amount of expression from the annotated AUG. This demonstrates that the ribosome needs all or part of the 5'UTR as a 'run up' to recognise the AUU codon. The hairpin identified in section 4.3.1 (Figure 4-2) immediately upstream of the AUU codon may indeed be used by the ribosome to increase recognition of the unusual AIC, which on its own is not strong enough to initiate translation and produce the middle isoform of GATAD1. Removing all of the 5'UTR results in expression from the annotated AUG. A small amount of leaky scanning also takes place causing expression from the internal +55 AUG.

Leaky scanning resulted in cells expressing several isoforms. This would produce skewed data when comparing the localisation of each individual isoform by immunofluorescence. To overcome this, internal mutagenesis of the remaining initiation codons downstream of the AUG mutation were made to produce constructs which should translate one single isoform when expressed in a cell line. Internal mutants were made whereby the -45 AUU was mutated to a UAC and the annotated AUG was mutated to a CUU. Although the internal mutations prevent any leaky scanning and translation from the AICs within the 5'UTR, a small amount of N-terminally truncated GATAD1 is still expressed from the internal +55 AUG which is not normally recognised. Once sequenced (Figure 6-3), a FLAG-probed immunoblot was used to analyse the expression of the internal FCS mutants (Figure 6-4).

A



B

Initiation Codon	Expected Size (kDa)
Δ -207	38.6
Δ -45	33.1
Δ +1	31.5

C

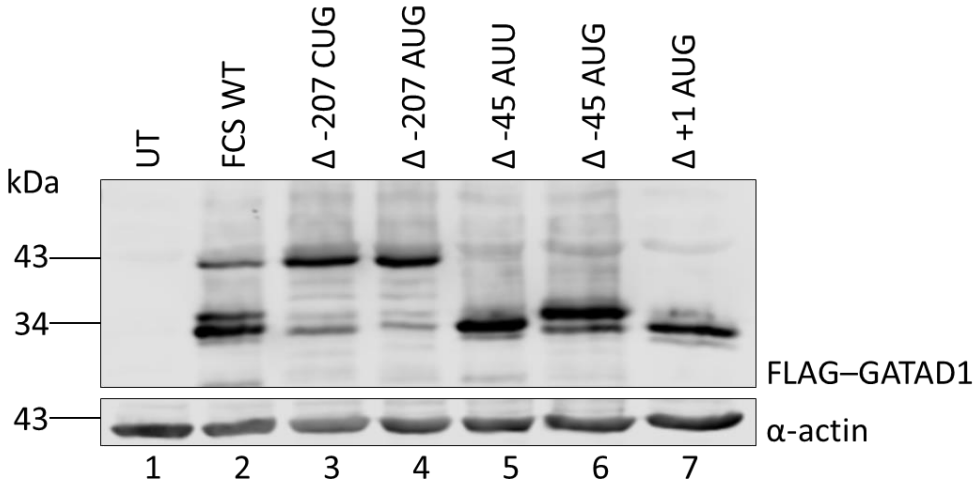


Figure 6-2: Expression of GATAD1 FCS Constructs

(A) Virtual representation of each of the GATAD1 FCS isoforms containing a C-terminal 3xFLAG-tag. Δ represents the 5'UTR truncation at each initiation codon. (B) The expected molecular weights of the protein isoforms containing 3x-FLAG-tag translated from each of the AICs as well as the canonical start site. (C) FLAG-probed Western blot showing expression of GATAD1 protein isoforms when N-terminal truncations and AUG-mutations were made to the mRNA sequence within the 5'UTR. 10 µg of transfected HeLa whole cell lysate was loaded on to a 10% SDS-PAGE gel.

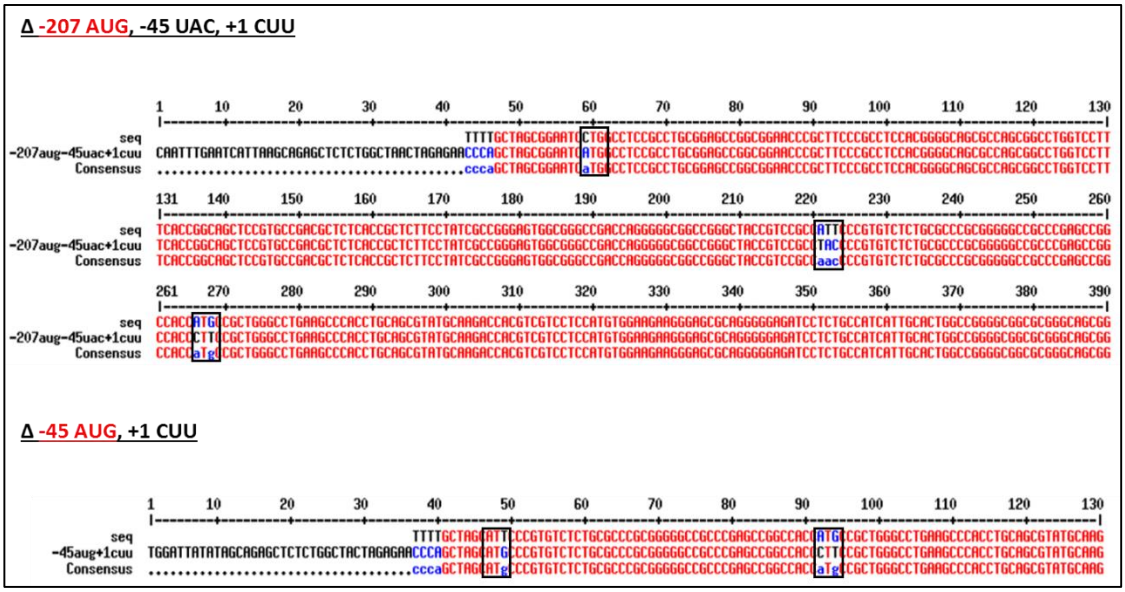


Figure 6-3: Sequence alignment of GATAD1 FCS Internal Mutant Plasmids

Sequence alignment of the GATAD1 mutants to wild-type Δ GATAD1 isoforms were carried out using Multalin, confirming that the correct mutations have been made at each internal initiation codon.

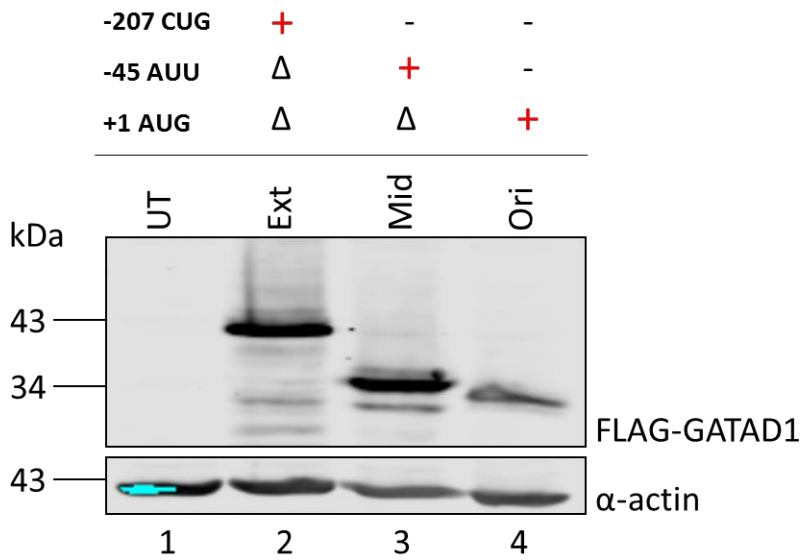


Figure 6-4: Expression of Isoform with Internal Mutations

FLAG-probed Western blot showing expression of GATAD1 protein isoforms when N-terminal truncations (represented by Δ) and AUG-mutations (represented by +) were made to each AIC, as well as internal mutations to the downstream initiation codons (represented by -) to prevent any leaky scanning from taking place. 10 µg of transfected HeLa whole cell lysate was loaded on to a 10% SDS-PAGE gel.

6.3.3 Predictions of Subcellular Localisation of GATAD1 Protein Isoforms

A prediction of the subcellular localisation of a protein can be carried out by analysing the N-terminal amino acid sequence to search for a signal peptide or targeting sequence. Two different prediction sites Psort (Nakai and Kanehisa, 1992) and TargetP (Emanuelsson et al., 2000) were able to predict a subcellular localisation for each isoform. Both searches predicted that the isoform translated from position -207 contained an N-terminal MTP, directing the largest isoform predominantly to the mitochondria. Both the -45 and annotated isoform were predicted to localise to the nucleus and cytoplasm (Table 6-1).

Table 6-1: Prediction of Isoform Subcellular Localisation

Analysis of the amino acid sequence of each isoform provided a subcellular localisation prediction. Psort used the *k*-nearest neighbour (*k*-NN) NN prediction algorithm, whereby the output of 23 sub programmes are normalised to provide a final localisation prediction. TargetP has a 90% accuracy in discriminating between proteins containing different signal sequences; mitochondrial target peptide (mTP) or signal peptide (SP) as well as providing predicted presequence length (TPlen) and also provides a reliability class (RC) 1-5, where 1 indicates the strongest prediction.

	Psort	TargetP
-207	56.5 %: mitochondrial 30.4 %: nuclear 4.3 %: vesicles of secretory system 4.3 %: cytoskeletal 4.3 %: cytoplasmic >> prediction for QUERY is mit (k=23)	Len mTP SP other Loc RC TPlen 338 0.850 0.034 0.151 M 2 40
-45	34.8 %: nuclear 34.8 %: cytoplasmic 21.7 %: mitochondrial 8.7 %: cytoskeletal >> prediction for QUERY is nuc (k=23)	Len mTP SP other Loc RC TPlen 284 0.335 0.083 0.526 _ 5 -
+1	34.8 %: nuclear 30.4 %: mitochondrial 26.1 %: cytoplasmic 4.3 %: cytoskeletal 4.3 %: peroxisomal >> prediction for QUERY is nuc (k=23)	Len mTP SP other Loc RC TPlen 269 0.251 0.071 0.699 _ 3 -

6.3.4 Immunofluorescence to Determine GATAD1 Isoform Localisation

Immunofluorescence was initially carried out on endogenous WT GATAD1, expressing all three isoforms in both HeLa cells and in CRISPR GATAD1-3xFLAG-tagged HEK293 cells. In both cell lines, endogenous GATAD1 localises to both the nucleus and the cytoplasm, with no obvious mitochondrial localisation (Figure 6-5).

Over-expression of FCS FLAG-tagged GATAD1 isoforms show a difference in subcellular localisation between the annotated isoform and the N-terminally extended isoforms when analysed by immunofluorescence (Figure 6-6). The untransfected control shows no FLAG fluorescence, confirming FLAG antibody specificity. Wild-type GATAD1 expressing all three isoforms, localises to both the nucleus as well as the cytoplasm and shows accumulation of GATAD1 around the nuclear envelope. Both N-terminally extended isoforms Ext and Mid, translated from -207 CUG and -45 AUU also localise to the nucleus and cytoplasm and show a strong signal around the nuclear envelope, whether translated from their annotated initiation codon or enhanced with an AUG codon (Figure 6-7). On the other hand, the annotated GATAD1 isoform localises only to the nucleus, with no signal at the nuclear envelope and minimal cytoplasmic localisation. The fluorescent signal intensities were quantified using ImageJ in order to confirm the localisation of each isoform (Figure 6-8). This was then also confirmed using Stimulated Emission Depletion Microscopy (STED) on a demo instrument, which was able to generate images with super-resolution. Images were acquired by scanning focused light over an area of the cell alongside a second STED laser beam which depletes fluorophores at the focal periphery. The fluorescence is sequentially collected pixel by pixel, overcoming the diffraction limited resolution of confocal microscopes. The accumulation of extended GATAD1 isoforms around the nuclear envelope is still visible, although not as intense (Figure 6-9). Also, extended GATAD1 appear to be localising to the nucleoli, which was not previously observed using widefield microscopy, however this could not be further investigated without having the instrument for longer.

In order to confirm that the differences in subcellular localisation observed were not due to human bias, blind quantification was carried out on transfected HeLa cells. This confirmed that the N-terminally extended isoforms were more cytoplasmic than annotated GATAD1. 74% of the Ext isoform and 66% of the Mid isoform localised to the cytoplasm as well as the nucleus, compared to just 16.2% of annotated GATAD1 (Figure 6-10).

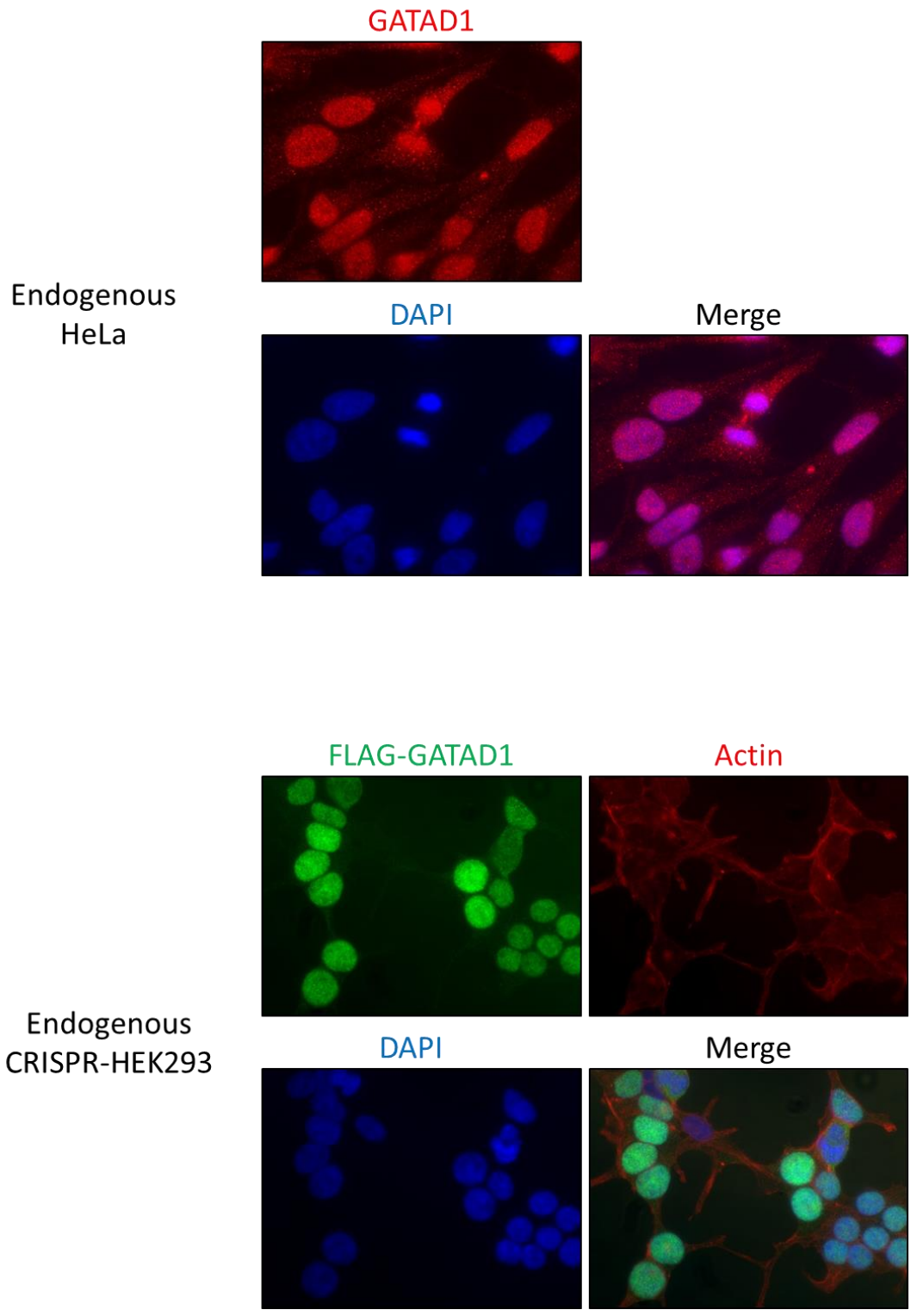


Figure 6-5: Endogenous GATAD1 Localisation

The localisation of endogenous GATAD1 within HeLa cells was observed by carrying out immunofluorescence using Sigma anti-GATAD1 produced in rabbit as the primary antibody, followed by anti-rabbit Alexa Fluor 555 as the secondary. GATAD1 in the CRISPR-HEK293 cells was probed using Sigma anti-FLAG produced in mouse as the primary antibody, followed by anti-mouse Alexa Fluor 488 as the secondary. The actin cytoskeleton was stained using Alexa Fluor 555 Phalloidin.

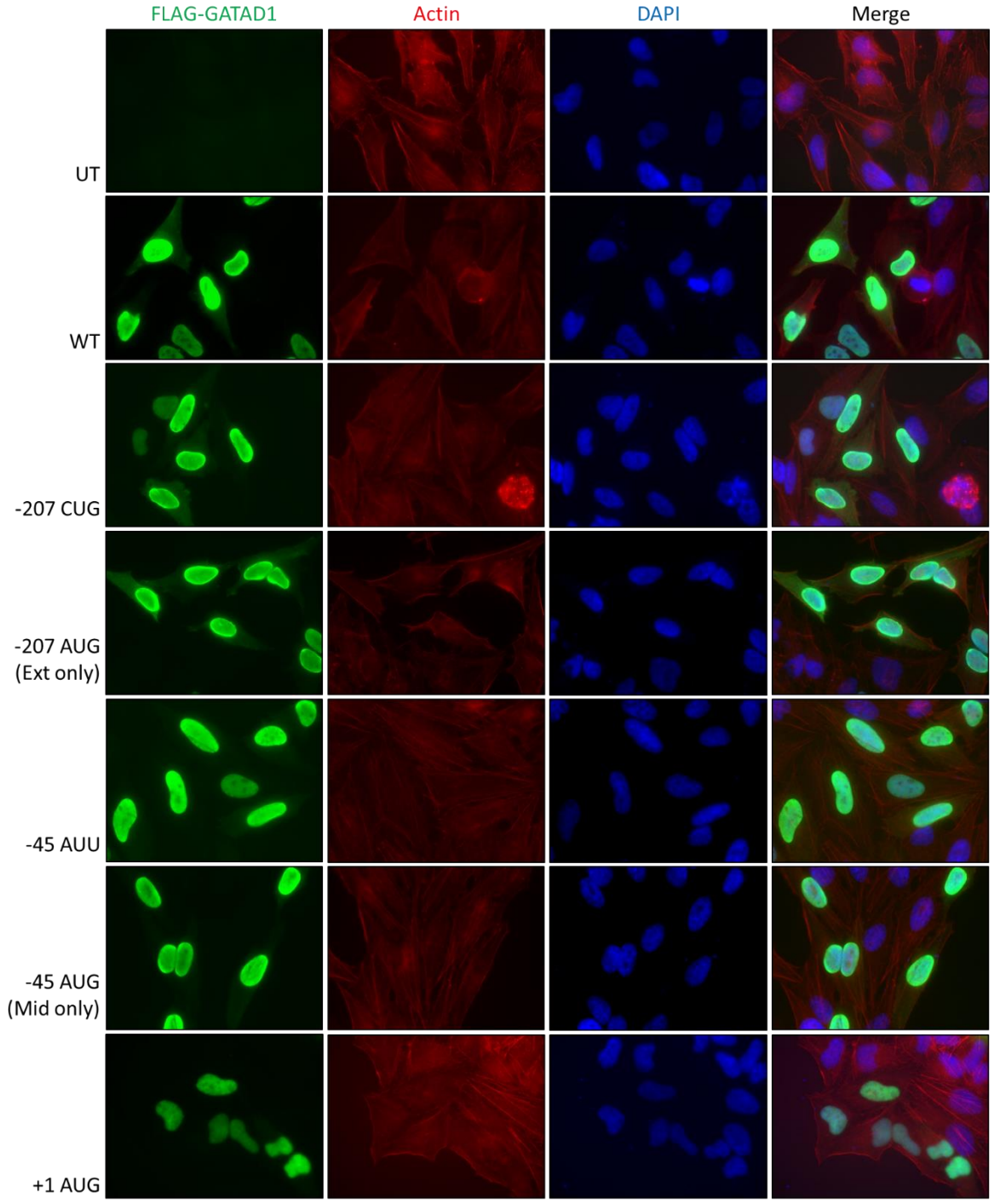


Figure 6-6: Subcellular Localisation of each GATAD1 Isoform

Immunofluorescence was carried out on each N-terminally truncated FCS Δ construct with its annotated initiation codon (-207 CUG, -45 AUU and +1 AUG), as well as N-terminally truncated constructs containing AUG-mutations to each AIC and internal mutations to downstream initiation codons (Ext only, Mid only) which prevent leaky scanning from taking place, (Figure 6-4). UT = untransfected and WT = wild-type. 3xFLAG-tagged GATAD1 was probed using Sigma anti-FLAG produced in mouse as the primary antibody, followed by anti-mouse Alexa Fluor 488 as the secondary. The actin cytoskeleton was stained using Alexa Fluor 555 Phalloidin, whilst DNA was stained with DAPI.

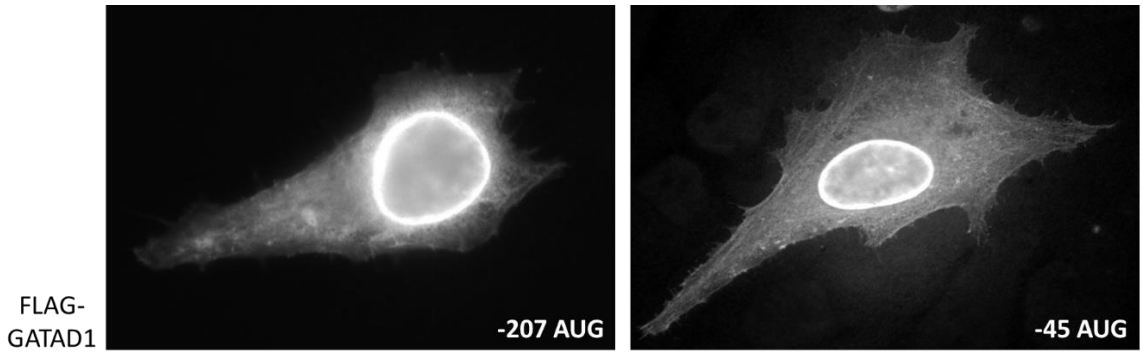
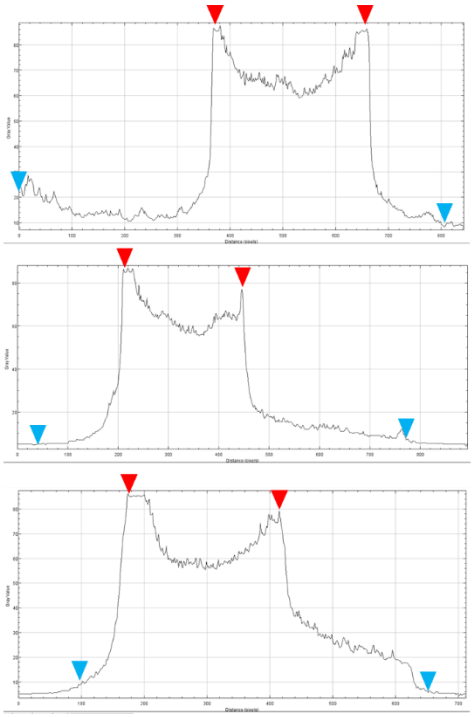


Figure 6-7: Subcellular Localisation of Extended GATAD1 Isoforms

Immunofluorescence was carried as in the previous figure. The FLAG 488 channel is shown here in black and white to show the lamin localisation more clearly for each of the GATAD1 extended isoforms, translated from -207 AUG and -45 AUG.

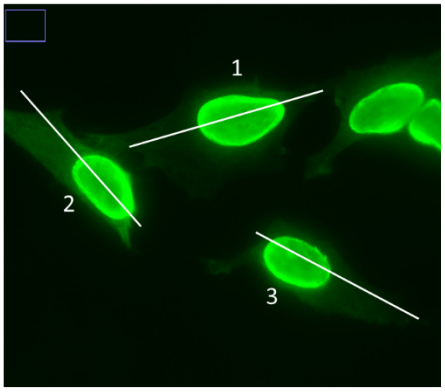
A



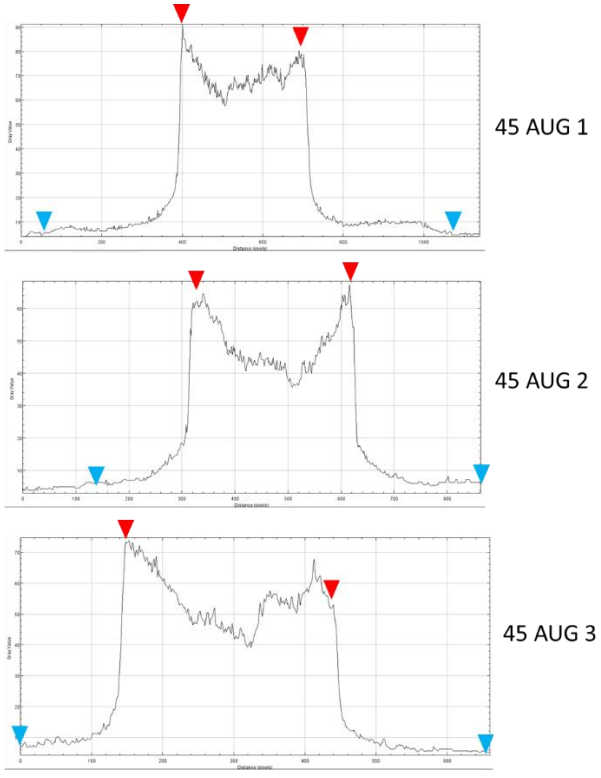
207 AUG 1

207 AUG 2

207 AUG 3



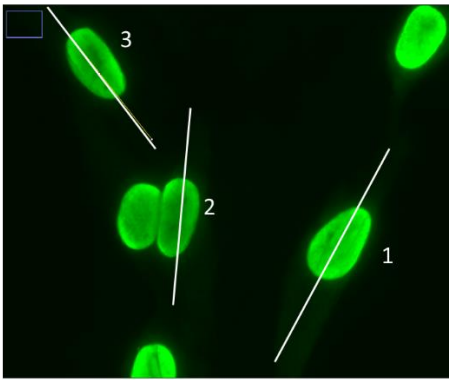
B



45 AUG 1

45 AUG 2

45 AUG 3



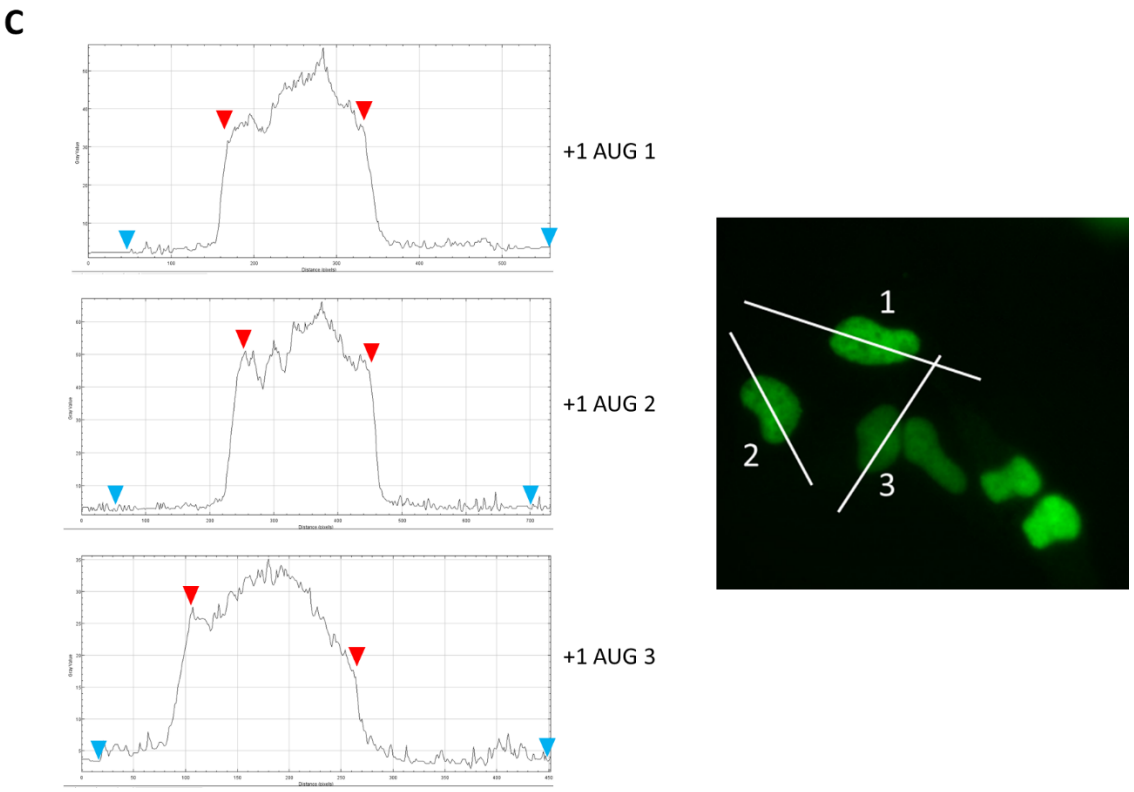


Figure 6-8: Fluorescence Intensities of each GATAD1 Isoform across the Cell

Plot profiles were analysed using ImageJ for three cells expressing each isoform, which show that there is an accumulation of Ext and Mid (panel A and B) at the nuclear envelope which is not present in the annotated isoform (panel C). Also, Ext has the most expression within the cytoplasm, followed by Mid, whereas expression of annotated GATAD1 within the cytoplasm is negligible. The blue arrowheads represent the cell boundary whilst the red arrowheads represent the nuclear boundary.

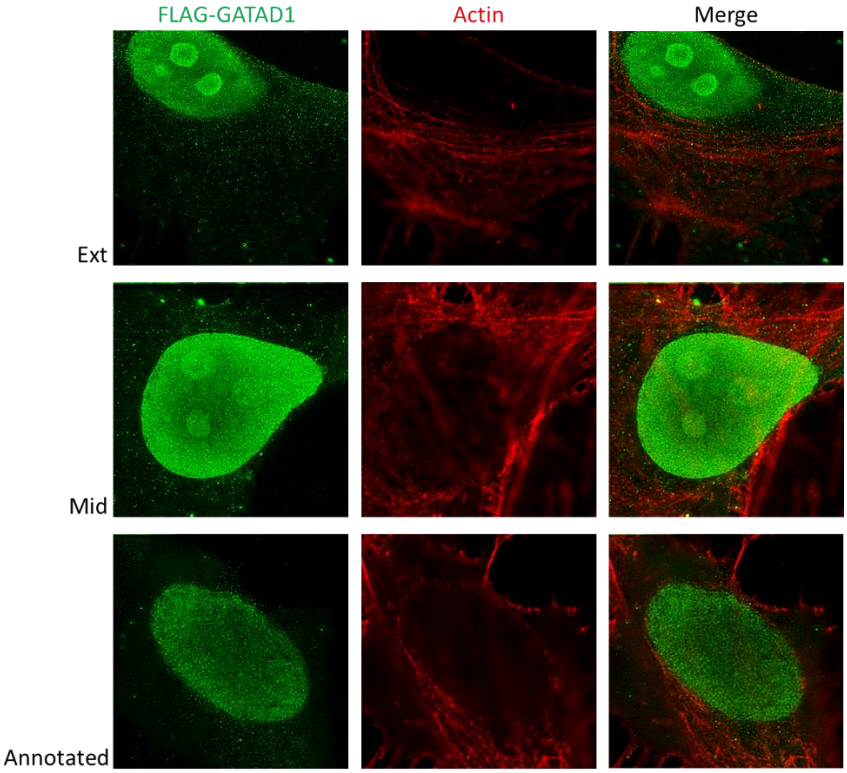


Figure 6-9: STED Imaging of GATAD1 Isoforms

3xFLAG-tagged GATAD1 was probed using Sigma anti-FLAG produced in mouse as the primary antibody, followed by anti-mouse Alexa Fluor 488 as the secondary and the actin cytoskeleton was stained using Alexa Fluor 555 Phalloidin.

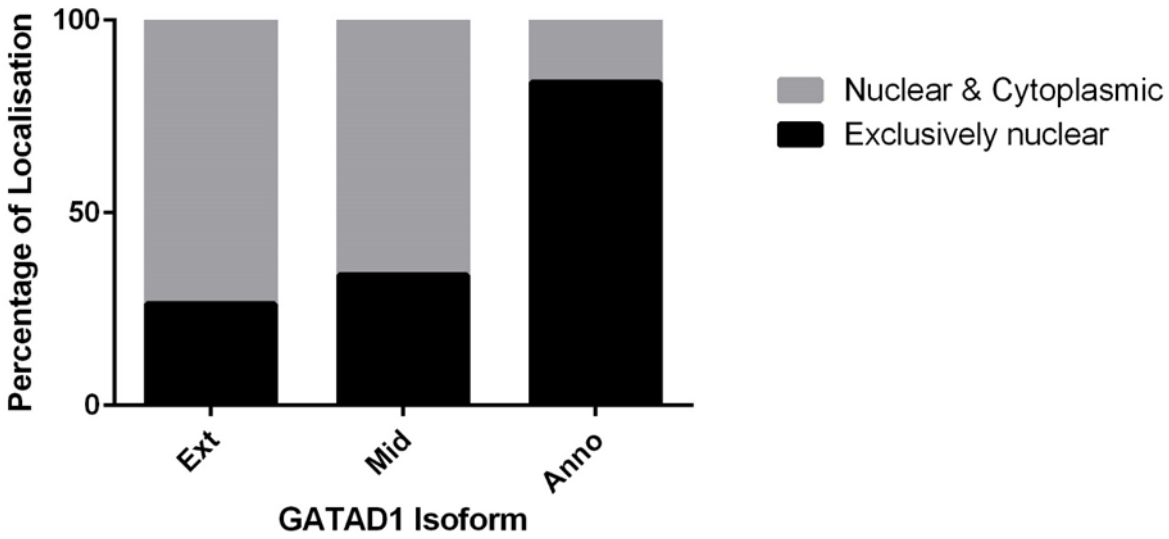


Figure 6-10: Extended Isoforms Are More Cytoplasmic Than Annotated GATAD1

Quantifications were based on the localisations of 209, 225 and 173 transfected HeLa cells (Ext, Mid, Anno respectively).

6.3.5 Localisation of Extended GATAD1 Isoforms

6.3.5.1 No co-localisation of Ext isoform with mitochondria as predicted

Two prediction softwares Psort and TargetP predicted that the Ext isoform translated from -207 CUG contained an undefined mitochondrial target peptide within the N-terminal extension. Immunofluorescence was carried out in order to see whether there was any co-localisation between Ext and CoxIV, a mitochondrial marker, however no co-localisation was observed under the conditions used, even when the Ext localisation looked unusual and potentially mitochondrial (Figure 6-11).

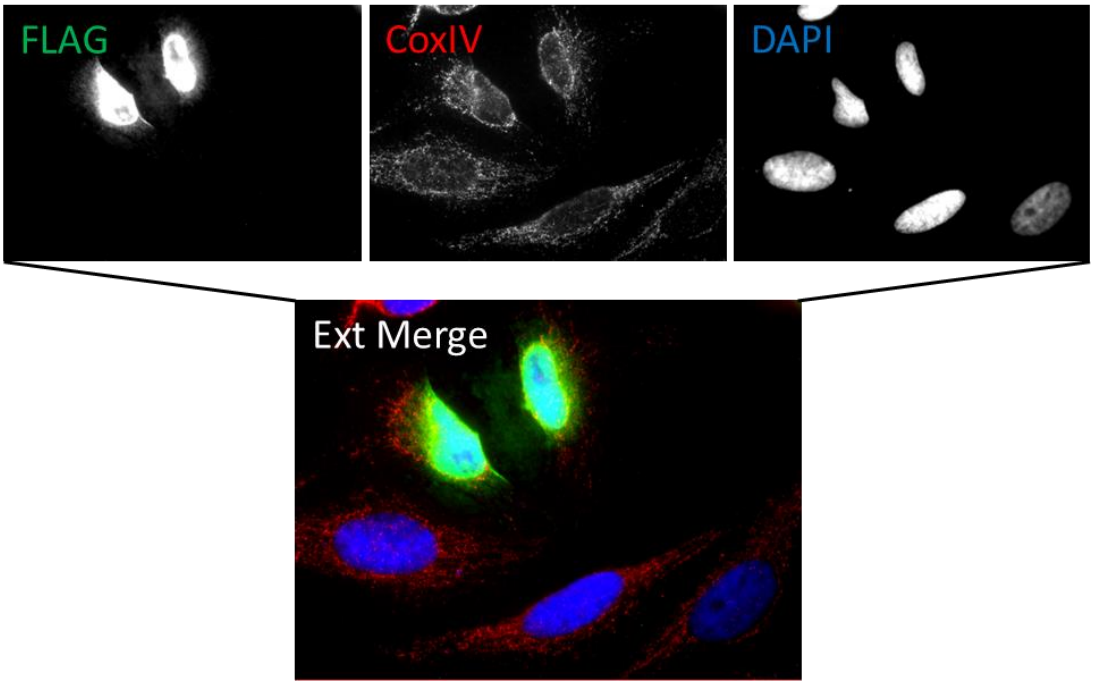


Figure 6-11: GATAD1 -207 Isoform and CoxIV Localisation

The Ext 3xFLAG-tagged GATAD1 isoform was probed using Sigma anti-FLAG produced in mouse as the primary antibody, followed by anti-mouse Alexa Fluor 488 as the secondary. The mitochondria was probed using a CoxIV marker produced in rabbit as the primary antibody, followed by anti-rabbit Alexa Fluor 555 as the secondary antibody, whilst DNA was stained with DAPI.

6.3.5.2 Extended isoforms co-localise with the nuclear envelope

Immunofluorescence showed that both N-terminally extended GATAD1 isoforms (Ext and Mid) had enriched localisation at the nuclear envelope (Figure 6-7). To confirm this, the nuclear localisation of Ext and Mid was compared to that of Lamin A/C, which forms part of the nuclear lamina on the interior leaflet of the nuclear envelope (Figure 6-12). Co-localisation was not possible since both FLAG and Lamin A/C antibody were raised in mouse, although tyramide signal amplification would have allowed sensitive double fluorescence immunostaining from the same host species (Toth and Mezey, 2007).

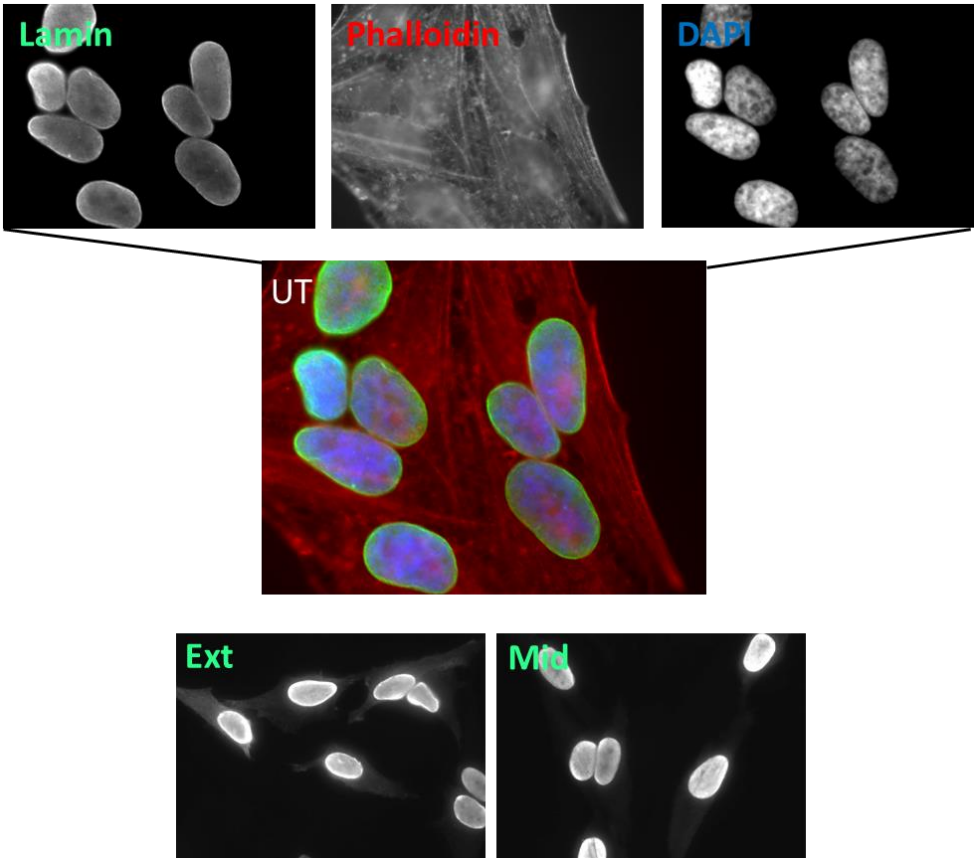


Figure 6-12: Lamin Localisation Compared to -207 AUG Localisation

Lamin was probed using CST Lamin A/C produced in mouse as the primary antibody, followed by anti-mouse Alexa Fluor 488 as the secondary (green signal emission), the actin cytoskeleton was stained with Alexa Fluor 555 Phalloidin (red signal emission) and DNA was stained with DAPI, (blue signal).

6.3.6 Signals Controlling Localisation of GATAD1 Isoforms

6.3.6.1 Nuclear Localisation Signals (NLSs)

Once translated in the cytoplasm, proteins are targeted to their final subcellular localisation through a range of signals, which are usually found at the N-terminus of the protein. Nuclear proteins are tagged for nuclear import with a NLS, consisting of short sequences of positively charged residues such as lysine or arginine. Since all three GATAD1 isoforms are nuclear, it was predicted that a NLS would be present within the annotated protein sequence.

Classical NLSs bind to the major binding pocket of importin α and are rich in basic amino acids, whether these are monopartite class 1 consisting of 4 consecutive basic amino acids, or class 2 which have 3 basic amino acids in consensus K(K/R)X(K/R) where X is any amino acid. A classical bipartite NLS consists of two clusters of basic residues separated by a 10-12 residue linker (K/R)(K/R)X10-12(K/R)3/5. The prediction software cNLS Mapper (Kosugi et al., 2009) was unable to identify any classical NLSs within the GATAD1 protein sequence (Figure 6-13).

There are also noncanonical NLS signals which can be recognised by either importin α or importin β. Eukaryotic NLS classes 3 and 4 bind the minor binding pocket of importin α with consensus sequence KRX(W/F/Y)XXAF and (P/R)XXKR(K/R) respectively, (Kosugi et al., 2009). Importin β-dependent NLSs have more variation and longer sequences than classical NLSs. Although a consensus sequence has not yet been established, many are rich in arginine residues. No non-canonical NLSs were identified within GATAD1 either.

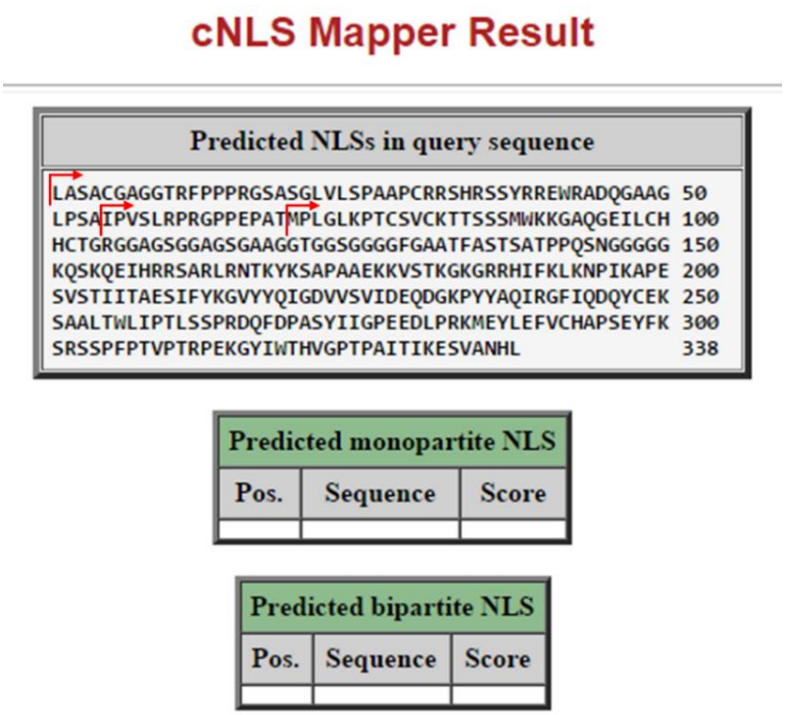


Figure 6-13: No Nuclear Localisation Signal within GATAD1

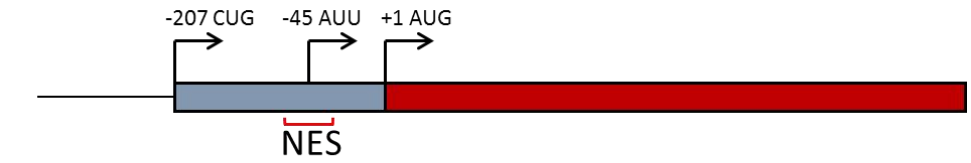
cNLS Mapper accurately predicts importin α-dependent NLS signals within protein sequences, based on nuclear import assays in budding yeast. The importin α/β pathway is highly conserved in eukaryotes therefore NLS predictions should be accurate. The N-terminus of the three GATAD1 isoforms are indicated by red arrows.

6.3.6.2 Nuclear Export Signals (NESs)

Many nuclear proteins are shuttled in and out of the nucleus by co-ordination of NLS and NES transport signals. Since the GATAD1 isoforms all have nuclear localisation (although no obvious NLS), and extended GATAD1 isoforms are also localised to the cytoplasm, the protein sequences were analysed for the presence of an NES.

The classical nuclear export pathway is mediated by the CRM1/exportin protein, whereby the CRM1-Ran-GTP complex binds directly to the NES, directing export across the nuclear membrane through the nuclear pore complex (NPC). Classical NESs are leucine-rich sequences containing large hydrophobic residues, with the consensus sequence L-X(2,3)-[LIVFM]-X(2,3)-L-X-[LI]. On the other hand, a non-classical nuclear export pathway is mediated by other importin β members such as Msn5, whereby export of proteins can be controlled by their phosphorylation status, (Kaffman et al., 1998).

The GATAD1 Ext protein sequence (which included Mid and Anno) was entered into NetNES 1.1 prediction software (Cour et al., 2004), which predicted the presence of a full NES within Ext and partial NES within the Mid isoform with a probability of approximately 0.6 (Figure 6-14). In order to investigate whether the potential NES with sequence LPSAIPVSL was functional, leucine-alanine mutations were made to Ext (Figure 6-15A). Although a partial NES exists in the Mid isoform, it would not be functional without the first hydrophobic leucine residue. When the predicted NES was made non-functional, Ext remained localised to the cytoplasm as well as the nucleus (Figure 6-15B). Ext and Mid are therefore not shuttling between the nucleus and cytoplasm. This was confirmed using Leptomycin B, an anti-fungal antibiotic which also acts as a nuclear export inhibitor by forming a covalent complex with CRM1, resulting in loss of NES recognition and subsequent nuclear export. Both Ext and Mid remained localised in the cytoplasm after treatment with 40 nM Leptomycin B for 2 hours (data not shown).



Mid start

#Seq-Pos-Residue	ANN	HMM	NES	Predicted
Sequence-50-G	0.072	0.000	0.000	-
Sequence-51-L	0.097	0.908	0.602	Yes
Sequence-52-P	0.082	0.908	0.603	Yes
Sequence-53-S	0.081	0.908	0.604	Yes
Sequence-54-A	0.071	0.908	0.603	Yes
Sequence-55-I	0.099	0.910	0.610	Yes
Sequence-56-P	0.075	0.910	0.611	Yes
Sequence-57-V	0.079	0.911	0.608	Yes
Sequence-58-S	0.069	0.905	0.599	Yes
Sequence-59-L	0.087	0.944	0.640	Yes
Sequence-60-R	0.085	0.043	0.000	-

Sequence - NetNES 1.1 prediction

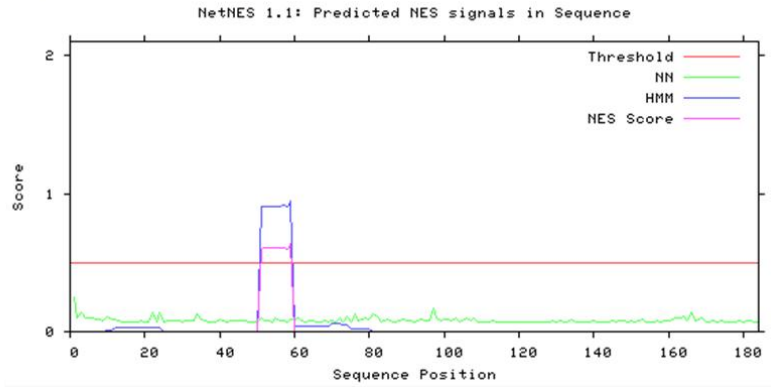


Figure 6-14: Full NES Present in Ext GATAD1

There is a classical leucine-rich NES within the Ext GATAD1 isoform, which spans the Mid start site. The ANN value recognises the most C-terminal hydrophobic position of the NES motif, but does not peak in this search, whilst the HMM assigns a score to hydrophobic positions within the motif, generating a final NES probability.

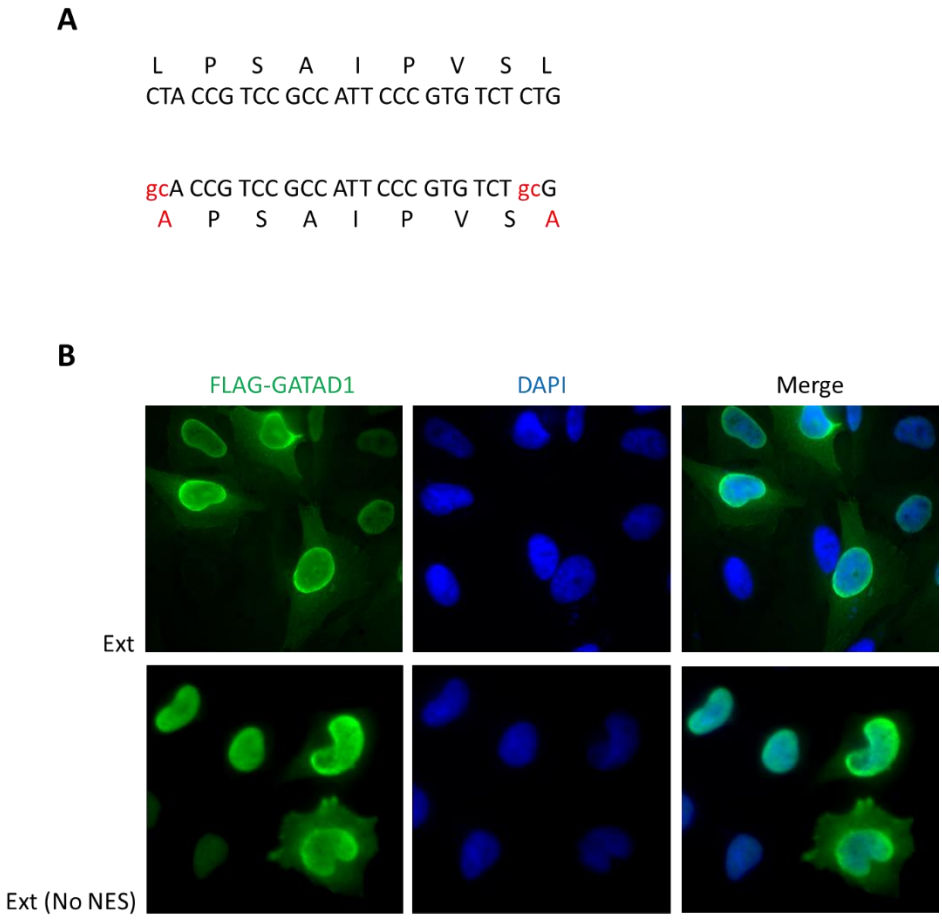


Figure 6-15: Predicted NES Is Not Functional

(A) The predicted NES would require leucine residues to bind to exportin. Alanine mutations should abolish the NES function, preventing export of the Ext isoform, if functional. (B) The Ext GATAD1 isoform remains localised to both the nucleus and the cytoplasm, when the NES has been mutated out. 3xFLAG-tagged Ext GATAD1 was probed using Sigma anti-FLAG produced in mouse as the primary antibody, followed by anti-mouse Alexa Fluor 488 as the secondary, whilst DNA was stained with DAPI. The actin signal from Phalloidin in this experiment was poor and was therefore not shown.

6.4 Summary of Main Findings

- N-terminally extended GATAD1 isoforms Ext and Mid have a more cytoplasmic localisation than the annotated isoform.
- Ext does not localise to the mitochondria, as the predicted by online software.
- No obvious NLS or NES to explain both nuclear and cytoplasmic localisation of GATAD1 isoforms.

6.5 Discussion

6.5.1 N-terminally Extended GATAD1 Isoforms Are Cytoplasmic and Nuclear

Both Psort and TargetP prediction softwares predicted that Ext GATAD1 isoform would localise to the mitochondria, however immunofluorescence showed no evidence for this. It may be possible that certain conditions such as hypoxia or cell stress may re-localise Ext to the mitochondria.

Both N-terminally extended GATAD1 isoforms Ext and Mid have enriched expression around the nuclear envelope, which is not seen with the annotated isoform. Although the localisation of the halo appeared very similar to the localisation of lamin on the inner leaflet of the nuclear membrane, this must be experimentally confirmed. Initial saponin permeabilisation experiments were carried out in order to determine whether the enrichment was present on the inner or outer leaflet of the nuclear envelope. Saponin is a detergent which selectively and reversibly permeabilises membrane cholesterol which is abundant in the plasma membrane, but not in the nuclear envelope, which saponin therefore cannot permeabilise. By selectively permeabilising the cell it would be possible to determine the exact localisation of the Ext and Mid isoforms around the nucleus. 0.1% saponin treatment permeabilised the nuclear membrane as well as the plasma membrane, therefore this must be optimised using lower concentrations (0.01%). On the other hand, digitonin may be used as an alternative, which is far more selective against the nuclear membrane.

There are several potential explanations for the enrichment of Ext and Mid at the nuclear envelope. Localisation of transcription factors to the inner nuclear envelope occurs in order to increase the efficiency of mRNA delivery to the cytoplasm once transcription has taken place in the nucleus. Direct association of transcription factors with membrane proteins or the nuclear lamina have been shown to sequester transcription factors away from the target chromatin, reducing access to their target genes (Heessen and Fornerod, 2007). Since GATAD1 is implicated in transcriptional

repression (Levy and Gozani, 2010), this method of sequestration will result in transactivation of target genes. The enrichment around the nuclear membrane was most apparent in cells expressing high levels of Ext and Mid, therefore sequestering the N-terminally extended GATAD1 away from the target chromatin may be an extra layer of transcriptional control.

Cytoplasmic N-terminally extended GATAD1 isoforms and their associated binding partners (including histone demethylase KDM5A), are away from their suggested functional target, the chromatin. KDM5A protein is also found in the cytoplasm and is enriched in several polysome profile fractions (Van Rechem et al., 2015), whilst another JmjC-demethylase KDM4A, also localises to the cytoplasm where it has a role in regulating protein translation, affecting translation initiation factor distribution (Van Rechem et al., 2015). On the other hand, lysine methylation of non-histone proteins can regulate both their stability and function, (Egorova et al., 2010). Therefore the cytoplasmic localisation of a H3K4 demethylase complex may indeed have a functional role, rather than sequestration from its functional chromatin target.

6.5.2 GATAD1 Isoforms Are Not Imported To Nucleus by NLS

GATAD1 forms part of a histone demethylase complex and so has a central functional role within the nucleus. However, no NLS has been identified within the amino acid sequence. Several mechanisms exist to import nuclear proteins without a NLS, including non-conventional nuclear transport mechanisms and a 'piggybacking' mechanism, which is used by another transcription factor with no NLS, MIER1α. An ELM2 domain was necessary and sufficient for nuclear localisation of MIER1α via binding to HDAC1/HDAC2 which were used to 'piggyback' into the nucleus (Li et al., 2013). MIER1α forms part of a co-repressor complex with HDAC1/2 once in the nucleus. Interestingly, the ELM2 (Egl-27 and MTA1 homology 2) domain is usually found in the N-terminus of a GATA binding domain (present in GATAD1) as well as associated with the ARID DNA binding domain (found in KDM5A). Although a domain search did not find an ELM2 domain within GATAD1, there may be a similar sequence allowing GATAD1 to 'piggyback' into the nucleus with HDAC1/2, which interact with the GATAD1 histone demethylase complex in the nucleus (Vermeulen et al., 2010).

6.5.3 N-terminally Extended GATAD1 Isoforms Are Not Exported to Cytoplasm by NES

There is no recognisable, functional NES within the N-terminally extended GATAD1 isoforms, therefore the proteins are not shuttled between the nucleus and cytoplasm. The cytoplasmic isoforms may not have been imported into the nucleus in the first instance, since there is no recognisable NLS and Ext/Mid isoforms were still present in the cytoplasm after Leptomycin B treatment.

Chapter Seven

Function of GATAD1 Isoforms

7. Function of GATAD1 Isoforms

7.1 Introduction

Alternative translation initiation takes place on the GATAD1 mRNA transcript, utilising an upstream CUG and AUU. This results in the translation of two N-terminally extended GATAD1 isoforms, which have a more cytoplasmic localisation than the nuclear annotated isoform. GATAD1 protein-protein interactions have been investigated using GATAD1 tagged with GFP using BAC TransgeneOmics to mimic the endogenous genomic context and protein levels in HeLa cells (Vermeulen et al., 2010). Quantitative SILAC-based GFP pulldowns followed by mass spectrometry analysis were then carried out to identify complexed proteins and their relative abundance. Several transcriptionally-repressive proteins were pulled-down with GATAD1 in similar ratios, including KDM5A (Jarid1A), EMSY, Sin3b and HDAC1/2. Other studies have also identified similar interactions. The Drosophila homolog of the Jarid1 family of proteins (Jarid1A, 1B, 1C) is Lid. The Lid complex has been shown to contain Sin3, EMSY and GATAD1 homologs (Lee et al., 2008). KDM5A has also been shown to complex with Sin3 proteins and HDAC1/2 in mice (van Oevelen et al., 2008). This evidence suggests that together, these proteins form an uncharacterised chromatin complex (Figure 7-1).

Annotated GATAD1 forms part of a histone demethylase complex comprising KDM5A (Jarid1A), which demethylates H3K4me3. H3K4me3 is associated with the transcription start site of actively transcribed genes and demethylation of this mark by KDM5A results in transcriptional repression (Vermeulen et al., 2010). The H3K4me3 mark increases transcription of associated genes by indirectly modifying chromatin structure. This allows DNA to uncoil from nucleosomes, exposing binding sites for transcription factors and RNA polymerases. H3K4me3 recruits positive chromatin remodelling factors to promote transcription and prevents binding of repressive transcription factors such as the NuRD complex (Nishioka et al., 2002).

In addition to methylation, H3K4 is also modulated by acetylation, which promotes transcription by changing the charge of the chromatin, directly opening the structure (Yan and Boyd, 2006). H3K4ac is also enriched at the promoters of actively transcribed genes and is located just upstream of the mutually exclusive H3K4me3 mark (Guillemette et al., 2011). This may explain the presence of HDAC1/2 within the novel GATAD1-containing chromatin complex, which deacetylate histones resulting in repression of transcription. HDAC1/2 along with Sin3 also form part of the Sin3 histone deacetylase complex, leading to gene silencing.

Together, the GATAD1-containing chromatin complex contains both histone demethylase enzymes as well as histone deacetylase enzymes and so could have the potential to repress translation through coupled histone demethylation and deacetylation on the H3K4 residue.

We showed that the KDM5A histone demethylase enzyme which is thought to complex with GATAD1, localises to the cytoplasm as well as the nucleus where its histone target is localised. Enrichment of KDM5A has been shown in several polysome profile fractions (Van Rechem et al., 2015). The N-terminally extended GATAD1 isoforms also localise to the cytoplasm as well as the nucleus, where their function away from the standard histone target remains unknown.

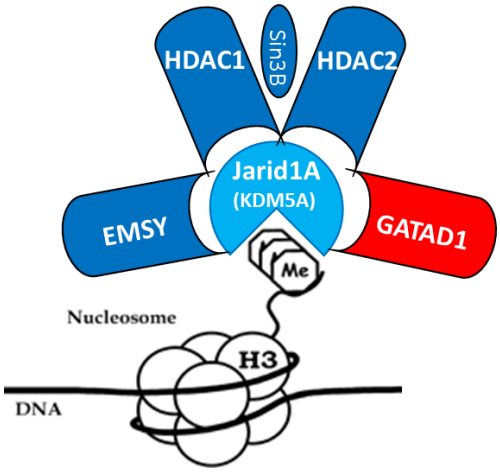


Figure 7-1: Depiction of GATAD1-Containing Chromatin Complex

Although the exact function of GATAD1 is unknown, it complexes with KDM5A as well as other transcriptionally repressive proteins, suggesting the formation of a regulatory chromatin complex.

7.2 Hypothesis and Aims

7.2.1 Hypothesis

Extended GATAD1 isoforms have an alternative non-chromatin mediated function in the cytoplasm.

7.2.2 Aims

The overall aim of this chapter was to determine whether the alternatively translated, N-terminally extended GATAD1 isoforms have a function in the cytoplasm, away from their regular chromatin target. This was carried out by:

- Investigating whether the extended isoforms remain complexed to KDM5A histone demethylase by carrying out co-IP experiments.
- Knocking-down endogenous GATAD1 before re-introducing each isoform and observing phenotypic changes.

7.3 Results

7.3.1 KDM5A Expression

7.3.1.1 KDM5A-HaloTag

GATAD1 forms a complex with KDM5A, forming a repressive transcription factor complex which functions in the nucleus. Co-IP assays were to be carried out on GATAD1 and KDM5A in order to ascertain whether N-terminally extended GATAD1 isoforms were still complexing with KDM5A when localised in the cytoplasm away from their functional target. Endogenous levels of KDM5A could not be detected using anti-Jarid1A (CST D28B10), therefore KDM5A-HaloTag (KDM5A-HaloTag human ORF in pFN21A – Promega FHC01704) was used (Figure 7-2A). Before the KDM5A-HaloTag clone was transfected, single NcoI and XhoI diagnostic digests were carried out which confirmed the presence of the correct gene (Figure 7-2C). Also, a KDM5A forward primer was designed to sequence part-way through the HaloTag into the KDM5A gene, whilst a reverse primer was designed to sequence back through the KDM5A gene, which produced a 784 bp PCR product again confirming that the correct gene was present (Figure 7-3). The primers were then used to sequence the plasmid and ensure that the gene was in the same frame as the HaloTag (Figure 7-4).

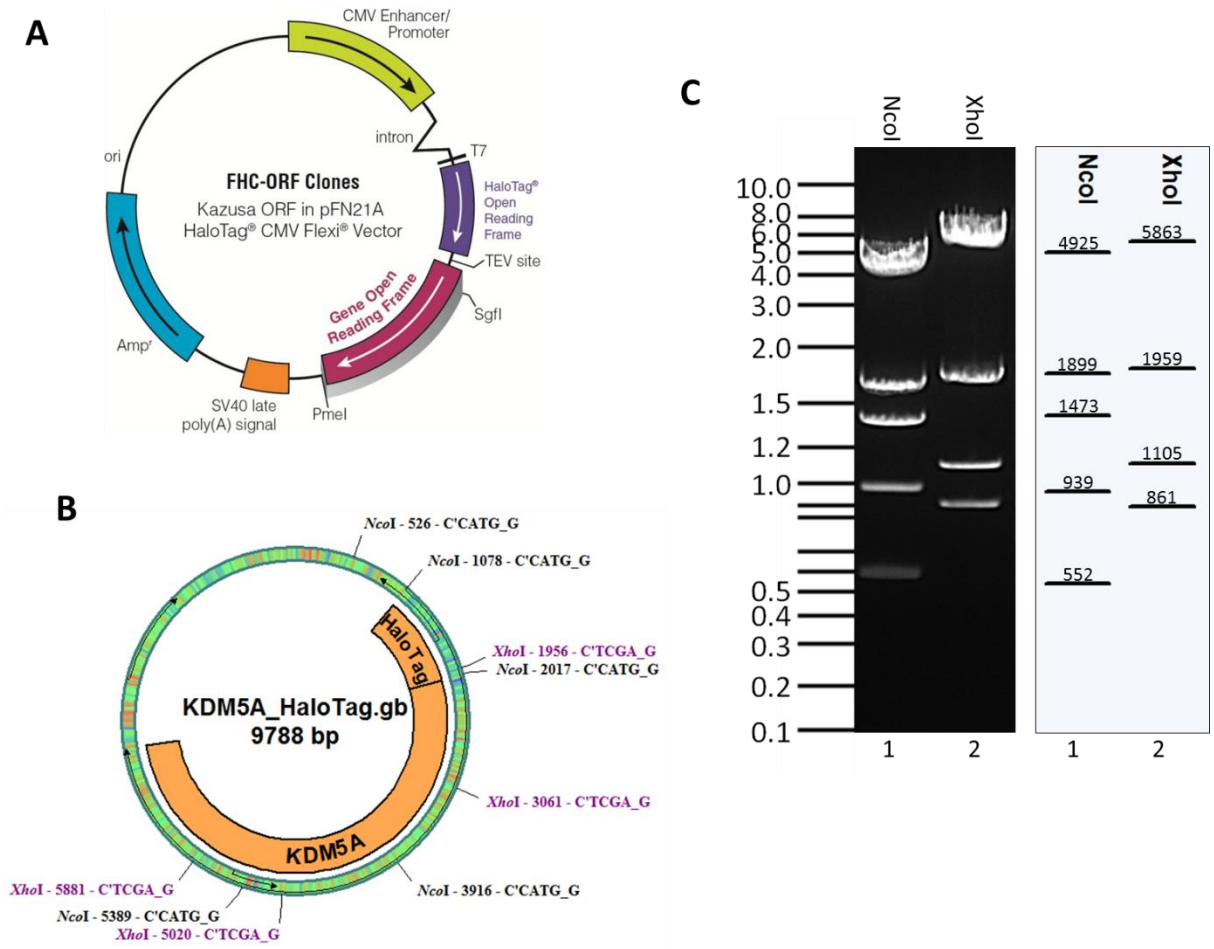
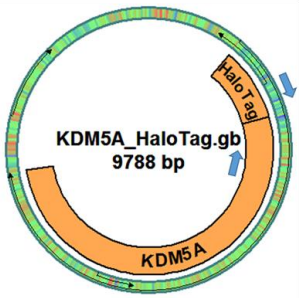


Figure 7-2: KDM5A-HaloTag

(A) Plasmid map of the Kazusa pFN21A vector with SgfI/PmeI cloning sites for the gene of interest (KDM5A) and N-terminal HaloTag. (B) Recombinant plasmid map showing the pFN21A vector containing KDM5A and the NcoI/XhoI restriction sites. (C) NcoI and XhoI diagnostic digests of KDM5A clone, run on a 0.8% agarose gel, confirming the presence of the correct gene in the correct vector.

Primer name	Primer sequence (5'-3')
HaloTag F	TCAGAACGTTTTTATCGAGGGTACG
KDM5A R	TTCTCTCTACCACAGGGATCTTCAG



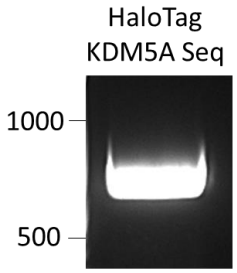


Figure 7-3: KDM5A-HaloTag Sequencing Primers PCR

Primers shown in the table amplified a region of approximately 784 bp, as predicted.

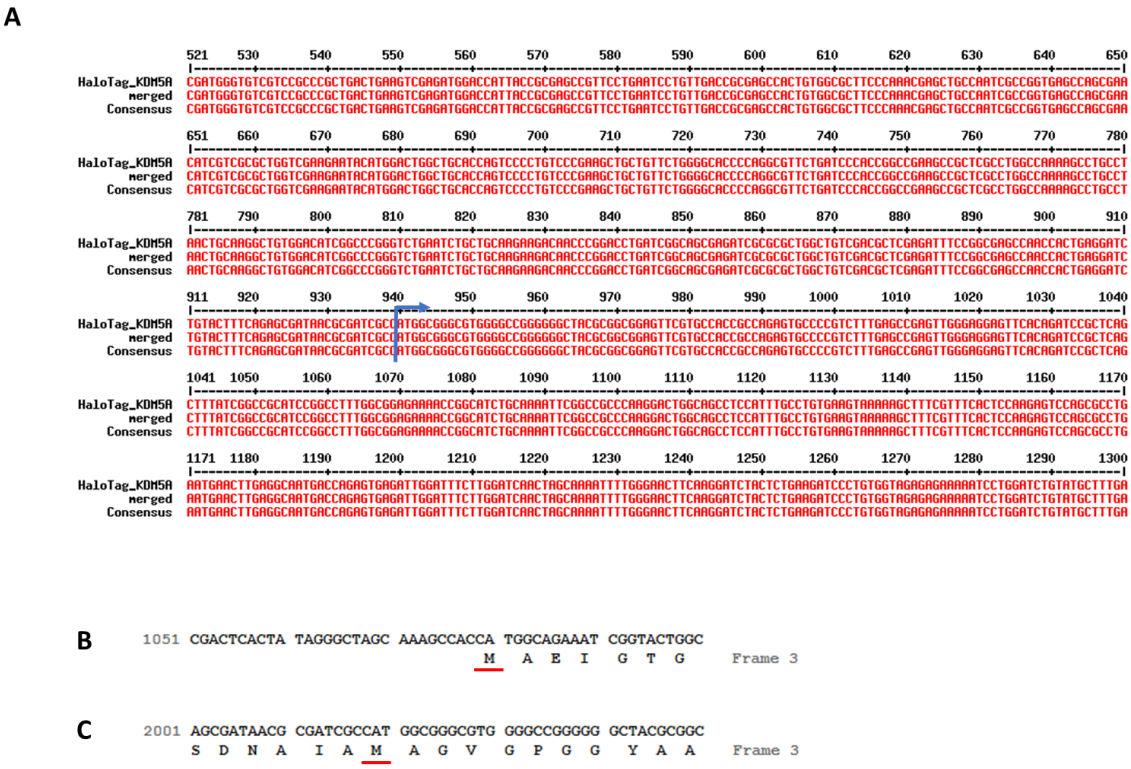


Figure 7-4: KDM5A Is In-Frame with HaloTag

(A) Sequencing files (HaloTag forward and KDM5A reverse) were merged using the Emboss merger tool (http://emboss.bioinformatics.nl/cgi-bin/emboss/merger) and aligned using Multalin. The files sequenced correctly (the blue arrow indicates the start of KDM5A. (B) The KDM5A gene is in-frame with the HaloTag (C), as shown.

7.3.1.2 KDM5A Is Rapidly Turned Over

Expression of the 242 kDa HaloTag-KDM5A in HeLa cells required optimisation in order to achieve a sharp band on a western blot. Firstly, the specificity of the anti-HaloTag antibody was confirmed using the pHT2 vector containing a HaloTag (Figure 7-5A).

The final conditions required to successfully blot for KDM5A-HaloTag were a double transfection on day 1 and 2 post-seeding, as well as an 18 hour MG132 treatment to a final concentration of 5 nM to inhibit proteasomal degradation. Cells were then lysed in RIPA buffer to extract all nuclear protein and samples were run on a 6% SDS-PAGE gel. A wet transfer was carried out to increase transfer efficiency of large proteins from gel to membrane (Figure 7-5B-E).

Immunoblotting for KDM5A-HaloTag prior to MG132 treatment resulted in a consistent smear above the 250 kDa KDM5A band. This is commonly observed when a protein is ubiquitinated prior to degradation by the proteasome. Two successive steps are involved in the Ubiquitin Proteasome Pathway (UPP); firstly, conjugation of ubiquitin to ε-lysines of a target protein takes place through several ATP-dependent enzymatic steps utilising E1, E2 and E3. The protein is then targeted to the proteasome for degradation. MG132 inhibits the proteasome, reducing the degradation of ubiquitinated proteins (Figure 7-6). Numerous potential ubiquitination sites are present along the KDM5A protein, suggesting KDM5A is rapidly turned over (Figure 7-7). An 18 hour treatment of MG132 prior to cell harvest stabilised KDM5A and eliminated the smear observed on the immunoblots (Figure 7-5).

Since MG132 treatment was required to efficiently assay KDM5A, MG132-mediated HeLa cytoxicity was measured by quantifying the amount of extracellular LDH as a measure of plasma membrane integrity (see section 2.20). The LDH assay is a simple colorimetric method using a indicator dye, which does not damage living cells and can be carried out directly on the cell culture medium. There is however high inherent LDH activity in animal sera in cell culture medium, which is corrected for in the toxicity calculations. Alternative assays used to detect dead cells include protease release or DNA staining which can only occur when cell membrane integrity is reduced. On the other hand, cytotoxicity can be measured by quantifying cell viability by using indicator dyes relying on NADH in metabolically active cells. An ATPase assay is another method of detecting cell viability, whereby ATP is measured using a luciferase assay, which has the highest sensitivity of the assays but requires lysed cells, preventing multiplexing experiments.

The LDH assay initially required calculation of the optimum number of cells required for the assay, ensuring that the LDH signal was in the linear range (Figure 7-8A). The optimum cell number was between 5000-10000 cells, therefore 7000 HeLa cells were seeded for each experimental assay. When compared to untreated transfected HeLa cells, MG132 treatment resulted in a small increase in cytotoxicity as expected (Figure 7-8B).

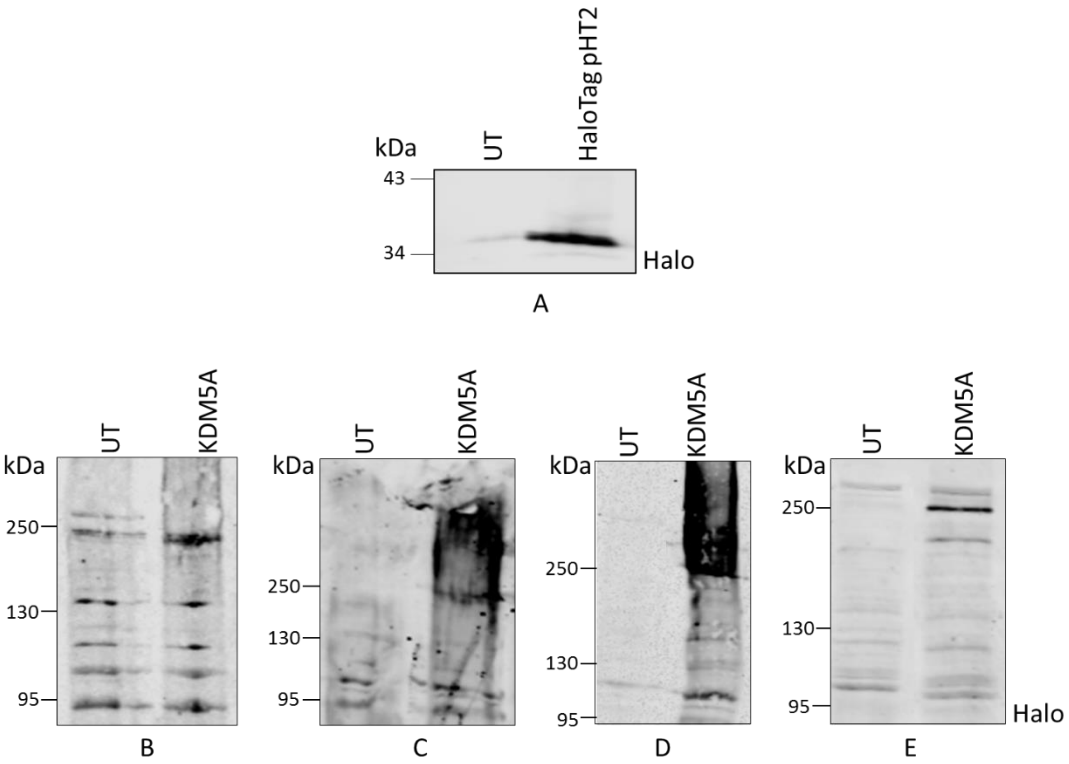


Figure 7-5: Conditions Required to Obtain KDM5A Western Blot

(A) The anti-HaloTag pAb (Promega) was able to detect the 33 kDa HaloTag protein, confirming antibody specificity, where UT = untransfected. (B) Cells were lysed using a gentle lysis buffer - M-PER (mammalian protein extraction reagent, Thermo). Whole cell lysate was then run on a 7.5% SDS-PAGE gel. This protocol did not successfully extract enough KDM5A, with a band also in the UT lane and a high amount of background. (C) Cells were lysed using a more stringent lysis buffer – RIPA, before following the protocol of A. This resulted in a messy blot and no sharp KDM5A band. (D) Cells were transfected two times, rather than the usual single transfection. Cells were then lysed using RIPA buffer and the whole cell lysate was run on a 6% SDS-PAGE gel. This protocol resulted in a more obvious, but smeared band around 250 kDa. (E) Protocol C was followed, with the addition of an 18 hour 5 nM MG132 treatment, resulting in a sharp KDM5A-Halo band at 250 kDa.

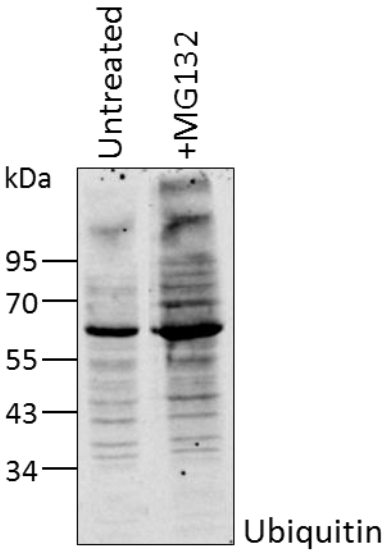
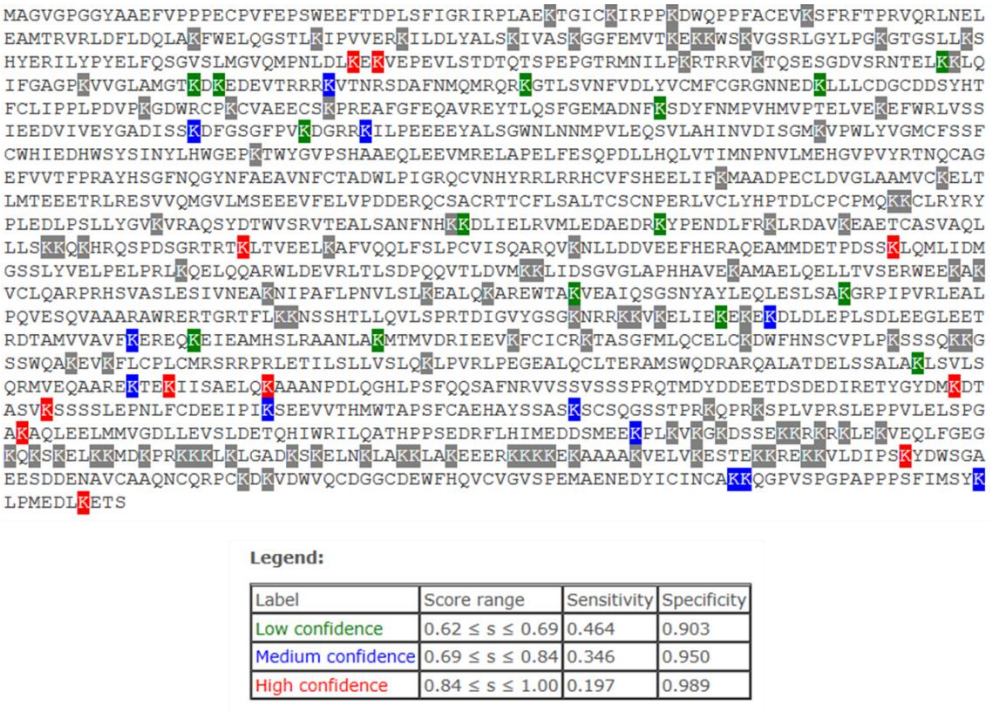


Figure 7-6: MG132 Treatment Increased Ubiquitinated Proteins

An 18 hour 5 nM MG132 treatment results in an increase in levels of ubiquitinated cellular proteins.



Legend:

Label	Score range	Sensitivity	Specificity
Low confidence	0.62 ≤ s ≤ 0.69	0.464	0.903
Medium confidence	0.69 ≤ s ≤ 0.84	0.346	0.950
High confidence	0.84 ≤ s ≤ 1.00	0.197	0.989

Figure 7-7: KDM5A Potential Ubiquitination Sites

UbPred prediction software (Radivojac et al., 2010) was able to identify numerous potential lysine ubiquitination sites with a high confidence throughout the KDM5A protein.

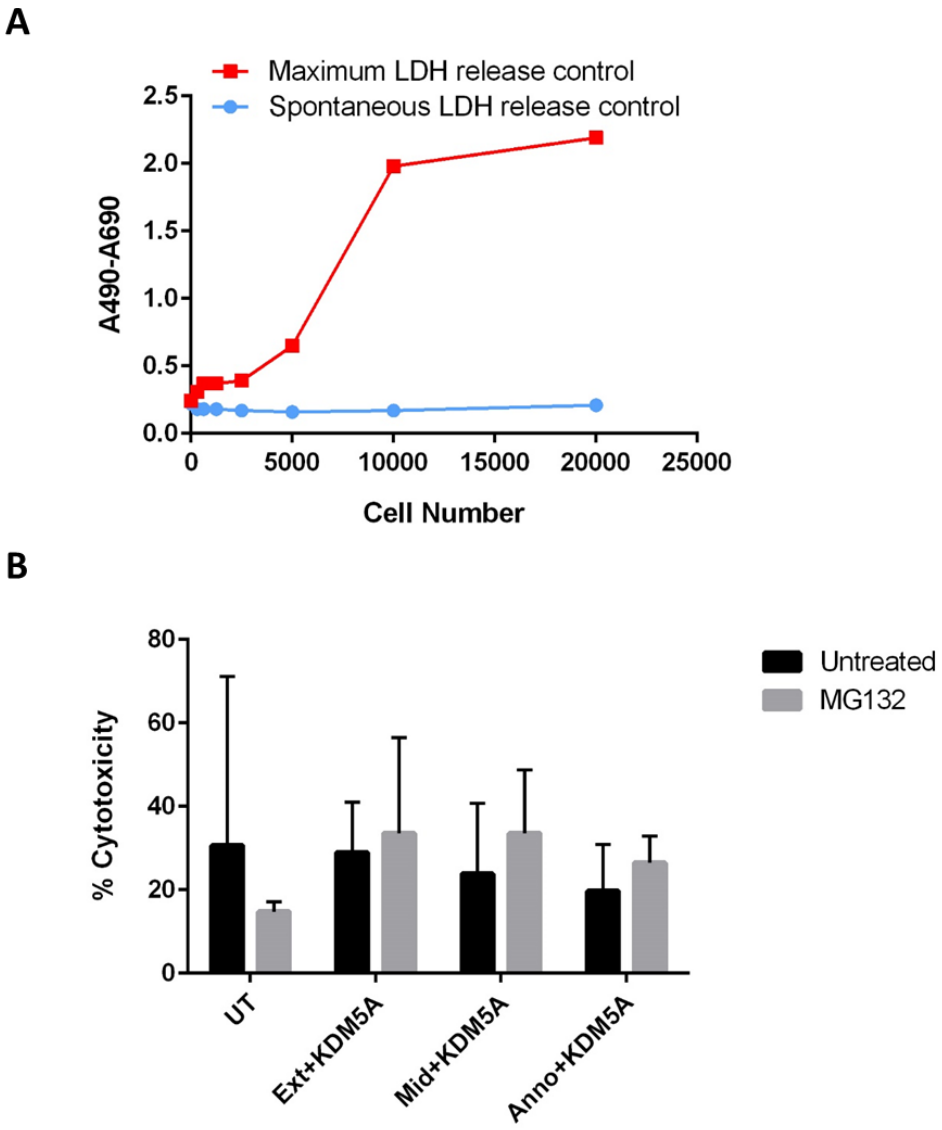


Figure 7-8: Determination of LDH Cytotoxicity of MG132 in HeLa cells

(A) Determination of optimum cell number for LDH cytotoxicity assay. The LDH signal must be within the linear range when carrying out the experimental LDH assays. The linear range and optimum cell seeding density is between approximately 5000-10000 cells. (B) HeLa cells (7000 cells per well) were seeded in a 96-well plate in DMEM supplemented with 10% FBS. Cells were transfected 24 hours post-seeding, followed by a second KDM5A transfection and 5 nM MG132 treatment 24 hours later, before incubating at 37ºC, 5% CO_2 for 18 hours, alongside untransfected cells. LDH cytotoxicity was measured using the Pierce LDH Cytotoxicity Assay Kit. Representative quantification from two independent experiments. Error bars indicate the standard deviation (n=2).

7.3.2 GATAD1/KDM5A Co-IP

Co-IP assays were carried out in order to determine whether the N-terminally extended GATAD1 isoforms were also complexing with KDM5A, away from their functional chromatin target. Initially, the Pierce Co-IP Kit was used to immobilise FLAG antibody onto the supplied agarose support. The co-IP was then carried out, using 500 µg of pre-cleared lysate, transfected with each GATAD1 isoform as well as KDM5A following the protocol described in section 7.3.1.2. Elution of the co-IP using the elution buffer provided was not efficient with very little GATAD1-FLAG detected when run on a Western blot. In a separate experiment, 20 µg FLAG peptide was used to compete off the complexed proteins and the same resin was subsequently boiled in 5x sample buffer for 5 minutes, to determine the efficiency of FLAG peptide elution. By comparing the intensity of the FLAG (bait) bands in each type of elution, it was calculated that 68% of Ext, 69% of Mid and 83% of Annotated GATAD1 remained bound to the resin following FLAG peptide elution and was eluted only via boiling of the resin in sample buffer, which results in elution of contaminating antibody fragments (Figure 7-9).

Control assays were carried out to ensure that both FLAG and Halo antibodies were specific to their respective tags and to rule out non-specific interactions with the resin matrix (Figure 7-10). Quenched antibody coupling resin without conjugated antibody did not pull down any FLAG or Halo-tagged protein. However, a control IP with FLAG lysates on Halo-conjugated resin resulted in a FLAG signal in the eluate totalling 34% of the input signal. Similarly, the reciprocal control IP with Halo lysates on FLAG-conjugated resin resulted in a Halo signal in the eluate totalling 18% of the input signal. Non-specific interactions are occurring on both antibodies which makes quantification difficult, however GATAD1-KDM5A co-IP experiments were still carried out as the non-specific binding accounted for only a small proportion of overall binding seen initially in Figure 7-9.

Immobilising all three FLAG-GATAD1 isoforms on Pierce Antibody Coupling Resin resulted in co-IP of Halo-KDM5A. Both Ext and annotated GATAD1 precipitated more than three times the amount of KDM5A than the Mid isoform (Figure 7-11). The reciprocal experiment whereby KDM5A was immobilised, resulted in co-IP of all three GATAD1 isoforms, again with less Mid than Ext and Annotated (Figure 7-12). The co-IP results suggest that all three GATAD1 isoforms complex with KDM5A, with Mid forming less efficiently than the other two isoforms.

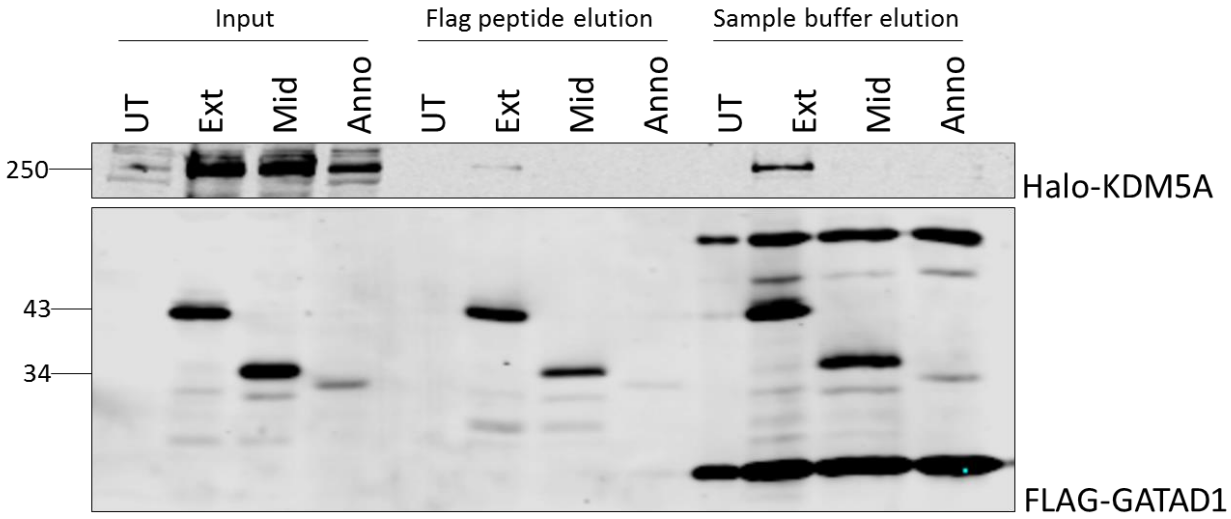


Figure 7-9: Elution of Proteins from Resin

Boiling the resin in 5x sample buffer for 5 minutes was the most effective way of eluting both bait (GATAD1) and prey (KDM5A) proteins, where UT = untransfected.

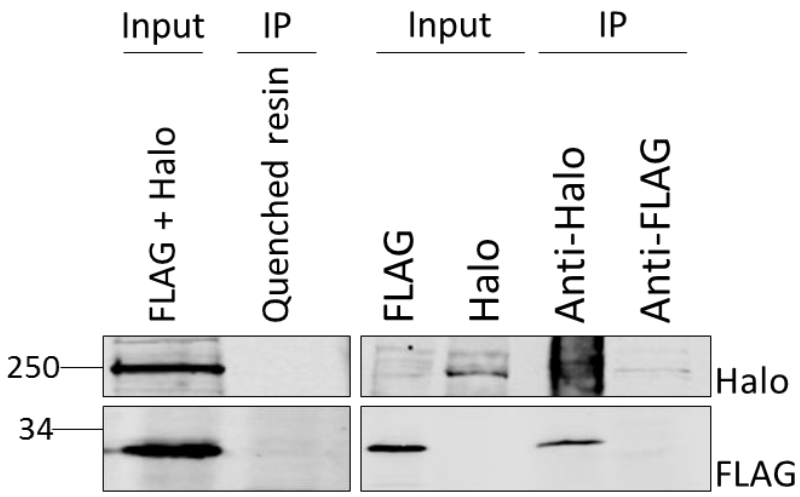


Figure 7-10: Control IP Assay

Quenched antibody coupling resin control IP involved adding 200 µL of Quenching Buffer to the Antibody Coupling Resin instead of FLAG/Halo antibody. 500 µg annotated GATAD1 and KDM5A-transfected lysate was used in the IP, following the standard Pierce protocol. The non-specific antibody control IP involved adding 500 µg Halo-transfected lysate to FLAG-immobilised resin and FLAG-transfected lysate to Halo-immobilised resin.

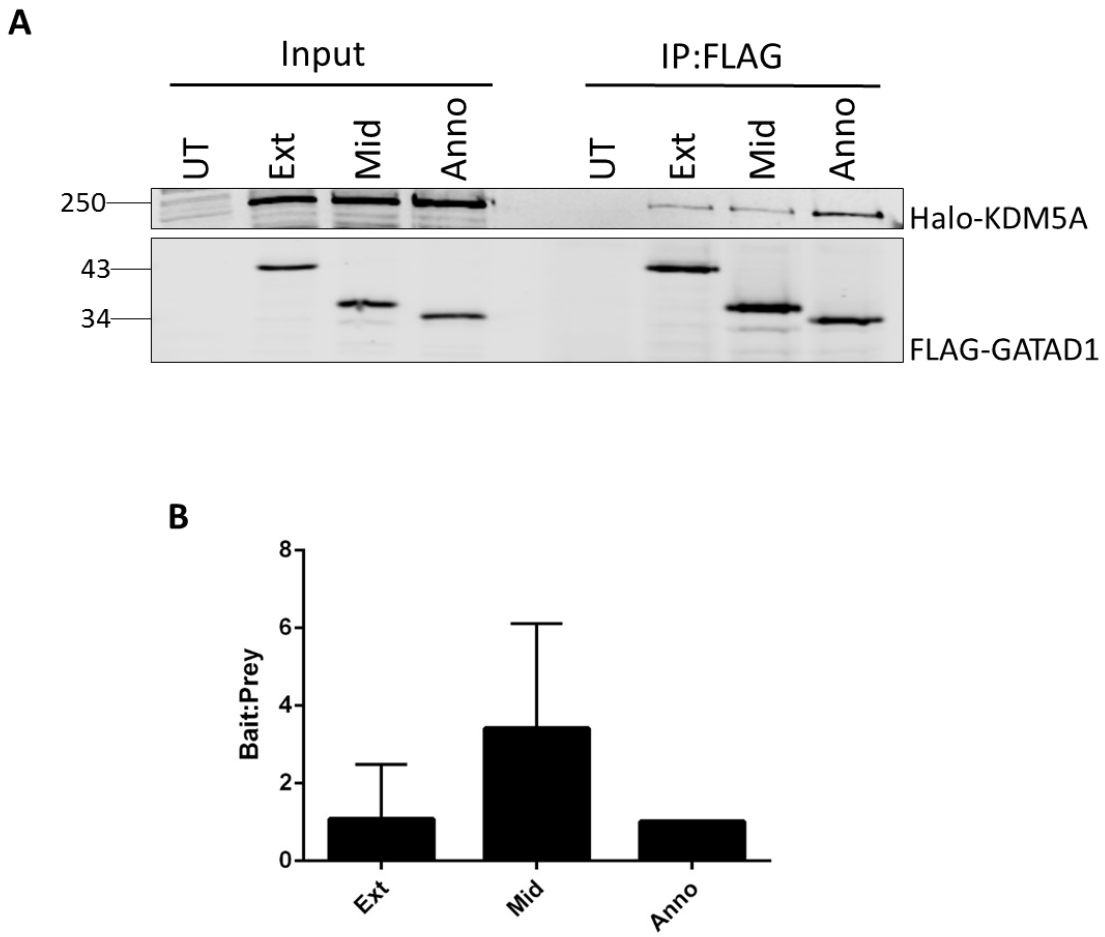


Figure 7-11: FLAG-GATAD1/Halo-KDM5A Co-IPs

(A) Immobilising the bait protein (FLAG-GATAD1) on Pierce Antibody Coupling Resin resulted in co-immunoprecipitation of Halo-KDM5A (prey protein) for all three GATAD1 isoforms and no co-immunoprecipitation of the untransfected control (UT). (B) Representative quantification from three independent experiments. Error bars indicate the standard deviation (n=3). The higher the bait:prey ratio, the less KDM5A is complexed with the GATAD1 isoform. This was calculated by dividing the bait:prey ratio in the IP by the bait:prey ratio in the input and normalising to Annotated GATAD1.

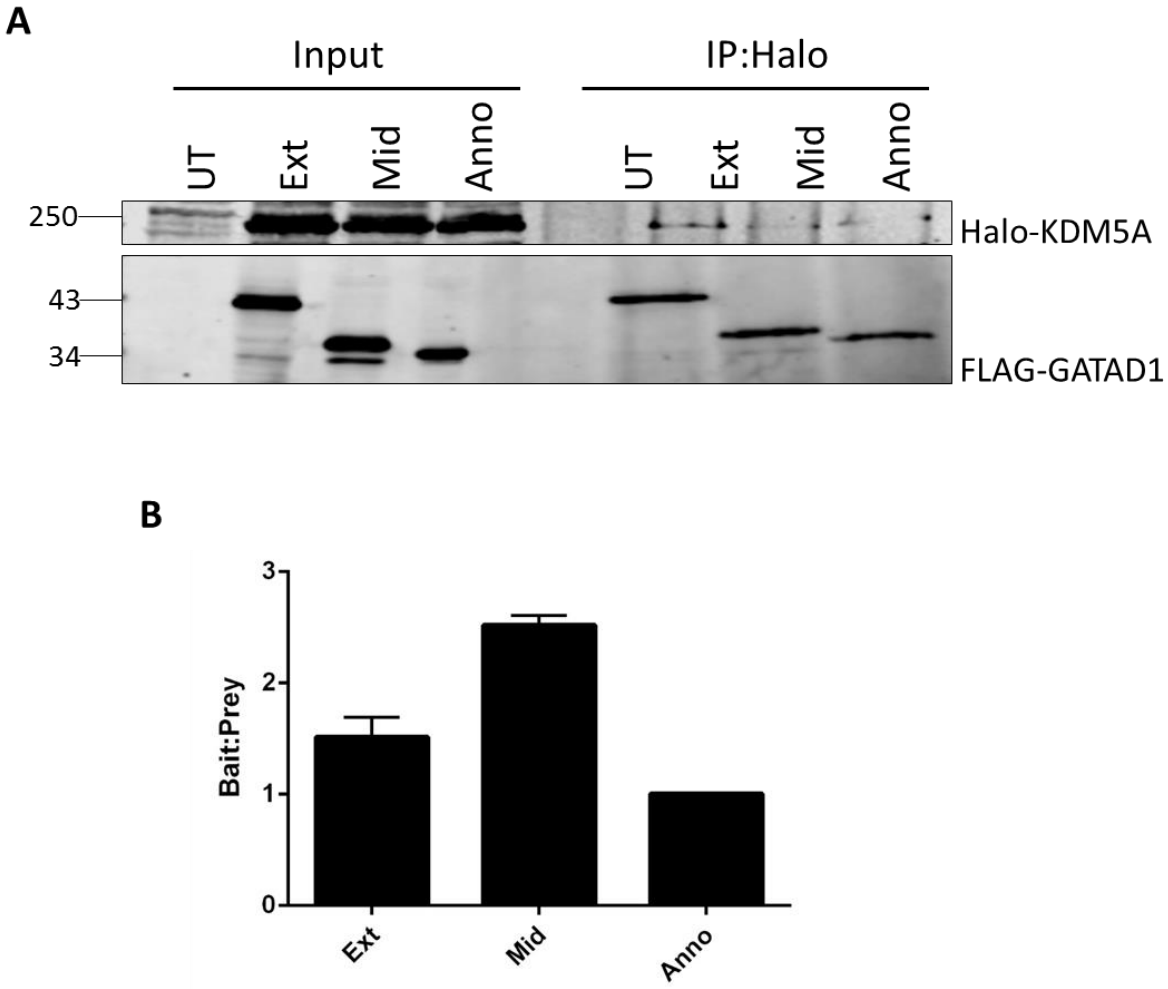


Figure 7-12: Halo-KDM5A/FLAG-GATAD1 Co-IPs

(A) Immobilising the bait protein (Halo-KDM5A) on Pierce Antibody Coupling Resin resulted in co-immunoprecipitation of all three FLAG-GATAD1 isoforms (prey protein) and no co-immunoprecipitation of the untransfected control (UT). (B) Representative quantification from two independent experiments. Error bars indicate the standard deviation (n=2). The higher the bait:prey ratio, the less GATAD1 isoform is complexed with KDM5A. This was calculated by dividing the bait:prey ratio in the IP by the bait:prey ratio in the input and normalising to Annotated GATAD1.

7.3.3 NanoBiT Protein:Protein Interaction System

The NanoBiT (NanoLuc Binary Technology) system (Promega), was used in order to confirm the specificity of the GATAD1-KDM5A interaction in live cells, since it was difficult to quantify the co-IP experiments due to poor antibody specificity. The Large BiT (LgBiT) and Small BiT (SmBiT) nanoluciferase subunits were fused to the proteins of interest at either the N or C-terminus. When expressed, successful PPI would result in the structural complementation of the two NanoLuc subunits, forming a functional enzyme that generates a luminescent signal (Figure 7-13). In order to determine the optimal orientation for the LgBiT:SmBiT interaction, GATAD1 was fused at both the N and C-terminus of SmBiT using NheI and XhoI sites, whilst KDM5A was fused in three fragments to the LgBiT C-terminus using both NheI and XhoI/SalI sites (resulting in the loss of the XhoI restriction site) (Table 7-1), since the full length 204 kDa KDM5A protein was unable to express efficiently. The boundaries of the three KDM5A fragments were between functional domains which also enabled an estimation of the position of the KDM5A-GATAD1 interaction site when the NanoBiT assay was carried out (Figure 7-14). Transfections were carried out in HeLa cells using combinations of LgBiT:SmBiT GATAD1/KDM5A pairs in order to determine the optimum orientation for PPI and further experiments (Table 7-2).

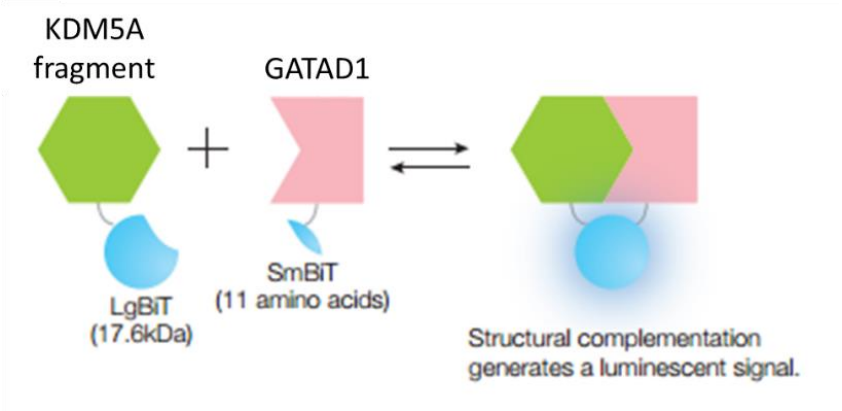


Figure 7-13: NanoBiT Protein-Protein Interaction System

Image from Promega technical manual – NanoBiT PPI System. Successful GATAD1/KDM5A PPI would result in a luminescent signal above the negative controls.

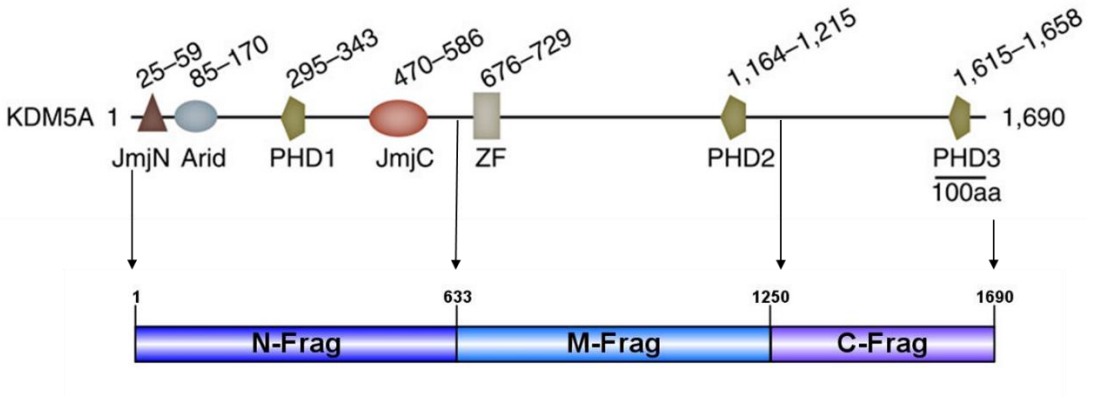


Figure 7-14: KDM5A Fragments and Domains

Domain map of KDM5A (Torres et al., 2015) and the three fragments of KDM5A used throughout the NanoBiT experiments. The JmjN and JmjC domains are found within the jumonji family of transcription factors and are thought to form a functional unit within the folded protein.

Table 7-1: Expression constructs made for GATAD1-KDM5A PPI

Each GATAD1 isoform was fused to SmBiT at both the N and C-termini. The KDM5A fragments were fused to LgBiT at the C-terminus only.

Ext GATA	Ext-SmBiT	**Mid GATA**	Mid-SmBiT	**Anno GATA**	Anno-SmBiT
	SmBiT-Ext		SmBiT-Mid		SmBiT-Anno

N-Frag	LgBiT-Nfrag	**M-Frag**	LgBiT-Mfrag	**C-Frag**	LgBiT-Cfrag

Table 7-2: Combinations of LgBiT/SmBiT Fusions Screened to Detect PPI

Transfections were carried out in HeLa cells using combinations of LgBiT:SmBiT GATAD1/KDM5A pairs in order to determine the optimum orientation for PPI and further experiments.

N-frag	Ext-SmBiT:LgBiT-Nfrag	SmBiT-Ext:LgBiT-Nfrag
	Mid-SmBiT:LgBiT-Nfrag	SmBiT-Mid:LgBiT-Nfrag
	Anno-SmBiT:LgBiT-Nfrag	SmBiT-Anno:LgBiT-Nfrag
M-frag	Ext-SmBiT:LgBiT-Mfrag	SmBiT-Ext:LgBiT-Mfrag
	Mid-SmBiT:LgBiT-Mfrag	SmBiT-Mid:LgBiT-Mfrag
	Anno-SmBiT:LgBiT-Mfrag	SmBiT-Anno:LgBiT-Mfrag
C-frag	Ext-SmBiT:LgBiT-Cfrag	SmBiT-Ext:LgBiT-Cfrag
	Mid-SmBiT:LgBiT-Cfrag	SmBiT-Mid:LgBiT-Cfrag
	Anno-SmBiT:LgBiT-Cfrag	SmBiT-Anno:LgBiT-Cfrag

The expression of the NanoBiT clones was initially verified in several ways, before carrying out NanoBiT assays. Initially, a Western blot was carried out using an antibody raised against the full length Nanoluciferase protein, which was therefore also capable of binding the LgBiT NanoLuc subunit. This antibody picked up several other non-specific bands, in both untreated and MG132-treated conditions, (Figure 7-15). In both cases, the LgBiT positive control expressed well at around 70 kDa. The LgBiT-KDM5A N, M and C-fragments were only present in the presence of MG132 and were expected to run at 90, 88 and 67 kDa, respectively. NanoBiT KDM5A still has a high turnover rate when expressed in smaller fragments, therefore the MG132 treatment was carried out for all NanoBiT assays.

In order to confirm the KDM5A expression observed by Western blotting and to verify the SmBiT-GATAD1-(Ext, Mid, Anno) isoform expression, the LgBiT/SmBiT nanoluciferase subunit present in the NanoBiT clones were swapped for full length nanoluciferase, in order to carry out a dual luciferase assay (Figure 7-16). As well as the nanoluciferase-containing plasmids, HeLa cells were also transfected with a firefly-containing plasmid as a transfection control. The Nano-Glo Dual Luciferase Reporter System (Promega) was used to measure the expression efficiency of both GATAD1 and KDM5A NanoBiT clones. All of the clones expressed (Figure 7-17), however the efficiency of each varied, which may have implications on the quantification of NanoBiT Assay results. The least efficiently expressed NanoBiT clone was 1.1C-KDM5A-Nfrag, which was still almost 5000x above background signal.

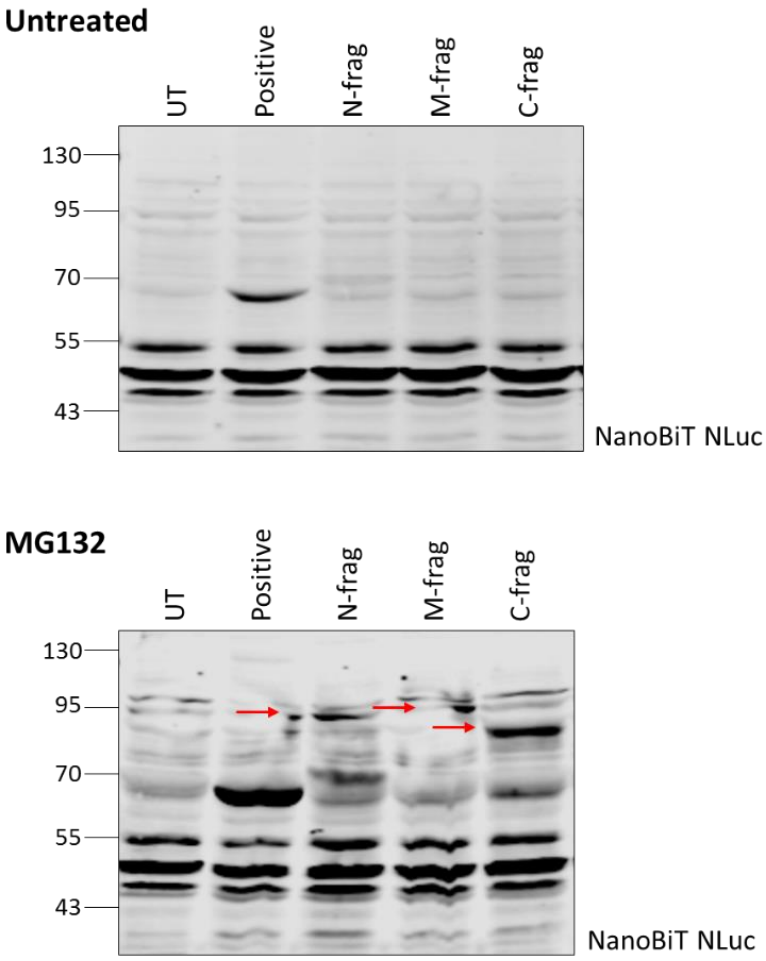


Figure 7-15: Verifying Expression of LgBiT-KDM5A NanoBiT Clones

LgBiT-KDM5A N, M and C fragments were expressed in HeLa cells which were untreated, or treated with 5 nM MG132 for 18 hours, alongside untransfected cells and the LgBiT positive NanoBiT control. Numerous non-specific bands are visible when blotting with the NanoBiT NLuc antibody, however KDM5A expression is observed with MG132 treatment, indicated by the red arrows.

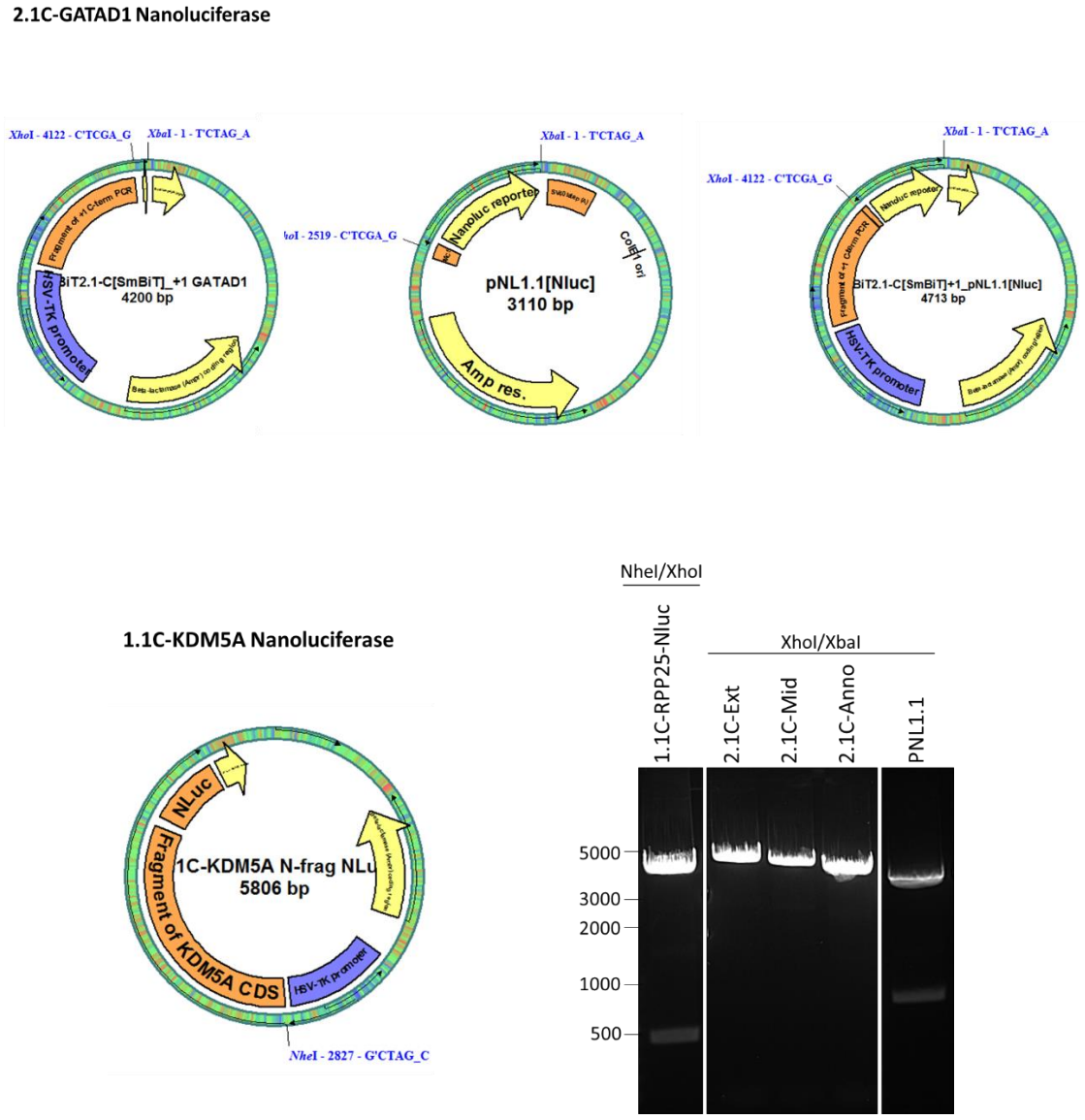


Figure 7-16: Exchange of SmBiT/LgBiT For Full Length Nanoluciferase

An XhoI/XbaI double digest of the C-terminal GATAD1 clones dropped out SmBiT, which was replaced with nanoluciferase from XhoI/XbaI-digested PNL1.1 vector (lower band). However, swapping LgBiT for nanoluciferase in the C-terminal KDM5A NanoBiT clones was not possible by a straight swap, since the XhoI site was lost when KDM5A was initially cloned with SalI. Instead, a C-terminal NanoLuc plasmid containing RPP25 gene and full length nanoluciferase (from Jo Cowan) was utilised. A NheI/XhoI double digest dropped out RPP25 (utilised upper band– 1.1C nanoluciferase vector), which was replaced with NheI/SalI-digested KDM5A fragment (N, M and C) PCRs.

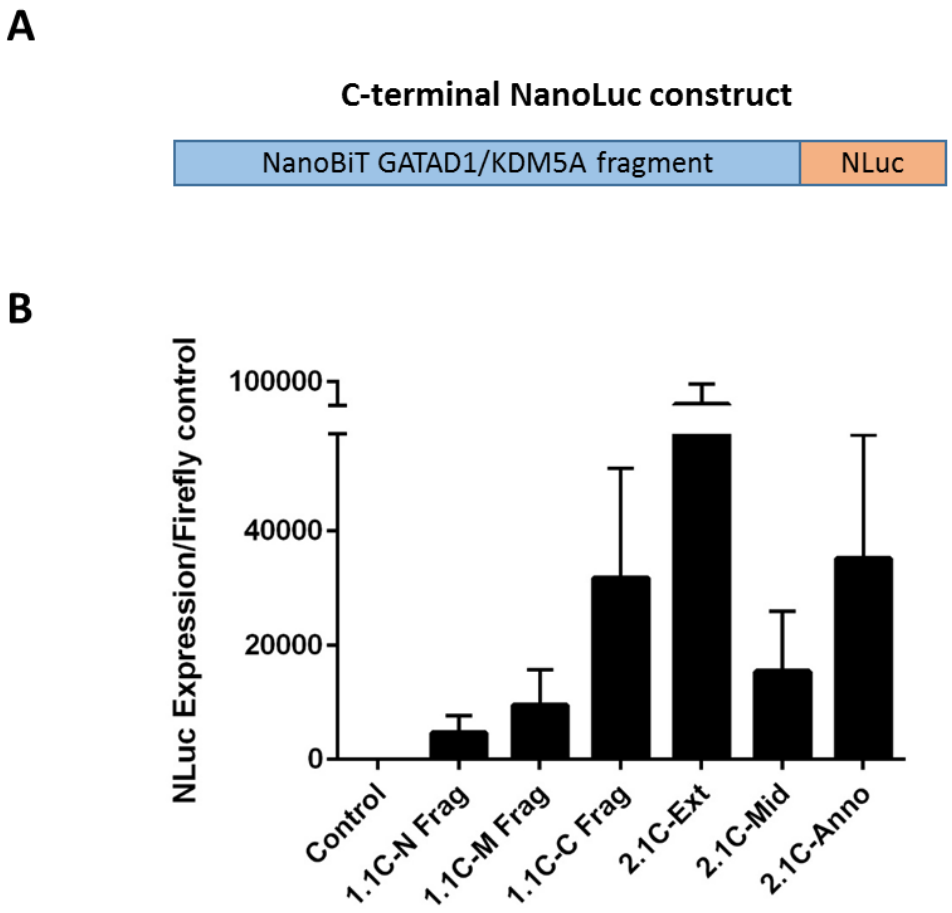


Figure 7-17: Confirming Expression of NanoBiT Clones Using Nanoluciferase

(A) Nanoluciferase was cloned at the C-terminus of the 1.1C KDM5A fragments and the 2.1C GATAD1 fragments, replacing the NanoBiT luciferase subunit, allowing analysis of expression by luciferase assay. (B) All clones expressed nanoluciferase, although with varying efficiencies. The KDM5A fragments are almost twice the size of the GATAD1 isoforms and therefore would expect to express less efficiently. Representative quantification from three independent experiments. Error bars indicate the standard deviation (n=3).

Spontaneous association of LgBiT:SmBiT within the cell may lead to false positives when analysing the NanoBiT interaction data. Therefore, a stable, ubiquitously expressed HaloTag (HT) negative control construct (SmBiT:HaloTag) was used. The HT control was co-transfected with each LgBiT KDM5A construct in order to quantify the signal from a spontaneously-associating non-interacting protein pair. Interaction data from each NanoBiT experiment was normalised to the HT control, which should avoid misinterpretation of a false GATAD1-KDM5A PPI. A positive control pair (PRKACA:PRKAR2A) were also used in each NanoBiT experiment, which have been optimised to strongly associate and express efficiently (Figure 7-15). The positive control luminescent signals obtained were extremely high (2-7 $x10^5$ RLU).

Both N and C-terminal GATAD1 fusion pairs were cloned and tested, in order to determine the optimal orientation for interaction of SmBiT and LgBiT upon GATAD1-KDM5A PPI. GATAD1 isoforms fused to the N-terminus of SmBiT (2.1N) gave the maximal fold signal increase over the respective fusion with SmBiT-HaloTag negative control (Figure 7-18). This suggests the optimal interaction and the 2.1N-GATAD1 constructs were therefore used in further NanoBiT assay repeats.

The GATAD1-KDM5A NanoBiT assay results appear to correlate with the co-IP data, where Mid GATAD1 did not interact as efficiently with KDM5A as Ext and Anno. The signal obtained from Mid interacting with all three KDM5A fragments was barely above the negative HT control, whereas both Ext and Anno gave a signal approximately 5x above the HaloTag control when interacting with both the N and C-terminus of KDM5A. As well as increasing the expression efficiency, using N, M and C-fragments of KDM5A also gives an indication of where the GATAD1 binding site(s) may be. Very little signal was obtained when co-expressing each GATAD1 isoform with KDM5A M-fragment, suggesting the middle section of KDM5A is not involved in GATAD1 binding. On the other hand, both the N and C-terminal regions of KDM5A appear to be involved in GATAD1 binding (Figure 7-19).

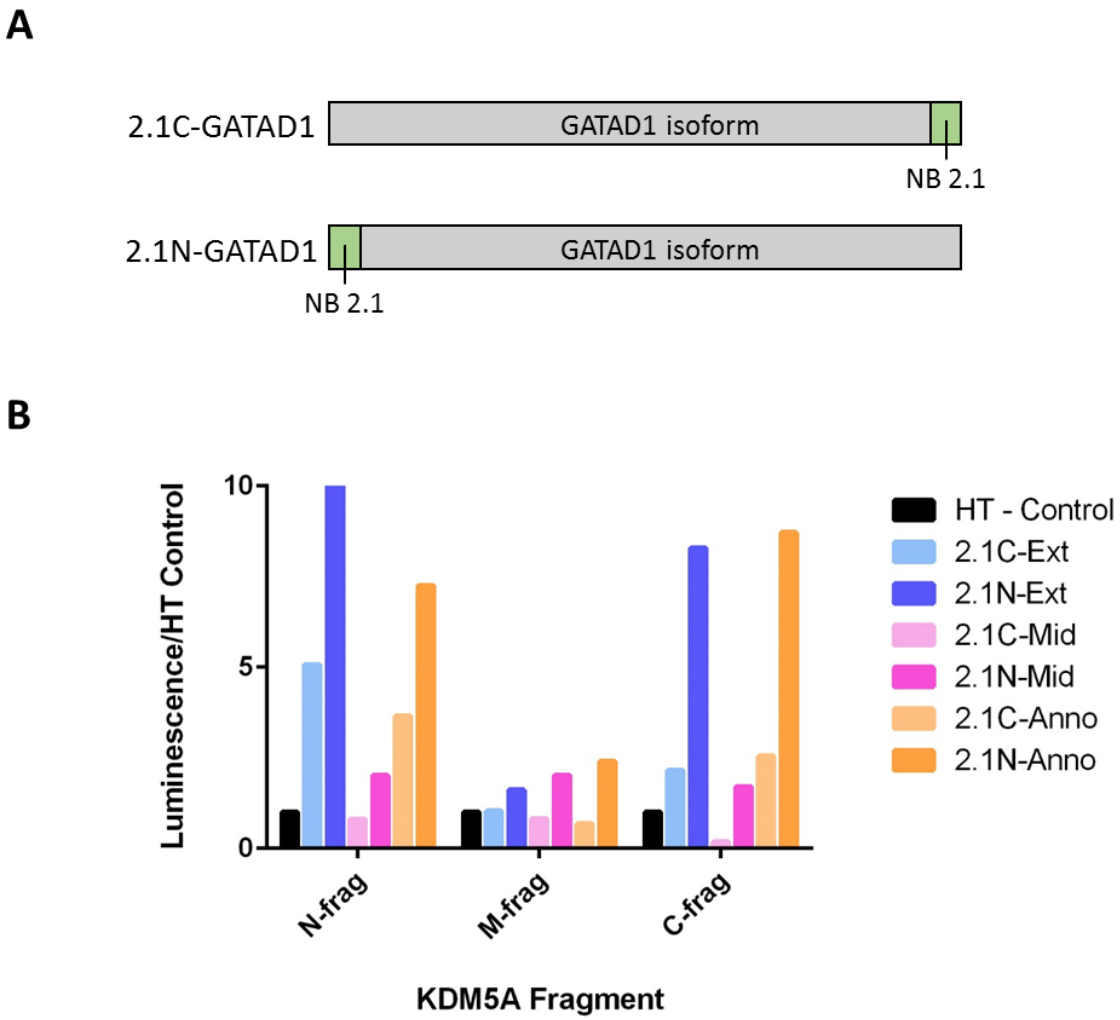


Figure 7-18: N-terminal GATAD1 SmBiT Has Optimal Interaction with KDM5A LgBiT

(A) The GATAD1 isoforms had the 2.1 NanoBiT luciferase subunit (Small BiT) cloned at both the N-terminus (dark shaded bars) and C-terminus (light shaded bars), in order to determine which orientation had optimal interaction with the KDM5A 1.1 (Large BiT) subunit. (B) An initial NanoBiT assay showed that the 2.1N GATAD1 constructs gave a higher signal above HT background compared to the 2.1C constructs.

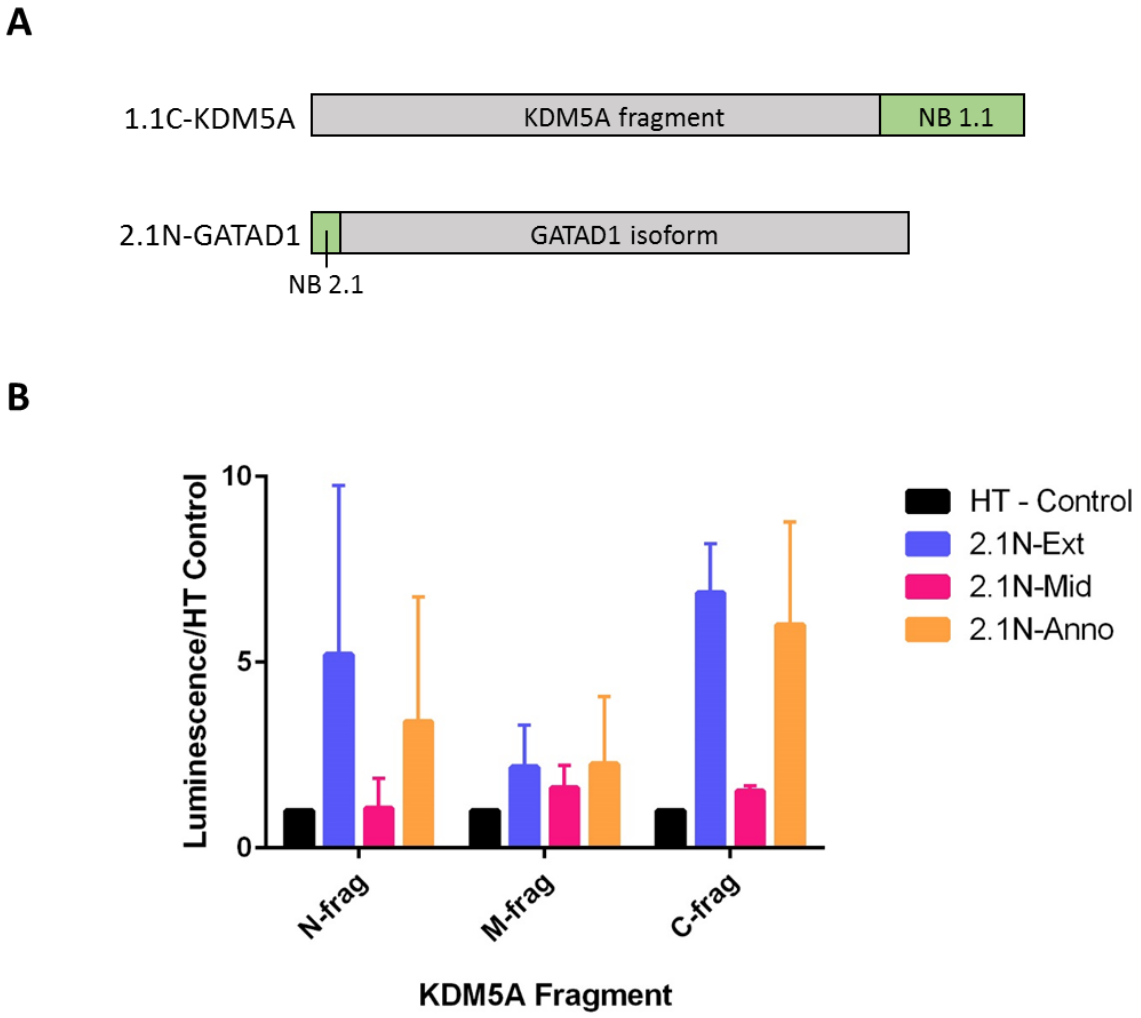


Figure 7-19: GATAD1-KDM5A NanoBiT Assay

(A) 1.1C KDM5A and 2.1N GATAD1 isoforms were used for the NanoBiT assays. (B) HeLa cells were transfected with combinations of NanoBiT constructs and treated with 5 nM MG132 for 18 hours. The NanoBiT assay was carried out 48 hours post-transfection according to section 2.21. Representative quantification from three independent experiments. Error bars indicate the standard deviation (n=3).

7.3.3.1 GATAD1 Does Not Self-Associate

GATA-1 modulates transcription of genes expressed in erythroid cells and is the founding member of the family of transcription factors containing highly conserved GATA-type zinc finger (ZnF) domains, which specifically bind the DNA sequence 5'-(A/T)GATA(A/G)-3'. GATA-1 has been shown to both self-associate and form dimers with other GATA-transcription factors in whole-cell extracts, an interaction which is mediated by its GATA ZnFs and is able to modulate transcription (Crossley et al., 1995). Since the GATA ZnFs are implicated in dimerization and PPIs as well as DNA-binding, we investigated whether GATAD1, which contains a single copy of the GATA ZnF domain (Figure 7-20), was also able to dimerise.

A NanoBiT assay was carried out by simultaneously transfecting both SmBiT and LgBiT of each GATAD1 isoform, in order to determine whether dimerization was taking place. The luminescence for all three GATAD1 isoforms was less than twice that of the LgBiT GATAD1 fusion co-expressed with the HaloTag negative control, indicating that no specific GATAD1 self-association is taking place (Figure 7-21).

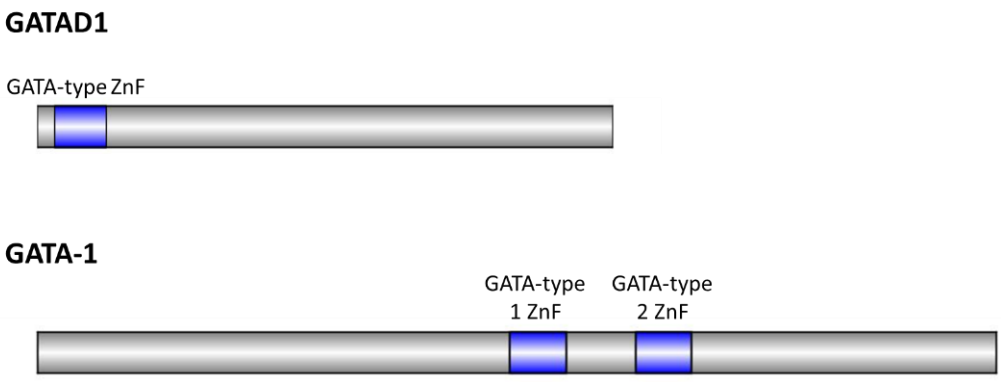


Figure 7-20: GATAD1 Contains a Single GATA-type ZnF

The 413 amino acid protein GATA-1 (UniProtKB P15976) contains two 25 amino acid GATA zinc fingers (positions 204-228 and 258-282). GATAD1- GATA zinc finger domain-containing protein 1 (UniProtKB Q8WUU5) is 269 amino acids and contains one single copy of the GATA zinc finger domain at its N-terminus (position 9-33).

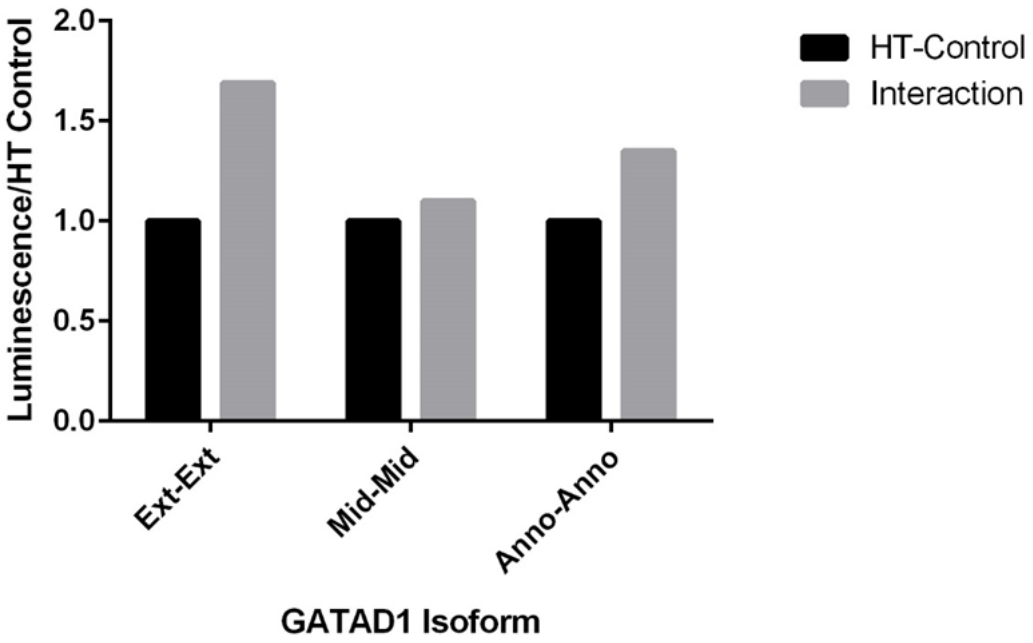


Figure 7-21: GATAD1 Does Not Self-Dimerise

HeLa cells were transfected with N-terminal LgBiT and SmBiT of each GATAD1 isoform, as well as HaloTag negative control SmBiT constructs with GATAD1 LgBiT. Cells were treated with 5nM MG132 for 18 hours. Luminescence readings have been normalised to the HT-control read.

7.3.4 GATAD1/KDM5A Co-localisation

Both KDM5A and N-terminally extended GATAD1 localise to the cytoplasm and the nucleus, where their function away from the standard histone target remains unknown. However, co-IPs have shown that all three GATAD1 isoforms complex with KDM5A (although Mid complexes less efficiently), suggesting these proteins form a cytoplasmic complex. Immunofluorescence of FLAG-GATAD1 isoforms and Halo-KDM5A confirm that they co-localise in both the nucleus and the cytoplasm (Figure 7-22).

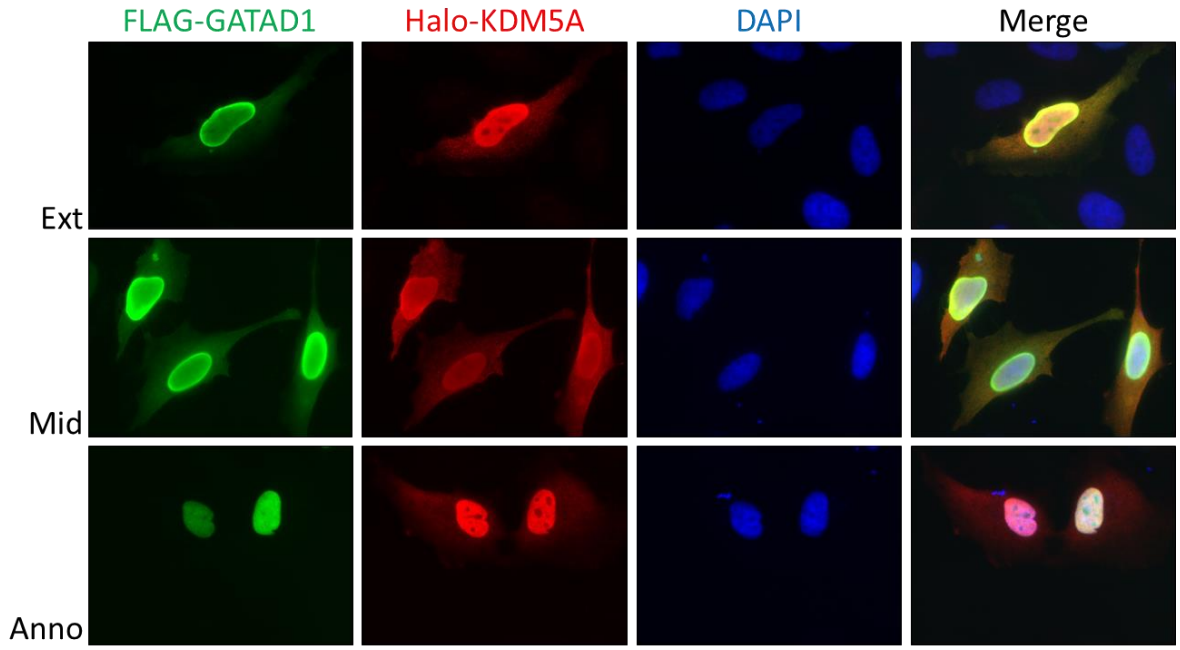


Figure 7-22: GATAD1 and KDM5A Co-Localise in HeLa Cells

Co-expression of Ext, Mid and Anno FLAG-GATAD1 isoforms with Halo-KDM5A was carried out, followed by immunofluorescence. 3xFLAG-tagged GATAD1 was probed using Sigma anti-FLAG produced in mouse as the primary antibody, followed by anti-mouse Alexa Fluor 488 as the secondary. Halo-KDM5A was probed using anti-Halo produced in rabbit as the primary antibody, followed by anti-rabbit Alexa Fluor 555 as the secondary, whilst DNA was stained with DAPI.

7.3.5 RNAi as a Tool for Analysis of GATAD1 Function

7.3.5.1 Determination of GATAD1 siRNA Knockdown

Small interfering RNA (siRNA) was used to transiently knockdown GATAD1 mRNA by utilising the RNA interference (RNAi) pathway (Figure 7-23). The psiCHECK-2 Vector (Promega) enabled determination of RNAi efficacy by monitoring target gene expression when fused downstream of a Renilla luciferase reporter. A 417 bp section of GATAD1 3'UTR surrounding the siRNA target site was cloned into the MCS (using XhoI/NotI) downstream of the Renilla luciferase stop codon (Figure 7-24). Initiation of RNAi in response to GATAD1 siRNA treatment resulted in degradation of fusion mRNA and decreased Renilla luciferase signal when normalised to Firefly luciferase transfection control reporter (Figure 7-25).

Western blot analysis of the GATAD1 protein levels following mRNA knockdown was not possible due to the poor antibody signal towards endogenous GATAD1, as well as difficulty in transfecting the FLAG-tagged GATAD1 CRISPR cell line. A psiCHECK assay was carried out to confirm that an RNA containing the siRNA target site would be knocked down, which involved quantifying the cleavage and degradation of a GATAD1 3'UTR (siRNA target)-Renilla reporter fusion, following siRNA treatment. mRNA knockdown of GATAD1 also resulted in a significant decrease in protein expression (Figure 7-26).

RT-qPCR of siRNA-treated HeLa cells enabled quantification of GATAD1 knockdown relative to β2M, which confirmed a 52% knockdown in mRNA levels, relative to a non-specific control siRNA (Figure 7-27A). Subsequent KDM5A mRNA levels decreased by 13% following GATAD1 knockdown (Figure 7-27B).

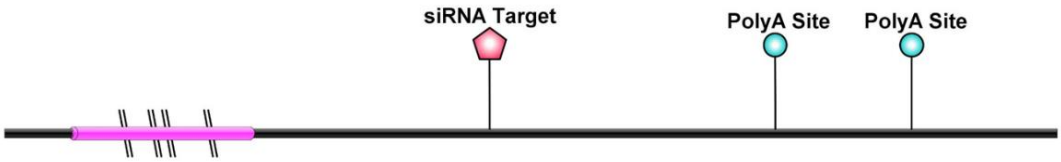


Figure 7-23: siRNA Target Site

A diagram of GATAD1 mRNA showing the siRNA target within the 3'UTR, in relation to the CDS (pink) and exon-exon boundaries (cut sequence).

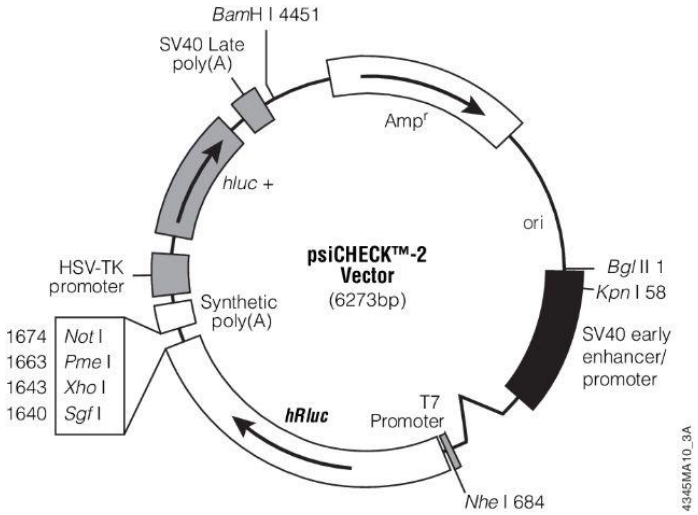


Figure 7-24: psiCHECK-2 Vector

The GATAD1 siRNA target was cloned in to the psiCHECK-2 vector downstream of hRluc.

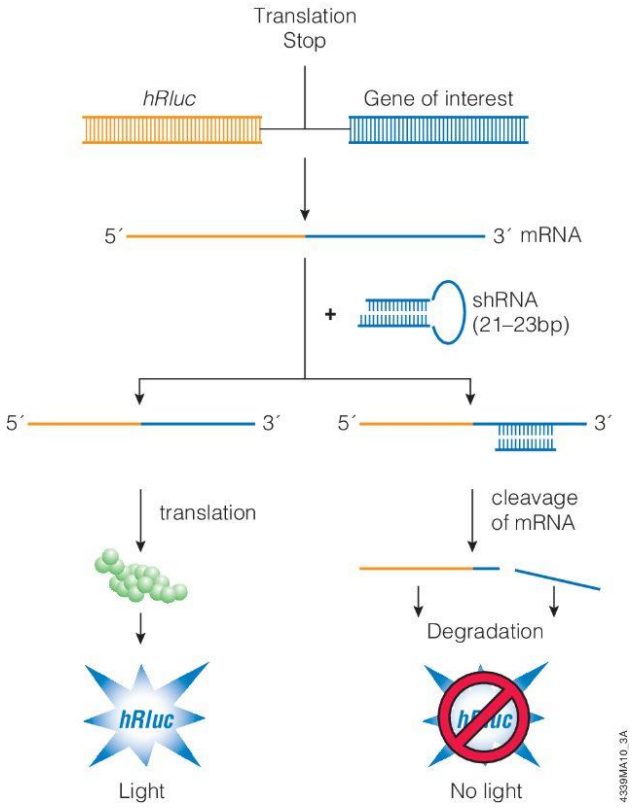


Figure 7-25: Mechanism of action of psiCHECK-2 Vector

siRNA treatment causes activation of RNAi, resulting in cleavage of fusion mRNA and decreased Renilla luciferase signal, when compared to the control siRNA.

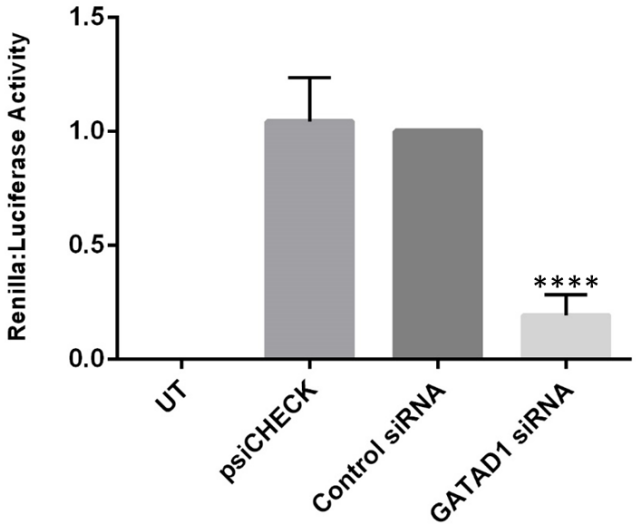


Figure 7-26: GATAD1 siRNA Targets Renilla-GATAD1 psiCHECK construct

Dual luciferase reporter assays were carried out by transfecting HeLa cells with psiCHECK2 vector containing a Renilla luciferase reporter fused to the 3'UTR of GATAD1, where UT = untransfected. Renilla signal was quantified 48 hours post-transfection, using firefly as an internal control. Quantification made using GraphPad Prism (p=<0.0001, n=4, Unpaired T-test). Error bars indicate the standard deviation.

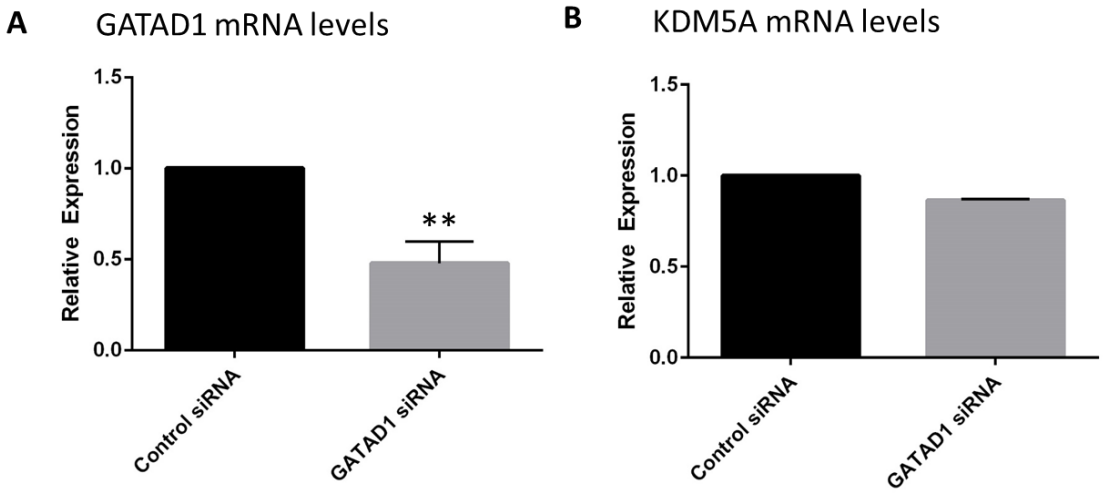


Figure 7-27: Detection of GATAD1 Knockdown by RT-qPCR

(A) HeLa cells were treated with a GATAD1 siRNA duplex, or with a non-silencing control siRNA at 5 nM for 48 hours. The amount of GATAD1 transcript was then analysed by RT-qPCR, relative to β2M control. Quantification made using GraphPad Prism (p=0.0017, n=3, Unpaired T-test). Error bars indicate the standard deviation. (B) Cells were treated with a control or GATAD1 siRNA as before. The amount of KDM5A transcript was then analysed by RT-qPCR, relative to β2M control. Error bars indicate standard deviation (n=2).

7.3.5.2 Effect of GATAD1 Isoforms on H3K4me3 Levels

GATAD1 is thought to form part of a chromatin complex which demethylates H3K4me3 via KDM5A. In order to determine whether the N-terminally extended isoforms were also able to form a functional transcription factor complex, GATAD1 was initially transiently knocked-down using siRNA. Ext, Mid and Anno were then added-back individually and the levels of H3K4me3 were quantified. Knockdown of GATAD1 was expected to result in an increase in H3K4me3, however Western blot quantification showed that there was no change in H3K4me3 (Figure 7-28). Each GATAD1 isoform was then individually added-back (over-expressed) into the system, which also had no significant effect on H3K4me3 levels, although the RNAi alone control should have been included in this experiment to achieve an accurate result (Figure 7-29). It was not necessary to transfect mutated versions of GATAD1 immune to the siRNA because the 3xFLAG-tag reporters used in previous chapters do not contain the 3'UTR siRNA target.

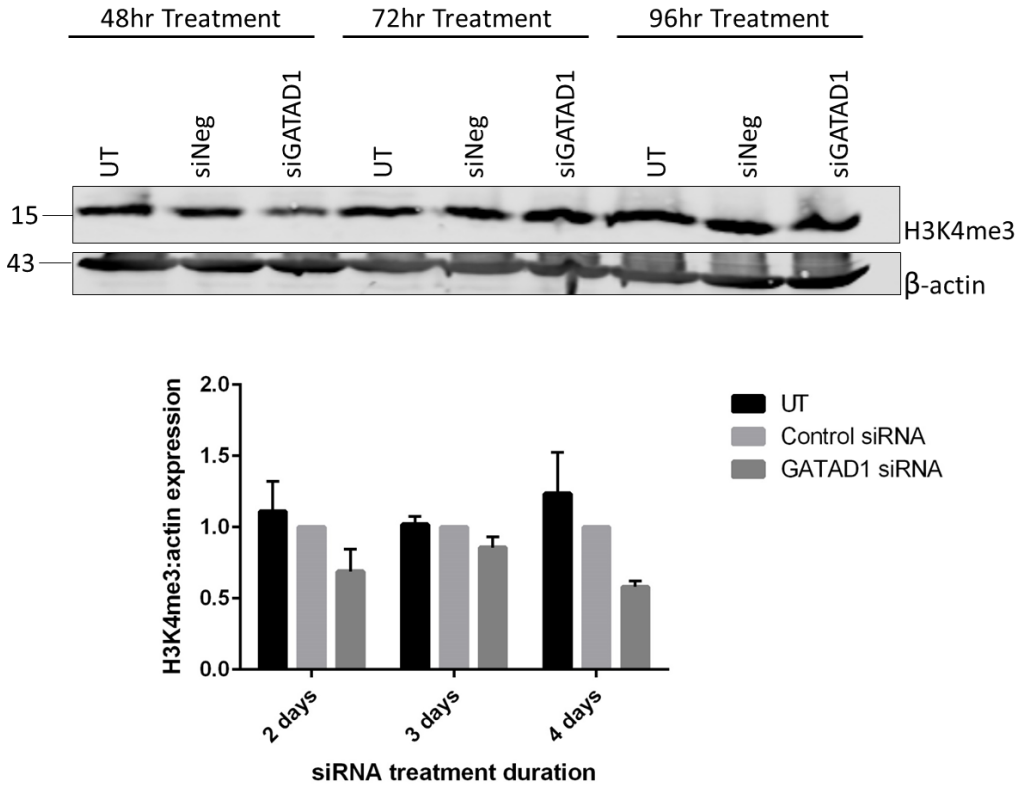


Figure 7-28: GATAD1 Knockdown Results in a Decrease in H3K4me3

A Western blot was carried out on HeLa cells which had undergone siRNA knockdown of GATAD1 with a treatment time of 2, 3 or 4 days alongside untreated cells (UT). Quantification of the bands relative to β-actin loading control resulted in a decrease in H3K4me3, relative to the negative control siRNA. Representative blot from two independent experiments. Error bars indicate the standard deviation (n=2).

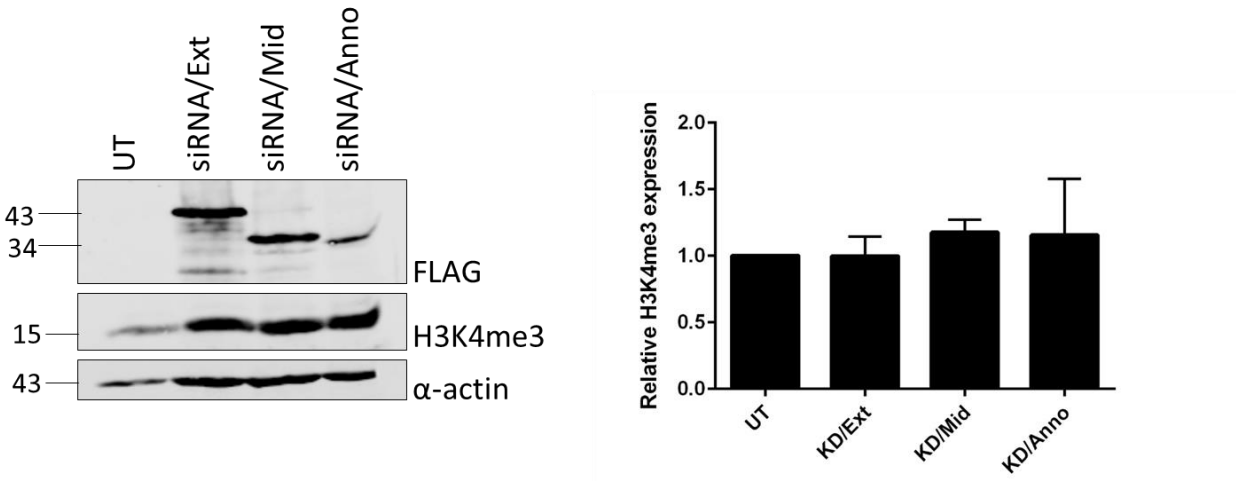


Figure 7-29: Over-expressing GATAD1 Rescues H3K4me3 Levels

GATAD1 was knocked-down in HeLa cells using siRNA targeting the 3'UTR. Cells were simultaneously transfected with GATAD1 plasmids Ext, Mid and Anno which expressed the 5'UTR and ORF only alongside untransfected cells (UT). Cells were harvested 48 hours post-transfection and whole cell lysates were run on a Western blot. Quantification of the bands relative to α-actin loading control resulted in a rescue of H3K4me3, relative to untreated HeLa lysate. Representative blot from three independent experiments. Error bars indicate the standard deviation (n=3).

7.4 Summary of Main Findings

- Both Ext and Anno GATAD1 isoforms complex with KDM5A complex more efficiently than the Mid isoform.
- The GATAD1-KDM5A interaction site has not previously been mapped. NanoBiT PPI assays have narrowed this down, showing that KDM5A has GATAD1 binding sites within both the N and C-terminal region of the protein.
- RNAi knockdown of GATAD1 suggests that GATAD1 isoforms may not play a role in demethylation of H3K4me3 as predicted.

7.5 Discussion

7.5.1 GATAD1 forms PPI with KDM5A

7.5.1.1 GATAD1-KDM5A Co-IP

In order to determine the potential function of the N-terminally extended GATAD1 isoforms, their interaction with KDM5A was investigated. Previous studies have confirmed that annotated GATAD1 complexes with KDM5A, a H3K4 demethylase, which forms part of a repressive chromatin complex (section 1.8.1). However, N-terminally extended GATAD1 isoforms have a more cytoplasmic localisation than annotated GATAD1, away from the chromatin target of the histone demethylase complex. Co-IPs were therefore carried out, immunoprecipitating both GATAD1 and KDM5A with their interacting proteins. Western blot analysis of both sets of co-IPs resulted in higher bait:prey ratios for the Mid GATAD1 isoform, suggesting less KDM5A-GATAD1 Mid complex was forming compared with the Ext and Anno isoforms which formed KDM5A complexes with a similar affinity.

Although the co-IP experiments give an indication of GATAD1 complex formation, there are drawbacks to quantifying the GATAD1-KDM5A PPI in this manner. Firstly, in order to obtain an initial KDM5A Western blot, HeLa cells had to be lysed with the stringent RIPA buffer (section 7.3.1). However, in order to maintain complex formation throughout the co-IP, a less stringent low-ionic strength lysis buffer was used, resulting in a poor KDM5A signal which was difficult to optimise and quantify. In addition to this, lysing the cells breaks down membranes preventing differentiation between complexes formed in the nucleus versus cytosolic complexes. Prior to

overexpressing Halo-KDM5A, an attempt to stabilise the endogenous KDM5A-FLAG-GATAD1 interaction was made by incorporating L-Photo-Leucine (4 mM) and L-Photo-Methionine (2 mM) into the proteins. The diazirine rings present in the amino acid derivatives allow UV cross-linking of protein complexes, which should be evident as an increase in molecular weight when run on SDS-PAGE and analysed by Western blot. No GATAD1 shift was observed using endogenous KDM5A (data not shown), however this experiment did not use MG132 to prevent the rapid protein turnover shown in section 7.3.1.2. It would therefore be interesting to repeat the photoreactive cross linking experiment using the same conditions as used in the co-IP experiments, to confirm the results (overexpressing Halo-KDM5A, double transfection and MG132 treatment).

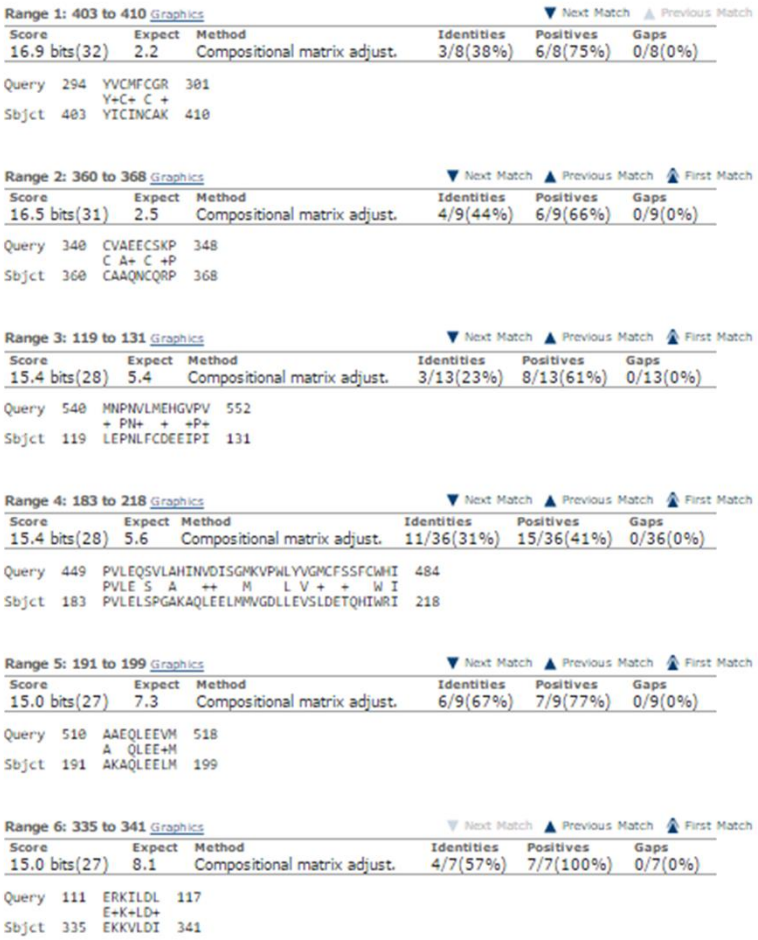
A further point to consider is non-specific binding to the IP antibodies (Figure 7-10). FLAG protein is being pulled down on the Halo-resin and Halo protein is being pulled down on the FLAG-resin. These are non-specific interactions occurring even when the lysate has been pre-cleared, which could be minimised in future experiments by increasing the ionic strength of the IP buffer, resulting in more stringent washes of the resin-bound immune complexes.

Future co-IP work would involve immunoblotting for other proteins which have been suggested to form part of the H3K4 chromatin complex, including HDAC1/2, Emsy and Sin3b (Vermeulen et al., 2010). Comparing the levels of these proteins complexing with each GATAD1 isoform may indicate whether the N-terminally extended isoforms are forming a functional histone demethylase complex in the nucleus, or whether they are forming another cytoplasmic complex with KDM5A, potentially involved in protein synthesis (Van Rechem et al., 2015). This could also be investigated by carrying out IP assays on separate nuclear and cytoplasmic fractions.

7.5.1.2 GATAD1/KDM5A NanoBiT PPI

The NanoBiT system was used as an alternative to co-IP assays, to confirm the GATAD1-KDM5A interaction. NanoBiT offers greater sensitivity and uses a non-lytic assay format which makes the data more biologically relevant. The NanoBiT assay provided data relating to each GATAD1 isoform's ability to complex with KDM5A, as well as the region of KDM5A in which the PPI was taking place which has not yet been mapped. The results confirmed the co-IP assays, in that both Ext and Anno GATAD1 isoforms were able to complex with KDM5A with a higher affinity than Mid, which gave a luminescent signal just above that of the negative control. KDM5A appears to have two GATAD1 binding sites, within both the N and C-terminus of the protein, suggesting it wraps around GATAD1. Deletion of sequence within the N and C-terminal region would give a more accurate prediction of the GATAD1-KDM5A binding-site. A BLASTp search of KDM5A N-terminus and C-terminus resulted in 6 regions of alignment which could be targeted for mutagenesis initially as potential GATAD1-KDM5A interaction sites (Figure 7-30). Although the N- and C-

termini of KDM5A are clearly different, a short motif may be enough for the GATAD1-KDM5A interaction.



Range 1: 403 to 410 Graphics — Next Match — Previous Match

Score	Expect	Method	Identities	Positives	Gaps
16.9 bits(32)	2.2	Compositional matrix adjust.	3/8(38%)	6/8(75%)	0/8(0%)

```
Query  294  YVCMFCGR  301
            Y+C+ C +
Sbjct  403  YICINCAK  410
```

Range 2: 360 to 368 Graphics — Next Match — Previous Match — First Match

Score	Expect	Method	Identities	Positives	Gaps
16.5 bits(31)	2.5	Compositional matrix adjust.	4/9(44%)	6/9(66%)	0/9(0%)

```
Query  340  CVAEECSKP  348
            C A+ C +P
Sbjct  360  CAAQNCQRP  368
```

Range 3: 119 to 131 Graphics — Next Match — Previous Match — First Match

Score	Expect	Method	Identities	Positives	Gaps
15.4 bits(28)	5.4	Compositional matrix adjust.	3/13(23%)	8/13(61%)	0/13(0%)

```
Query  540  MNPNVLMEHGVPV  552
            + PN+  +  +P+
Sbjct  119  LEPNLFCDEEIPI  131
```

Range 4: 183 to 218 Graphics — Next Match — Previous Match — First Match

Score	Expect	Method	Identities	Positives	Gaps
15.4 bits(28)	5.6	Compositional matrix adjust.	11/36(31%)	15/36(41%)	0/36(0%)

```
Query  449  PVLEQSVLAHINVDISGMKVPWLYVGMCFSSFCWHI  484
            PVLE S  A    ++   M    L V +  +   W I
Sbjct  183  PVLELSPGAKAQLEELMMVGDLLEVSLDETQHIWRI  218
```

Range 5: 191 to 199 Graphics — Next Match — Previous Match — First Match

Score	Expect	Method	Identities	Positives	Gaps
15.0 bits(27)	7.3	Compositional matrix adjust.	6/9(67%)	7/9(77%)	0/9(0%)

```
Query  510  AAEQLEEVM  518
            A  QLEE+M
Sbjct  191  AKAQLEELM  199
```

Range 6: 335 to 341 Graphics — Next Match — Previous Match — First Match

Score	Expect	Method	Identities	Positives	Gaps
15.0 bits(27)	8.1	Compositional matrix adjust.	4/7(57%)	7/7(100%)	0/7(0%)

```
Query  111  ERKILDL  117
            E+K+LD+
Sbjct  335  EKKVLDI  341
```

Figure 7-30: Potential GATAD1-KDM5A Interaction Sites

There are six regions of alignment following a BLASTp search of KDM5A N-terminus as the subject, with the C-terminus as the query, using the KDM5A fragement sequences utilised by the NanoBiT assays (Figure 7-14).

The NanoBiT data suggests that Mid complexes with KDM5A with a lower efficiency than Ext and Anno, possibly even not at all. This is difficult to explain, since the localisation of Mid is similar to Ext, which is still able to complex and there are no domains within Mid that are not also present in Ext. This leads to questions surrounding the expression of Mid within the NanoBiT system, which is approximately 10x lower than Ext (Figure 7-17). KDM5A also has a lower expression than GATAD1, most likely due to being twice the size. The NanoBiT vectors use the HSV-TK promoter, which provides low-level expression in HeLa cells in order to minimise nonspecific background signal. Exchanging the HSV-TK for a CMV promoter would provide >100-fold higher levels of expression, although this could lead to increased non-specific associations. Exchanging for a SV40 promoter would show whether Mid does complex with KDM5A when expressed well, whilst

confirming Ext and Anno interactions, which should have a luminescent signal >10-fold higher than the negative HaloTag control fusions if they are specific PPIs.

7.5.1.3 GATAD1 Does Not Self-Associate

Other GATA-type zinc finger-containing transcription factors including GATA-1, have been shown to self-associate as well as dimerise with other GATA-transcription factors. The GATA-ZnFs are implicated in PPIs as well as DNA-binding, therefore a NanoBiT assay was carried out on individual GATAD1 isoforms to determine whether they were dimerising in the cell. The luminescent signal for each GATAD1 isoform pair was just above the negative control, suggesting no such dimerization takes place with GATAD1 (Figure 7-21). As mentioned in 7.3.3.1, GATAD1 contains only a single N-terminal zinc finger, unlike the GATA-transcription factors which contain two related (50% amino acid homology) ZnFs (N and C-terminal). The C-terminal GATA ZnF is necessary for DNA binding, whilst the N-terminal ZnF is implicated in stability of the DNA-protein interaction, and interactions with other transcription factors including Fog, Sp1 and EKLF (Yang and Evans, 1992). Both the N and C-terminal ZnFs are necessary for self-association of GATA-1, specifically a 25 amino acid subdomain immediately downstream of each ZnF. Mutation of either subdomain significantly reduces self-association of GATA-1, (Mackay et al., 1998). GATAD1 is unable to self-associate since it does not contain the C-terminal ZnF sequence which is also necessary for dimerization.

7.5.2 Function of GATAD1 within KDM5A complex

GATAD1 was thought to form part of a repressive chromatin complex which demethylates H3K4me3 through KDM5A, therefore the ability of each GATAD1 isoform to form a functional complex was investigated. siRNA-mediated endogenous knockdown of GATAD1 was initially carried out, showing a significant decrease in GATAD1 RNA levels. The knockdown could have been further optimised in several ways. Delivery of the siRNA into cultured cells could be optimised by using alternative transfection reagents (such as siPORT NeoFX transfection reagent, Thermo Fisher), cell densities and higher siRNA concentrations. siRNA transfection efficiency can be monitored by fluorescently labelling siRNA, which can also be used to correlate transfection with down-regulation of the target protein.

If GATAD1 contributed towards H3K4 demethylation as predicted, then siRNA knockdown should result in increased H3K4me3 levels. However, no change in H3K4me3 was observed, suggesting that a compensatory mechanism may be taking place with another histone demethylase acting on the same histone mark; losing GATAD1 from the KDM5A system may render the whole complex inert, allowing another complex to take over at the H3K4me3 mark. Adding-back each

GATAD1 isoform individually by overexpression also had no significant effect on H3K4me3 levels, although the GATAD1 knockdown alone should also have been included in this experiment to make informed conclusions.

Measuring the levels of H3K4me3 as a level of GATAD1 function may not be accurate, since there are numerous other histone demethylase enzymes also acting on the same histone mark, including KDM2B (FBXL10), KDM5B, KDM5C and KDM5D. An alternative method of investigating the function of each GATAD1 isoform would be to use chromatin immunoprecipitation (ChIP). ChIP would determine whether all three isoforms were forming a functional chromatin complex within the nucleus, or whether the extended isoforms had a different function. Since the NanoBiT assay suggested that Mid does not complex with KDM5A as efficiently as Ext and Anno, it would be interesting to investigate the implications of this on the function of the protein.

8. Final Discussion

Alternative translation initiation using in-frame non-AUG codons is one mechanism used by the cell to translate multiple protein isoforms from a single mRNA message, increasing the size and complexity of the proteome. Until recently, little was known about the frequency of alternative translation events or their functional significance. There are now numerous *in silico* bioinformatic databases available to predict aTIS, as well as advances in next generation sequencing enabling quantitative analysis of *in vivo* translation by ribosome profiling. However, each of these prediction methods have limitations which are discussed in detail later. It is therefore important to combine sequence analysis with experimental validation using cell culture to identify alternative translation initiation sites, which allows further investigation surrounding the regulation of translation and functional significance of the N-terminally extended/truncated protein isoforms. During this thesis, GATAD1 was identified as a candidate gene through an experimentally-informed pipeline which enabled the identification of novel aTIS and aspects of their regulation, as well as consequences of expressing the N-terminally extended GATAD1 isoforms.

I have confirmed that alternative translation initiation takes place within the GATAD1 transcript to generate N-terminally extended isoforms (Figure 8-1). A commonly used alternative initiation codon CUG was identified at position -207 as well as an unusual AUU at position -45. The AICs identified here differed from those predicted by both bioinformatic databases as well as ribosome profiling data. Downstream of the AUU initiation codon there is a GC-rich sequence which encourages translation to take place, potentially forming a secondary structure which compensates for the suboptimal codon and weak Kozak consensus. In addition, sequences coding for polyproline residues downstream of both initiation sites appear to encourage translation from these positions, through ribosomal stalling and increased recognition of suboptimal initiation codons. Other factors regulating alternative translation initiation within GATAD1 include oxidative stress, which is a failsafe mechanism allowing unusual initiation mechanisms when the major canonical cap-dependent translation initiation mechanism is compromised. In addition, initiation factor eIF1 is responsible for regulating the stringency of start codon selection within GATAD1, contributing to differential isoform expression. Further experiments were attempted with the CRISPR-tagged line that was generated to confirm the usage of AICs from the endogenous gene locus (Chapter 3), however I had little success in overexpressing factors as well as knocking down endogenous FLAG-tagged GATAD1. This may be because clonal expansion resulted in a cell line that needed a modified or different approach to transfection, which could be explored in future. I have also identified that the N-terminal eORF on both alternative GATAD1 isoforms causes relocalisation when compared to the annotated protein. In both cases, the N-terminal extension results in a more cytoplasmic/nuclear

envelope distribution. Using the NanoBiT PPI system, I have narrowed down the interaction site between GATAD1 and KDM5A which has not previously been mapped. The Mid-GATAD1 isoform translated from AUU interacts less efficiently with KDM5A than the Ext and annotated isoforms, although it is also expressed less efficiently (Figure 7-17) therefore further work would need to be carried out to ensure that the expression was not having an effect on Mid-KDM5A complex formation. However, both Ext and annotated GATAD1 interact with both the N- and C-terminus of KDM5A. Further work could identify whether the extended GATAD1 isoforms have an alternative function to the annotated protein and whether the GATAD1-KDM5A complex formed by Ext is functional as a chromatin reader. SNPs were also identified surrounding the -207 CUG, which dramatically altered the expression pattern of GATAD1 in the cell, having potential impact on health and disease, which would be interesting to investigate further once the function of the GATAD1 isoforms have been confirmed.

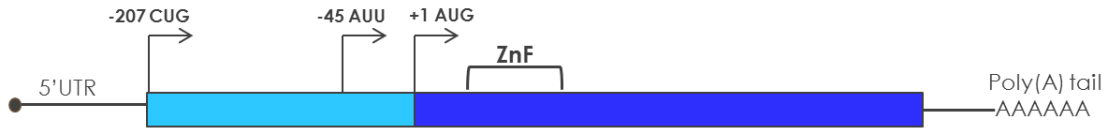


Figure 8-1: Alternative Translation Initiation in GATAD1

We have identified two upstream non-AUG codons used to initiate translation within the GATAD1 mRNA transcript, a CUG at position -207 and an AUU at position -45, with respect to the annotated AUG.

Non-AUG initiation contributes towards generating a more complex proteome, along with alternative splicing and alternative promoter usage. Although translation from non-AUGs is generally inefficient, cis-acting structures and signals within the mRNA itself as well as trans-acting factors encourage translation to take place from sub-optimal AICs within certain mRNAs, including GATAD1. Alternative translation may therefore be used to modulate gene expression in order to prevent overexpression of such regulatory factors (Rogozin et al., 2001). Often, N-terminally extended isoforms differ from the annotated protein in various ways; they may be functionally different such as the two pore domain potassium channels TREK1/2, which show altered ion selectivity (Thomas et al., 2008), have different subcellular localisations such as BAG-1 (Packham et al., 1997), or retain the same function with different levels of efficiency such as eIF4GI (Coldwell and Morley, 2006). In the case of GATAD1, the N-terminal extensions relocalise both extended isoforms, whilst Mid cannot complex with KDM5A as efficiently as the other two isoforms, suggesting that it may potentially have an alternative function, or may be a redundant version used

to regulate translation of the other isoforms. Translation initiation from an upstream CUG and AUU codon therefore regulates the expression of annotated GATAD1, avoiding overexpression of the repressive transcription factor which would be likely to result in detrimental effects on the transcriptome. However, in certain cell stress conditions such as oxidative stress, translation is switched to favour non-AUG GATAD1 AICs, suggesting the levels of the three GATAD1 isoforms can be regulated by exogenous signals.

A ribosome profiling study in HEK293 cells (Lee et al., 2012) identified a single upstream CUG at position -144 within GATAD1, which was also predicted by the ExTATIC macro. We were able to confirm through mutagenesis and cell culture that this CUG was not used to initiate translation *in vivo* in numerous cell lines, including HEK293. The GWIPS (Genome Wide Information on Protein Synthesis) genome browser (Michel et al., 2014) combines published ribosome profiling datasets. Peaks of ribosome density can be seen within GATAD1 at positions -144 CUG and -45 AUU as well as the annotated AUG (Figure 8-2). No peak is seen at -207 CUG although we were able to confirm translation from this position in cell culture. There are clear differences between the GATAD1 initiation sites we have experimentally confirmed and those observed in ribosome profiling, which raises questions around the accuracy of ribosome profiling aggregates. Indeed, ribosome profiling reliability is questionable, with variation in sequence determinants of footprint densities between datasets, resulting in data which is difficult to aggregate (O'Connor et al., 2016). In addition, ribosome profiling uses translation inhibitors such as cycloheximide, which can cause artefacts in the data as a result of analysing translation initiation under stress conditions, which has the potential to alter the use of AICs, (Gerashchenko and Gladyshev, 2014). The Lee et al ribosome profiling study which identified the -144 CUG used LTM (lactimidomycin) to stall the 80S ribosome at initiation sites. LTM is a CHX-like drug which uses a similar mechanism to block the translocation step in elongation, suggesting that the initiation sites identified in this dataset may not be reliable. The discrepancies observed in the ribosome profiling data confirms that we are validated in our way of identifying potential non-AUG upstream AICs via sequence homology and experimental confirmation using mutagenesis and cell culture to test the usage of the AICs.

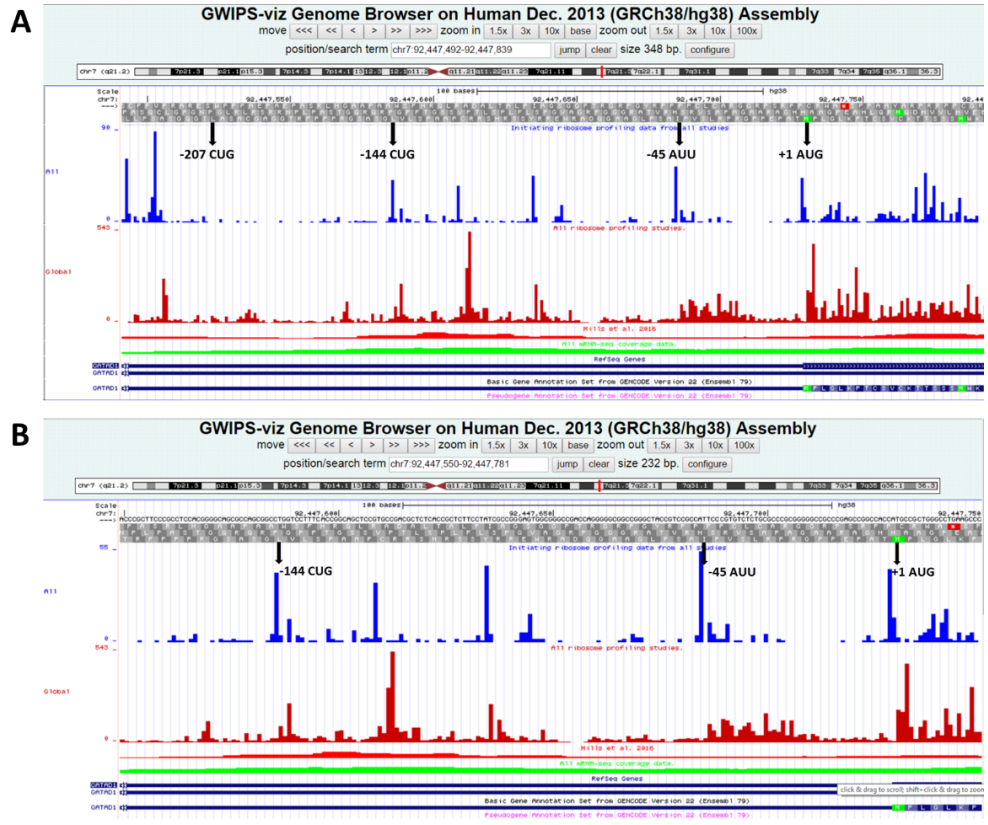


Figure 8-2: GATAD1 GWIPS Data

(A) GATAD1 (frame 3) full ribosome profiling track from initiating ribosomes (blue) or elongating ribosomes (red) aggregated from all studies compiled up to March 2017. (B) Ribosome profiling tracks zoomed to base level, where peaks can be seen at -144 CUG, -45 AUU and +1 AUG.

An updated conservation screen was carried out using a tBLASTn search of the GATAD1 translated 5'UTR (Figure 8-3A), which illustrates how many species have been sequenced since the original data mining took place (Figure 3-2). The translated 5'UTR sequence is conserved from the -45 AUU but not from the -207 CUG or -144 CUG. This gives convincing evidence to suggest the functional importance of the Mid GATAD1 isoform.

An interesting point to consider is why the ribosome scans through a CUG (-144) within a strong context and preferentially initiates translation at an unusual AUU (-45) codon within a relatively weak Kozak consensus. Mutagenesis confirmed that the GC-rich sequence downstream of the AUU was encouraging translation from the sub-optimal AIC. In order to analyse whether a hairpin or secondary structure is indeed present downstream of the AUU, a structural alignment of the 5'UTR sequence 105 bp upstream of the annotated AUG was carried out from multiple species, using the LocARNA tool (Will et al., 2012) (Figure 8-3B). There is a conserved hairpin structure downstream of the AUU codon which must be encouraging translation since silent mutations to the sequence corresponding to the 3' stem of the structure reduced translation from the AUU (Figure 4-3).

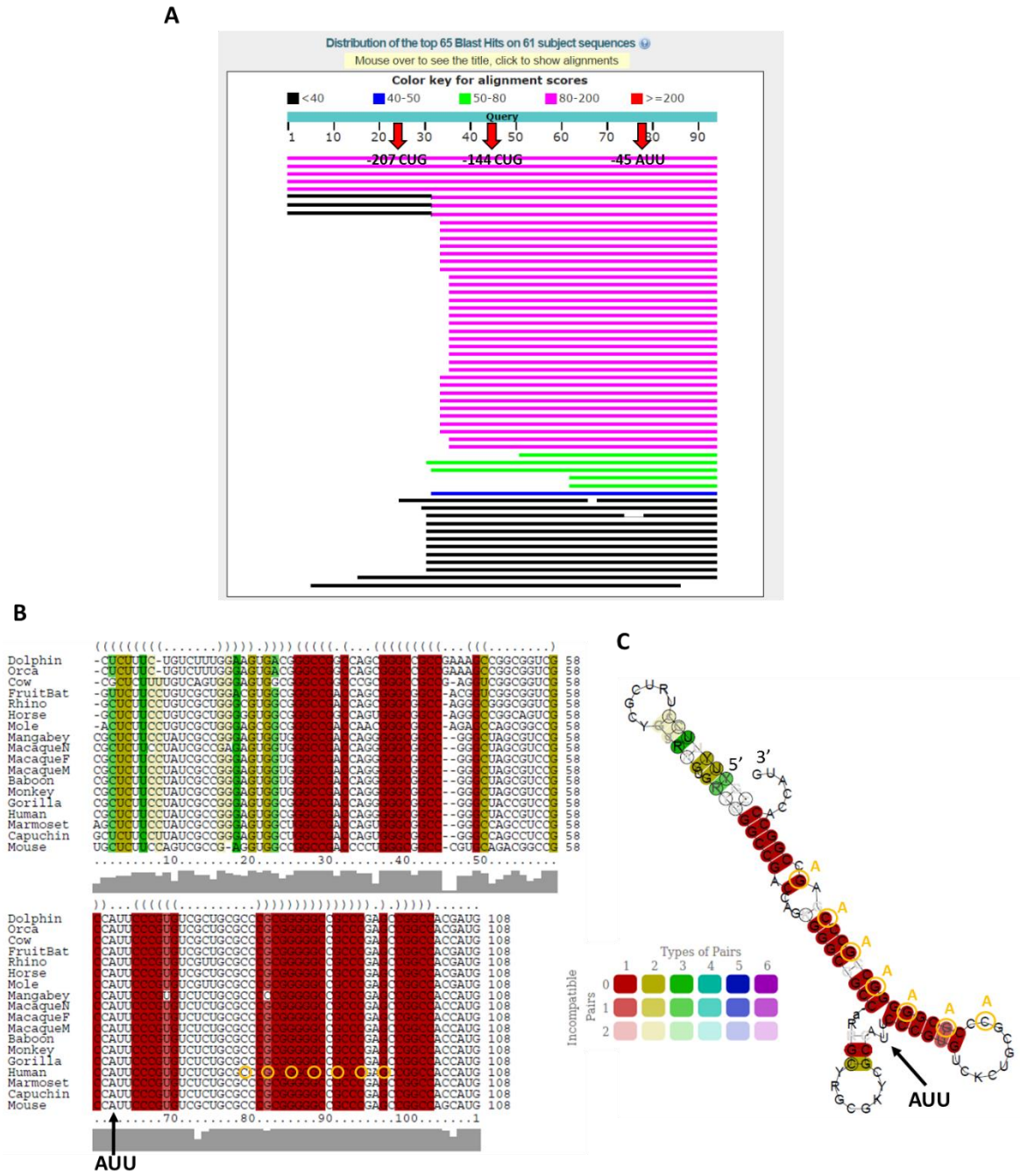


Figure 8-3: GATAD1 5'UTR Alignment and Structural Prediction

(A) GATAD1 conservation between species shown by carrying out a tBLASTn search of the 5'UTR. Further species have been sequenced since the original data mining took place. (B) An alignment of the sequence immediately downstream of the annotated AUG using http://rna.informatik.uni-freiburg.de. The AUU is completely conserved and the bases downstream of the AUU that were mutated in chapter four are circled in yellow. (C) GATAD1 structural prediction; compatible base pairs are coloured according to the key which indicates the structural conservation of the pair. Bases involved in the silent mutations made in chapter four are shown in yellow.

Further experiments which would increase our understanding of GATAD1 alternative translation regulation as well as the function of the subsequent N-terminally extended isoforms, include DMS-MaPseq, as discussed in section 4.5.2. This technique would enable the experimental confirmation of structures within the 5'UTR of GATAD1, which may be influencing translation from sub-optimal AICs. It would be of particular interest to analyse structure formed from the GC-rich sequence downstream of the AUU which is encouraging translation from this unusual AIC and to determine whether this is a hairpin or G-quadruplex structure. On the other hand, the regulatory sequence downstream of the AUU may harbour a RNABP site, which may be identified by carrying out an RNA pull-down. It would also be interesting to investigate the effects of the alternative translation initiation factor eIF2A on the translation of alternative GATAD1 isoforms, as well as further work to clarify the effect of hypoxia on alternative translation initiation. Cells could be cultured in a hypoxia chamber within a standard CO_2 incubator, at oxygen levels which mimic oxygen levels in tumour cells, which are typically between 0.1-4%, depending on the type and progression of the tumour (McKeown, 2014).

Further investigations into the function of the N-terminally extended GATAD1 isoforms could begin with an optimised co-IP experiment, whereby complexes are cross-linked prior to being isolated, as discussed in section 7.5.1.1. This would allow the use of higher stringency wash buffers which should prevent the non-specific binding observed in Figure 7-10. Also, isolating the cytoplasmic lysate from the nuclear fraction prior to carrying out the co-IP would help to maintain physiological interactions and should give an indication of whether the extended isoforms interact with different proteins once localised to the cytoplasm. Identification of other proteins within the complex would also extend our understanding of the function of the complexes formed. The most accurate way of identifying interacting proteins would be to carry out mass spectrometry-coupled co-IP. This would ideally be carried out on the GATAD1-CRISPR-3xFLAG cell line to maintain physiological relevance, which can be modified (by mutating the initiation sites that are not required) by further rounds of CRISPR to translate Ext, Mid or Anno respectively. Mass-spec can then be carried out on purified GATAD1 isoforms and their complexed proteins, from both the nuclear and cytoplasmic fraction.

In addition, ChIP-seq using the previously mentioned CRISPR cell lines would enable identification of individual GATAD1 isoform interactions with DNA, indicating any differences in function between the N-terminally extended isoforms and annotated GATAD1. Since analysing levels of the H3K4me3 mark as a way to assay the individual contributions of the GATAD1 isoforms in H3K4me3 demethylation was unreliable, analysing levels of downstream genes such as KISS/TAC3 (Lomniczi et al., 2015) by DNA microarray in the presence of individual GATAD1 isoforms would also elucidate the function of each GATAD1 isoform more clearly. Finally, the generation of GATAD1 transgenic mice would enable extensive analysis of GATAD1 function *in vivo*. By generating a conditional GATAD1 knock-out mouse as well as mice expressing only Ext,

Mid and Anno, the function of each isoform and its resultant phenotype can be distinguished. Similarly, generating mice expressing each of the two SNPs surrounding the -207 CUG AIC would identify the impact on gene function and phenotype, which may be advantageous within certain populations; for example, all submitted occurrences to the 1000 Genome project of the SNP preventing translation of the Ext isoform of GATAD1 (rs192745223) were from the African population, which suggests that Ext has a function disadvantageous to this population. Finally, it would be interesting to engineer conditional tissue-specific transgenic mice, whereby each GATAD1 isoform is individually expressed in the retina to investigate the contribution of each isoform to retinal development. GATAD1 (ODAG) was found to be highly expressed in mice at postnatal day 2 (P2), downregulated at by P10 and undetectable by P14. By understanding the function of each isoform, the contribution of alternative translation initiation to expanding the proteome in this case may be better understood.

9. References

Accari, S. L. & Fisher, P. R. 2015. Emerging Roles of JmjC Domain-Containing Proteins. *Int Rev Cell Mol Biol,* 319(165-220.

Ahringer, J. 2000. NuRD and SIN3 - histone deacetylase complexes in development. *Trends in Genetics,* 16(8)**,** pp 351-356.

Allmang, C. & Krol, A. 2006. Selenoprotein synthesis: UGA does not end the story. *Biochimie,* 88(11)**,** pp 1561-71.

Anderson, D. M., Anderson, K. M., Chang, C. L., Makarewich, C. A., Nelson, B. R., McAnally, J. R., Kasaragod, P., Shelton, J. M., Liou, J., Bassel-Duby, R. & Olson, E. N. 2015. A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell,* 160(4)**,** pp 595-606.

Aspden, J. L., Eyre-Walker, Y. C., Phillips, R. J., Amin, U., Mumtaz, M. A., Brocard, M. & Couso, J. P. 2014. Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *Elife,* 3(e03528.

Ayoubi, T. A. & Van De Ven, W. J. 1996. Regulation of gene expression by alternative promoters. *Faseb j,* 10(4)**,** pp 453-60.

Baierlein, C. & Krebber, H. 2010. Translation termination New factors and insights. *Rna Biology,* 7(5)**,** pp 548-550.

Becerra, S. P., Rose, J. A., Hardy, M., Baroudy, B. M. & Anderson, C. W. 1985. Direct Mapping of Adeno-Associated Virus Capsid Protein-B and Protein-C - A Possible ACG Initiation Codon. *Proceedings of the National Academy of Sciences of the United States of America,* 82(23)**,** pp 7919-7923.

Betts, L. & Spremulli, L. L. 1994. Analysis of the Role of the Shine-Dalgarno Sequence and mRNA Secondary Structure on the Efficiency of Translational Initiation in the Euglena gracilis Chloroplast atpH mRNA. *The Journal of Biological Chemistry,* 269(42)**,** pp 26456-63.

Blobel, G. 1980. Intracellular protein topogenesis. *Proc Natl Acad Sci U S A,* 77(3)**,** pp 1496-500.

Boeck, R. & Kolakofsky, D. 1994. Position +5 and +6 Can Be Major Determinants of the Efficiency of Non-AUG Initiation Codons for Protein Synthesis. *Embo Journal,* 13(15)**,** pp 3608-3617.

Boulikas, T. 1993. Nuclear localization signals (NLS). *Crit Rev Eukaryot Gene Expr,* 3(3)**,** pp 193-227.

Brackertz, M., Boeke, J., Zhang, R. & Renkawitz, R. 2002. Two highly related p66 proteins comprise a new family of potent transcriptional repressors interacting with MBD2 and MBD3. *Journal of Biological Chemistry,* 277(43)**,** pp 40958-40966.

Briske-Anderson, M. J., Finley, J. W. & Newman, S. M. 1997. The influence of culture time and passage number on the morphological and physiological development of Caco-2 cells. *Proc Soc Exp Biol Med,* 214(3)**,** pp 248-57.

Browne, G. J. & Proud, C. G. 2002. Regulation of peptide-chain elongation in mammalian cells. *European Journal of Biochemistry,* 269(22)**,** pp 5360-5368.

Bugaut, A. & Balasubramanian, S. 2012. 5′-UTR RNA G-quadruplexes: translation regulation and targeting. *Nucleic Acids Research,* 40(

Byrd, M. P., Zamora, M. & Lloyd, R. E. 2002. Generation of multiple isoforms of eukaryotic translation initiation factor 4GI by use of alternate translation initiation codons. *Mol Cell Biol,* 22(13)**,** pp 4499-511.

Cao, J. & Geballe, A. P. 1995. Translational inhibition by a human cytomegalovirus upstream open reading frame despite inefficient utilization of its AUG codon. *J Virol,* 69(2)**,** pp 1030-6.

Chan, S. W. & Hong, W. J. 2001. Retinoblastoma-binding protein 2 (Rbp2) potentiates nuclear hormone receptor-mediated transcription. *Journal of Biological Chemistry,* 276(30)**,** pp 28402-28412.

Chappell, S. A., Edelman, G. M. & Mauro, V. P. 2004. Biochemical and functional analysis of a 9-nt RNA sequence that affects translation efficiency in eukaryotic cells. *Proceedings of the National Academy of Sciences of the United States of America,* 101(26)**,** pp 9590-9594.

Chicas, A., Kapoor, A., Wang, X. W., Aksoy, O., Evertts, A. G., Zhang, M. Q., Garcia, B. A., Bernstein, E. & Lowe, S. W. 2012. H3K4 demethylation by Jarid1a and Jarid1b contributes to retinoblastoma-mediated gene silencing during cellular senescence. *Proceedings of the National Academy of Sciences of the United States of America,* 109(23)**,** pp 8971-8976.

Coldwell, M. J., deSchoolmeester, M. L., Fraser, G. A., Pickering, B. M., Packham, G. & Willis, A. E. 2001. The p36 isoform of BAG-1 is translated by internal ribosome entry following heat shock. *Oncogene,* 20(30)**,** pp 4095-100.

Coldwell, M. J. & Morley, S. J. 2006. Specific isoforms of translation initiation factor 4GI show differences in translational activity. *Mol Cell Biol,* 26(22)**,** pp 8448-60.

Coldwell, M. J., Sack, U., Cowan, J. L., Barrett, R. M., Vlasak, M., Sivakumaran, K. & Morley, S. J. 2012. Multiple isoforms of the translation initiation factor eIF4GII are generated via use of alternative promoters, splice sites and a non-canonical initiation codon. *Biochem J,* 448(1)**,** pp 1-11.

Colgan, D. F. & Manley, J. L. 1997. Mechanism and regulation of mRNA polyadenylation. *Genes Dev,* 11(21)**,** pp 2755-66.

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A. & Zhang, F. 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science,* 339(6121)**,** pp 819-23.

Cong, L. & Zhang, F. 2015. Genome engineering using CRISPR-Cas9 system. *Methods Mol Biol,* 1239(197-217.

Cour, T. l., Kiemer, L., Mølgaard, A., Gupta, R., Skriver, K. & Brunak, S. 2004. Analysis and prediction of leucine-rich nuclear export signals. *Protein Engineering, Design and Selection,* 17(6)**,** pp 527-36.

Crick, F. 1970. Central Dogma of Molecular Biology. *Nature,* 226(1198)**,** pp 561-563.

Crossley, M., Merika, M. & Orkin, S. H. 1995. Self-Association of the Erythroid Transcription Factor GATA-1 Mediated by Its Zinc Finger Domains. *Molecular & Cellular Biology,* 15(5)**,** pp 2448-56.

Das, S. & Maitra, U. 2001. Functional significance and mechanism of eIF5-promoted GTP hydrolysis in eukaryotic translation initiation. *Prog Nucleic Acid Res Mol Biol,* 70(207-31.

de Ruijter, A. J., van Gennip, A. H., Caron, H. N., Kemp, S. & van Kuilenburg, A. B. 2003. Histone deacetylases (HDACs): characterization of the classical HDAC family. *Biochem J,* 370(Pt 3)**,** pp 737-49.

Derry, M. C., Yanagiya, A., Martineau, Y. & Sonenberg, N. 2006. Regulation of poly(A)-binding protein through PABP-interacting proteins. *Cold Spring Harbor Symposia on Quantitative Biology,* 71(537-543.

Dever, T. E. 2002. Gene-specific regulation by general translation factors. *Cell,* 108(4)**,** pp 545-556.

Dever, T. E., Feng, L., Wek, R. C., Cigan, A. M., Donahue, T. F. & Hinnebusch, A. G. 1992. Phosphorylation of initiation factor 2 alpha by protein kinase GCN2 mediates gene-specific translational control of GCN4 in yeast. *Cell,* 68(3)**,** pp 585-96.

DiTacchio, L., Le, H. D., Vollmers, C., Hatori, M., Witcher, M., Secombe, J. & Panda, S. 2011. Histone lysine demethylase JARID1a activates CLOCK-BMAL1 and influences the circadian clock. *Science,* 333(6051)**,** pp 1881-5.

Drabkin, H. J. & Rajbhandary, U. L. 1998. Initiation of protein synthesis in mammalian cells with codons other than AUG and amino acids other than methionine. *Molecular and Cellular Biology,* 18(9)**,** pp 5140-5147.

Eckmann, C. R., Rammelt, C. & Wahle, E. 2011. Control of poly(A) tail length. *Wiley Interdiscip Rev RNA,* 2(3)**,** pp 348-61.

Egorova, K. S., Olenkina, O. M. & Olenina, L. V. 2010. Lysine Methylation of Nonhistone Proteins Is a Way to Regulate Their Stability and Function. *Biochemistry-Moscow,* 75(5)**,** pp 535-548.

Emanuelsson, O., Nielsen, H., Brunak, S. & Heijne, G. v. 2000. Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence. *J. Mol. Biol.,* 300(4)**,** pp 1005-16.

Fabian, M. R., Sonenberg, N. & Filipowicz, W. 2010. Regulation of mRNA translation and stability by microRNAs. *Annu Rev Biochem,* 79(351-79.

Fekete, C. A., Mitchell, S. F., Cherkasova, V. A., Applefield, D., Algire, M. A., Maag, D., Saini, A. K., Lorsch, J. R. & Hinnebusch, A. G. 2007. N- and C-terminal residues of eIF1A have opposing effects on the fidelity of start codon selection. *Embo Journal,* 26(6)**,** pp 1602-1614.

Fernandez, I. S., Bai, X. C., Murshudov, G., Scheres, S. H. & Ramakrishnan, V. 2014. Initiation of translation by cricket paralysis virus IRES requires its translocation in the ribosome. *Cell,* 157(4)**,** pp 823-31.

Frottin, F., Martinez, A., Peynot, P., Mitra, S., Holz, R. C., Giglione, C. & Meinnel, T. 2006. The proteomics of N-terminal methionine cleavage. *Molecular & Cellular Proteomics,* 5(12)**,** pp 2336-2349.

Gandin, V., Miluzio, A., Barbieri, A. M., Beugnet, A., Kiyokawa, H., Marchisio, P. C. & Biffo, S. 2008. Eukaryotic initiation factor 6 is rate-limiting in translation, growth and transformation. *Nature,* 455(7213)**,** pp 684-U81.

Gao, X., Wan, J., Liu, B., Ma, M., Shen, B. & Qian, S.-B. 2015. Quantitative profiling of initiating ribosomes in vivo. *Nature Methods,* 12(2)**,** pp 147-53.

Gavel, Y. & von Heijne, G. 1990. Cleavage-site motifs in mitochondrial targeting peptides. *Protein Eng,* 4(1)**,** pp 33-7.

Gerashchenko, M. V. & Gladyshev, V. N. 2014. Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic Acids Res,* 42(17)**,** pp.

Goping, I. S., Millar, D. G. & Shore, G. C. 1995. Identification of the human mitochondrial protein import receptor, huMas20p. Complementation of delta mas20 in yeast. *FEBS Lett,* 373(1)**,** pp 45-50.

Gorlich, D. & Kutay, U. 1999. Transport between the cell nucleus and the cytoplasm. *Annu Rev Cell Dev Biol,* 15(607-60.

Gradi, A., Imataka, H., Svitkin, Y. V., Rom, E., Raught, B., Morino, S. & Sonenberg, N. 1998. A novel functional human eukaryotic translation initiation factor 4G. *Mol Cell Biol,* 18(1)**,** pp 334-42.

Grollman, A. P. 1967. Inhibitors of protein biosynthesis. II. Mode of action of anisomycin. *J Biol Chem,* 242(13)**,** pp 3226-33.

Grunert, S. & Jackson, R. J. 1994. The immediate downstream codon strongly influences the efficiency of utilization of eukaryotic translation initiation codons. *Embo Journal,* 13(15)**,** pp 3618-3630.

Grzenda, A., Lomberk, G., Zhang, J.-S. & Urrutia, R. 2009. Sin3: Master scaffold and transcriptional corepressor. *Biochimica Et Biophysica Acta-Gene Regulatory Mechanisms,* 1789(6-8)**,** pp 443-450.

Guillemette, B., Drogaris, P., Lin, H.-H. S., Armstrong, H., Hiragami-Hamada, K., Imhof, A., Bonneil, E., Thibault, P., Verreault, A. & Festenstein, R. J. 2011. H3 Lysine 4 Is Acetylated at Active Gene Promoters and Is Regulated by H3 Lysine 4 Methylation. *Plos Genetics,* 7(3)**,** pp.

Hashimoto, N. N., Carnevalli, L. S. & Castilho, B. A. 2002. Translation initiation at non-AUG codons mediated by weakened association of eukaryotic initiation factor (eIF) 2 subunits. *Biochemical Journal,* 367(Pt 2)**,** pp 359-368.

Hayakawa, T. & Nakayama, J. 2011. Physiological Roles of Class I HDAC Complex and Histone Demethylase. *Journal of Biomedicine and Biotechnology,* 2011(129383)**,** pp.

Hayakawa, T., Ohtani, Y., Hayakawa, N., Shinmyozu, K., Saito, M., Ishikawa, F. & Nakayama, J. 2007. RBP2 is an MRG15 complex component and down-regulates intragenic histone H3 lysine 4 methylation. *Genes to Cells,* 12(6)**,** pp 811-826.

Heessen, S. & Fornerod, M. 2007. The inner nuclear envelope as a transcription factor resting place. *Embo Reports,* 8(10)**,** pp 914-919.

Heifetz, A., Keenan, R. W. & Elbein, A. D. 1979. Mechanism of action of tunicamycin on the UDP-GlcNAc:dolichyl-phosphate Glc-NAc-1-phosphate transferase. *Biochemistry,* 18(11)**,** pp 2186-92.

Hellen, C. U. T., Sarnow, P. 2001. Internal ribosome entry sites in eukaryotic mRNA molecules. *Genes and Development,* 15(13)**,** pp 1593-1612.

Hinnebusch, A. G. 2006. eIF3: a versatile scaffold for translation initiation complexes. *Trends Biochem Sci,* 31(10)**,** pp 553-62.

Hinnebusch, A. G. 2011. Molecular Mechanism of Scanning and Start Codon Selection in Eukaryotes. *Microbiology and Molecular Biology Reviews,* 75(3)**,** pp 434-467.

Hou, J., Wang, Z. L., Yang, L. N., Guo, X. M. & Yang, G. 2014. The function of EMSY in cancer development. *Tumor Biology,* 35(6)**,** pp 5061-5066.

Hussain, T., Llacer, J. L., Fernandez, I. S., Munoz, A., Martin-Marcos, P., Savva, C. G., Lorsch, J. R., Hinnebusch, A. G. & Ramakrishnan, V. 2014. Structural changes enable start codon recognition by the eukaryotic translation initiation complex. *Cell,* 159(3)**,** pp 597-607.

Ingolia, N. T., Lareau, L. F. & Weissman, J. S. 2011. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell,* 147(4)**,** pp 789-802.

Islam, A., Richter, W. F., Lopez-Bigas, N. & Benevolenskaya, E. V. 2011. Selective targeting of histone methylation. *Cell Cycle,* 10(3)**,** pp 413-424.

Ivanov, I. P., Firth, A. E., Michel, A. M., Atkins, J. F. & Baranov, P. V. 2011. Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. *Nucleic Acids Research,* 39(10)**,** pp 4220-4234.

Jackson, R. J., Hellen, C. U. T. & Pestova, T. V. 2010. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nature Reviews,* 10(113-127.

Jennings, M. D., Zhou, Y., Mohammad-Qureshi, S. S., Bennett, D. & Pavitt, G. D. 2013. eIF2B promotes eIF5 dissociation from eIF2*GDP to facilitate guanine nucleotide exchange for translation initiation. *Genes Dev,* 27(24)**,** pp 2696-707.

Jjingo, D., Conley, A. B., Yi, S. V., Lunyak, V. V. & Jordan, I. K. 2012. On the presence and role of human gene-body DNA methylation. *Oncotarget,* 3(4)**,** pp 462-474.

Kaffman, A., Rank, N. M., O'Neill, E. M., Huang, L. S. & O'Shea, E. K. 1998. The receptor Msn5 exports the phosphorylated transcription factor Pho4 out of the nucleus. *Nature,* 396(6710)**,** pp 482-6.

Kikin, O., D'Antonio, L. & Bagga, P. S. 2006. QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Research,* 34(

Kim, E., Goren, A. & Ast, G. 2008. Alternative splicing: current perspectives. *Bioessays,* 30(1)**,** pp 38-47.

Ko, L. J. & Engel, J. D. 1993. DNA-Binding Specificities of the GATA Transcription Factor Family. *Molecular and Cellular Biology,* 13(7)**,** pp 4011-4022.

Kochetov, A. V., Palyanov, A., Titov, II, Grigorovich, D., Sarai, A. & Kolchanov, N. A. 2007. AUG_hairpin: prediction of a downstream secondary structure influencing the recognition of a translation start site. *Bmc Bioinformatics,* 8(318)**,** pp.

Kosugi, S., Hasebe, M., Matsumura, N., Takashima, H., Miyamoto-Sato, E., Tomita, M. & Yanagawa, H. 2009. Six Classes of Nuclear Localization Signals Specific to Different Binding Grooves of Importin . *Journal of Biological Chemistry,* 284(1)**,** pp 478-85.

Kozak, M. 1987a. An analysis of 5'-noncoding sequences from 699 vertebrate messenger-RNAs. *Nucleic Acids Research,* 15(20)**,** pp 8125-8148.

Kozak, M. 1987b. Effects of intercistronic length on the efficiency of reinitiation by eucaryotic ribosomes. *Mol Cell Biol,* 7(10)**,** pp 3438-45.

Kozak, M. 1989. Context effects and inefficient initiation at non-AUG codons in eukaryotic cell-free translation *Molecular and Cellular Biology,* 9(11)**,** pp 5073-5080.

Kozak, M. 1990. Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. *PNAS,* 87(21)**,** pp 8301-5.

Kozak, M. 1991. A short leader sequence impairs the fidelity of initiation by eukaryotic ribosomes. *Gene Expr,* 1(2)**,** pp 111-5.

Kwun, H. J., Toptan, T., Ramos da Silva, S., Atkins, J. F., Moore, P. S. & Chang, Y. 2014. Human DNA tumor viruses generate alternative reading frame proteins through repeat sequence recoding. *Proc Natl Acad Sci U S A,* 111(41)**,** pp E4342-9.

Labbadia, J. & Morimoto, R. I. 2013. Huntington's disease: underlying molecular mechanisms and emerging concepts. *Trends Biochem Sci,* 38(8)**,** pp 378-85.

Lachner, M. & Jenuwein, T. 2002. The many faces of histone lysine methylation. *Curr Opin Cell Biol,* 14(3)**,** pp 286-98.

Landry, J. J., Pyl, P. T., Rausch, T., Zichner, T., Tekkedil, M. M., Stutz, A. M., Jauch, A., Aiyar, R. S., Pau, G., Delhomme, N., Gagneur, J., Korbel, J. O., Huber, W. & Steinmetz, L. M. 2013. The genomic and transcriptomic landscape of a HeLa cell line. *G3 (Bethesda),* 3(8)**,** pp 1213-24.

Laurencikiene, J., Kallman, A. M., Fong, N., Bentley, D. L. & Ohman, M. 2006. RNA editing and alternative splicing: the importance of co-transcriptional coordination. *Embo Reports,* 7(3)**,** pp 303-307.

Lee, M. G., Wynder, C., Bochar, D. A., Hakimi, M. A., Cooch, N. & Shiekhattar, R. 2006. Functional interplay between histone demethylase and deacetylase enzymes. *Mol Cell Biol,* 26(17)**,** pp 6395-402.

Lee, N., Erdjument-Bromage, H., Tempst, P., Jones, R. S. & Zhang, Y. 2008. The H3K4 Demethylase Lid Associates with and Inhibits Histone Deacetylase Rpd3. *Molecular & Cellular Biology,* 29(6)**,** pp 1401-10.

Lee, S., Liub, B., Leec, S., Huangd, S.-X., Shend, B. & Qian, S.-B. 2012. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *PNAS,* 109(37)**,** pp 2424-32.

LeFebvre, A. K., Korneeva, N. L., Trutschl, M., Cvek, U., Duzan, R. D., Bradley, C. A., Hershey, J. W. B. & Rhoads, R. E. 2006. Translation initiation factor eIF4G-1 binds to eIF3 through the eIF3e subunit. *Journal of Biological Chemistry,* 281(32)**,** pp 22917-22932.

Levy, D. & Gozani, O. 2010. Decoding Chromatin Goes High Tech. *Cell,* 142(6)**,** pp 844-846.

Li, S. N., Paterno, G. D. & Gillespie, L. L. 2013. Nuclear Localization of the Transcriptional Regulator MIER1 alpha Requires Interaction with HDAC1/2 in Breast Cancer Cells. *Plos One,* 8(12)**,** pp.

Lin, Y. C., Boone, M., Meuris, L., Lemmens, I., Van Roy, N., Soete, A., Reumers, J., Moisse, M., Plaisance, S., Drmanac, R., Chen, J., Speleman, F., Lambrechts, D., Van de Peer, Y., Tavernier, J. & Callewaert, N. 2014. Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. *Nat Commun,* 3(5)**,** pp 4767.

Linder, P. & Jankowsky, E. 2011. From unwinding to clamping - the DEAD box RNA helicase family. *Nat Rev Mol Cell Biol,* 12(8)**,** pp 505-16.

Lomniczi, A., Wright, H., Castellano, J. M., Matagne, V., Toro, C. A., Ramaswamy, S., Plant, T. M. & Ojeda, S. R. 2015. Epigenetic regulation of puberty via Zinc finger protein-mediated transcriptional repression. *Nat Commun,* 6(10195)**,** pp.

Loughran, G., Sachs, M. S., Atkins, J. F. & Ivanov, I. P. 2012. Stringency of start codon selection modulates autoregulation of translation initiation factor eIF5. *Nucleic Acids Research,* 40(7)**,** pp 2898-2906.

Lovett, P. S. & Rogers, E. J. 1996. Ribosome regulation by the nascent peptide. *Microbiol Rev,* 60(2)**,** pp 366-85.

Lytton, J., Westlin, M. & Hanley, M. R. 1991. Thapsigargin inhibits the sarcoplasmic or endoplasmic reticulum Ca-ATPase family of calcium pumps. *J Biol Chem,* 266(26)**,** pp 17067-71.

Mackay, J. P., Kowalski, K., Fox, A. H., Czolij, R., King, G. F. & Crossley, M. 1998. Involvement of the N-finger in the Self-association of GATA-1. *The Journal of Biological Chemistry,* 273(46)**,** pp 30560-7.

Majumdar, R. & Maitra, U. 2005. Regulation of GTP hydrolysis prior to ribosomal AUG selection during eukaryotic translation initiation. *Embo Journal,* 24(21)**,** pp 3737-46.

Marino-Ramirez, L., Kann, M. G., Shoemaker, B. A. & Landsman, D. 2005. Histone structure and nucleosome stability. *Expert Rev Proteomics,* 2(5)**,** pp 719-29.

Marintchev, A., Edmonds, K. A., Marintcheva, B., Hendrickson, E., Oberer, M., Suzuki, C., Herdy, B., Sonenberg, N. & Wagner, G. 2009. Topology and Regulation of the Human eIF4A/4G/4H Helicase Complex in Translation Initiation. *Cell,* 136(3)**,** pp 447-460.

McKeown, S. R. 2014. Defining normoxia, physoxia and hypoxia in tumours-implications for treatment response. *Br J Radiol,* 87(1035)**,** pp 20130676.

Merrick, W. C. 2003. Initiation of protein biosynthesis in eukaryotes. *Biochemistry and Molecular Biology Education,* 31(6)**,** pp 378-385.

Michel, A. M., Fox, G., A, M. K., De Bo, C., O'Connor, P. B., Heaphy, S. M., Mullan, J. P., Donohue, C. A., Higgins, D. G. & Baranov, P. V. 2014. GWIPS-viz: development of a ribo-seq genome browser. *Nucleic Acids Res,* 42(

Montealegre, M. C., Rosa, S. L. L., Roh, J. H., Harvey, B. R. & Murraya, B. E. 2015. The Enterococcus faecalis EbpA Pilus Protein: Attenuation of Expression, Biofilm Formation, and Adherence to Fibrinogen Start with the Rare Initiation Codon ATT. *American Society of Microbiology,* 6(3)**,** pp 467.

Morris, D. R. & Geballe, A. P. 2000. Upstream open reading frames as regulators of mRNA translation. *Molecular and Cellular Biology,* 20(23)**,** pp 8635-8642.

Mueller, P. P. & Hinnebusch, A. G. 1986. Multiple upstream AUG codons mediate translational control of GCN4. *Cell,* 45(2)**,** pp 201-7.

Nakai, K. & Kanehisa, M. 1992. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics,* 14(4)**,** pp 897-911.

Nishioka, K., Chuikov, S., Sarma, K., Erdjument-Bromage, H., Tempst, P. & Reinberg, D. 2002. Set9, a novel histone H3 methyltransferase that facilitates transcription by precluding histone tail modifications required for heterochromatin formation. *Genes & Development,* 16(4)**,** pp 479-89.

O'Connor, P. B., Andreev, D. E. & Baranov, P. V. 2016. Comparative survey of the relative impact of mRNA features on local ribosome profiling read density. *Nat Commun,* 4(7)**,** pp 12915.

Omichinski, J. G., Clore, G. M., Schaad, O., Felsenfeld, G., Trainor, C., Appella, E., Stahl, S. J. & Gronenborn, A. M. 1993. NMR structure of a specific DNA complex of Zn-containing DNA-binding domain structure of GATA-1. *Science,* 261(5120)**,** pp 438-446.

Packham, G., Brimmell, M. & Cleveland, J. L. 1997. Mammalian cells express two differently localized Bag-1 isoforms generated by alternative translation initiation. *Biochem J,* 15(328)**,** pp 807-13.

Passmore, L. A., Schmeing, T. M., Maag, D., Applefield, D. J., Acker, M. G., Algire, M. A., Lorsch, J. R. & Ramakrishnan, V. 2007. The eukaryotic translation initiation factors eIF1 and eIF1A induce an open conformation of the 40S ribosome. *Molecular Cell,* 26(1)**,** pp 41-50.

Patel, J., McLeod, L. E., Vries, R. G. J., Flynn, A., Wang, X. & Proud, C. G. 2002. Cellular stresses profoundly inhibit protein synthesis and modulate the states of phosphorylation of multiple translation factors. *European Journal of Biochemistry,* 269(12)**,** pp 3076-85.

Paulin, F. E. M., Campbell, L. E., O'Brien, K., Loughlin, J. & Proud, C. G. 2001. Eukaryotic translation initiation factor 5 (elF5) acts as a classical GTPase-activator protein. *Current Biology,* 11(1)**,** pp 55-59.

Pavlov, M. Y., Watts, R. E., Tan, Z., Cornish, V. W., Måns Ehrenberg & Forster, A. C. 2008. Slow peptide bond formation by proline and other N-alkylamino acids in translation. *PNAS,* 106(1)**,** pp 50-54.

Peabody, D. S. 1989. Translation initiation at non-AUG triplets in mammalian cells. *Journal of Biological Chemistry,* 264(9)**,** pp 5031-5035.

Pestova, T. V. & Kolupaeva, V. G. 2002. The roles of individual eukaryotic translation initiation factors in ribosomal scanning and initiation codon selection. *Genes & Development,* 16(22)**,** pp 2906-2922.

Pestova, T. V., Lomakin, I. B., Lee, J. H., Choi, S. K., Dever, T. E. & Hellen, C. U. T. 2000. The joining of ribosomal subunits in eukaryotes requires eIF5B. *Nature,* 403(6767)**,** pp 332-335.

Pisarev, A. V., Hellen, C. U. T. & Pestova, T. V. 2007. Recycling of eukaryotic posttermination ribosomal complexes. *Cell,* 131(2)**,** pp 286-299.

Preiss, T. & Hentze, M. W. 1999. From factors to mechanisms: translation and translational control in eukaryotes. *Current Opinion in Genetics & Development,* 9(5)**,** pp 515-521.

Radivojac, P., Vacic, V., Haynes, C., Cocklin, R., Mohan, A., Heyen, J., Goebl, M. & Lakoucheva, L. 2010. Identification, analysis, and prediction of protein ubiquitination sites. *Proteins,* 78(2)**,** pp 365-80.

Reuter, K., Biehl, A., Koch, L. & Helms, V. 2016. PreTIS: A Tool to Predict Non-canonical 5' UTR Translational Initiation Sites in Human and Mouse. *Plos Computational Biology,* 12(10)**,** pp e1005170.

Rogers, G. W., Richter, N. J., Lima, W. F. & Merrick, W. C. 2001. Modulation of the helicase activity of eIF4A by eIF4B, eIF4H, and eIF4F. *Journal of Biological Chemistry,* 276(33)**,** pp 30914-30922.

Rogozin, I. B., Kochetov, A. V., Kondrashov, F. A., Koonin, E. V. & Milanesi, L. 2001. Presence of ATG triplets in 5' untranslated regions of eukaryotic cDNAs correlates with a 'weak' context of the start codon. *Bioinformatics,* 17(10)**,** pp 890-900.

Rozovsky, N., Butterworth, A. C. & Moore, M. J. 2008. Interactions between eIF4AI and its accessory factors eIF4B and eIF4H. *Rna,* 14(10)**,** pp 2136-48.

Sasaki, T., Watanabe, W., Muranishi, Y., Kanamoto, T., Aihara, M., Miyazaki, K., Tamura, H., Saeki, T., Oda, H., Souchelnytskyi, N., Souchelnytskyi, S., Aoyama, H., Honda, Z., Furukawa, T., Mishima, H. K., Kiuchi, Y. & Honda, H. 2009. Elevated Intraocular Pressure, Optic Nerve Atrophy, and Impaired Retinal Development in ODAG Transgenic Mice. *Investigative Ophthalmology & Visual Science,* 50(1)**,** pp 242-248.

Schofield, J. 2016. Control of translation initiation and neuronal subcellular localisation of mRNAs by G-quadruplex structures. *PhD Thesis*.

Sehgal, S. N. 1995. Rapamune (Sirolimus, rapamycin): an overview and mechanism of action. *Ther Drug Monit,* 17(6)**,** pp 660-5.

Sendoel, A., Dunn, J. G., Rodriguez, E. H., Naik, S., Gomez, N. C., Hurwitz, B., Levorse, J., Dill, B. D., Schramek, D., Molina, H., Weissman, J. S. & Fuchs, E. 2017. Translation from unconventional 5′ start sites drives tumour initiation. *Nature,* 541(7638)**,** pp 494-99.

Sheikh, M. S. & Fornace, A. J. 1999. Regulation of translation initiation following stress. *Oncogene,* 18(45)**,** pp 6121-6128.

Shi, Y. & Shi, Y. 2004. Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. 119(7)**,** pp 941-53.

Shi, Y. & Whetstine, J. R. 2007. Dynamic regulation of histone lysine methylation by demethylases. *Molecular Cell,* 25(1)**,** pp 1-14.

Sonenberg, N. & Hinnebusch, A. G. 2009. Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. *Cell,* 136(4)**,** pp 731-745.

Starck, S. R., Tsai, J. C., Chen, K., Shodiya, M., Wang, L., Yahiro, K., Martins-Green, M., Shastri, N. & Walter, P. 2016. Translation from the 5' untranslated region shapes the integrated stress response. *Science,* 351(6272)**,** pp 3867.

Starosta, A. L., Lassak, J., Peil, L., Atkinson, G. C., Virumae, K., Tenson, T., Remme, J., Jung, K. & Wilson, D. N. 2014. Translational stalling at polyproline stretches is modulated by the sequence context upstream of the stall site. *Nucleic Acids Research,* 42(16)**,** pp 10711-9.

Stewart, J. D., Cowan, J. L., Perry, L. S., Coldwell, M. J. & Proud, C. G. 2015. ABC50 mutants modify translation start codon selection. *The Biochemical journal,* 467(2)**,** pp 217-29.

Stratmann, A. & Haendler, B. 2011. The histone demethylase JARID1A regulates progesterone receptor expression. *Febs Journal,* 278(9)**,** pp 1458-1469.

Takahashi, K., Maruyama, M., Tokuzawa, Y., Murakami, M., Oda, Y., Yoshikane, N., Makabe, K. W., Ichisaka, T. & Yamanaka, S. 2005. Evolutionarily conserved non-AUG translation initiation in NAT1/p97/DAP5 (EIF4G2). *Genomics,* 85(3)**,** pp 360-371.

Theis, J. L., Sharpe, K. M., Matsumoto, M. E., Chai, H. S., Nair, A. A., Theis, J. D., de Andrade, M., Wieben, E. D., Michels, V. V. & Olson, T. M. 2011. Homozygosity Mapping and Exome Sequencing Reveal GATAD1 Mutation in Autosomal Recessive Dilated Cardiomyopathy. *Circulation-Cardiovascular Genetics,* 4(6)**,** pp 585-U44.

Thomas, D., Plant, L. D., Wilkens, C. M., McCrossan, Z. A. & Goldstein, S. A. 2008. Alternative translation initiation in rat brain yields K2P2.1 potassium channels permeable to sodium. *Neuron,* 58(6)**,** pp 859-70.

Torres, I. O., Kuchenbecker, K. M., Nnadi, C. I., Fletterick, R. J., Kelly, M. J. S. & Fujimori, D. G. 2015. Histone demethylase KDM5A is regulated by its reader domain through a positive-feedback mechanism. *Nature Communications,* 6(6204)**,** pp.

Toth, Z. E. & Mezey, E. 2007. Simultaneous visualization of multiple antigens with tyramide signal amplification using antibodies from the same species. *J Histochem Cytochem,* 55(6)**,** pp 545-54.

Touriol, C., Bornes, S., Bonnal, S., Audigier, S., Prats, H., Prats, A. C. & Vagner, S. 2003. Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons. *Biology of the Cell,* 95(3-4)**,** pp 169-178.

Tsuruga, T., Kanamoto, T., Kato, T., Yamashita, H., Miyagawa, K. & Mishima, H. K. 2002. Ocular development-associated gene (ODAG), a novel gene highly expressed in ocular development. *Gene,* 290(1-2)**,** pp 125-130.

Tzani, I., Ivanov, I. P., Andreev, D. E., Dmitriev, R. I., Dean, K. A., Baranov, P. V., Atkins, J. F. & Loughran, G. 2016. Systematic analysis of the PTEN 50 leader identifies a major AUU initiated proteoform. *The Royal Society,* 6(5)**,** pp.

van Oevelen, C., Wang, J., Asp, P., Yan, Q., Kaelin, W. G., Jr., Kluger, Y. & Dynlacht, B. D. 2008. A role for mammalian Sin3 in permanent gene silencing. *Mol Cell,* 32(3)**,** pp 359-70.

Van Rechem, C., Black, J. C., Boukhali, M., Aryee, M. J., Graslund, S., Haas, W., Benes, C. H. & Whetstine, J. R. 2015. Lysine Demethylase KDM4A Associates with Translation Machinery and Regulates Protein Synthesis. *Cancer Discovery,* 5(3)**,** pp 255-263.

Varshavsky, A. 2011. The N-end rule pathway and regulation by proteolysis. *Protein Science,* 20(8)**,** pp 1298-1345.

Vermeulen, M., Eberl, H. C., Matarese, F., Marks, H., Denissov, S., Butter, F., Lee, K. K., Olsen, J. V., Hyman, A. A., Stunnenberg, H. G. & Mann, M. 2010. Quantitative Interaction

Proteomics and Genome-wide Profiling of Epigenetic Histone Marks and Their Readers. *Cell,* 142(6)**,** pp 967-980.

Villa, N., Do, A., Hershey, J. W. B. & Fraser, C. S. 2013. Human Eukaryotic Initiation Factor 4G (eIF4G) Protein Binds to eIF3c, -d, and -e to Promote mRNA Recruitment to the Ribosome. *Journal of Biological Chemistry,* 288(46)**,** pp 32932-32940.

Wang, G. G., Song, J., Wang, Z., Dormann, H. L., Casadio, F., Li, H., Luo, J.-L., Patel, D. J. & Allis, C. D. 2009. Haematopoietic malignancies caused by dysregulation of a chromatin-binding PHD finger. *Nature,* 459(7248)**,** pp 847-U6.

Wang, X. M., Flynn, A., Waskiewicz, A. J., Webb, B. L. J., Vries, R. G., Baines, I. A., Cooper, J. A. & Proud, C. G. 1998. The phosphorylation of eukaryotic initiation factor eIF4E in response to phorbol esters, cell stresses, and cytokines is mediated by distinct MAP kinase pathways. *Journal of Biological Chemistry,* 273(16)**,** pp 9373-9377.

Waskiewicz, A. J., Johnson, J. C., Penn, B., Mahalingam, M., Kimball, S. R. & Cooper, J. A. 1999. Phosphorylation of the cap-binding protein eukaryotic translation initiation factor 4E by protein kinase Mnk1 in vivo. *Mol Cell Biol,* 19(3)**,** pp 1871-80.

Wegrzyn, J. L., Drudge, T. M., Valafar, F. & Hook, V. 2008. Bioinformatic analyses of mammalian 5'-UTR sequence properties of mRNAs predicts alternative translation initiation sites. *Bmc Bioinformatics,* 9(232)**,** pp.

Wilkinson, K. A., Merino, E. J. & Weeks, K. M. 2006. Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature Protocols,* 1(3)**,** pp 1610-6.

Will, S., Joshi, T., Hofacker, I. L., Stadler, P. F. & Backofen, R. 2012. LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. *Rna,* 18(5)**,** pp 900-14.

Williams, D. D., Pavitt, G. D. & Proud, C. G. 2001. Characterization of the initiation factor eIF2B and its regulation in Drosophila melanogaster. *Journal of Biological Chemistry,* 276(6)**,** pp 3733-3742.

Woolstenhulme, C. J., Guydosh, N. R., Green, R. & Buskirk1, A. R. 2015. High Precision Analysis of Translational Pausing by Ribosome Profiling in Bacteria Lacking EFP. *Cell Reports,* 11(1)**,** pp 13-21.

Wulff, B. E. & Nishikura, K. 2010. Substitutional A-to-I RNA editing. *Wiley Interdiscip Rev RNA,* 1(1)**,** pp 90-101.

Xia, X. H. 2007. The+4G Site in Kozak Consensus Is Not Related to the Efficiency of Translation Initiation. *Plos One,* 2(2)**,** pp.

Xiaoling Ma, J. L., Brian Brost, Wenjun Cheng, Shi-Wen Jiang 2014. Decreased expression and DNA methylation levels of GATAD1 in preeclamptic placentas. *Cellular Signalling,* 26(5)**,** pp 959-967.

Yan, C. & Boyd, D. D. 2006. Histone H3 acetylation and H3 K4 methylation define distinct chromatin regions permissive for transgene expression. *Mol Cell Biol,* 26(17)**,** pp 6357-71.

Yang, H.-Y. & Evans, T. 1992. Distinct Roles for the Two cGATA-1 Finger Domains. *Molecular & Cellular Biology,* 12(10)**,** pp 4562-70.

Yu, Y., Marintchev, A., Kolupaeva, V. G., Unbehaun, A., Veryasova, T., Lai, S.-C., Hong, P., Wagner, G., Hellen, C. U. T. & Pestova, T. V. 2009. Position of eukaryotic translation initiation factor eIF1A on the 40S ribosomal subunit mapped by directed hydroxyl radical probing. *Nucleic Acids Research,* 37(15)**,** pp 5167-82.

Yuan, Y., Hilliard, G., Ferguson, T. & Millhorn, D. E. 2003. Cobalt Inhibits the Interaction between Hypoxia-inducible Factor-alpha and von Hippel-Lindau Protein by Direct Binding to Hypoxia-inducible Factor-alpha. *Journal of Biological Chemistry,* 278(

Zhou, X., Sun, H., Chen, H. B., Zavadil, J., Kluz, T., Arita, A. & Costa, M. 2010. Hypoxia Induces Trimethylated H3 Lysine 4 by Inhibition of JARID1A Demethylase. *Cancer Research,* 70(10)**,** pp 4214-4221.

Zu, T., Gibbens, B., Doty, N. S., Gomes-Pereira, M., Huguet, A., Stone, M. D., Margolis, J., Peterson, M., Markowski, T. W., Ingram, M. A., Nan, Z., Forster, C., Low, W. C., Schoser, B., Somia, N. V., Clark, H. B., Schmechel, S., Bitterman, P. B., Gourdon, G., Swanson, M. S., Moseley, M. & Ranum, L. P. 2011. Non-ATG-initiated translation directed by microsatellite expansions. *Proc Natl Acad Sci U S A,* 108(1)**,** pp 260-5.

Zubradt, M., Gupta, P., Persad, S., Lambowitz, A. M., Weissman, J. S. & Rouskin, S. 2016. DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nature Methods,* 14(1)**,** pp 75-82.

10. Supplementary Data

10.1 Primers

Primers for GATAD1 mutagenesis – identifying AICs

GATAD1 clone F (part CDS) NheI	TTTTTTGCTAGCGATCCCTTTCCCAGTCCTGCTTCC
GATAD1 clone R (part CDS) XhoI	TTTTTTCTCGAGGGCGGGATCAAATTGGTCTCTGGG
CUG-AUG (-207) F	CGGGCCAGGGAATCATGGCGTCCGC
CUG-AUG (-207) R	GCGGAGGCCATGATTCCCTGGCCCG
CUG-UAC (-207) F	CGGGCCAGGGAATCTACGCCTCCGCCTGCGG
CUG-UAC (-207) R	CCGCAGGCGGAGGCGTAGATTCCCTGGCCCG
CUG-AUG (-144) F	CAGCGCCAGCGGCATGGTCCTTTCACC
CUG-AUG (-144) R	GGTGAAAGGACCATGCCGCTGGCGCTG
CUG-UAC (-144) F	GCAGCGCCAGCGGCTACGTCCTTTCACCGGC
CUG-UAC (-144) R	GCCGGTGAAAGGACGTAGCCGCTGGCGCTGC
AUU-AUG (-45) F	CAGAGACACGGGCATGGCGGACGGTAG
AUU-AUG (-45) R	CTACCGTCCGCCATGCCCGTGTCTCTG
AUU-UAC (-45) F	GGGCTACCGTCCGCCTACCCCGTGTCTCTGCGC
AUU-UAC (-45) R	GCGCAGAGACACGGGGTAGGCGGACGGTAGCCC
GUG-AUG (-39) F	GCGGGCGCAGAGCCATGGGAATGGCGGA
GUG-AUG (-39) R	TCCGCCATTCCCATGGCTCTGGCGCCCGC
GUG-UAC (-39) F	CGTCCGCCATTCCCTACTCTCTGCGCCCGCG
GUG-UAC (-39) R	CGCGGGCGCAGAGAGTAGGGAATGGCGGACG
CUG-AUG (-33) F	CCGCCATTCCCGTGTCTATGGGCCCGCGGG
CUG-AUG (-33) R	CCCGCGGGCCCATAGACACGGGAATGGCGG
CUG-UAC (-33) F	GTCCGCCATTCCCGTGTCTTACCGCCCGCGGG
CUG-UAC (-33) R	CCCGCGGGCGGTTAAGACACGGGAATGGCGGAC
AUG-CUU (+1) F	GAGCCGGCCACCCTTCCGCTGGGCCTG
AUG-CUU (+1) R	CAGGCCCAGCGGAAGGGTGGCCGGCTC
AUG-CUU (+55) F	ACCACGTCGTCCTCCCTTTGGAAGAAGGGAGCG
AUG-CUU (+55) R	CGCTCCCTTCTTCCAAAGGGAGGACGACGTGGT

Primers for GATAD1_3xFLAG CRISPR cloning/screening

CRISPR GATAD1 Guide 9 F (into BbsI)	CACCgTAAAACTGGGTTTCCAGGCC
CRISPR GATAD1 Guide 9 R (into BbsI)	AAACGGCCTGGAAACCCAGTTTTAg
GATAD1 CRISPR screen F	AATGCCAGGTTGCTCCCAGAGTGGT
GATAD1 CRISPR screen R	GCCCTATCCCTTTTAATTTTCTAAG
FLAG CRISPR screen F	gacagccGACTACAAAGACCA

qPCR primers

GATAD1 F (eurofins)	CCCCGTCGCTACTAAAAATAC
GATAD1 R (eurofins)	CTCACTACAACCTCCACCTC
FLAG F eurofins	CAAGGATGACGATGACAAGTAG
FLAG R eurofins	AGGGGCAAACAACAGATGG
KDM5A F	AGCAAGACCAGCAAATAACAG
KDM5A R	ACTACAACCCCACATCCTC
β actin F	CTCGCCTTTGCCGATCC
β actin R	CATCATCCATGGTGAGCTGG
B2M F	CCCCCACTGAAAAAGATGAG
B2M R	ATCCAATCCAAATGCGGC

Primers for mutagenesis of secondary structure surrounding GATAD1 AICs

US 45 AUU F	GCCGACCAGGGAGCAGCCGGGCTACC
US 45 AUU R	GGTAGCCCGGCTGCTCCCTGGTCGGC
DS 45 AUU F	CCCGTGTCTCTGCGACCACGAGGACCACCAGAACCGGCCACCATGCC
DS 45 AUU R	GGCATGGTGGCCGGTTCTGGTGGTCCTCGTGGTCGCAGAGACACGGG
SD 45 AUU F	CCAGGGGGCGGCAGGACTACCATCAGCCATTCCCGTG
SD 45 AUU R	CACGGGAATGGCTGATGGTAGTCCTGCCGCCCCCTGG

Primers for GATAD1 proline-alanine mutants

207 Pro-Ala F	GAACCCGCTTCGCGGCTGCACGGGGCAGC
207 Pro-Ala R	GCTGCCCCGTGCAGCCGCGAAGCGGGTTC
45 Pro-Ala F	CCGCGGGGGGCGGCCGAGCCGG
45 Pro-Ala R	CCGGCTCGGCCGCCCCCCGCGG

Primers for GATAD1 SNP mutagenesis

SNP CUG-CCG (-207) F	CAGGCGGAGGCCGGGATTCCCTGGC
SNP CUG-CCG (-207) R	GCCAGGGAATCCCGGCCTCCGCCTG
SNP AUC-AUG (-210) F	CGGGCCAGGGAATGCTGGCCTCC
SNP AUC-AUG (-210) R	GGAGGCCAGCATTCCCTGGCCCG

Primers for full length GATAD1 cloning

GATAD1 FCS R XhoI	TTTTCTCGAGCAAATGGTTGGCAACTGATTC
GATAD1 -207 CUG F NheI	TTTTGCTAGCGGAATCCTGGCCTCCGCCTGCGGAGCC
GATAD1 -207 AUG F NheI	TTTTGCTAGCGGAATCATGGCCTCCGCCTGCGGAGCC
GATAD1 -45 AUU F NheI	TTTTGCTAGCATTCCCGTGTCTCTGCGCCCGCGGGGG

GATAD1 -45 AUG F NheI	TTTTGCTAGCATGCCCGTGTCTCTGCGCCCGCGGGGG
GATAD1 +1 AUG F NheI	TTTTGCTAGCGGCACCATGCCGCTGGGCCTGAAGCCC
GATAD1 +1 CUG F NheI	TTTTGCTAGCGGCGCCCTGGCGCTGGGCCTGAAGCCC
GATAD1 STOP (NO FLAG) R XhoI	TTTTCTCGAGTTACAAATGGTTGGCAACTGATTCC

Primers for GATAD1 NES mutagenesis

NES -207 L-A 1 F	GGGCGGCCGGGGCACCGTCCGCCA
NES -207 L-A 1 R	TGGCGGACGGTGCCCCGGCCGCCC
NES -207 L-A 1+2 F	CCGCCATTCCCGTGTCTGCGCGCCCGCG
NES -207 L-A 1+2 R	CGCGGGCGCGCAGACACGGGAATGGCGG
NES -45 L-A F	CTAGCATGCCCGTGTCTGCGCGCCCGCG
NES -45 L-A R	CGCGGGCGCGCAGACACGGGCATGCTAG

Primers for KDM5A sequencing

KDM5A seq HaloTag F	TCAGAACGTTTTTATCGAGGGTACG
KDM5A seq HaloTag R	TTCTCTCTACCACAGGGATCTTCAG

Primers for GATAD1/KDM5A NanoBiT cloning and sequencing

C-term NB KDM5A N-frag F NheI	TTTTTTGCTAGCcATGGCGGGCGTGGGGCCGGG
C-term NB KDM5A N-frag R SalI	TTTTTTGTCGACccGGCAGCCAGCCCCACATCTA
C-term NB KDM5A M-frag F NheI	TTTTTTGCTAGCcATGGTCTGCAAAGAATTGACTC
C-term NB KDM5A M-frag R SalI	TTTTTTGTCGACccTGTCAAACACTGCAGGGCCT
C-term NB KDM5A C-frag F NheI	TTTTTTGCTAGCcATGGGTGCTATGAGTTGGCA
C-term NB KDM5A C-frag R SalI	TTTTTTGTCGACccACTGGTCTCTTTAAGATCCT
C-term GATAD1 -207 F NheI	TTTTTTGCTAGCcATGGCCTCCGCCTGCGGAGCCG
C-term NB GATAD1 -45 F NheI	TTTTTTGCTAGCcATGGCCCGTGTCTCTGCGCCC
C-term NB GATAD1 +1 F NheI	TTTTTTGCTAGCcATGGCGCTGGGCCTGAAGCCC
C-term NB GATAD1 R XhoI	TTTTTTCTCGAGccCAAATGGTTGGCAACTGATTC
N-term NB GATAD1 -207 F XhoI	TTTTTTCTCGAGcggtATGGCCTCCGCCTGCGGAGCCG
N-term NB GATAD1 -45 F XhoI	TTTTTTCTCGAGcggtATGGCCCGTGTCTCTGCGCCC
N-term NB GATAD1 +1 F XhoI	TTTTTTCTCGAGcggtATGGCGCTGGGCCTGAAGCCC
N-term NB GATAD1 R NheI	TTTTTTGCTAGCCctaCAAATGGTTGGCAACTGATTC
NanoBiT 1.1C_2.1C_2.1N F	CTTCGCATATTAAGGTGACGCGTGTGGCCT
NanoBiT 1.1N seq F	CGAGTAACCA TCAACAGTGG GAGTTCCGGT
KDM5A 750-1600 seq F	TTGGCAATGGGAACAAAAGATAAAGA
KDM5A 1550-2400 seq F	TGGAGGAGGTGATGAGAGAGCT
KDM5A 2350-3200 seq F	TGTAAAAGAAGCTGAGACCTGTGCT
KDM5A 3175-4070 seq F	GGAGTGGCAAAAATAGGAGGAAAAA
NanoBiT 1.1N_2.1N KDM5A seq R	CGGCCGCCCCGACTCTAGAAGATCTGCTAG

1.1C KDM5A seq R	CACCTCCCTGTTCAAGGACTTGGTCCAGGT
2.1C KDM5A seq R	GGTTTGTCCAAACTCATCAATGTATCTTAT
SV40 R seq	TGTGGTTTGTCCAAACTCATC

Primers for GATAD1 psiCHECK cloning

psiCHECK 3'UTR F XhoI	TTTTCTCGAGAGAATTGGTTGAACCCAGGAG
psiCHECK 3'UTR R NotI	TTTTGCGGCCGCAAGCAGCCAAAAAGCTGTCA
hRenSalI R (for psiCHECK seq)	ccccccgtcgacAGATCCTCACACAAAAAACC