Post-transcriptional regulation of HIV-1 *asp*: Potential control by a series of short open reading frames

Michael Salvatore Barbagallo

Doctor of Philosophy

School of Applied Sciences and Engineering

Monash University
Australia

August 2013

TABLE OF CONTENTS

ABSTRACT v

STATEMENT OF AUTHENTICITY vii

ACKNOWLEDGEMENTS viii

LIST OF ABBREVIATIONS ix

LIST OF TABLES xii

LIST OF FIGURES xiii

CHAPTER 1 – LITERATURE REVIEW 1

1.1 The HIV-1 negative sense transcript asp and its sORFs 1

1.2 Negative sense transcripts 7

1.3 The control of gene expression by sORFs 8

1.4 Splicing and RNA interference 11

1.5 The mechanism of translation and the scanning model of initiation 13

1.6 The control of translation 16

1.6.1 Kozak context 16

1.6.2 Non-AUG initiation codons 16

1.6.3 Leader sequence 17

1.6.4 The poly(A) tail, m^7G cap and 3′-UTR 17

1.6.5 Global control of translation 18

1.6.6 Other factors affecting translation 19

1.6.7 Translational control by sORFs 20

1.7 Alternative models of translational control 21

1.7.1 Leaky scanning 21

1.7.2 Termination reinitiation 24

1.7.3 Internal ribosome entry sites (IRES) 27

1.7.4 Ribosomal shunting 30

1.8 Summary of the effects of sORFs on translation 33

1.9 Hypothesis of this study 34

1.10 Aims of this study 34

CHAPTER 2 – GENERAL MATERIALS AND METHODS 36

2.1 Reagents 36

2.2 Cell culture 36

2.2.1 Bacterial cell culture and media 36

2.2.2 Mammalian cell culture and media 36

2.3 Plasmids 36

2.4 Plasmid amplification 40

2.4.1 Preparation of competent cells 40

2.4.2 Transformation of competent cells 40

2.4.3 Plasmid preparation 41
2.4.4 Plasmid DNA extraction 41

2.5 Analysis of DNA *42*
2.5.1 Agarose gel electrophoresis of DNA 42
2.5.2 Spectrophotometric analysis of DNA 42
2.5.3 Qubit™ Quant-It™ DNA assay 43
2.5.4 Sequencing 43

2.6 Transient transfection of eukaryotic cells *44*
2.6.1 Preparation of eukaryotic cells for transfection 44
2.6.2 Transient transfection of eukaryotic cells by the calcium phosphate method 44
2.6.3 Activation of eukaryotic cells with sodium butyrate and trichostatin A 45

2.7 Reporter gene assays *45*
2.7.1 Preparation of cell lysates 45
2.7.2 Specific activity of enhanced green fluorescent protein 45
2.7.2.1 Fluorescence analysis 45
2.7.2.2 Bradford assay 46

2.8 Data analysis *46*
2.8.1 t-test for the statistical analysis of the means of two populations 46

2.9 Reverse-transcriptase PCR *46*
2.9.1 RNA isolation 47
2.9.2 Analysis of RNA 47
2.9.2.1 Agarose gel electrophoresis of RNA 47
2.9.2.2 Spectrophotometric analysis of RNA 48
2.9.2.3 Qubit™ Quant-It™ RNA assay 48
2.9.3 Synthesis of cDNA 48
2.9.4 RT-PCR conditions 49
2.9.5 Extraction and purification of PCR products 50
2.9.6 Sub-cloning of PCR products 50

2.10 Northern blotting *51*
2.10.1 Generation of DIG-labelled EGFP, RFP and GAPDH probes by PCR 51
2.10.2 Blotting and development of northern blots 52
2.10.3 Autoradiography and densitometry analysis of northern blots 53

CHAPTER 3 – *IN SILICO* ANALYSES OF ASP AND THE sORF REGION **54**

3.1 General introduction *54*

3.2 Aims *55*

3.3 Materials and methods *56*
3.3.1 Conservation of ASP across strains and subtypes of HIV and associated species 56
3.3.2 Investigation of *asp*-like sequences in other retroviruses 59
3.3.3 Conservation of sORFs across strains and subtypes of HIV-1 60
3.3.4 Comparison of codon usage preference within sORF and *asp* coding sequences 60

3.4 Results *61*
3.4.1 Conservation of *asp* and associated sORFs across the atrains and subtypes of HIV and its ancestral SIVs 61
3.4.2 Conservation of ASP amino acid sequence and structural features across the strains and subtypes of HIV and its ancestral SIVs 64
3.4.2 Investigation of *asp*-like ORF in other retroviruses 68

3.4.3 Conservation of sORFs across strains and subtypes of HIV-1 68

3.5 Discussion *73*

3.6 Conclusions *76*

CHAPTER 4 – PRELIMINARY STUDIES OF THE sORF REGION AND THE CONTROL OF DOWNSTREAM GENE EXPRESSION **77**

4.1 General introduction *77*

4.2 Aims *78*

4.3 Materials and methods *79*

4.3.1 Reporter gene constructs 79

4.3.2 Investigation of sORF effects on reporter expression 84

4.3.3 Investigation of transcriptional activators on reporter expression 84

4.3.4 Co-transfection studies with pDsRed-N1 84

4.3.5 Investigation of active sORF initiation codons on reporter expression 85

4.4 Results *86*

4.4.1 Investigation of sORF inhibition 86

4.4.2 Investigation of sORF inhibition in the presence of the transcriptional activators TSA and NaBut. 88

4.4.3 Investigation of transcript abundance 91

4.4.4 Mutational analysis of sORF initiation codons on downstream expression 95

4.5 Discussion *98*

4.6 Conclusions *101*

CHAPTER 5 – CHARACTERISATION OF THE sORF TRANSCRIPT AND THE ROLE OF SPLICING IN GENE EXPRESSION **102**

5.1 General introduction *102*

5.2 Aims *103*

5.3 Materials and methods *104*

5.3.1 Reporter gene constructs 104

5.3.2 Transient transfections and reporter gene assays 109

5.3.3 Transcript analysis 109

5.3.4 Quantitative real-time PCR analysis 109

5.4 Results *111*

5.4.1 Alternative splicing of the sORF region and conservation of splice sites 111

5.4.2 Analysis of Spliced Variants by real-time PCR 124

5.4.3 Effects of mutating the SA 1 and SD 1 motifs 127

5.4.5 Re-examination of sORF effects by mutation of initiation codons 130

5.5 Discussion *136*

5.6 Conclusions *139*

CHAPTER 6 – POTENTIAL FOR THE TRANSLATIONAL CONTROL OF GENE EXPRESSION IN THE HIV-1 *asp* sORF REGION **140**

6.1 General introduction *140*

6.2 Aims *144*

6.3 Materials and methods *145*

6.3.1 RNA secondary structure prediction by MFOLD 145

6.3.2 Plasmid constructs 145
6.3.3 Probes 145
6.3.4 *In vitro* transcription 146
6.3.5 Probe labelling with [γ^{32}P]ATP 146
6.3.6. Toeprinting assay 147
6.3.6.1 Probe hybridisation and in vitro translations 147
6.3.6.2 Primer extension 147
6.3.7 Dideoxy-cycle sequencing reactions 148
6.3.8 Electrophoresis and autoradiography 148

6.4 Results *149*
6.4.1 Prediction of mRNA secondary structures 149
6.4.2 Optimisation of primer extension and ribosomal binding conditions 152
6.4.3 Toeprint analysis of ribosomal stalling at sORF I, II, III, IV, V, VI and VI_{alt} initiation codons 158

6.5 Discussion *173*

6.6 Conclusions *176*

CHAPTER 7 – GENERAL DISCUSSION **177**

7.1 Summary of findings *177*

7.1.1 A potential mechanism for regulation of *asp* expression by upstream sORFs 171

7.2 General discussion *182*

7.3 Concluding remarks *185*

REFERENCES **187**

APPENDIX 1- MULTIPLE SEQUENCE ALIGNMENT OF SPLICED VARIANTS **209**

APPENDIX 2- HIV-1 SUBTYPE B SEQUENCES **210**

APPENDIX 3- MFOLD PREDICTIONS **211**

APPENDIX 4- PUBLICATIONS **213**

ABSTRACT

The positive sense strand of the HIV-1 genome encodes nine different proteins. These include structural proteins (Gag, Pol and Env), regulatory proteins (Tat and Rev) as well as accessory proteins (Vpu, Vpr, Vif and Nef). In addition to the nine positive sense genes, a negative sense gene, *asp*, has been identified opposite in orientation to *env.* Bioinformatic analyses suggest that *asp* encodes a hydrophobic, membrane associated protein of 189 aa. Negative sense transcription, regulated by LTR sequences, has been observed early in HIV-1 infection *in vitro*. However the mechanism of *asp* expression and function of the putative ASP protein still remain unclear. In some viral strains a series of six short open reading frames (sORFs) positioned upstream of the *asp* gene have the potential to regulate *asp* expression. This thesis examines the role of these sORFs in control of expression of downstream genes.

All subtypes of HIV-1 were examined to detect the negative sense *asp* ORF, and to identify potential regulatory sequences. A series of strongly conserved upstream sORFs was identified. The sORF series was particularly well conserved amongst the A, B, C and D clade strains with sORFs I, V and VI being highly conserved across all the subtypes examined. This potential control region from HIV-1NL4-3, containing six sORFs, was cloned upstream of the reporter gene EGFP. Expression by transfection of HEK293 cells indicated that the introduction of this sORF region inhibits EGFP reporter expression; analysis of transcripts revealed no significant change in levels of EGFP mRNA, suggesting that regulation occurs post-transcription. RT-PCR analysis of transcripts further demonstrated that the upstream sORF region undergoes alternative splicing *in vitro*. The most abundant product (Spliced Variant 1) is spliced to remove sORFs I to V, leaving only the in-frame sORF VI. Sequence analysis revealed the presence and high conservation of typical splice donor and acceptor site motifs. Spliced Variant 2, containing sORFs I, II and VI; utilised a lesser well conserved donor in conjunction with the splice acceptor common to Spliced Variant 1. Cloning of Spliced Variant 2 enabled the detection of a third product, Spliced Variant 3. This Spliced Variant (3)

presented an alternate initiation codon for sORF VI, designated VI_{alt}. While Spliced Variants 1 and 2 inhibited, to varying extents, downstream expression; Spliced Variant 3 permitted expression. Mutation of the highly conserved splice donor and acceptor sites modulates, but does not fully relieve, inhibition of reporter EGFP production. These data were further supported by sequential mutation of the sORF initiation codons in which, to varying levels, each mutation alleviated the inhibitory nature of the sORF series, suggesting a translational mechanism for the control of *asp* expression. Toeprinting analysis of the sORF region also revealed the potential for ribosomes to initiate at sORFs I, II, IV, VI and VI_{alt}, yet only weak toeprints were observed for sORF III and sORF V. Initiation at a cryptic CUG codon located 14 nucleotides downstream from sORF VI_{alt} was also detected. These data suggest that the leaky scanning and/or termination reinitiation mechanisms of translation account for the mode of translation across the sORF transcript, and that sORF VI, alone, inhibits downstream translation. Upstream sORFs engage the ribosome, facilitating subsequent initiation downstream at sORF VI. Alternative splicing determines the presence or absence of upstream sORFs, therefore the efficiency of recognition of the sORF VI initiation codon and the degree of inhibition of downstream gene expression.

These findings suggest a complex mechanism, involving both splicing and translational control, modulate *asp* gene expression. The strong conservation of *asp* and its sORFs across all HIV-1 subtypes suggests that the *asp* gene product may have a role in the pathogenesis of HIV-1 and requires tight regulation. This study promotes further examination of the negative sense transcript and its function in the HIV-1 viral life cycle.

STATEMENT OF AUTHENTICITY

I certify that this thesis, except with the Research Graduate School Committee's approval, contains no material which has been accepted for the award of any other degree or diploma in any university or other institution. I affirm that to the best of my knowledge, this thesis contains no material previously published or written by another person, except where due reference is made in the text of this thesis.

Michael S Barbagallo
August 2013

Michael S Barbagallo
August 2013

ACKNOWLEDGEMENTS

The completion of this thesis could not have occurred without the support, guidance and assistance of the following people.

First and foremost to my supervisors Associate Professor Jennifer Mosse and Associate Professor Nicholas Deacon from the School of Applied Sciences and Engineering, Monash University, Gippsland Campus. Your support, ideas, suggestions and ability to act as life mentors I am deeply grateful.

To the team of past and present researchers from the Molecular Biology Laboratory within the School of Applied Sciences and Engineering, Monash University, Gippsland Campus; Mr. Zoon Chan, Ms. Jing Jing Khoo, Ms. Teagan Guanaccia, Ms. Fatin Nabila Shaari and Ms. Jacinta Hansen. Your work alongside mine has made the journey a rewarding experience and the wealth of ideas, support and company we have endured together I am truly thankful. Special thanks to Mrs. Kate Birch, your endless ideas and help particularly in the early stages of this thesis I am grateful.

To the laboratory assistants; Mrs. Mary Lambe-Donelly, Mrs. Catherine Chambers, Mrs. Dorota Adamowicz, Mrs. Rachel Rachiele, Mrs. Justine Barrett, Dr. Ben Webb and Mrs. Lisa Lee your tireless support is acknowledged.

To the friendly staff of the MICROMON sequencing facility at Monash University, Clayton Campus for all the sequencing electrophoresis and to Ms. Louise Carolan and Dr. Karen Laurie from the Victorian Infectious Diseases Reference Laboratory, Melbourne for assistance in conducting the qRT-PCR experiments.

To my dearest family, thank you for your love and support during the entire length of my studies. Thick and thin you were and will always be there for me, I could not have completed any of this without you.

LIST OF ABBREVIATIONS

aa	Amino acid
asp	Antisense gene
ASP	Antisense protein
bp	Base pairs
cDNA	Complimentary deoxyribonucleic acid
CMV	Cauliflower Mosaic Virus
DMEM	Delbecco's Modified Eagle Medium
DNA	Deoxyribonucleic acid
D-PBS	Delbecco's Phosphate Buffered Saline
dsDNA	Double stranded deoxyribonucleic acid
EGFP	Enhanced Green Fluorescent Protein
eIF	Eukaryotic Initiation Factor
FIV	Feline Immunodeficiency Virus
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
gp120	Envelope glycoprotein 120
gp41	Envelope glycoprotein 41
GTP	Guanosine triphosphate
HEK293	Human embryonic kidney 293 cells
HFV	Human foamy virus
HIV-1	Human Immunodeficiency Virus type 1
hnRNA	Heterogeneous nuclear ribonucleic acid
HTLV-1	Human T-lymphotropic virus type 1
IRES	Internal ribosome entry site
kb	Kilo base
LTR	Long terminal repeat
m^7G	7-methylguanosine
MCS	Multiple cloning site
MDDCs	Monocyte-derived dendritic cells
MDMs	Monocyte-derived macrophages
Met-tRNA	Methionyl-transfer ribonucleic acid
miRNA	Micro ribonucleic acid
MLV	Murine Leukemia Virus

MMTV	Mouse Mammary Tumor Virus
M-PMV	Mason-Pfizer Monkey Virus
mRNA	Messenger ribonucleic acid
NaBut	Sodium butyrate
ncRNA	Non-coding ribonucleic acid
NMD	Nonsense mediated decay
nt	Nucleotide
ORF	Open reading frame
PABP	Poly (A) binding protein
PCR	Polymerase chain reaction
PMA	Phorbol myristate acetate
PMBCs	Peripheral blood mononuclear cells
RAR	Retinoic acid receptor
RF	Reading frame
RFP	Red Fluorescent Protein
RISC	Ribonucleic acid-induced silencing complex
RNA	Ribonucleic acid
RNAi	Ribonucleic acid interference
rpm	Revolutions per minute
RT	Reverse transcriptase
SA	Splice acceptor
SD	Splice donor
SE	Standard error
SIV	Simian immunodeficiency virus
SOB	Super optimal broth
SOC	Super optimal medium with catabolite repression
sORF	Short open reading frame
sRNA	Short ribonucleic acid
STE	Sodium Chloride Tris-EDTA
STEs	Stabilizer elements
TAR	Trans-activation response element
TBE	Tris borate EDTA
TE	Tris EDTA
TSA	Trichostatin A

TURBS	Termination upstream ribosomal binding site
UTR	Untranslated region
Wt	Wild type

LIST OF TABLES

Table 2.1	Reverse Transcriptase-PCR primers	49
Table 2.2	PCR primers used to generate DIG-Labelled probes for northern blotting	52
Table 3.1	HIV-1 sequences used in this study	57
Table 3.2	HIV-2 sequences used in this study	58
Table 3.3	SIV sequences used in this study	58
Table 3.4	Other retrovirus sequences used in this study	59
Table 3.5	Conservation of *asp* amongst selected HIV-1 sequences	62
Table 3.6	Conservation of ASP amongst select SIV sequences	63
Table 3.7	Analysis of the number of transmembrane domains present in all ASP amino acid sequences	67
Table 3.8	Analysis of the Kozak initiation strength of each sORF and the percentage homology of the sORF nucleotide sequence to the consensus sequence	70
Table 3.9	Comparison of codon usage frequencies within sORF and *asp* sequences	72
Table 4.1	Site directed mutagenesis primers for the knock-out of each sORF initiation codon	81
Table 4.2	PCR primers used to generate DIG-Labelled RFP probe for Northern blotting	85
Table 5.1	Site directed mutagenesis primers for the knock-out of splicing donor 1 and splicing acceptor 1	105
Table 5.2	In-Fusion™ primers used to generate the various Spliced Variant constructs	105
Table 5.3	TaqMan® probes used for individual Variant TaqMan® quantitative real-time PCR assays	110
Table 6.1	Deoxyoligonucleotide toeprinting probes	145
Table 6.2	Comparison of Gibbs Free Energy between predicted secondary structures	150
Table 6.3	HIV-1 conservation of cryptic CUG codon located downstream of the sORF VI AUG initiation codon	170
Table 6.4	Summary of toeprinting assay data with Kozak context of each sORF	171
Table A.1	HIV-1 B sequences used in this study	210

LIST OF FIGURES

Figure 1.1	The organization of the HIV-1 proviral genome and location of *asp*	6
Figure 1.2	Simplified scheme of translation initiation	15
Figure 1.3	The mechanism of leaky scanning	23
Figure 1.4	The mechanism of termination reinitiation	25
Figure 1.5	The IRES mechanism	29
Figure 1.6	The ribosomal shunt model	31
Figure 2.1	Schematic of base plasmids used throughout this study	38
Figure 2.2	Sequence of the sORF region from HIV-1 NL4-3	39
Figure 3.1	Amino acid sequence alignment of ASP sequences and conservation of sequence motifs	65
Figure 3.2	Hydrophilicity profile of the consensus sequence generated from the multiple sequence alignment with transmembrane region highlighted	66
Figure 4.1	Schematic of base plasmid used throughout this study	80
Figure 4.2	Schematic of plasmids used throughout this study to investigate the effect of upstream sORFs on gene expression	83
Figure 4.3	Effect of HIV-1 *asp* upstream sORFs I-VI on gene expression with shorter and longer intercistronic distances	87
Figure 4.4	The effect of cell activators NaBut and TSA on EGFP expression in the presence of HIV-1 *asp* upstream sORFs I-VI	90
Figure 4.5	Effect of HIV-1 *asp* upstream sORFs I-VI on pDsRed expression in *trans* by co-transfection of reporter constructs with an equal mass of pDsRed-N1	92
Figure 4.6	Effect of HIV-1 *asp* upstream sORFs I-VI and pDsRed expression on EGFP transcript abundance by co-transfection of reporter constructs with an equal mass of pDsRed-N1	93
Figure 4.7	Effect of HIV-1 *asp* upstream sORFs I-VI and EGFP expression on RFP transcript abundance by co-transfection of reporter constructs with an equal mass of pDsRed-N1	94
Figure 4.8	Effect of HIV-1 *asp* upstream sORFs I-VI and mutation of sORF initiation codons on the reporter EGFP	96
Figure 4.9	Effect of HIV-1 *asp* upstream sORFs I-VI and mutation of sORF initiation codons on EGFP transcript abundance	97

Figure 5.1 Schematic of base plasmids used throughout this study to investigate the effect of each sORF on gene expression 106

Figure 5.2 Schematic of base plasmids used throughout this study to investigate the effect of each sORF on gene expression 107

Figure 5.3 Schematic of base plasmids used throughout this study to investigate the effect of each sORF on gene expression within Variant 3 108

Figure 5.4 RT-PCR analysis HIV-1 *asp* upstream sORFs I-VI 113

Figure 5.5 RT-PCR analysis HIV-1 *asp* upstream sORFs I-VI Spliced Variants 114

Figure 5.6 Multiple sequence alignment of unspliced HIV-1 *asp* upstream sORF region and Spliced Variants 1-3 115

Figure 5.7 Summary of splicing events within the sORF region 117

Figure 5.8 Effect of HIV-1 *asp* upstream sORFs I-VI Spliced Variants on the reporter EGFP gene expression, transcript abundance and analysis of Spliced Variants by RT-PCR 120

Figure 5.9 Conservation of splice donor and acceptor motifs in subtype B HIV sequences 123

Figure 5.10 Abundance of HIV-1 *asp* upstream sORFs I-VI Spliced Variants 126

Figure 5.11 Effect of mutating the major splice donor and acceptor motifs within the HIV-1 *asp* upstream sORFs I-VI region on the reporter EGFP gene expression, transcript abundance and analysis of Spliced Variants by RT-PCR 129

Figure 5.12 Re-examination of the effect of unspliced HIV-1 *asp* upstream sORFs I-VI by mutation of sORF initiation codons on the reporter EGFP and transcript abundance 132

Figure 5.13 Examination of the effect of sORFs I-VI Spliced Variant 1, the predominant HIV-1 *asp* upstream sORF configuration, by mutation of sORF initiation codons on the reporter EGFP and transcript abundance 135

Figure 6.1 Schematic of the toeprinting technique 142

Figure 6.2 Secondary structure prediction for sORF IV 151

Figure 6.3 Comparison of ribosomal binding temperatures and primer extension times in toeprinting reactions 155

Figure 6.4 Optimisation of ribosomal binding conditions in toeprinting reactions 157

Figure 6.5 Toeprinting analysis of initiating AUG codons of sORFs I and II 160

Figure 6.6 Toeprinting analysis of initiating AUG of sORF III 162

Figure 6.7 Toeprinting analysis of initiating AUG of sORF IV 164

Figure 6.8 Toeprinting analysis of initiating AUG of sORF V 167

Figure 6.9 Toeprinting analysis of initiating AUGs of sORFs VI and VI_{alt} 169

Figure 6.10 Summary of sequence features and translational characteristics of the sORF region of HIV-1NL4-3 used in the toeprinting assays 172

Figure 7.1 Summary of the characteristics of the sORF region, its associated spliced products and translational events in HIV-1 NL4-3 181

Figure A.1 Secondary structure predictions for sORFs I to V 211

Figure A.2 Secondary structure predictions for sORFs VI and VI_{alt} 212

CHAPTER 1 – LITERATURE REVIEW

1.1 The HIV-1 negative sense transcript *asp* and its sORFs

The positive sense strand of the HIV-1 genome encodes nine different proteins. These include structural proteins (Gag, Pol and Env), regulatory proteins (Tat and Rev) as well as accessory proteins (Vpu, Vpr, Vif and Nef) (Schwartz *et al*., 1992). All of these proteins can be produced from a single RNA transcript by employing a variety of mechanisms. Alternative splicing and regulation of translation have both been reported to control gene expression in HIV-1, ensuring that genes are expressed at the correct time and at the correct level. In the early stages of infection, the completely spliced transcripts (encoding Tat, Rev and Nef) are transported to the cytoplasm for translation. Tat a *trans*-activator, is created by a multiple splicing event. Rev is also a *trans*-activator protein, which facilitates the differential expression of other proteins by control at the post-transcriptional level (Steffy and Wong-Staal, 1991). However, translation of *rev* occurs through reinitiation after translation of *tat* (Steffy and Wong-Staal., 1991). Accumulation of Rev permits the transport of incompletely spliced transcripts (encoding Env and the accessory proteins Vif, Vpr and Vpu) and unspliced transcripts (encoding Gag and Pol) (Stoltzfus, 2009). Transcripts for the structural proteins, Gag and Pol, contain *cis*-acting repressor sequences present in the 3'-UTR that function to block gene expression; however blocking is relieved in the presence of Rev (Cochrane *et al.*, 1991). Env is produced via a discontinuous ribosome-scanning mechanism while Vpu, encoded by the same transcript, is produced via the leaky-scanning model of translation (Anderson *et al*., 2007). Interestingly a minimal sORF, consisting of only a start and stop codon, has been implicated in the regulation of Env expression (Krummheuer *et al*., 2007).

Along with the positive sense genes encoded on the HIV-1 genome, a negative sense open reading frame, *asp,* has been identified opposite *env* (Miller, 1988; Bukrinsky and Etkin, 1990) (Figure 1.1, A). A second antisense gene, located within the LTR, has also been identified in HIV-1(Ludwig *et al*., 2006). Negative strand ORFs opposite *env* have also been described in

HTLV-1 (Larocca *et al.*, 1989), HTLV-2 (Halin *et al.*, 2009) and FIV (Briquet *et al.*, 2001). The negative sense HTLV-I HBZ protein has been shown to modulate viral transcription (Gaudray *et al.*, 2002) and enhance viral persistence (Arnold *et al.*, 2006). Two alternatively spliced variants of the HTLV-I HBZ RNA, which encode different HBZ isoforms, have been characterised (Cavanagh *et al.,* 2006); both isoforms downregulate Tax mediated viral transcription (Lamasson *et al.*, 2007; Yoshida *et al.*, 2008). The antisense protein of HTLV-2, APH-2, also suppresses Tax mediated viral transcription (Halin *et al.*, 2009).

The negative sense ORF, *asp*, is positioned opposite the gp120/gp41 junction of the *env* gene in HIV-1, and has been shown to produce both RNA and protein products (Vanhee-Brossollet *et al*., 1995). The *asp* ORF is highly conserved amongst all strains and subtypes of HIV-1 but absent from HIV-2 sequences (Miller, 1988; Bukrinsky and Etkin, 1990; Briquet and Vaquero, 2002) consistent with observations that endogenously expressed *asp* RNA inhibits the replication of HIV-1 but not HIV-2 (Tagieva and Vaquero, 1997). *In vitro* studies have also indicated that the ASP protein is recognised by antibodies present in the sera of HIV^+ individuals (Vanhee-Brossollet *et al.*, 1995) however the HIV-1 ASP protein has not yet been detected *in vivo*.

Recent *in vitro* studies confirm localisation of ASP to the plasma membrane (Clerc *et al.*, 2011) while earlier studies indicated that the ASP protein is also present in viral particles released from HIV-1 infected cells (Briquet and Vaquero, 2002), suggesting that ASP may play a pivotal role in the life cycle of the virus. Preliminary analysis of the presumed 189 amino acid ASP sequence predicts two highly hydrophobic transmembrane regions, a cysteine rich-region and a proline repeat motif. Similar cys-rich domains in HIV-1 Tat, have been observed to form a zinc finger that promotes the binding of heavy metals and dimerization of the protein (Frankel *et al*., 1988, Greene, 1990), while a similar cys-rich motif functions as a transcription repressor or activator binding site that regulates transcription of the human Oct-4 gene (Nordhoff *et al*., 2001). The proline repeat sequence motif is similar to the PxxP repeat sequence of the HIV-1 Nef protein (Picard *et al*., 2002; Greenway *et al.*, 2003)

and the ORF-3 protein in Hepatitis E virus (Ray *et al*., 1992; Korkaya *et al.*, 2001). In both of these proteins the PxxP region has been shown to interact with cellular protein kinases.

ASP tagged with the FLAG epitope at the C-terminal was detected at 14 days post-infection of both monocyte derived- macrophages (MDMs) and dendritic cells (MDDCs) (Laverdure *et al*., 2012), however this group was not able to detect ASP in activated T lymphocytes. Recently Torresilla *et al*. (2013) confirmed that ASP does not localise to the nucleus and demonstrated that ASP induces autophagy in transfected cells. Using an overexpression vector, this study detected low levels of ASP by western blot early after transfection, followed by the detection of increasingly large ASP aggregates that may trap other cellular proteins and/or induce autophagy. Interestingly these experiments utilised a codon optimised version of ASP, which the authors argue might also disrupt inhibitory sequences that affect the nuclear export of the transcripts and/or the formation of secondary structures that could otherwise inhibit ASP translation (Torresilla *et al*., 2013). However this was not investigated in the expression experiments conducted. HIV-1 has been shown to interact with the autophagic pathway, with different responses in different cell types (Espert and Biard-Piechaczyk, 2009). The study conducted by Torresilla *et al*. (2013) indicates that ASP has an essential role in HIV-1 replication in monocytic cells, where ASP-induced autophagy is associated with increased viral yield. ASP induction of autophagy may also explain why ASP is so difficult to detect *in vivo.*

Early studies conducted by Bukrinsky and Etkin (1990) detected three polyadenylated negative sense *asp* RNA transcripts (1.6, 1.1 and 1.0kb) in acutely infected H9 cells; these transcripts were present early in infection (day 3) but were not detected later (on days 5 or 7). Michael *et al*. (1994b) confirmed the presence of a negative sense transcript in tissue cultures and in PMBCs isolated from HIV-1 infected patients; sequence analysis predicted a full-length transcript of 2.3kb. Landry *et al*. (2007) provided further evidence for negative sense transcription in HIV-1, identifying an alternative poly(A) signal that would produce a 4.1kb transcript. However were unable to detect

this transcript by northern blot. More recently, an extended analysis confirmed the presence of a 2.6kb *asp* transcript in acutely and chronically infected cells, transcribed from the U3 region of the 3′LTR (Kobayashi-Ishihara *et al*., 2012). This study also reported the localisation of the *asp* negative sense transcript in the nuclei of infected cells (MAGIC-5A, OM10.1, PBMC, ACH-2 and Molt-4 cell lines) and found that its expression was able to substantially suppress the replication of HIV-1 for a prolonged period of time. Given that negative sense expression in HIV-1 inhibits positive sense expression, negative sense expression must be strongly controlled, as negative sense transcripts are rarely detected in HIV infected cells (Bukrinsky and Etkin, 1990; Michael *et al*., 1994b; Landry *et al*., 2007).

The mechanism of regulation of *asp* expression remains unclear to date. *In vitro* studies have confirmed that transcription of the negative sense ORF occurs early in infection, controlled by long terminal repeat (LTR) sequences (Pereira *et al.*, 2000; Bentley *et al.*, 2004). Sites that bind the transcription factors, Sp1, NFκB and Lef-1 (Benkirane *et al*., 1998; Kharroubi *et al*., 1998; Pereira *et al*., 2000; de la Fuente *et al*., 2002; Wortman *et al*., 2002) are vital for negative sense transcription. Upregulation by the negative sense promoter is affected by the transcription factors NFκB, Lef-1, RBF1 and RBF2, while the Ets-1, c-Myb and COUP-TF sequences present in the modulatory region do not affect the negative sense promoter (Peeters *et al*., 1996; Bentley *et al*., 2004). The role of the transcriptional activator, Tat, in negative sense transcription is yet to be fully elucidated. Michael *et al*. (1994a) and Bentley *et al*. (2004) both suggested that Tat simultaneously upregulates positive sense transcription and downregulates negative sense transcription, so may ‘switch off’ expression of negative sense genes soon after infection. As more complete transcripts encoding Tat are produced, the binding of TAR increases and negative sense transcription is reduced in favour of positive sense gene expression (Martin and Green, 1992; Jeang *et al*., 1999; Tang *et al*., 1999; Marzio *et al*., 2002). However, studies by Landry *et al*. (2007) suggest that Tat acts to upregulate negative sense transcription. Despite the absence of TAR from these transcripts, Tat has been shown to regulate Sp1 binding, and an Sp1 site has been proposed for negative sense transcription (Peeters *et al*.,

1996). More recently, the notion that Tat was a regulator of antisense transcription was addressed using a dual LTR expression vector (Laverdure *et al*., 2012). This report concluded that negative sense expression is not directly altered by Tat. Their work also established that high levels of negative sense transcript were detected 14 days post infection in MDDC cells and that approximately 75% of these cells did not express the positive sense Gag protein, supporting an inverse correlation between negative sense and positive sense transcription.

In NL4-3 six short open reading frames (sORFs), denoted I to VI, have been identified upstream of the *asp* gene opposite to *env* (Figure 1.1, B). In preliminary experiments, these sORFs were shown to reduce downstream gene expression by approximately 90% in HEK293 and HeLa cells (Yap, Vardarli and Deacon, unpublished results). The intention of this study is to further characterise this sORF region and the potential mechanism by which it could regulate the expression of *asp*. Thus the following sections of this chapter review the possible mechanisms by which sORFs may regulate gene expression.

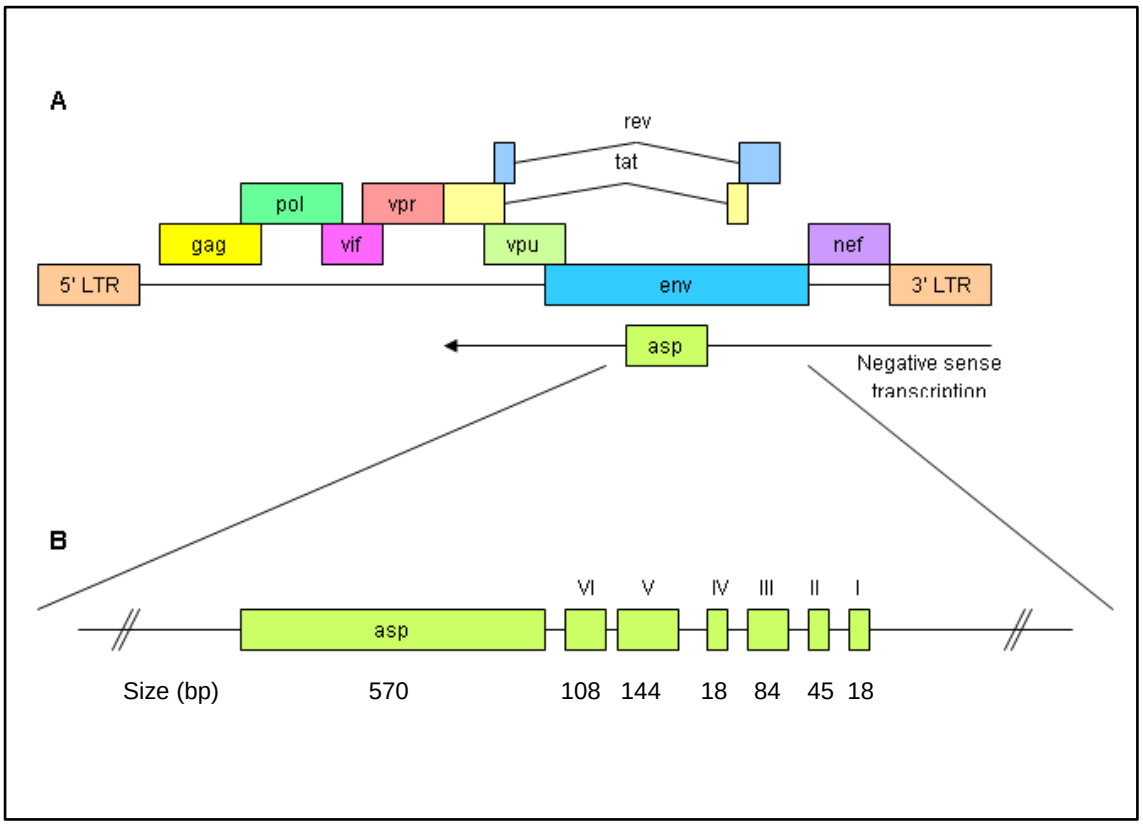


Figure 1.1 The organization of the HIV-1 proviral genome and location of *asp*. (A) The HIV-1 genome, positive sense genes, *gag, pol and env* (structural proteins), *tat* and *rev* (regulatory proteins) and *vpu, vpr, vif, nef* (accessory proteins) indicated. The negative sense gene, *asp* is indicated below, its sequence complementary to part of *env* and spanning the junction of gp120 and gp41. Negative sense transcription is controlled by the 3'-LTR. (B) The HIV-1 NL4-3 sORFs located upstream of *asp* ranging in size from 6 to 48 codons in length, excluding termination codons.

1.2 Negative sense transcripts

Negative sense transcripts have been shown to serve important functions in a number of gene systems. These include the rat Rev-ErbAα, a member of the T3/steroid hormone receptor family (Lazar *et al*., 1990), human ASM-1, complementary to the human c-*myc* proto-oncogene (Celano *et al*., 1992), HBZ from HTLV-1 (Arnold *et al*., 2006; Cavanagh *et al*., 2006; Gaudray *et al*., 2002), and *ebna* from the Epstein-Barr virus (Prang *et al*., 1995). Negative sense transcripts have also been identified in Herpes Simplex Virus 1 (Lagunoff and Roizman, 1994) and Pseudorabies Virus (Vlcek *et al*., 1990).

Negative sense transcripts exist in two forms, *cis*- and *trans*-encoded negative sense transcripts. The *cis*-encoded transcript is transcribed from the same locus as its complementary sense transcript, while the *trans*-encoded negative sense transcript does not originate from the same locus as its sense transcript (Vanhee-Brossollet and Vaquero, 1998). There are two potential functions for the negative sense transcript; it may either be used as a template for translation or may be used in the regulation of the sense gene. In the first instance, the transcript contains an ORF and is transported into the cytoplasm where translation occurs. For example, the Rev-ErbAα gene sequence is complementary to part of the gene encoding, C-ErbA, but is not involved in the regulation of C-ErbA expression (Lazar *et al*., 1989; Lazar *et al*., 1990). In the latter case, the complementarity between the negative sense transcript and the sense transcript may promote hybridisation, thus affecting the expression of the sense transcript (Wagner and Simons, 1994; Kumar and Carmichael, 1998). For example, the basic fibroblast growth factor (bFGF) gene is regulated by a negative sense transcript (Li *et al*., 1996), the negative sense transcript EBNA1 regulates the sense transcript BZLF1 (Prang *et al*., 1995), and the *c-myc* proto-oncogene is regulated by a negative sense transcript (Celano *et al*., 1992).

1.3 The control of gene expression by sORFs

Eukaryotic mRNA transcripts generally produce one functional protein per mature transcript. The initiation of translation in eukaryotes is dependent upon the successful recognition of the 5′ cap, which engages the pre-initiation complex to begin scanning the mRNA until an AUG initiation codon is located (Kozak, 1986). However, mechanisms that enable upstream sORFs to be skipped allow initiation to occur at a downstream ORF. Viral genomes, however, rely upon unique mechanisms to store and express maximum genetic information in minimal space, so often produce more than one functional protein per transcript.

Kozak's original model for the mechanism of translation postulates that the first AUG always begins the open reading frame (ORF) and thus is read and translated in the process. In most cases this is true; however an increasing number of exceptions are reported in the literature. Many pre-mRNAs consist of multiple ORFs, which allow the pre-mRNA to encode multiple proteins via the use of different reading frames and alternative splicing (Kozak, 1978; Kos *et al*., 2002). The occurrence of short upstream open reading frames (sORFs) or upstream AUGs (uAUGs) has been shown to affect the efficiency of translation in many examples. These include the suppressor of cytokine signalling 1 protein (SOCS-1) (Schluter *et al*., 2000), human β_1-adrenergic receptor gene (Zimmer *et al*., 1994; Evanko *et al*., 1998), cathelicidin (Wu *et al*., 2002), HER-2 (*neu*, *erb*B-2) receptor (Child *et al*., 1999a), the human major vault protein (Holzmann *et al*., 2001), the huntingtin gene (Lee *et al*., 2002), rat V_{1b} vasopressin receptor (Nomura *et al*., 2001), CCAAT/enhancer-binding protein α and β (Lincoln *et al*., 1998), *bcl-2* (Harigai *et al*., 1996), the corticotropin releasing hormone type 1 receptor (Xu *et al*., 2001), S-adenosylmethionine decarboxylase (Mize *et al*., 1998), the *lck* proto-oncogene (Marth *et al*., 1988), and the HER-2 oncogene (Child *et al*., 1999b).

Functional roles for short RNAs have been demonstrated in eukaryotes. These RNAs, typically <200nt long, may silence gene expression by targeting specific gene sequences (Brosnan and Voinnet, 2009). These transcripts may represent non-protein coding functions, including vital regulatory roles

(Mattick, 2005). It is estimated that 15 to 53% of 5'-UTRs (depending upon the organism) contain sORFs (Pesole *et al.*, 2000; Rogozin *et al*., 2001). The majority of sORFs identified to date are regulatory and do not appear to express functional protein products. Genes controlled by sORFs include proto-oncogenes, transcription factors, DNA binding proteins, receptors, immune and inflammation mediators, genes involved in signal transduction and growth factors (Kozak, 1991a). Several sORFs occur in the viral leader sequences of CMV (Pooggin *et al*., 2000), the rice tungro bacilliform virus (Futterer *et al*., 1996), and the avian sarcoma-leukosis retroviruses (Moustakas *et al*., 1993; Donze *et al*., 1995).

A host of post-transcriptional processing events are responsible for ensuring the nascent transcript is prepared for translation in the cytoplasm. Initially the pre-mRNA must be capped in order to define the boundary of the initial exon. In this process the enzyme guanylyl transferase adds GTP, while guanine-7-methyl transferase catalyses the addition of a methyl group. The transcript is cleaved at the 3´ end, allowing the addition of the poly(A) tail by poly(A) polymerase. This permits the removal of the 3´ intron, transport of the transcript to the cytoplasm and stabilizing the transcript via resistance to degradation by 5´ exonucleases. The cap-binding protein eIF4E is crucial for translation initiation. The binding of eIF4E to the 7 methylguanosine (m^7G) cap facilitates the recruitment of other translation initiation factors and the 40S ribosomal subunit (Scheper and Proud, 2002). In the nucleus, eIF4E promotes the transport of the transcript to the cytoplasm (Strudwick and Borden, 2002; Mangus *et al*., 2003).

In some instances the peptide product produced by the translating ribosome, alone or in conjunction with another cellular component or environmental agent, may have an effect on the function of the translating ribosome (Lovett and Rogers, 1996). The encoded peptide may cause the ribosome to stall and thus block further scanning of the transcript. The peptide may also affect the elongation or termination phases of translation by interacting with the peptide release factor or peptidyltransferase centre. If the peptide release factor is blocked by the peptide, the peptide chain itself cannot be released and thus

the ribosome stalls. However if the peptidyltransferase centre is blocked during elongation, the ribosome will also stall mid-way through translation of the ORF (Janzen *et al*., 2002; Raney *et al*., 2002). A number of examples of translation inhibition by a sORF encoded peptide have been reported and include the well-studied human cytomegalovirus gpUL4 (Schleiss *et al*., 1991; Degnin *et al*., 1993; Cao and Geballe, 1994, Cao and Geballe, 1995; Alderete *et al*., 2001), β2.7 (Bergamini *et al*., 1998), mouse RARβ2 (Reynolds *et al*., 1996), mouse vigilin (Rohwedel *et al*., 2003), the ubiquitously expressed mammalian CHOP (Jousse *et al*., 2001), the yeast CPA1 gene (Werner *et al*., 1987; Delbecq *et al*., 1994) and more recently the sORF of CDKN1A (Kim *et al*., 2012).

The stability of the mRNA transcript can vary quite dramatically in different cell types and under different states (Linz *et al*., 1997). The nonsense mediated decay (NMD) pathway is one mechanism by which the level of intact mRNA transcripts in the cell can be regulated. In this process, mRNA transcripts are removed from the mRNA pool by the selection of transcripts that prematurely terminate translation, due to either a frame shift or nonsense mutation. NMD also directly regulates sub-sets of normal transcripts, for example those involved in brain development and function (Ngyuyen *et al*., 2012). This pathway has been found to regulate the expression of approximately 10 to 20% of the transcriptome (Conti and Izaurralde, 2005; Mata *et al*., 2005; Weischenfeldt *et al*., 2012).

Some research has suggested that the presence of sORFs in the transcript can play a vital role in the regulation of the stability and thus the turnover of the mRNA by NMD (Ruiz-Echevarria and Peltz, 2000). Transcripts containing mutations that have created a ‘new’ sORF that was not originally present in the natural transcript become more susceptible to degradation by the NMD pathway (Ruiz-Echevarria *et al*., 1998; Baker and Parker, 2004).

In contrast, mRNA transcripts containing natural sORFs are not affected by the NMD pathway, as these sequences contain ‘stabilizer elements’ (STEs) that inhibit NMD. STE sequences have been identified in a number of sORF

systems including; yeast *GCN4* sORF4 (Oliveira and McCarthy, 1995; Ruiz-Echevarria and Peltz, 2000), and yeast YAP2 sORFs 1 and 2 (Vilela *et al*., 1999). The STEs interact with Pub1p, an RNA binding protein that prevents de-capping of the transcript, essential for subsequent degradation via the NMD pathway (Ruiz-Echevarria and Peltz, 2000). The Rous sarcoma virus also contains stability elements immediately downstream of the *gag, pol and src* sequences that mask the termination codons from the NMD machinery (Withers and Beemon, 2010).

1.4 Splicing and RNA interference

It is believed that up to 94% of human genes are spliced, with particular emphasis on certain types of genes (Ward and Cooper, 2010), including cell surface receptors and others required in the nervous and immune systems. Splicing events may also create new sORFs within viral leader sequences, as has been reported in CMV (Kiss-Laszlo *et al*., 1995).

Following transcription by polymerase II, splicing may remove non-coding sequences (introns) from the nascent mRNA transcript and fuse the protein coding sequences (exons) together. The sequence that directs the removal of the intron includes the first two residues (GU) of the upstream intron (splice donor) and the last two residues (AG) of the downstream intron (splice acceptor). Consensus sequences include a polypryimidine tract immediately upstream of the splice acceptor site and an A residue at the branch point of the intron, located approximately 30 nucleotides upstream of the 3´ splice position (Coolidge *et al*., 1997). During splicing, the spliceosome directs formation of an unusual 2´-5´ bond between the upstream exon and the A residue, resulting in the lariat loop. This allows the 5´ end of the downstream exon to attack the 3´ splice site in order to bring together the up- and downstream exons, effectively removing the intron (Smith *et al*., 1989; Smith and Valcarcel, 2000).

The introns removed from the transcript may be rapidly degraded by exonucleases. However, the discovery of RNA interference (RNAi) indicates

that some of these small microRNA fragments (miRNAs) may base pair with complementary sequences thus preventing gene expression (Aravin and Tuschl, 2005; Sullivan and Ganem, 2005). Translation is blocked and the mRNA transcript may be degraded after being incorporated into an RNA-induced silencing complex (RISC) (Filipowicz *et al*., 2005). A host of reports have shown the inhibition of gene expression by RNAi in HIV-1, with five potential miRNAs being predicted for the virus (Bennasser *et al*., 2004; Lee and Rossi, 2004; Yeung *et al*., 2005; Provost *et al*., 2006). In rare circumstances miRNAs have been shown to upregulate translation. For example the miR-122 miRNA promotes translation of the Hepatitis C virus via direct interaction with the 5′-UTR; this response is cell specific (Henke *et al*., 2008; Goergen and Niepmann, 2012).

Alternative splicing sees the use of different splice donor and acceptor sites. For example if exons are termed 1-2-3-4-5, some of the possibilities that could arise include 1-2-5, 2-4-5, 1-3-5 and so on. However the order of exons is always retained, so arrangements such as 2-5-3 are not possible. Examples of alternative splicing are many and include removal of *gag* and *pol* sequences from the *env* mRNA and the production of the regulatory proteins Tat and Rev in HIV-1 (Purcell and Martin, 1993). Similarly, the alternative splicing within the mouse retinoic acid receptor β produces three isoforms with different 5′-UTR sequences (Zelent *et al*., 1991).

Riboswitches have been shown to control bacterial gene expression levels. More than 20 different riboswitch classes have been identified in prokaryotes, yet their presence in eukaryotes is limited (Wachter, 2010). To date, the only riboswitches identified in eukaroytes modulate splicing activity within filamentous fungi and plants. One such example is the splicing control of *NMT1* mRNA which encodes a protein involved in TPP metabolism in the filamentous fungus *N.crassa* (Cheah *et al*., 2007). In this example, three different riboswitches act to regulate gene expression, one activates, while the remaining two repress expression. In this example a riboswitch senses TPP and is able to modulate transcript products with alternate 3′ UTR lengths, ultimately leading to variations in protein production (Wachter *et al*., 2007).

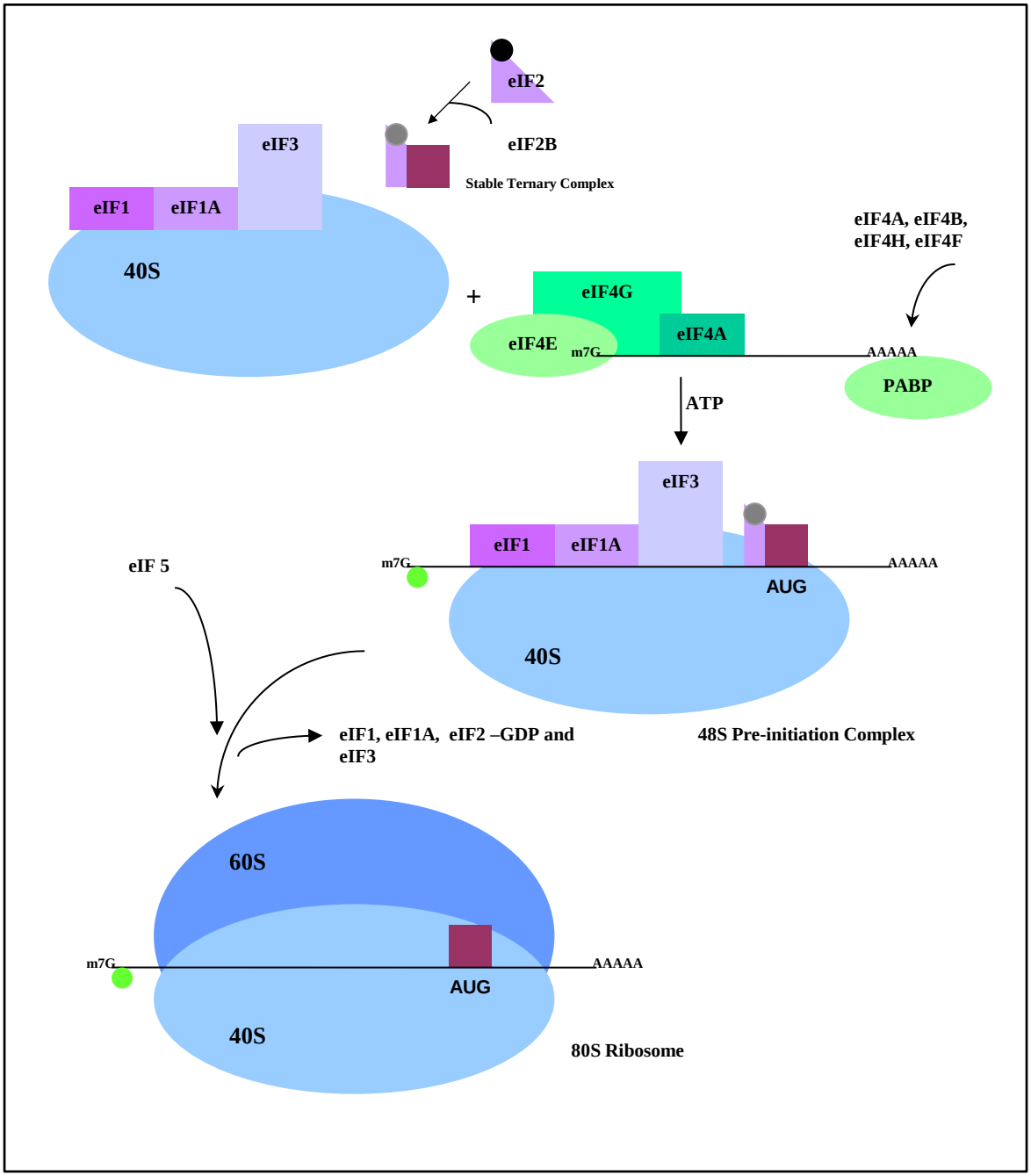
1.5 The mechanism of translation and the scanning model of initiation

Translation is tightly regulated at a series of discrete stages, all of which may affect the levels of gene expression; mRNA level secondary structures and/or upstream open reading frames may also affect translation. A large body of work has been devoted to the study of these control mechanisms, and our understanding is increasing rapidly in this area.

The scanning model of translation begins with the eukaryotic initiation factor, eIF2, binding to the initiating Met-tRNA in a GTP dependent manner to form a stable ternary complex (Figure 1.2). This complex, along with the initiation factors eIF1, eIF1A, and eIF3, then binds with the 40S ribosomal subunit (Preiss and Hentze, 2003). The 7-methylguanosine (m^7G) cap-binding protein (eIF4E), with initiation factors eIF4G and eIF4A, recognises the mRNA m^7G cap to form the cap-binding complex, otherwise known as eIF4F. The 3′ poly(A) tail is bound by the poly(A)-binding protein (PABP), allowing the 40S complex to bind to the mRNA transcript (Dever, 1999). The eIF4A, together with eIF4B and eIF4F, unwinds any secondary structures located within the 5′UTR, while the eIF4G and PABP enhance binding of eIF4E to the m^7G cap.

Kozak's model for translation then supposes that the 40S ribosome scans the 5'-untranslated region (5'-UTR) until the first initiation codon (AUG) is reached, where it stops to form the 48S pre-initiation complex (Kozak, 1991b). The 40S subunit has been reported to be capable of scanning distances as large as 1700 nucleotides without loss in efficiency (Berthelot *et al*., 2004). Once the start codon is located, the initiation factor eIF5 interacts with the pre-initiation complex, stimulating eIF2-GTP hydrolysis and releasing eIF2 and other eIFs. This action permits the attachment of the 60S ribosomal subunit to form the functional 80S ribosomal complex (Calkhoven *et al*., 2002). While the AUG triplet serves as the initiation codon of the ORF in the majority of cases, it is important to note that non-AUG initiation codons may be used some rare instances (Boeck and Kalakofsky, 1994).

Figure 1.2 Simplified scheme of translation initiation. The process of translation begins with eIF2 (triangle) binding to the initiator, Met-tRNA (maroon box) in a GTP dependent manner (black circle). The initiation factors; eIF1, eIF1A, and eIF3 also bind with the 40S ribosomal subunit. The m^7G cap-binding protein (eIF4E) along with eIF4G and eIF4A form the cap-binding complex (green circle). The 5′-cap of the mRNA is recognized by the eIF4E (a subunit of eIF4F) and the 3′ poly(A) tail is bound by the poly(A)-binding protein (PABP). The eIF4G (a subunit of eIF4F) enables the 43S complex to bind to the mRNA transcript (black line). The eIF4A, together with eIF4B and eIF4F unwinds secondary structures present within the 5′UTR, while the eIF4G and PABP enhance the binding of eIF4E to the m^7G cap. The ribosome scans the 5'-UTR for the first AUG (thereby forming the 48S pre-initiation complex). The initiation factor, eIF5 stimulates eIF2-GTP hydrolysis to release eIF2 and other eIFs to permit the joining of the 60S ribosomal subunit, forming the 80S ribosomal complex (Adapted from Dever, 1999; Myasnikov *et al*., 2009).


eIF2
eIF2B
Stable Ternary Complex
eIF3
eIF1
eIF1A
40S
+
eIF4A, eIF4B,
eIF4H, eIF4F
eIF4G
eIF4E
m7G
eIF4A
AAAAA
PABP
ATP
eIF3
eIF1
eIF1A
m7G
AAAAA
AUG
40S
eIF 5
eIF1, eIF1A, eIF2 –GDP and
eIF3
48S Pre-initiation Complex
60S
m7G
AAAAA
AUG
40S
80S Ribosome

1.6 The control of translation

1.6.1 Kozak context

The efficiency of recognition of the first AUG codon by the scanning ribosome is affected by the sequences immediately surrounding the codon. Analysis of 699 vertebrate mRNAs revealed a consensus sequence (now known as 'Kozak's consensus') for the initiation of translation to be GCC**R**CCAUG**G** (Kozak, 1984b, Kozak, 1987a). The most highly conserved nucleotides include: the R (either A or G) at position -3 (where A of the initiation codon is assigned +1) and the G at position +4. A sequence including one of these important nucleotides is considered adequate, while a sequence displaying both is considered strong. It is important to note that a pyrimidine substitution at position -3 also sensitises translation to changes in positions -1, -2 and +4 of the consensus (Kozak, 1986). Further work has progressed to link one or two copies of GCC preceding the -3 position with improved efficiency of initiation (Kozak, 1987b). More recent analysis has revealed that there may be potential elements within the 30 nucleotides upstream of the AUG codon which also affect the initiation process (Shabalina *et al*., 2004) and much work is required to identify these elements, some of which are discussed in the following.

1.6.2 Non-AUG initiation codons

Translation may also be initiated at non-AUG codons (Peabody, 1989; Kozak, 2002) including CUG, ACG and GUG (Kozak, 2002). Generally the use of non-AUG codons occurs when only one nucleotide within the triplet is different and the strong Kozak context is maintained. Positions +4, +5 and +6 consisting of the bases, G, A and U respectively in the consensus, produce a strong, non-AUG initiation codon (Boeck and Kalakofsky, 1994; Grunert and Jackson, 1994). Factors that directly affect the eIF2 or eIF5 initiation factors increase the use of non-AUG codons. An increase in hydrolysis of GTP bound to eIF2 of the 48S pre-initation complex, has been shown to affect the stringency in start site selection in *S. cerevisiae* (Huang *et al*., 1997). The loss of eIF1 from the 48S pre-initiation complex has also been observed to affect

the capacity for AUG discrimination (Pestova and Kolupaeva, 2002). The use of non-AUG triplets may serve useful functions in the regulation of gene expression. For example, the generation of multiple isoforms by the recognition of different start sites has previously been reported for fibroblast growth factor 2 (FGF2) and vascular endothelial growth factor (VEGF), both encoded on the same mRNA (Touriol *et al*., 2003).

1.6.3 Leader sequence

The 48S pre-initiation complex spans a region of 27 nucleotides around the AUG initiation codon (Lazarowitz and Robertson, 1977). As a result the length of the leader sequence may play a significant role in the regulation of gene expression. A short leader sequence would be expected to lower the stability of the pre-initiation complex with the mRNA transcript. Leader sequences of less than 8-12 nucleotides have been reported to result in leaky-scanning of the transcript, where the initiation codon is occasionally skipped; in contrast a longer sequence increases the efficiency of ribosomal initiation (Kozak, 1991c). This has been observed in the satellite tobacco necrosis virus (STNV), where mRNA transcripts in which the 600 nt leader sequences were deleted were translated less efficiently than their full length counterparts (Danthinne *et al*., 1993).

1.6.4 The poly(A) tail, m^7G cap and 3′-UTR

The 3′-UTR, poly(A) tail and the m^7G cap present on eukaryotic mRNAs play vital roles in the efficiency of translation. For instance the poly(A) tail and m^7G cap may both enhance the translation of the mRNA by the joining of eIF4E to the m^7G cap, stimulating the attachment of the 60S ribosomal subunit (Borman *et al*., 2000; Kahvejian *et al*., 2005). This enhancement is brought about by the circularization of the mRNA transcript via the interaction of the poly(A) binding protein (PABP) factor with the poly(A) tail (Imataka *et al*., 1998; Derry *et al*., 2006).

Despite the importance of the poly(A) tail and m^7G cap in the initiation of translation, there are some instances where sequences lacking these features may still be efficiently translated. Many of these examples include viral mRNAs. One such example is the Dengue transcript, a positive sense RNA virus of the *Favivirus* genus which does not possess a poly(A) tail. In this instance the PABP is able to interact with a stem-loop structure within the 3′-UTR, allowing circularisation of the transcript for the initiation of translation (Polacek *et al*., 2009; Friebe and Harris, 2010). Another novel mechanism has been reported in the Barley yellow dwarf virus, a positive sense RNA virus whose mRNA is both uncapped and lacking a poly(A) tail. In this example, the transcript contains a translational enhancer and complementary stem-loop structures in the 3′ and 5′-UTRs; the enhancer region recruits eIF4F which can then promote movement from the 3′ to the 5′-UTR by complementary stem-loop circularisation, thus the formation of the 48S pre-initiation complex for translation (Treder *et al*., 2008).

1.6.5 Global control of translation

Global or whole cell control can occur as a direct result of cell state or environment. For example phosphorylation events specifically involving the translation initiation factors, or their regulating partners, may subsequently affect the translational activity within the cell. For instance, the initiation factor eIF2α is responsible for the conversion of inactive eIF2-GDP to the active eIF2-GTP. When eIF2α is phosphorylated under cell stress, the lack of active eIF2-GTP prevents pre-initiation complex formation (Harding *et al*., 2000). In contrast the phosphorylation of eIF4E in the herpes simplex virus type 1 does not reduce translational gene expression but is crucial for successful gene expression (Walsh and Mohr, 2004). This effect has also been observed in synthesis of the transforming growth factor β (TGFβ) protein, where blockage of eIF4E phosphorylation subsequently prevents expression and thus also prevents mesangial cell hypertrophy (Das *et al*., 2013).

1.6.6 Other factors affecting translation

Other factors that may affect gene expression at the level of translation include the blockage of the scanning ribosomal subunit by RNA-binding proteins, such as the spliceosomal human U1A proteins that may form RNA-protein complexes in the 5′-UTR to block the translation initiation process (Stripecke *et al*., 1994). The polypyrimidine tract-binding protein binds to both the 5′ and 3′ ends of the Hepatitis C virus; this binding has been shown to inhibit the translational machinery (Ito and Lai, 1999). The sex-lethal protein (SXL) has been shown to bind to an upstream sORF on the *msl-2* mRNA in Drosophilia, which subsequently prevents the scanning ribosome from reaching the downstream ORF (Medenbach *et al*., 2011). CUGBP1 and calreticulin RNA-binding proteins compete for the same site on the p21 mRNA; the binding of CUGBP1 increases the translation of p21 by recruitment of eIF2 (Timchenko *et al*., 2006), while binding of calreticulin has the opposite effect and blocks translational machinery via secondary structure stabilisation within the 5′-UTR of the transcript (Iakova *et al*., 2004).

Viral genomes may also use recoding events, including frame shifting and read-through, to extend the coding capacity of their small genomes. In the first instance (frame shifting) the scanning ribosome may slip forward or backward by one nucleotide, thus changing the reading frame of the translating ribosome (Atkins and Gesteland, 1996). It is believed that the efficiency of frame shifting may be enhanced by the presence of either a pseudoknot or hairpin downstream, which causes the ribosome to pause or slowdown at the shift site (Mathews, 1996). In the case of read-through, a translating ribosome may pass through a stop codon and continue translation. This mechanism is used by many viruses to generate a carboxy-terminally extended protein (Mathews, 1996).

1.6.7 Translational control by sORFs

The sORF, as previously mentioned, is a small open reading frame located upstream of the main start codon. An investigation of 4870 human genes found that approximately half displayed sORFs in the 5'-UTR (Yamashita *et al*., 2003). One well-characterized example of sORF-controlled gene expression is that of yeast *GCN4*. In this example, a series of four sORFs in the leader of the *GCN4* mRNA, inhibit *GCN4* expression in a termination-reinitiation manner under non-starved conditions (Mueller and Hinnebusch, 1986; Hinnebusch, 1997). However, under conditions of starvation, the resulting phosphorylation of eIFs reduces the number of complexes able to reinitiate translation after termination of sORF 1. This results in the scanning of sORFs 2 to 4 by the pre-initiation complex. The functional ribosomal complex may, therefore, reinitiate while scanning between sORF 4 and the *GCN4* initiation codon, bypassing the inhibitory effects of sORFs 2 to 4 (Gaba *et al*., 2001).

There are three possible ways in which a sORF can be organised with respect to the main ORF; these include: (1) the non-overlapping sORF, (2) the overlapping sORF and (3) the in-frame sORF, with no intercistronic region between the sORF stop codon and the start codon of the main ORF (Geballe, 1996). In the first two instances the sORFs are differentiated from the main ORF via the position of the termination codon relative to the initiation codon of the main ORF. In both of these cases any product (if produced) from the sORF will be completely different to that of the main ORF (Geballe, 1996). The in-frame sORF may result in two different products being created, depending upon which AUG initiation codon is utilized (either that of the sORF of that of the main ORF). Expression of these bi-functional genes can be achieved by two mechanisms; when transcription is initiated from two alternative promoters or through the process of leaky scanning (Kozak, 1991b). In the first instance, two different mRNA transcripts can be produced, initiating at each AUG initiation codon. Examples of these genes include Invertase (SUC2, yeast) (Carlson *et al*., 1983) and α-Isopropylmalate synthase (LEU4, yeast) (Belzer *et al*., 1988). In leaky scanning, the ribosome

may miss either the first or second AUG initiation codon to create long and short isoforms. Genes that produce proteins via this mechanism, include *vpu* and *env* from HIV-1 (Schwartz *et al*., 1990) and *pre-S* and $p24^S$ from the Hepatitis B virus (Persing *et al*., 1985).

1.7 Alternative models of translational control

With increased knowledge of translation and mechanisms of gene regulation, Kozak's original model for the mechanism of translational control was reviewed (Kozak, 1989b). Two of Kozak's original hypotheses on the translation mechanism still stand today (Wang and Rothnagel, 2004). These include: (1) Leaky scanning, where the ribosome may not recognise, and thus bypass, a sORF to initiate at the downstream AUG, and (2) Termination reinitiation whereby, after termination from the sORF, the ribosome may remain associated with the mRNA and reinitiate at a downstream ORF (Kozak, 1984a). In the following sections some alternative models of translational control are considered, many of which are the subject of ongoing study, particularly amongst the highly structured viral 5′-UTRs (Gale *et al*., 2000).

1.7.1 Leaky scanning

The majority of AUG codons (95-97%) initiating the main ORF of an mRNA display a strong or adequate sequence context. The frequency of these strong-adequate context AUG codons drops to 43-63% amongst sORFs (Suzuki *et al*., 2000; Meijer and Thomas, 2002). Experimental data shows that sORFs with weak sequence context are often "missed" by the ribosome and thus permit the translation of the main (or downstream) ORF (Figure 1.3). This 'by-pass' or leaky scanning results in an increased level of translation initiating at the main ORF. Wang and Rothnagel (2004) also showed that when the sORFs are in strong sequence context, a significant level of initiation can occur at the main ORF when the leader sequences are lengthy (>94 nt) and do not contain any downstream secondary structures. Similarly if the main ORF is mutated to an adequate sequence context (rather than a strong

context), the translation of the main ORF is reduced to the point where only the sORFs themselves are translated (Futterer *et al*., 1997; Gaba *et al.*, 2001).

It is important to note that factors other than sequence context may result in leaky scanning. For example, the presence of downstream secondary structures may improve the efficiency of initiation at a weak AUG initiation codon by slowing the ribosome (Kozak, 1990). Initiation at the second AUG initiation codon may also be blocked by ribosomes that have stalled at an upstream AUG, while the first AUG initiation codon may also be missed if it is situated too close to the 5' end of a preceding ORF to be recognised by the ribosome efficiently (Sedman et al., 1990; Kozak, 1991b; Kozak, 1995). Translation of an upstream sORF may stall the ribosome, preventing it initiating at the downstream ORF (Wang et al., 1998). In these circumstances two (or more) in-frame AUG codons located extremely close together may produce multiple proteins from the same mRNA transcript. This has been observed with the translation of the NA and NB glycoproteins of influenza B virus, situated only 4 nucleotides apart (Williams and Lamb, 1989) and both p69 and p206 located only 7 nucleotides apart on the turnip yellow mosaic virus transcript (Matsuda and Dreher, 2006).

Leaky scanning (especially in higher eukaryotes) may also occur in situations where non-AUG codons are utilized. There are many instances in which non-AUG start codons are utilized to produce two forms of a protein (by initiating at the non-AUG or the first AUG) (Kozak, 1999). Some of these include; the mouse Krox-24 gene (Lemaire et al., 1990), the mouse int-2 gene (Acland et al., 1990) and the MuLV cell surface antigen (Prats et al., 1989).

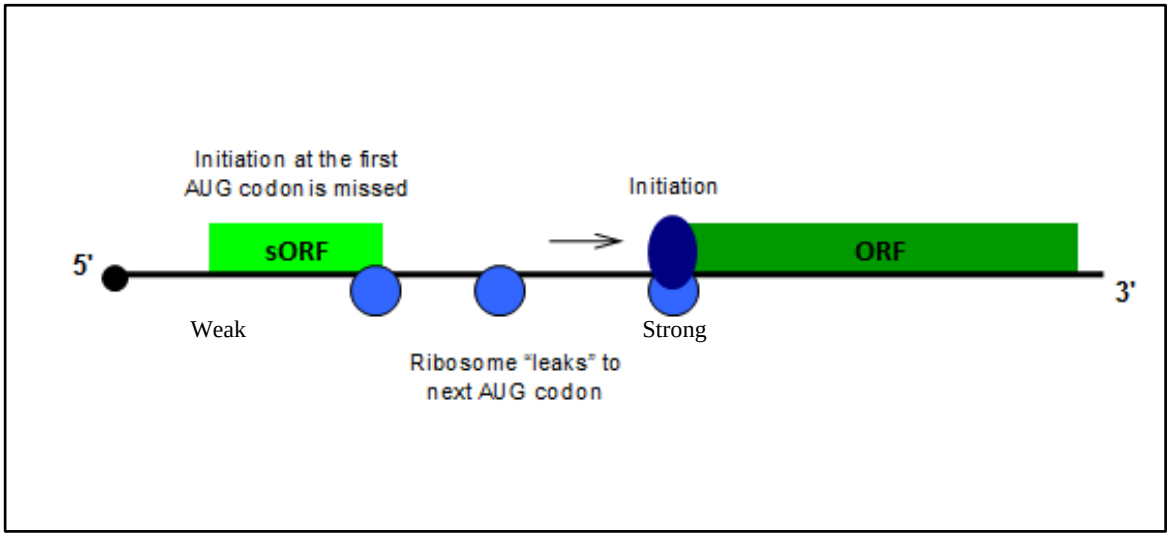


Figure 1.3 The mechanism of leaky scanning. In this instance some ribosomes may initiate at the first AUG (weak context), however due to its stronger context most ribosomes bypass the first AUG to initiate at a downstream AUG (strong context).

1.7.2 Termination reinitiation

In contrast to leaky scanning, where the “weaker” AUG may be missed by the ribosome and translation begins at the “stronger” initiation codon, the mechanism of termination reinitiation allows translation from both the downstream and the upstream AUG initiation codons when both are in adequate sequence context (Figure 1.4) (Liu *et al*., 1984; Peabody and Berg, 1986). However, the ability of the ribosome to remain associated with the transcript and to reinitiate at a downstream AUG is affected by a number of factors such as: the time required translating the sORF, the intercistronic sequence and the intercistronic length (Kozak, 2001a).

The intercistronic distance is an important factor in determining the ability of the ribosome to reinitiate at a downstream AUG, as has been demonstrated in a number of examples including HIV-1 *tat* (Luukkonen *et al*., 1995), yeast *GCN4* (Grant *et al*., 1994) and the ORF36 encoded Kaposi’s sarcoma-associated herpesvirus (KSHV) protein kinase (Kronstad *et al*., 2013). While some vital translation factors remain associated with the scanning 40S ribosomal subunit after termination (eg. the tRNA, GDP and eIF2 complexes) other factors must be replenished if initiation at a downstream site is to take place. A longer intercistronic distance (at least 80 nucleotides) is required for the 40S ribosomal subunit to replenish these vital translation factors and thus be able to initiate at the downstream AUG (Kozak, 1987c).

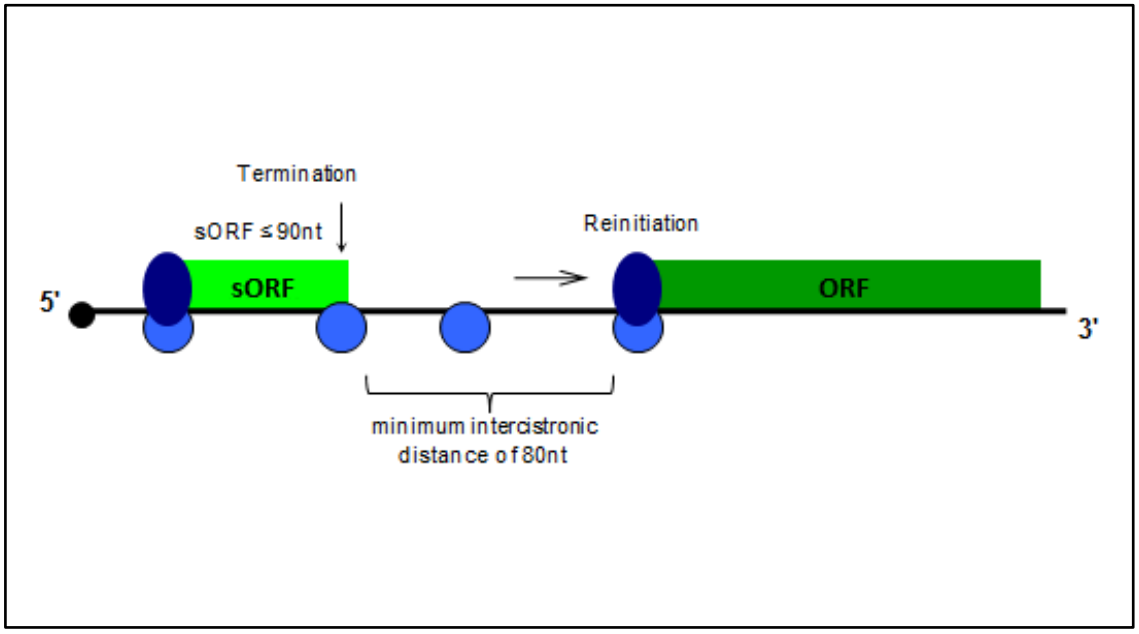


Figure 1.4 The mechanism of termination reinitiation. Once the 80S ribosome has scanned the first AUG and reached the stop codon some or all of the 40S ribosomal subunits (light blue circles) stay associated with the transcript (black line) to reinitiate at downstream AUGs.

The efficiency of reinitiation at a downstream initiation codon may be affected by sequence, particularly those that allow secondary structure formation. One such example is the vertebrate *AdoMetDC* gene where, the peptide translated by a sORF (encoding the hexapeptide MAGDIS) is able to directly interact with the translating ribosome to repress *AdoMetDC* translation (Hill and Morris, 1993). However, the *R* gene in maize, containing one sORF, is unaffected by peptide sequence. In this instance ribosomes that translate the upstream sORF reinitiate inefficiently due to the intercistronic sequences downstream from the sORF termination codon, which contains multiple stop codons (Wang and Wessler, 1998). Similarly, experiments with the sequence immediately downstream of the sORFs in *GCN4* demonstrated that reinitiation could be altered depending upon the sequence of the intercistronic region (Grant *et al*., 1995). In particular, reinitiation could be impaired in this system by stable base pairing between the nucleotides following the stop codon of upstream sORF1 with the translational complex or sequences within the *GCN4* mRNA transcript, prior to release of the mRNA transcript. This base pairing results in the dissociation of the ribosome due to the arrest in ribosomal scanning, after peptide release (Grant and Hinnebusch, 1994). Similar observations have also been noted in the human cytomegalovirus gpUL4 (gp48) transcript (Cao and Geballe, 1996) and the cauliflower mosaic virus 35S RNA leader (Ryabova and Hohn, 2000).

As well as the length and sequence of the intercistronic region, the length of the sORF has also been observed to affect the efficiency of reinitiation of the ribosome at a downstream AUG codon. It is believed that the factors that allow the 40S subunit to continue scanning post sORF translation are slowly depleted during the formation of the peptide chain, thus resulting in the dissociation of the ribosome from the mRNA transcript (Johansen *et al*., 1984; Peabody and Berg, 1986; Jackson and Wickens, 1997). For example eIF3, a factor released during 80S complex formation, is able to bring about efficient post-termination reinitiation by the repositioning of a new met-tRNA (Roy *et al*., 2010). A shorter sORF (less than 90 nucleotides) loses few, if any, vital translation factors, thus allowing the reinitiation of translation at a downstream AUG (Poyry *et al*., 2003).

Ribosomes have also been shown to reinitiate after translation of much longer sORFs. It is believed that, after release of the 60S ribosomal unit, a termination upstream ribosomal binding site (TURBS) is able to hold and stabilise the 40S unit allowing sufficient time for initiation factors to be recruited and translation to be reinitiated (Poyry *et al.*, 2007; Luttermann and Meyers, 2009). This mechanism is observed in the plant viral reinitiation factor transactivator-viroplasmin (TAV), which has been shown to promote the stability of the scanning ribosome via a dependant interaction with the target-of-rapamycin protein kinase (TOR) (Schepetilnikov *et al*., 2011). Other examples include the expression of the M1 and BM2 ORFs of the influenza B virus (Powell *et al*., 2011) and the expression of the VP2 protein of the calicivirus murine norovirus (Napthine *et al.*, 2009).

The possible role of sequences surrounding the termination codon and their effects on termination reinitiation has also been highlighted. The study of the sORFs 1 and 4 of yeast GCN4 have shown either AU or GC rich sequences may inhibit reinitiation (Grant and Hinnebusch, 1994). The dependence upon sequences surrounding the termination-reinitiation site, including the reinitiating AUG site preference, and the requirement for a pseudo knot structure closely followed by the upstream sORF termination signal, were also investigated in the *Helminthosporium victoriae* virus 190S (Li *et al*., 2011). A similar study within *GCN4* speculated that a 5′ enhancer structure folds progressively as sORF 1 is being translated. This enhancer structure contains four reinitiation-promoting elements that are able to directly interact with the eIF3a to stabilise the 40S ribosomal unit to allow for successful reinitiation at the downstream initiation codon (Munzarová *et al*., 2011).

1.7.3 Internal ribosome entry sites (IRES)

The IRES mechanism is a cap-independent mechanism in which an IRES structure is located near an internal AUG codon, allowing the ribosome to bypass any upstream inhibition to begin translation (Pelletier and Sonenberg, 1988; Kozak, 2001b). While the structural elements are not fully conserved, an IRES consists of a group of base paired stem-loop structures

(approximately 400-500 nucleotides in length) with several small unpaired sequence motifs that act as sites for RNA-binding proteins, which in turn are recognised by the ribosome and its initiation factors (Jackson, 1996; Kozak, 2003; Pisarev *et al*., 2005). This then allows the ribosome to begin translation of the next downstream AUG codon, with an optimal distance of 14 nucleotides from the base of the structure (Kozak, 1990) (Figure 1.5).

Many examples of IRES mediated translational control have emerged in the literature, the majority being of viral origin; examples include the shrimp white spot syndrome virus (Han and Zhang, 2006), the hepatitis C virus (Murata *et al*., 2005; Fraser and Doudna, 2007), and HIV-1 *gag* (Buck *et al*., 2001). Examples of non-viral origin include; the HAP4 gene in *Saccharomyces cerevisiae* (Seino *et al*., 2005) and the human PDGF2/c-*sis* gene (Rao *et al*., 1988). In short viral mRNAs, IRES mediated initiation of translation provides an added advantage. Initially the ribosome has the ability to bypass any upstream sequences, which may be vital in other aspects of the lifecycle of the virus. As there is also no need for the presence of a 5' cap structure to facilitate ribosomal scanning, this eliminates the need for the virus to replicate in the nucleus where mRNA capping occurs (Mathews, 1996) a benefit for many viruses.

The structural elements and sequences of IRES vary widely, as do their individual requirements for specific initiation factors or other translation factors. For instance the Hepatitis A virus requires that eIF4E must be associated with initiation factor eIF4G for ribosomal complex formation and the mRNA is uncapped (Ali *et al*., 2001). More recent studies have suggested that this requirement for eIF4G is dependent upon other factors. For example, Redondo *et al.* (2012) found that the cleavage of eIF4G, together with inactivation of eIF2, promoted IRES mediated translation. While much remains to be understood of IRES mediated translational control, current data suggest strong parallels between mechanisms of IRES-mediated translation and m^7G cap dependent translation.

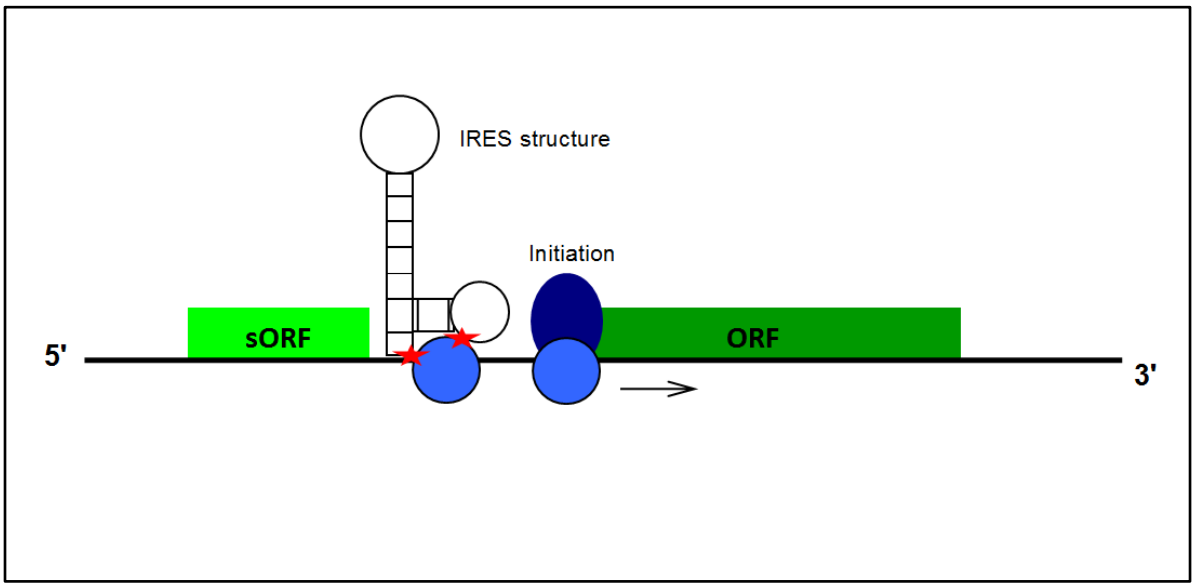


Figure 1.5 The IRES mechanism. The IRES structure (approximately 400-500 nucleotides in length) is composed of several small unpaired sequence motifs (red stars). These unpaired motifs bind RNA proteins, which promote the ribosome (light blue circle) to begin translation of the next downstream AUG codon.

1.7.4 Ribosomal shunting

The ribosomal shunt model postulates that, when a scanning ribosome comes to a hairpin loop structure, instead of unwinding it as the scanning model predicts, the ribosome “shunts” across the structure (Figure 1.6) (Jackson, 1996). The ribosomal shunt is cap dependant and the same ribosome is capable of translating both the up and downstream ORFs as it is shunted across the transcript (Hemmings-Mieszczak *et al*., 2000). The key features that ensure the ribosome shunts across the secondary structure have been partly defined. For instance the primary sORF must be between 2 and 15 codons in length and the distance between the sORF termination codon and the base of the secondary structure should be between 5 to 10 nucleotides (Dominguez *et al*., 1998; Hemmings-Mieszczak *et al*., 2000).

While little is understood about the shunting mechanism, research has shown that efficient termination and peptide release from the upstream ORF are required in order for the ribosome to shunt to the next ORF. The length of the sORF, location of the secondary structure relative to the 5' cap and the context of the termination codon are vital (Hemmings-Mieszczak *et al*., 2000). The context of the downstream ORF is not considered to be as important, as the ribosome is slowed during the shunt, enabling recognition of the start codon (Pooggin *et al*., 2000). A number of examples of ribosomal shunt mediated translational control are emerging, again the majority being from viruses. Such examples include; the Y proteins from the Sendai virus (Latorre *et al*., 1998), the E1 protein from the human papilloma virus type 18 (Remm *et al*., 1999), the tricistronic S1 mRNA of avian reovirus (Racine and Duncan, 2010) and the most well studied cauliflower mosaic virus RNA leader (Pooggin *et al*., 2000). Interestingly, the minimal sORF implicated in the regulation of Env expression by Krummheuer *et al*. (2007) exhibits potential shunting qualities. The authors propose that the 40S ribosome may pause at the minimal sORF, allowing it to interact with an RNA structure consistent with the ribosomal shunting model (Krummheuer *et al*., 2007).

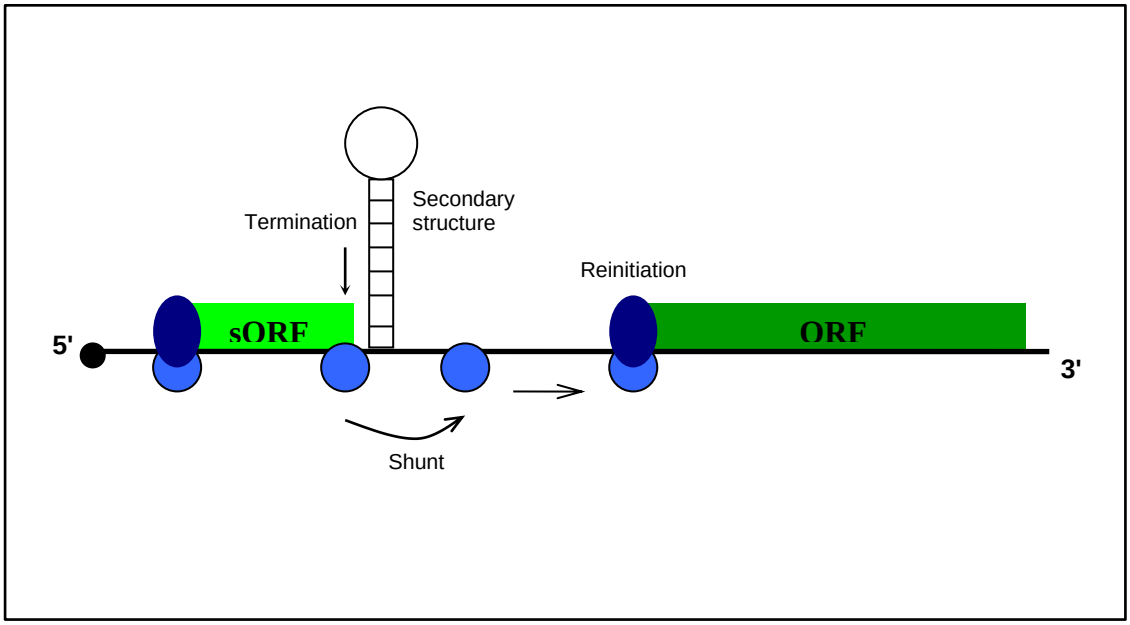


Figure 1.6 The ribosomal shunt model. Once the 80S ribosome (two large blue circles) has completed translation of the sORF (light green box) the 40S ribosomal subunit stalls at the base of the secondary structure and "shunts" across the structure, to resume scanning the mRNA transcript (thick line) and reinitiate at the downstream ORF.

A developing understanding of the fine interactions that occur in the ribosomal shunt model is emerging in the literature. Such an example includes the transactivator (TAV) protein which recruits the reinitiation supporting protein (RISP) (Thiebeauld *et al*., 2009); the RISP allows the 40S ribosomal unit to remain associated with the transcript for initiation downstream. Other examples include the adenovirus *hsp70* mRNA which harbours many secondary structures within the 5′-UTR; these secondary structures may contain binding sites able to anchor the 18S rRNA for efficient shunting (Yueh and Schneider, 2000). Further work has shown that the adenovirus 100k protein is able to aid in the shunt by its ability to bind the 5′ non-coding region and form a complex with eIF4G and the poly(A)-binding PABP protein, thus recruiting the 40S ribosomal subunit (Xi *et al*., 2004; Xi *et al*., 2005). In contrast, the translation shunt for the Y1 and Y2 proteins of the Sendai virus P/C mRNA does not display the typical shunt donor site, a secondary structure, an AUG codon or any other supposed 'essential' features (Latorre *et al*., 1998; de Breyne *et al*., 2003). Mapping of these shunting signals is proving difficult, as has been discussed in shunting studies of the duck hepatitis B virus (Cao and Tavis, 2011), the rice tungro spherical virus and rice tungro bacilliform virus (Pooggin *et al*., 2012).

Alternatives to the ribosomal shunt model of translation initiation include ribosomal tethering and ribosomal clustering. In the tethering model, the 40S ribosomal complex tethers to the 5′ mRNA m^7G cap structure (Chappell *et al*., 2006a). During Met-tRNA and initiation codon base pairing some initiation codons along the mRNA transcript can be missed, possibly by the ability to form loop structures that can mask these codons (Chappell *et al*., 2006a). Similarly the clustering model of translation initiation postulates that the 40S ribosomal complex is recruited by specific sequences along the mRNA transcript, thus permitting initiation at upstream AUG codons. It is proposed that a specific sequence, identified in the 5′ leader of the Gtx mRNA, is capable of binding the 18S rRNA to promote ribosome binding (Chappell *et al*., 2000; Chappell *et al*., 2006b). A novel model whereby repression (slowing) of ribosomes could occur has also been proposed. In this study, the

researchers present the role of the DEAD-box RNA helicase Dhh1, that represses ribosome translation independently of initiation factors eIF4E and eIF3b. This repression leads to the accumulation of ribosomes along the transcript (Sweet *et al*., 2012). The authors propose this model would allow for a rapid halt in mRNA translation with the added benefit of storing the mRNA transcript for rapid translation with ribosomal complexes pre-loaded on the transcript, or permit mRNA de-capping and thus degradation of the transcript (Sweet *et al*., 2012).

1.8 Summary of the effects of sORFs on translation

The presence of sORFs has been proposed to regulate translation in a number of different ways. In the first instance the position of the sORF relative to the 5'-UTR determines if the ribosome will remain associated with the mRNA; the efficiency of translation will be reduced if the sORF is positioned closer to the 5' cap (Sedman *et al*., 1990). The ribosome may remain associated with the mRNA and continue scanning if the interaction between eIF4F and the ribosome is maintained. In particular sORFs displaying poor AUG initiation codon contexts result in the ribosome 'missing' the initiation codon, otherwise known as 'leaky-scanning' (Poyry *et al*., 2003). The ribosome may reinitiate at a downstream initiation codon depending on the position of the termination codon of the previous sORF (Peabody and Berg, 1986) or if the intercistronic sequence is lengthy (Kozak, 1987c). The ribosome may also stall during elongation or termination resulting in the blockage of further ribosomal scanning of the mRNA. The presence of a 14 nucleotide, downstream stem-loop secondary structure (hairpin) may enhance initiation due to the slowing of ribosomal scanning (Kozak, 1990). Finally, sORF translation may also affect gene expression by altering the ability of the scanning ribosome to reach and initiate at downstream AUG initiation codons (for example via peptide interactions with the ribosomal complex) (Morris and Geballe, 2000).

1.9 Hypothesis of this study

The hypothesis of this study is that the HIV-1 *asp* sORF region (sORFs I to VI, cloned into the pEGFP-N1 plasmid) can modulate downstream EGFP gene expression. This modulation can occur by a series of splicing events and the control of sORF translation.

1.10 Aims of this study

The hypothesis of this study will be addressed by the following objectives:

Objective 1:
To assess the possible use of the series of six sORFs as modulators of asp expression by exploring the conservation of the region amongst viral strains/subtypes. High level conservation of the number, position, size and sequence of these sORF series may suggest their importance in the regulation of *asp* expression. The conservation of the sORF region will be examined in (A) all the HIV-1 subtypes, including randomly selected sequences from the strains/subtypes A, B, C, D, E, F, G, H, J, K, N and O; (B) some SIV sequences. The number of sORFs present will be examined and their nucleotide sequence conservation assessed, and (C) the relative strength of the initiation codon of each sORF according to Kozak's consensus will be determined.

Objective 2:
To investigate the effect the sORF region plays on downstream expression in a reporter construct where variants of the asp sORF region have been cloned upstream of the reporter EGFP and assessing (A) effects of the sORF region on downstream gene expression; (B) the effect of the intercistronic distance between sORF VI and the reporter EGFP; (C) the effect of transcriptional activation on expression; and (D) assessing the relative roles of each sORF by a series of mutations in each respective sORF AUG codon.

Objective 3:

To characterise the sORF transcript and investigate the potential for splicing events to modulate expression. The sORF transcript will be characterised by (A) RT-PCR and sequence analysis, followed by (B) cDNA cloning of any spliced variants to assess their relative contributions to expression. This will also be examined by (C) mutational analysis within the sORF AUG codons and of observed splice donor and acceptor motifs. (D) The levels of any spliced variants within the pool will be examined by Real-Time PCR to allow relative expression levels to be determined.

Objective 4:

To investigate the ability of the ribosomal unit to initiate translation at each respective sORF using the toeprinting assay and thus determine the relative importance of each sORF and suggest the potential mechanism by which the sORFs are translated.

CHAPTER 2 – GENERAL MATERIALS AND METHODS

2.1 Reagents

All reagents used throughout this investigation were of analytical or molecular biology grade, except where indicated in the text.

2.2 Cell culture

2.2.1 Bacterial cell culture and media

Escherichia coli K-12, strain JM-101 was grown in nutrient rich media. Either SOB broth (20g/L tryptone, 5g/L yeast extract, 0.5g/L sodium chloride, 2.5mM potassium chloride, 20mM magnesium sulphate) or SOB agar (as above with 1.5% (w/v) Bacto^{TM}-Agar) was used. Selection of pEGFP-construct transformed *E.coli* was completed by the addition of kanamycin sulphate (30µg/mL) to the media after being cooled to ~50°C. All media were stored at 4°C.

2.2.2 Mammalian cell culture and media

Human Embryonic Kidney (HEK) 293 cells were gown in Dulbecco's Modified Eagles Medium (DMEM) supplemented with 10% (v/v) heat-inactivated Newborn Calf Serum, 2mM L-glutamine, 100U/mL penicillin and 100µg/mL streptomycin (all from Invitrogen), in a humidified incubator at 37°C with 5% carbon dioxide.

2.3 Plasmids

Plasmid constructs as depicted in Figure 2.1 below were previously constructed by Dr Nicholas Deacon and Ms Fee Yee Wong (Macfarlane Burnet Institute, Victoria, Australia). These constructs were built to contain the sORFs upstream of HIV-1 *asp* (NL-43 strain) inserted into the multiple cloning site of the pEGFP-N1 vector (Clontech) before the reporter gene enhanced green florescent protein (EGFP) with a distance of 92bp between the AUG codon of EGFP and the stop codon of sORF VI. Within the virus (NL-43

strain) the stop codon of sORF VI is situated 48bp from the AUG codon of ASP, making these two elements inframe with one another. To more closely mimic this distance between the AUG codon of EGFP and the stop codon of sORF VI the pEGFP (sORF I-VI Wt) construct was derived from pNL4-3 DNA by subcloning the sORF segment (position 8798 to 7980 comprising the start codon of sORF I to the stop codon of sORF VI respectively) into pGEM®-T Easy vector system (Promega) with the addition of an Age I site by the PCR primers; Forward: 5'-CT**ACCGGT**CTCCAGGCAAGAATCC-3' and Reverse: 5'-CT**ACCGGT**CTTGCCACCCATCTTATAGC-3', Age I site shown in bold. The sORF segment was cloned into the Age I site of the pEGFP-N1 vector (Clontech) such that the stop codon of sORF VI sits 63bp upstream of the ATG codon of the reporter EGFP. Unless otherwise stated all studies were conducted using the latter construct.

The pEGFP-N1 plasmid, without insertion in the multiple cloning site, was used as a positive control, while the negative control (pEN1), supplied by Dr Damien Purcell (Department of Microbiology and Immunology, The University of Melbourne, Victoria, Australia), was constructed by excision of the reporter EGFP gene with Age I and Not I, end filling by Klenow and re-circularization of the “EGFP-empty” plasmid.

A variety of plasmid constructs were derived from this set of plasmid by mutation and/or subsequent cloning experiments. The details of each plasmid construct created for these experiments are presented in their respective chapters.

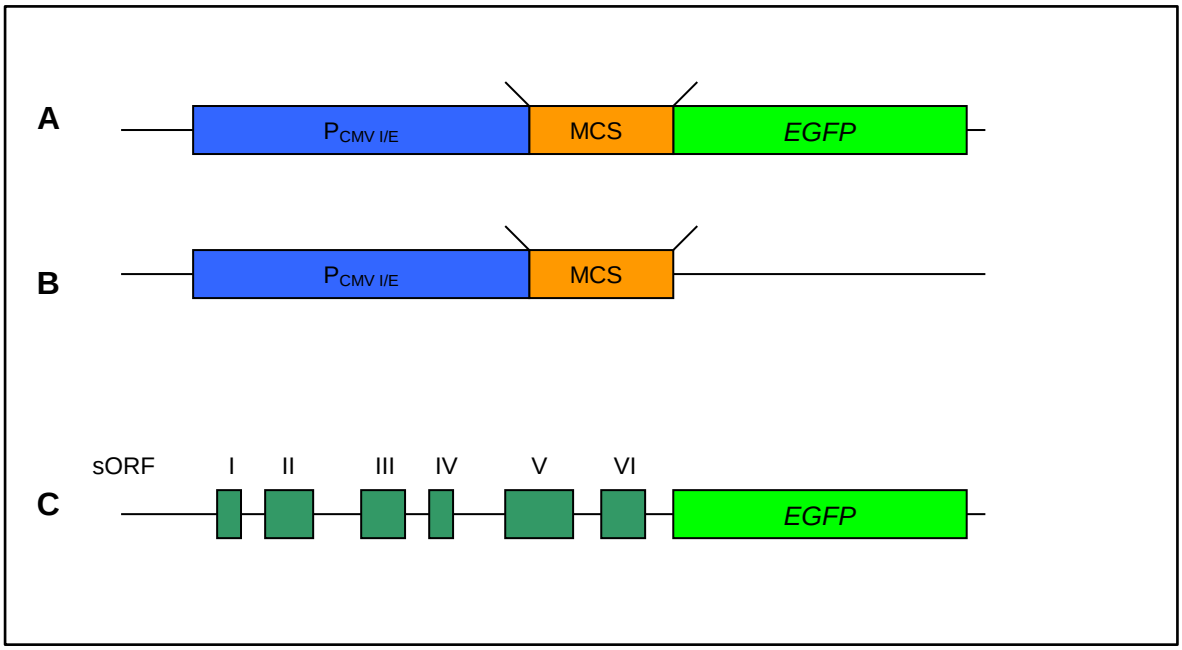


Figure 2.1 Schematic of base plasmids used throughout this study. (A) base pEGFP-N1 plasmid used as positive control, (B) pE-N1 plasmid with EGFP removed as negative control and (C) pEGFP (sORF I-VI Wt) plasmid with sORFs I – VI sub-cloned in the MCS.

Figure 2.2 Sequence of the sORF region from HIV-1 NL4-3. The sORFs I through to VI are highlighted in yellow from top to bottom respectively. Start and stop codons of each sORF are underlined with an internal in-frame ATG shown for sORF VI.

2.4 Plasmid amplification

2.4.1 Preparation of competent cells

The calcium chloride method for the transformation of *E.coli* was used to amplify plasmid DNA (Mandel and Higa, 1970). In order to prepare the competent cells *E.coli* was streaked onto SOB agar and incubated overnight at 37°C. A large loop full of the culture was picked and inoculated into 100mL of SOB broth in a 2L conical flask. The flask was incubated in a rotary shaker at 37°C at 95 rpm for ~4 hrs until the optical density of the culture at 600nm reached ~0.4. Once the culture reached this optical density it was rapidly cooled on ice for 10 min. The culture was then transferred to 50mL Falcon™ tube (BD Biosciences) and centrifuged at 3000rpm (Sorvall GSA rotor) at 4°C for 10 min. The supernatant was removed from the pelleted cells and resuspended in 33mL RF1 buffer (55mM $MnCl_2$, 15mM $CaCl_2$, 10mM potassium acetate, 10mM rubidium chloride, 10% (v/v) glycerol, pH 5.8). The cell suspension was incubated on ice for 1 hr and then centrifuged as previously detailed. The cell pellet was then resuspended in 7mL of RF2 buffer (10mM rubidium chloride, 3mM MOPS, 70mM $CaCl_2$, 10% (v/v) glycerol, pH 6.8). The competent cells were aliquoted, snap frozen in liquid nitrogen and stored at -70°C.

2.4.2 Transformation of competent cells

For each plasmid, 100µL of calcium chloride competent *E.coli* cells was mixed with ~100ng of DNA (in H_2O) in 10mL polypropylene tubes (Falcon) and incubated on ice for 30 min.

The *E.coli* cells were heat shocked to promote the uptake of the plasmid DNA by incubating the mix in a 42°C water bath for 50 sec. The cells were then immediately placed on ice and allowed to cool for 2 min. Following this 900µL of warm SOC broth was added and the cells allowed to recover at 37°C with gentle agitation (50 rpm) for 1 hr. An aliquot of 200µL of the transformed cells was then plated on SOB agar containing the appropriate antibiotic for

selection (for pEGFP-N1 plasmids, 30µg/mL kanamycin sulphate) and incubated overnight.

2.4.3 Plasmid preparation

A single colony of transformed *E.coli* was picked and inoculated into 50mL of SOB broth containing the appropriate antibiotic and incubated overnight at 37°C with gentle agitation (50 rpm).

Glycerol stocks of transformed *E.coli* were prepared by mixing 1.5mL 60% (v/v) glycerol with 500µL of overnight broth culture and frozen at -70°C.

2.4.4 Plasmid DNA extraction

Overnight broth cultures of transformed *E.coli* were harvested by centrifugation at 5000rpm (Sorvall GSA rotor) at 4°C for 10 min. Pelleted cells were then washed in 20mL of ice-cold STE buffer (10mM Tris, 1mM EDTA, 100mM NaCl, pH 8) before being centrifuged as previously detailed to ensure complete removal of culture medium.

The DNA was extracted from the washed cells using the JETStar Plasmid Purification Kit: MaxiPrep (Genomed) or using the Wizard® Plus SV Miniprep DNA Purification System (Promega) depending upon the scale of the culture and the mass of DNA required. Volumes were scaled accordingly for smaller preparations. Briefly the cells were resuspended in 10mL of Re-suspension buffer (50mM Tris, 10mM EDTA, pH 8 containing 1mg RNase A) and lysed by the addition of 10mL Lysis buffer (200mM NaOH, 1% (w/v) SDS) for 5 min. The cell lysis was stopped by the addition of 10mL of Neutralization buffer (3.1M potassium acetate, pH 5.5). The suspension was mixed by gently inverting the tube and then centrifuged at 9000rpm (Sorvall GSA rotor) at room temperature for 10 min.

The column was washed with 30mL of Equilibration buffer (600mM NaCl, 100mM sodium acetate, 0.15% (v/v) TritonX-100, pH 5) after which the DNA suspension was filtered onto the column through two layers of lint free tissue

in order to reduce protein contamination of the column. The column was then washed with 60mL of Wash buffer (800mM NaCl, 100mM sodium acetate, pH 5) before DNA was eluted with 10mL Elution buffer (1250mM NaCl, 100mM Tris, pH 8.5). For large scale preparations only, the eluted DNA was precipitated by the addition of 0.7% (v/v) volume isopropanol and incubated overnight at -20°C. The precipitated DNA was collected by centrifugation at 20,000rpm (Sorvall, SS-34 rotor) for 30 min at 4°C. DNA was washed with 70% ethanol and finally with absolute ethanol and centrifuged as detailed above. Air dried pellets were resuspended in 500µL of TE buffer (10mM Tris, 1mM EDTA, pH 7.5) and stored at -20°C.

2.5 Analysis of DNA

All DNA stocks were subsequently analysed for purity and concentration, using techniques described below.

2.5.1 Agarose gel electrophoresis of DNA

DNA samples were resolved by electrophoresis for qualitative examination on 1.5% (w/v) agarose gels in 0.5x TBE buffer (45mM Tris-borate, 1mM EDTA, pH 8.3) at ~75 Volts. Agarose gels contained 0.5µg/mL ethidium bromide for the visualisation of DNA under ultra violet irradiation and photography (Dolphin – DOC). All electrophoretic separations were monitored by the addition of loading dye (0.25% (w/v) bromophenol blue, 0.25% (w/v) xylene cyanol FF, and 15% (w/v) ficoll) to the DNA samples. All DNA samples were electrophoretically separated alongside markers of known size in order to estimate the molecular sizes of DNA samples.

2.5.2 Spectrophotometric analysis of DNA

The concentration and purity of DNA was checked by spectroscopy. Samples (10µL) were diluted in 990µL TE Buffer before absorbance was scanned at 260 and 280nm. The mass of DNA in the sample was calculated, using the extinction coefficient for dsDNA at 260nm = 0.020 $(\mu g/ml)^{-1}$ cm^{-1}, while the quality of DNA was determined by the ratio of the absorbance 260/280, where ratios between 1.8 to 2.1 were considered acceptable.

2.5.3 Qubit™ Quant-It™ DNA assay

The Quant-It™ Assay, which relies on the labelling of nucleic acids or proteins with a fluorescent dye to accurately determine concentration, was performed according to the manufacturer's instructions. Briefly, 198μL of the Quant-It™ dilution buffer was mixed with 1μL of the Quant-It™ reagent. 1μL of the test sample was added and mixed, incubated at room temperature for 2 mins (DNA or RNA) or 15 mins (proteins), spun briefly to remove air bubbles and placed in the Qubit™ device and the fluorescence recorded. This reading is plotted against a set of standards and the concentration of the sample is calculated by the instrument.

2.5.4 Sequencing

Plasmid DNA was sequenced according to the protocols of the ABI PRISM® Big Dye™ Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) in order to determine the integrity of the insert. The reaction mixture was established in a 200μL thin walled PCR tube containing: 300ng plasmid DNA, 4pmol Primer (details listed in each chapter), 3.5μL 5 x Reaction Buffer, 1μL BigDye Premix (Applied Biosystems) and made up to a total volume of 20μL. The reactions were then placed in a pre-heated (96ºC), Mastercycler® Gradient thermal cycler (Eppendorf) and incubated for 1 min. The PCR cycle was then initiated with a programme of 96ºC for 10 sec, 50ºC for 5 sec, 60ºC for 4 min for 30 cycles before cooling to 4ºC.

Sequence products were purified by addition of the 20μL reaction mixture to 3μL 3M sodium acetate (pH 4.6), 62.5μL absolute ethanol and 14.5μL water, which was mixed and allowed to precipitate at room temperature for 15 min before being pelleted by centrifugation at 12,000 x g for 20 min. The pellet was washed twice with 200μL 70% ethanol before being dried at 80ºC for 2 min. Dried samples were stored in the dark.

Capillary sequencing electrophoresis was carried out by the MICROMON DNA Sequencing Facility at the Department of Microbiology, Faculty of Medicine, Monash University, Clayton.

2.6 Transient transfection of eukaryotic cells

2.6.1 Preparation of eukaryotic cells for transfection

Cells were grown in 75cm^2 tissue culture flasks (BD Biosciences) and harvested for transfection when 80-90% confluent by the addition of 1mL Trypsin-EDTA (0.25% Trypsin, 1mM EDTA, Invitrogen) and incubated for 1 min at room temperature. Trypsin treatment was stopped by the addition of fresh medium (DMEM) and the cells collected. Cells were seeded at a density of 1.6 x 10^5 cells per well into 24-well culture plates (1 cm^2), volume adjusted to 1mL with fresh medium if required and incubated overnight. Cell growth medium was replaced 4 hrs prior to transfection.

2.6.2 Transient transfection of eukaryotic cells by the calcium phosphate method

Cells were transiently transfected using the calcium phosphate technique with the ProFection® Mammalian Transfection System (Promega) according to the manufacturer's instructions. In this technique the DNA is co-precipitated with calcium phosphate and the precipitate is finely spread over the cells (Graham and van der Eb, 1973; Wigler *et al*., 1977). Briefly, 1μg DNA was mixed with 12.3μL 2M $CaCl_2$ and total volume adjusted to 100μL with H_2O. Where detailed, co-transfections with the plasmid pDs-Red were conducted by mixing 0.5μg DNA with 0.5μg of DNA to be tested. This DNA mix was added drop wise to 100μL 2x HBS buffer (50mM HEPES, 280mM NaCl, 1.5mM Na_2HPO_4, pH 7.1) while gently agitating with a vortex mixer. The mix was then incubated at room temperature for 30 min before being briefly vortexed and added drop wise to the plated cells. Cells were returned to the incubator for either 48 hrs (non-activated) or 24 hrs (activated) as detailed below before harvest.

2.6.3 Activation of eukaryotic cells with sodium butyrate and trichostatin A

Where indicated, cells were activated with either sodium butyrate (NaBut) or trichostatin A (TSA). NaBut (Sigma) solution was prepared in H_2O and TSA (Sigma) was prepared in 50% ethanol. Cells that had been transiently transfected 24 hrs prior were removed from the incubator and NaBut or TSA were added to the cells drop wise to a final concentration of 5mM and 450nM respectively. Inactivated controls treated with H_2O and ethanol (50%v/v) were also prepared. The plates were mixed gently at 100rpm for 2 mins on a rotary shaker (Ratec) and returned to the incubator for 24 hrs before cell harvest.

2.7 Reporter gene assays

In order to quantify the amount of EGFP expressed in the cells were lysed and the intensity of green fluorescence measured. Specific activities were normalised by measurement of the total protein content of the lysates using the Bradford assay.

2.7.1 Preparation of cell lysates

Cells were harvested by removal of the medium and washing twice in Dulbecco's Phosphate Buffered Saline (D-PBS, Invitrogen). The cells were then lysed and removed from the surface of the plates by the addition of 100μl/cm^2 Reporter Lysis Buffer (Promega). Cells were scraped and plates freeze-thawed (at -20°C and room temperature) twice to facilitate lysis before lysate was transferred to microfuge tubes. Samples were centrifuged at maximum speed for 5 mins at room temperature to pellet cell debris and the supernatant removed for analysis.

2.7.2 Specific activity of enhanced green fluorescent protein

2.7.2.1 Fluorescence analysis

Intensity of green fluorescence of cell lysates provided a direct measure of levels of EGFP expression. Fluorescence was measured using the FLUOstar

OPTIMA micro-plate reader (BMG Labtech) in black 96-well plates at an excitation of 485nm and emission of 520nm.

2.7.2.2 Bradford assay

The Bradford assay was used to measure total protein in order to normalise the fluorescence. The Bradford assay was performed by the addition of 5μL lysate to 50μL Bradford reagent (Sigma) in clear 96-well plates. The samples were mixed by pipetting and incubated at room temperature for ~20 mins before absorbance at 595nm was measured using the FLUOstar OPTIMA micro-plate reader (BMG Labtech). Bovine serum albumin (BSA) standards were used to calibrate protein concentration.

2.8 Data analysis

2.8.1 t-test for the statistical analysis of the means of two populations

In order to observe the differences (if any) between the positive control (pEGFP-N1) and the test samples were of any significance the t-test for two population means was used.

In this case the t-test hypothesizes that there is no difference between the positive control and the test population (Expression of EGFP in the positive control is equal to that of EGFP expression in the test sample). Thus a p-value greater than 0.01 indicates that this hypothesis is true and that no significant difference between the two populations is observed. In contrast, if a p-value less than 0.01 indicates that the test hypothesis is false and a significant difference between the two populations is observed.

2.9 Reverse-transcriptase PCR

Unless otherwise stated all materials and reagents utilised within this section were carefully cleaned and treated to eliminate RNase activity. All water was treated with diethylpyrocarbonate (DEPC) at a concentration of 0.1%, incubated at room temperature overnight before residual activity was

destroyed by autoclaving at 121°C/15min. Glassware was soaked in 1M NaOH overnight before being rinsed in DEPC treated water. Pipettes were cleaned with RNaseAWAY (MedProbe) and filter tips were used to minimise RNase activity.

2.9.1 RNA isolation

The SV Total RNA Isolation System (Promega) was used to isolate RNA from the transfected cells according to the manufacturer's instructions. To prepare the cells the media was aspirated and the cells were rinsed with PBS. Cells were lysed by the addition of RNA Lysis Buffer™ and cell lysates passed through a 20G needle to shear genomic DNA. Each sample was heated for 3 mins at 70^{o}C, centrifuged at 13,000 rpm for 10 mins and the supernatant collected. Absolute ethanol was added and the solution was transferred to the spin column provided and centrifuged at 13,000rpm for 1 min. The column was washed by the addition of RNA Wash Solution™, and centrifuged for 1 min. DNA was digested with the addition of prepared DNase1 (50units) mixed with 0.09M magnesium chloride and Yellow Core Buffer™. Samples were incubated at room temperature for 30 mins. After this time, DNase Stop Solution™ was added to inactivate the DNase1. The column was rinsed twice with RNA wash™ solution. The RNA was eluted in 100µL of nuclease free water and stored at -70^{o}C until required.

2.9.2 Analysis of RNA

All RNA stocks were subsequently analysed for purity and concentration, using techniques described below.

2.9.2.1 Agarose gel electrophoresis of RNA

RNA samples were resolved by electrophoresis for qualitative examination on 1% (w/v) agarose gels with 7% formaldehyde in 1x formaldehyde running buffer (100mM 4-morpholinepropanesulfonic acid, 40mM sodium acetate, 5mM EDTA, pH 7.0) at ~75 Volts. Agarose gels contained 0.5µg/mL ethidium bromide for the visualisation of RNA under ultraviolet irradiation and photography (Dolphin – DOC). All electrophoretic separations were monitored

by the addition of loading dye (50% formamide, 6% formaldehyde, 7% glycerol, 0.5% (w/v) bromophenol blue in 1 x formaldehyde running buffer) to the RNA samples. All RNA samples were electrophoretically separated alongside markers of known size in order to allow for estimation of the sizes of RNA bands. Un-degraded high quality RNA samples showed clear bands at both 1.9kb and 5kb corresponding to the 18S and 28S ribosomal RNA components respectively.

2.9.2.2 Spectrophotometric analysis of RNA

The concentration and purity of RNA was checked by spectroscopy. Samples (10µL) were diluted in 990µL TE Buffer before absorbance was scanned at 260 and 280nm. The mass of RNA in the sample was calculated, using the extinction coefficient for ssRNA at 260nm = 0.025 $(\mu g/ml)^{-1}$ cm^{-1}, while the quality of RNA was determined by the ratio of the absorbance 260/280, where ratios between 1.8 to 2.1 were considered acceptable.

2.9.2.3 Qubit™ Quant-It™ RNA assay

The Quant-It™ Assay relies on the labelling of nucleic acids or proteins with a fluorescent dye which can then be quantitated by fluorescence spectroscopy. In this technique minute quantities of sample can be used to accurately determine concentration. The assay was followed according to the manufactures instructions as detailed in 2.5.3 of this chapter.

2.9.3 Synthesis of cDNA

The Superscript III™ First-Strand Synthesis System (Invitrogen) was used according to the manufacturer's instructions to generate cDNA from the extracted total RNA. Essentially a total mass of 50ng of total RNA was used as template and combined with 1µL 10mM dNTPs and 0.5µg oligo dT and the total volume adjusted to 10µL with RNase free water. Negative control samples (-OLI) did not contain any oligo dT. The samples were heated to 65°C for 5 mins in order to denature secondary structures and snap chilled on

ice. The reaction buffer was added to a final concentration of 1 x (200mM Tris-HCl, pH 8.3, 500mM KCl) followed by 2µL 0.1M 1,4-dithiothreitol and 1µL RNaseOUT™ Recombinant Ribonuclease Inhibitor (40units) in a total volume of 19µL adjusted with nuclease-free water. The reaction was initiated with 1µL SuperScript III RT (50units) and incubated at 50°C for 50 mins. Negative control samples (-RT) did not contain any SuperScript III RT. Reactions were terminated by incubation at 80°C for 5 mins followed by digestion of the template RNA strand via addition of 1µL RNase H (2units) and incubation at 37°C for 20 mins. All samples were stored at -20°C until required.

2.9.4 RT-PCR conditions

Samples of the cDNA were PCR amplified using Forward and Reverse RT primers depicted in Table 2.1. Reactions were established in the presence of 1.25 Units Taq DNA polymerase (New England Biolabs), 1x ThermoPol buffer, 20µM dNTP and 20µM of each primer. PCR conditions were as follows; initial denaturation of 94°C for 1 min followed by 30 cycles of denaturation (94°C for 30 sec), annealing (62°C for 30 sec for RT and M13 specific primers; or 58°C for 30 sec for GAPDH specific primers) and extension (72°C for 1 min) with a final extension of 72°C for 5 min. RT-PCR amplification reactions were controlled for DNA contamination (RNA sample with no RT), auto-priming (cDNA synthesis in the presence of RT with no primer) and general transcript integrity with GAPDH specific primers (sequences provided in Table 2.1). Amplified products were resolved via agarose gel electrophoresis as described in section 2.5.1.

Table 2.1 Reverse Transcriptase-PCR primers.

Probe	Sequence (5′ - 3′)	T_M (°C)
RT For	GCTACCGGACTCAGATCTCGAGC	61
RT Rev	CCTGGCTGTGGAAAGATACC	54
GAPDH For	TGCACCACCAACTGCTTAGC	54
GAPDH Rev	GGCATGGACTGTGGTCATGAG	56
M13 For	GTAAAACGACGGCCAG	54
M13 Rev	CAGGAAACAGCTATGAC	54

2.9.5 Extraction and purification of PCR products

PCR products were excised from the gel and extracted using the Wizard® SV Gel and PCR Clean up System (Promega) according to the manufacturer's instructions. Briefly after the addition of Membrane Binding Solution™ to the gel fragment, the agarose was dissolved by heating to 65°C for ~10 mins. The mixture was then applied to the spin column and centrifuged at 13,000rpm for 1 min after which the flow through was discarded. The spin column was rinsed twice with the Membrane Wash Solution™ with a final centrifugation at 13,000rpm for 2 min to dry the membrane. The DNA was eluted in 50µL nuclease-free water and stored at -20°C.

2.9.6 Sub-cloning of PCR products

The extracted and purified PCR products were sub-cloned for sequencing using the pGEM®-T Easy Vector System (Promega). This vector replies upon the incorporation of an adenosine overhang into PCR products during cycling conditions. The vector has a single thymidine overhang that promotes complementary base pairing and thus ligation efficiency is strong. Ampicillin resistance and α-complementation screening properties are used to identify recombinant colonies. Briefly the extracted and purified PCR product was ligated to 50ng of the ®-T Easy Vector with 3units of T4 DNA Ligase in the Rapid Ligation Buffer (60mM Tris-HCl (pH 7.8), 20mM $MgCl_2$, 20mM DDT, 2mM ATP and 10% polyethylene glycol) in a total volume of 10µL at 4°C overnight. The ligation reaction mixes were transformed into competent *E.coli* cells as detailed in section 2.4.2 of this chapter before being plated onto LB agar with 50µg/mL ampicillin, 0.1mM isopropyl-β-D-thiogalactopyranoside (IPTG) and 60µg/mL 5-bromo-4-chloro-3-indolyl-beta-D-galactopyranoside (X-GAL). Plates were incubated at 37°C overnight.

The resulting white colonies were amplified in 5mL LB broth at 37°C overnight before plasmid DNA was harvested using the Wizard® Plus SV Miniprep DNA Purification System (Promega) as described in section 2.4.4. Colonies were then screened for presence of insert by PCR, using the standard conditions detailed in section 2.9.4 before being sequenced using the ABI PRISM® Big

Dye™ Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) as detailed in section 2.5.4 with the M13 Forward and Reverse primers listed in Table 2.1.

2.10 Northern blotting

The method used in this investigation is an adaptation of that previously published (Augood *et al*., 1990; Maggiolini *et al*., 1999) and involves the direct detection of RNA species by incorporation of digoxigenin-11-dUTP (DIG-dUTP) by PCR. Quantitation of transcript abundance was conducted by dot-blot and detection of EGFP transcripts, normalised to GAPDH transcripts after stripping the membrane of the EGFP probe. Densitomitry analysis was conducted by comparison of samples to EGFP and GAPDH standards. As detailed in section 2.9 all materials and reagents were carefully cleaned and treated to eliminate RNase activity.

2.10.1 Generation of DIG-labelled EGFP, RFP and GAPDH probes by PCR

In order to generate a probe for Northern blotting, EGFP, RFP and GAPDH specific primers (Table 2.2) were used to amplify a small (~50-100bp) region of the pEGFP-N1 plasmid with the PCR DIG Probe Synthesis Kit (Roche) according to the manufacturer's instructions. Briefly the reactions were prepared with 50pg pEGFP-N1 cDNA, 0.1µM each primer, 20µM dNTP mix, 1x DIG PCR Buffer with $MgCl_2$, 20µM PCR DIG Mix (20µM each dATP,dCTP, dGTP, 13µM dTTP and 7µM DIG-11-dUTP, alkali-labile, pH 7.0) and 2.6units Expand High Fidelity Enzyme Mix in a total volume of 50µL. The reactions were amplified with the following cycling conditions; Initial denaturation of 94°C for 2 min, followed by 30 cycles of 94°C for 30 sec, annealing at 62°C for 30 sec for EGFP (58°C for 30 sec for GAPDH) and an extension at 72°C for 1 min, a final extension of 72°C for 7 min was performed following the cycling program. The integrity/concentration of the probes was checked as detailed in section 2.5 of this chapter.

Table 2.2 PCR primers used to generate DIG-Labelled probes for northern blotting.

Probe	Sequence (5′ - 3′)	T_M(°C)
EGFP For	CCTACGGCGTGCAGTGCTTCAGC	62
EGFP Rev	CGGCGAGCTGCACGCTGCCGTCCTC	69
GAPDH For	TGCACCACCAACTGCTTAGC	54
GAPDH Rev	GGCATGGACTGTGGTCATGAG	56

2.10.2 Blotting and development of northern blots

A total of 1µg total RNA per sample was used for blotting onto positively charged nylon membrane (Roche). The membranes were pre-wetted in DEPC-treated water before being assembled into the Bio-Dot® Microfiltration Apparatus (Bio-Rad). The wells of the apparatus were pre-treated with 0.5mL cold 10mM NaOH, 1mM EDTA solution before the samples were applied and rinsed twice with the same solution. The membrane was washed briefly in 2x SSC, 0.1% SDS and then baked under vacuum at 120°C for 30 mins to fix the samples to the membrane. Known amounts of DIG-labelled control RNA (Roche) were utilised as standards for the blot.

The membrane was pre-hybridized in 15mL DIG Easy Hyb Solution at 50°C for 30 mins with gentle agitation. The solution was then replaced with 200ng of denatured probe and left to hybridize at 50°C overnight. Following the hybridization the membrane was washed twice (5 mins at room temperature with gentle agitation) with Low Stringency Buffer (2x SSC, 0.1% SDS) and then twice (15 mins at 50°C with gentle agitation) with High Stringency Buffer (0.1x SSC, 0.1% SDS).

Chemiluminescent detection of DIG labelled samples was conducted using the DIG proprietary reagents for chemiluminescent development (Roche). Briefly the membrane was soaked in Washing Buffer™ for 2 mins at room temperature followed by 30 mins of Blocking Solution™ at room temperature with gentle agitation. The membrane was incubated at room temperature for

30 mins with 1500mU Anti-Digoxigenin-AP in Blocking solution and then washed twice (15 mins at room temperature) in Washing Buffer before development. The membrane was drained and covered between two acetate sheets for development with 1mL of a 1:100 dilution CSPD reagent in 1x Detection Buffer (Roche). The membrane was incubated at room temperature for 5 mins before autoradiography.

2.10.3 Autoradiography and densitometry analysis of northern blots

The developed membranes were sealed in acetate sheets and carefully cleaned for autoradiography. The membrane was exposed for 15-25 mins against Lumi-Film X-ray Film (Roche) before being developed in BTX Fixer Solution (Kodak), washed in water and fixed in BTX Fixer Solution (Kodak), washed and allowed to air dry. The message abundance was measured using the densitometry tool of the Dolphin-DOC image system (Wealtec), and the total EGFP message was normalised against the total GAPDH message. EGFP message abundance is presented as transcript relative to GAPDH message and were statistically analysed as detailed in section 2.8.

CHAPTER 3 – *IN SILICO* ANALYSES OF ASP AND THE sORF REGION

3.1 General introduction

Early studies conducted by Miller (1988) predicted the negative sense ORF, now known as *asp*, spanning the gp120/gp41 junction of *env*. Miller's work also showed the high level conservation of this ORF across 12 strains of HIV-1, which was later confirmed by *in silico* analyses that demonstrated the conservation of the *asp* ORF amongst 550 genomic and *env* HIV-1sequences and its absence from HIV-2 sequences (Bukrinsky and Etkin, 1990; Briquet and Vaquero, 2002). These findings are consistent with work that showed endogenously expressed *asp* RNA was able to inhibit the replication of HIV-1 but not HIV-2 (Tagieva and Vaquero, 1997).

The putative amino acid sequence encodes a highly hydrophobic protein with two predicted transmembrane helices (Miller, 1988). Further *in silico* analysis of the putative ASP protein sequence revealed a cysteine rich-region and a proline repeat motif (Clerc *et al*., 2011). Similar cys-rich motifs have also been observed in Oct-4 (Nordhoff *et al*., 2001) and HIV-1 Tat (Frankel *et al*., 1988, Greene, 1990). The proline repeat sequence motif is similar to the PxxP repeat sequence of the HIV-1 Nef protein (Picard *et al*., 2002) and the ORF-3 protein in Hepatitis E virus (Ray *et al*., 1992) which have been shown to interact with cellular protein kinases.

While negative sense transcription has been reported in HIV-1 (Michael *et al*., 1994b; Vanhee-Brossollet *et al*., 1995), the reported size of the negative sense transcripts encoding the putative ASP varies from 1.0 to 4.1kb (Bukrinsky and Etkin, 1990; Michael *et al*., 1994b; Landry *et al*., 2007; Kobayashi-Ishihara *et al*., 2012). Given that transcription of *asp* is initiated and regulated by the 3′LTR and the length of the proposed 5′UTR sequence is large, no studies of the region upstream of *asp* which may provide insight into the possible function and/or regulation of *asp*/ASP expression have been conducted. This chapter examines this region and proposes that *asp*/ASP regulation may occur by a series of sORFs present within the 5′UTR.

3.2 Aims

The work presented in this chapter examines the conservation of the HIV-1 ASP/*asp* sequences, including the sequences of six sORFs in the *asp* 5′UTR region, across the strains and subtypes of HIV-1, HIV-2 and SIV. The following *in silico* investigations were undertaken:

1. Detection of the *asp* ORF and its associated sORFs across the various strains and subtypes of HIV-1.
2. Detection of an *asp*-like ORF and associated sORFs in HIV-2, SIV and a selection of other retroviruses.
3. Re-investigation of the amino acid sequence conservation of the transmembrane domains, cysteine rich regions and proline repeat motifs as previously reported in the literature.
4. Examination of the nucleotide sequence conservation of the sORFs amongst the various subtypes of HIV-1.
5. Investigation of the conservation of the Kozak initiation strength of the sORFs amongst the various subtypes of HIV-1.
6. Investigation of the frequency of rare codons within sORF sequences.

3.3 Materials and methods

3.3.1 Conservation of ASP across strains and subtypes of HIV and associated species

A total of 37 HIV-1 complete genome sequences spanning the subtypes, A, B, C, D, E, F, G, H, J, K, N, O and U were randomly selected and downloaded from The HIV Sequence Database (http://hiv-web.lanl.gov) as detailed in Table 2.1. In addition, 5 strains of HIV-2 spanning the subtypes A, B and G, 8 SIV_{CPZ} sequences, 2 $SIV_{African\ Green\ Monkey}$ sequences, 2 $SIV_{Mandrill}$ sequences and 2 $SIV_{MAC/SMM}$ sequences were also collected to determine the presence/absence and relative similarity of ASP and the *asp* gene sequence (Sequence data Tables 3.2 and 3.3).

All sequences were scanned for ORFs using the Translation Overview tool in DNAMAN (Lynnon BioSoft) in all six reading frames with a line length of 8, minimum ORF length of 6 amino acids and width of 4. The location of the *env* gene was noted and any large ORF (~570bp) detected opposite was designated *asp*. The reading frame, number of encoded amino acids and position of the ORF were noted and the sequences collected. The DNA sequences collected were then translated into amino acid sequences using the Translation tool in DNAMAN. The resultant amino acid sequences were subjected to a multiple sequence alignment using Vector NTI (Invitrogen) with the following parameters: Gap opening penalty of 10, Gap extension penalty of 0.05, Gap separation penalty range of 8 and a % Identity for alignment delay of 40; and the features conserved across the different strains identified. The consensus sequence generated for ASP as a result of the multiple sequence alignment was analyzed using the Hydrophilicity profile tool in DNAMAN.

Table 3.1 HIV-1 sequences used in this study.

Name	Subtype/ Group	Accession	Source	Author	Reference
SE6594	A	AF069672	Sweden	Carr, JK	*AIDS* **13**(14):1819-26 (1999)
SE8538	A	AF069669	Sweeen	Carr, JK	*AIDS* **13**(14):1819-26 (1999)
U455	A	M62320	Uganda	Oram, JD	*ARHR* **6**(9):1073-8 (1990)
UGO37	A	U51190	Uganda	Gao, F	*J Virol* **70**(3):1651-67 (1996)
NL43	B	AF003887	U.S.A	Fang, G	*J Acquir Immune Defic Syndr Hum Retrovirol* **12**(4):352-7 (1996)
OYI	B	M26726	Gabon	Huet, T	*AIDS* **3**(11):707-15 (1989)
MN	B	M17449	U.S.A	Gurgo, G	*Virology* **164**(2):531-6 (1988)
ACH1	B	U34604	Netherlands	Guillon, C	*ARHR* **11**(12):1537-41 (1995)
BRU	B	K02013	France	Wain-Hobson, S	*Cell* **40**(1):9-17 (1985)
CAM1	B	D10112	Brazil	Saurya, S	*ARHR* **19**(1):73-6 (2003)
HXB2	B	K03455	France	Wong-Staal, F	*Nature* **313**(6000):277-84 (1985)
92BR025	C	U52953	Brazil	Gao, F	*J Virol* **70**(3):1651-1667 (1996)
ETH2220	C	U46016	Ethiopia	Salminen, MO	*ARHR* **12**(14):1329-39 (1996)
94UG114	D	U88824	Uganda	Gao, F	*J Virol* **72**(7):5680-98 (1998)
ELI	D	K03454	D.R.C	Alizon, M	*Cell* **46**(1):63-74 (1986)
NDK	D	M27323	D.R.C	Spire, B	*Gene* **81**(2):275-84 (1989)
92NGO83	E	U88826	Nigeria	Gao, F	*J Virol* **72**(7):5680-98 (1998)
93TH253	E	U51189	Thailand	Gao, F	*J Virol* **7**0(10):7013-29 (1996)
FIN9363	F	AF075703	Finland	Laukkanen, T	*Virology* **269**(1):95-104 (2000)
VI850	F	AF077336	Belgium	Laukkanen, T	*Virology* **269**(1):95-104 (2000)
CM53657	F	AF377956	Cameroon	Carr, JK	*Virology* **296**(1):168-81 (2001)
MP257	F	AJ249237	Cameroon	Peeters, M	*ARHR* **16**(2):139-51 (2000)
DRCBL	G	AF084963	Belgium	Debyser, Z	*ARHR* **14**(5):453-9 (1998)
HH8793	G	AF061641	Kenya	Salminen, MO	*ARHR* **8**(9):1733-42 (1992)
X558	G	AF423760	Spain	Delgado, E	*J Acquir Immune Defic Syndr* **29**(5):536-43 (2002)
V1991	H	AF190127	Belgium	Janssens, W	*AIDS* **14**(11):1533-43 (2000)
V1997	H	AF190128	Belgium	Janssens, W	*AIDS* **14**(11):1533-43 (2000)
SE9173	J	AF082395	D.R.C	Laukkaren, T	*ARHR* **15**(3):293-9 (1999)
SE92809	J	Af082934	Sweden	Laukkaren, T	*ARHR* **15**(3):293-9 (1999)
MP535	K	AJ249239	Cameroon	Peeters, M	*ARHR* **16**(2):139-51 (2000)
DJO0131	Group N	AY532635	Cameroon	Bodelle, P	*ARHR* **20**(8):902-908 (2004)
YBF30	Group N	AJ006022	Cameroon	Simon, F	*Nat Med* **4**(9):1032-7 (1998)
YBF106	Group N	AJ27137	Cameroon	Ayouba, A	*AIDS* **14**(16):6223-5 (2000)
ANT70	Group O	L20587	Belgium	Vanden Haesevelde, M	*J Virol* **68**(3):1586-96 (1994)
SEMP1300	Group O	AJ302647	Senegal	Vergne, L	*J Clin Microbiol* **38**(11)3919-25 (2000)
83CD003Z3	Group U	AF286236	D.R.C	Gao, F	*ARHR* **17**(12):1217-22 (2001)
CU68	Group U	AY894993	Cuba	Casado, G	Submitted **18 JAN 2005** *Patogenia Viral Instituto de Salud Carlos III*

Table 3.2 HIV-2 sequences used in this study.

Name	Subtype	Accession	Source	Author	Reference
BEN	A	M30502	Germany	Kirchhoff, F	*Virology* **177**(1):305-11 (1990)
ALI	A	AF082339	Guinea-Bissau	Azevedo-Pereira, JM	*Unpublished*
EHO	B	U27200	Cote d'Ivoire	Rey-Cuille, MA	*Virology* **202**(1):471-6 (1994)
D205	B	X61240	Ghana	Kreutz, R	*ARHR* **8**(9):1619-29 (1992)
ABT96	G	AF208027	Cote d'Ivoire	Brennan, CA	*ARHR* **13**(5):401-4 (1997)

Table 3.3 SIV sequences used in this study.

Name	Type	Accession	Source	Author	Reference
TAN1	CPZ	AF447763	Tanzania	Santiago, ML	*J Virol* **77**(3):2233-2242 (2003)
GAB2	CPZ	AF382828	Gabon	Bibollet-Ruche, F	*ARHR* **20**(12):1377-81 (2004)
ANT	CPZ	U42720	D.R.C	Vanden Haesevelde, M	*Virology* **22**1(2):246-50 (1996)
CAM3	CPZ	AF115393	Cameroon	Corbet, S	*J Virol* **74**(1):529-34 (2000)
CAM5	CPZ	AJ271369	Cameroon	Muller-Trutwin, MC	*J Med Primatol* **29**(3-4):166-72 (2000)
GAB	CPZ	X52154	Gabon	Huet, T	*Nature* **345**(6273):356-9 (1990)
ZUS	CPZ	AF103818	U.S.A	Gao, F	*Nature* **397**(6718):436-41 (1999)
VER9063	AGM	L40990	Kenya	Hirsch, VM	*J Virol* **69**(2):955-67 (1995)
SAB1C	AGM	U04005	Senegal	Jin, MJ	*EMBO J* **13**(12):2935-47 (1994)
STM	MAC/SMM	M83293	U.S.A	Novembre, FJ	*Virology* **186**(2):783-7 (1992)
MNE8	MAC/SMM	M32741	U.S.A	Kimata, JT	*J Virol* **72**(1):245-56 (1998)
MNDGB1	MND	M27470	Gabon	Tsujimoto, H	*Nature* **341**(6242):539-41 (1989)
5440	MND	AY159322	Unknown	Hu, J	*J Virol* **77**(8):4867-4880 (2003)

3.3.2 Investigation of *asp*-like sequences in other retroviruses

Complete genome sequences of seven retroviruses were randomly selected and collected from the NCBI database (http://www.ncbi.nlm.gov). The sequences selected included the retroviruses; Avian leukemia virus (ALV), Murine Leukemia virus (MLV), Mouse mammary tumor virus (MMTV), Mason-Pfizer monkey virus (M-PMV), Human T-cell leukemia virus (HTLV) and Human foamy virus (HFV) as detailed in Table 3.4 below. One representative sequence of each of virus was selected and scanned for ORFs using the Translation Overview tool in DNAMAN (Lynnon BioSoft) in all six reading frames with a line length of 8, minimum ORF length of 6 amino acids and width of 4. The location of the *env* gene was noted and thus the presence of any large ORF opposite was noted.

Table 3.4 Other retrovirus sequences used in this study.

Name	Accession	Host	Author	Reference
ALV	M37980	Fowl	Bieth, E	*Nucleic Acids Res* **20**(2):367 (1992)
MMTV	AF033807	Mouse	Petropoulos, CJ	*Retroviruses*: **757**, CSHL, USA (1997)
M-PMV	AF033815	Monkey	Petropoulos, CJ	*Retroviruses*: **757**, CSHL, USA (1997)
MLV	AF033811	Monkey	Petropoulos, CJ	*Retroviruses*: **757**, CSHL, USA (1997)
RSV	AF033808	Chicken	Petropoulos, CJ	*Retroviruses*: **757**, CSHL, USA (1997)
HTLV	AF033817	Human	Petropoulos, CJ	*Retroviruses*: **757**, CSHL, USA (1997)
HFV	Y07725	Human	Rethwilm, A	Submitted **30 AUG 1996** *Institut fuer Virologie und Immunbiologie*

3.3.3 Conservation of sORFs across strains and subtypes of HIV-1

All of the sequences upstream from the *asp* ORF were collected and scanned for ORFs using the Translation Overview tool in DNAMAN in all three minus reading frames with a line length of 8, minimum ORF length of 2 and width of 0. The number, position, reading frame and size of the sORFs were observed for each sequence. The resultant upstream nucleotide sequences were subjected to a multiple sequence alignment using Vector NTI (Invitrogen) with the following parameters: Gap opening penalty of 10, Gap extension penalty of 0.05, Gap separation penalty range of 8 and a % Identity for alignment delay of 40; and the features conserved across the different strains identified. The nucleotide conservation with respect to the consensus sequence generated was observed for each sORF. The Kozak consensus of each of the sORF initiation codon regions was assessed according to the consensus; GCC**R**CCAUG**G** (Kozak, 1984b, Kozak, 1987a), where the most highly conserved nucleotides are the R (either A or G) at position -3 (where A of the initiation codon is assigned +1) and the G at position +4. An initiation codon with both of these important nucleotides was considered strong; a sequence displaying one of these nucleotides was considered adequate, while an initiation codon without these two features was designated weak.

3.3.4 Comparison of codon usage preference within sORF and *asp* coding sequences

Comparison of codon usage preferences within the coding sequence of each sORF was examined. The frequencies (x100) of each codon, in highly expressed human genes were examined according to Haas *et al*. (1996). Frequencies in codon usage were denoted; frequencies >21, frequencies 11-20 and frequencies 0-10 and the percentages of each group calculated.

3.4 Results

3.4.1 Conservation of *asp* and associated sORFs across the atrains and subtypes of HIV and its ancestral SIVs

Over 85% of the 37 sequences selected for analysis in this study displayed a sequence similar to *asp* (>80% sequence homology to the consensus, derived from all sequences examined); consistent with a previous report that the *asp* ORF was well conserved amongst 550 genomic and *env* sequences of HIV-1 from both primary and laboratory sequences (Briquet and Vaquero, 2002), however comparisons across the subtypes were not mentioned in that study. Of all the HIV-1 sequences selected only five did not display an *asp* ORF opposite the *env* gene due to the lack of the AUG initiation codon for the *asp* ORF. Two of these were from subtype A sequences, two from subtype O and one from subtype U.

The reading frames of each *asp* ORF vary across all three negative frames, while the size of putative protein also varies markedly from 108 to 190aa. This data is consistent with that of Miller (1988), who examined 12 sequences from HIV-1 (mainly from subtypes A and B), where *asp* is highly conserved. This is also in agreement with reports from Briquet and Vaquero (2002). Reading frame and size were not examined in these studies.

As well as the negative sense ORF for *asp* the upstream region has been observed to contain a number of sORFs (Deacon *et al*., *unpublished*). The number and conservation of these sORFs was also examined. While all sequences possessing an *asp* ORF also displayed sORFs, the number of sORFs varied across the strains of HIV-1 (Table 3.5). None of the HIV-2 sequences displayed an *asp*-like ORF (data not shown).

Table 3.5 Conservation of *asp* amongst selected HIV-1 sequences. The reading frame (RF) and number of encoded amino acids (AA) noted for each sequence. Strains highlighted in green do not display a PxxP motif and the sORF immediately upstream of *asp* is not in frame. The strain highlighted in pink does not display the immediate upstream in frame sORF. The strain highlighted yellow does not display the PxxP motif, but has an in-frame upstream sORF.

HIV-1 Strain	Subtype/Group	RF	AA	Position (nt)	sORFs
SE6594	A	-3	144	6603-7035	3
SE8538	A	-	-	-	-
U455	A	-	-	-	-
UGO37	A	-2	158	6752-7226	5
NL43	B	-1	189	7363-7930	6
OYI	B	-3	185	6924-7497	3
MN	B	-2	157	7493-7964	5
ACH1	B	-1	186	7396-7959	5
BRU	B	-3	189	6966-7533	6
CAM1	B	-3	190	7371-7941	6
HXB2	B	-2	189	7373-7940	6
92BR025	C	-2	186	6719-7277	5
ETH2220	C	-2	182	6770-7316	5
94UG114	D	-1	188	6694-7258	6
ELI	D	-3	188	6912-7476	4
NDK	D	-1	108	7120-7444	5
92NG083	E	-3	179	6729-7266	2
93TH253	E	-1	185	7369-7924	3
FIN9363	F	-3	114	6894-7236	4
VI850	F	-2	175	6692-7217	4
CM53657	F	-1	161	6718-7201	3
MP257	F	-2	158	6749-7223	5
DRCBL	G	-1	182	7333-7879	4
HH8793	G	-1	178	6778-7312	4
X558	G	-1	186	7811-7369	5
V1991	H	-1	108	7534-7858	1
V1997	H	-3	109	6903-7230	3
SE9173	J	-3	190	6684-7254	4
SE92809	J	-2	142	6821-7247	4
MP353	K	-2	109	6773-7100	4
DJO0131	Group N	-2	135	6959-7364	1
YBF30	Group N	-3	138	7074-7488	1
YBF106	Group N	-3	136	7035-7443	1
ANT70	Group O	-	-	-	-
SEMP1300	Group O	-	-	-	-
83CD003Z3	Group U	-3	109	7089-7416	2
CU68	Group U	-	-	-	-

The majority of the SIV sequences failed to display an *asp*-like ORF (Table 3.6); *asp*-like sequences were seen in SIV_{CPZ} sequences only. Of the SIV_{CPZ} strains, 3/7 displayed an ORF for *asp* (Table 3.6) and these sequences were collected and translated into amino acid sequences and incorporated into the multiple sequence alignment (Figure 3.1) in order to assess their similarity to the HIV-1 strains.

Table 3.6 Conservation of ASP amongst select SIV sequences. The reading frame (RF) and amino acid size (AA) noted for each sequence. The strain highlighted in pink does not display the immediate upstream in frame sORF.

SIV	Type	RF	AA	Position (nt)	sORFs
TAN1	CPZ	-	-	-	-
GAB2	CPZ	-	-	-	-
ANT	CPZ	-	-	-	-
CAM3	CPZ	-3	140	6933-7353	6
CAM5	CPZ	-3	138	7245-7659	4
GAB	CPZ	-	-	-	-
ZUS	CPZ	-1	150	7495-7945	2
VER9063	AGM	-	-	-	-
SAB1C	AGM	-	-	-	-
STM	MAC/SMM	-	-	-	-
MNE8	MAC/SMM	-	-	-	-
MNDGB1	MND	-	-	-	-
5440	MND	-	-	-	-

3.4.2 Conservation of ASP amino acid sequence and structural features across the strains and subtypes of HIV and its ancestral SIVs

In order to observe the relative conservation of the putative protein produced from the proposed *asp* ORF, all the nucleotide sequences collected were translated into amino acid sequences and subjected to a multiple sequence alignment. A number of amino acid sequence motifs were observed (Figure 3.1 A).

The ASP sequences presented a high degree of similarity with the majority of the sequence conserved above 50%. This is also highlighted in the relative similarity plot (Figure 3.1 B) where a large number of peaks representing high similarity are observed, particularly at the N-terminal of the putative protein. The unusual cystine rich region and PxxP motif located close to the N-terminal of the sequence are also highly conserved.

A hydrophilicity profile was produced in order to check that the two transmembrane domains predicted in 1988 by Miller were conserved amongst these sequences. A profile of the consensus sequence generated indicated that only one transmembrane domain was conserved (Figure 3.2), consistent with the high conservation of the N-terminal portion of the sequence.

Each sequence was examined independently to determine the presence of transmembrane domains amongst the different clade viruses. Half (16/32) of the sequences examined displayed only one transmembrane domain (that being the most highly conserved domain, closest to the N-terminal at amino acid position 10-26). Of the remaining sequences, 14 displayed both transmembrane domains (at amino acid positions 10-26 and 59-75), while two displayed a third very weak transmembrane domain at the C-terminal (amino acid position 152-168). It is interesting to note that the majority of sequences displaying 2-3 transmembrane domains are either B or C clade viruses, the major infecting strains.

Figure 3.1 Amino acid sequence alignment of ASP sequences and conservation of sequence motifs. (A) Amino acid alignment of 32 HIV-1 and 3 SIV_{CPZ} (highlighted in red) Asp sequences. The consensus sequence, derived from all sequences, is shown at the bottom with the conservation of the Cys-rich and PxxP motifs highlighted in red. D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; Y, Tyr. Unhighlighted amino acids represent non-similar residues. Blue amino acids represent consensus residues derived from a block of similar residues. Green amino acids represent a block of weakly similar residues. (B) Relative similarity plot of the entire amino acid alignment whereby a value (and thus peak) of 1.00 (line) represent 100% similarity of a residue.

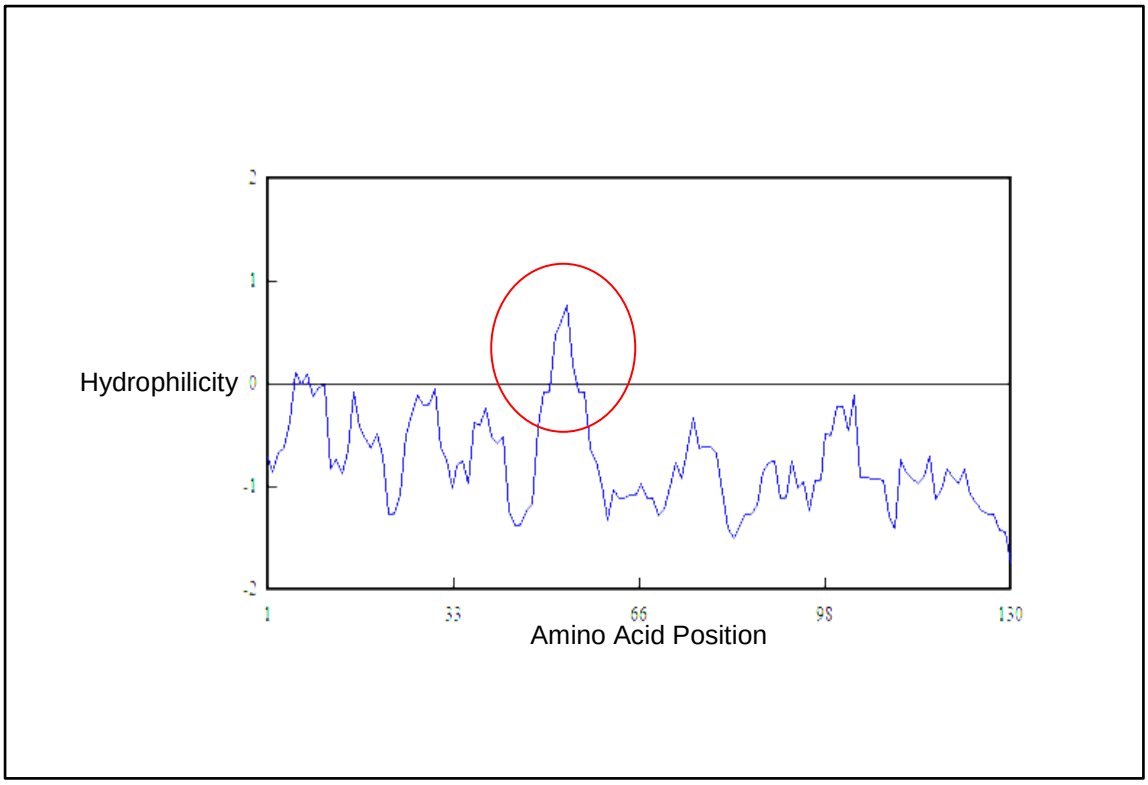


Figure 3.2 Hydrophilicity profile of the consensus sequence generated from the multiple sequence alignment with transmembrane region highlighted.

Table 3.7 Analysis of the number of transmembrane domains present in all ASP amino acid sequences.

HIV-1 Strain	Subtype/Group	Domains
SE6594	A	1
UGO37	A	1
NL43	B	2
OYI	B	2
MN	B	2
ACH1	B	2
BRU	B	3
CAM1	B	1
HXB2	B	2
92BR025	C	2
ETH2220	C	2
94UG114	D	2
ELI	D	3
NDK	D	1
92NG083	E	1
93TH253	E	1
FIN9363	F	1
VI850	F	1
CM53657	F	1
MP257	F	2
DRCBL	G	1
HH8793	G	1
X558	G	1
V1991	H	2
V1997	H	1
SE9173	J	2
SE92809	J	1
MP353	K	1
DJO0131	Group N	2
YBF30	Group N	2
YBF106	Group N	2
83CD003Z3	Group U	1

3.4.2 Investigation of *asp*-like ORF in other retroviruses

Two groups have previously reported the existence of similar antisense genes in *env* containing retroviruses, including the non-primate FIV (Briquet *et al*., 2001) and HTLV-1 (Arnold *et al*., 2006; Cavanagh *et al*., 2006; Ludwig *et al*., 2006). Several other retrovirus sequences were selected to check for the presence of any other large ORFs opposite *env* that may be similar to *asp*. No such ORFs were present in the 7 sequences examined.

3.4.3 Conservation of sORFs across strains and subtypes of HIV-1

The conservation of upstream sORFs amongst the HIV-1 sequences that displayed an *asp* ORF was assessed. The number of sORFs varied from 1 through to 6 (Table 3.5). Subtypes B and C displayed a more consistent number of sORFs (5 to 6) which varied in size from 6 to 48aa. Alignment of these sequences revealed the high conservation of this upstream region amongst the majority of the sequences examined (Table 3.8).

Multiple sequence alignment (not shown) of the region upstream of the *asp* ORF (nts 8742 to 7932 in HIV-1 NL4-3) in 31 HIV-1 sequences, representing all subtypes, revealed a series of well conserved sORFs that displayed nucleotide sequence conservation >80% with reference to the consensus sequence (Table 3.8). The major infecting HIV-1 subtypes (B, C and D) displayed 5-7 sORFs (typically 6 sORFs), while fewer sORFs (3-5) were observed in the rarer subtypes: (H, K and N). Lower levels of conservation (50-60%) were also observed in the sequences of the rarer subtypes, N and F. While sequences of sORFs were highly conserved, an AUG initiation codon was not always displayed (indicated by the absence of highlighting in Table 3.8); this was more common in the less predominant subtypes (E, F, G, H, J, K and groups N and U).

The sequences of sORFs I, V, VI and VI_{alt} (a sORF initiated at an alternative AUG codon present within sORF VI) were highly conserved amongst all subtypes, with strong conservation of the initiation codon (Table 3.8). One

sequence (CAM1) which did not display the initation codon for sORF VI, displayed only the alternate initiation codon, VI_{alt}. The sequences of sORFs II, III, IV and V were less well conserved (between 63-96%).

The translation initiation context of each sORF was assessed as weak, adequate or strong by comparison to the Kozak consensus sequence (Kozak, 1986). The sequence context of sORFs I (26/31 sequences), V (21/31), VI (25/31) and VI_{alt} (17/31) are typically adequate, consistent with other sORF regulating systems, such as that of yeast GCN4 (Hinnebusch, 1997). Interestingly sORFs II and IV display a weak initiation codon context in most HIV subtypes, but a strong initiation context is observed in B and C clade strains.

Table 3.8 Analysis of the Kozak initiation strength of each sORF and the percentage homology of the sORF nucleotide sequence to the consensus sequence. Shadings represent the strength of the initiation codon; red, yellow, green represent initiation codons of weak, adequate, strong context respectively and absence of shading reflects absence of initiation codon.

		sORF						
Subtype	**Strain**	**I**	**II**	**III**	**IV**	**V**	**VI**	**VI$_{alt}$**
A	SE6594	90	90	87	95	86	89	98
A	UG037	90	88	88	90	85	92	98
B	NL43	90	88	91	90	88	87	95
B	OYI	81	83	90	95	89	91	95
B	MN	95	88	86	95	87	89	93
B	ACH1	95	83	88	95	87	90	95
B	BRU	95	90	91	90	88	87	95
B	CAM1	100	88	86	95	90	89	98
C	92BR025	76	79	94	95	85	91	100
C	ETH2220	67	90	94	95	84	92	100
D	94UG114	81	90	90	95	83	88	90
D	ELI	81	83	90	95	87	88	93
D	NDK	81	88	90	95	88	86	88
E	92NG083	86	90	81	95	88	95	100
E	93TH253	89	73	77	95	82	83	88
F	FIN9363	86	96	90	95	84	94	95
F	VI850	86	92	94	95	88	94	98
F	CM53657	95	88	87	95	80	57	-
F	MP257	86	88	92	95	81	53	-
G	DRCBL	86	90	89	95	83	94	100
G	HH8793	86	90	86	95	85	93	98
G	X558	76	85	86	90	87	93	98
H	V1991	90	94	-	-	-	-	-
H	V1997	90	94	92	95	84	91	95
J	SE9173	90	98	86	95	85	93	98
J	SE92809	90	98	87	95	85	93	98
K	MP535	81	96	90	95	88	91	100
Group N	DJO 0131	57	75	75	81	63	67	74
Group N	YBF30	57	81	77	81	65	66	74
Group N	YBF100	57	77	83	81	63	66	74
Group U	83CD003Z3	90	98	85	87	84	95	98

It is interesting to note the variation in initiation codon strength of each respective sORF. Of all the sORFs, initiation codons of sORFs I, V and VI consistently displayed (~80%) an adequate sequence context. The sequence context of the initiation codon of sORF III is also adequate when it is present. The sORFs II and IV generally display a weak initiation codon context, except amongst the B and C clade strains where a strong initiation context is observed. Several sequences, while presenting quite high conservation of the sORF sequence (77-98%) did not present an initiation codon sequence. The majority of the sequences displaying this trait included the rarer clade viruses; H, K, N and U and primarily involved sORFs II, III and IV (rather than sORFs I, V and VI, the more highly conserved sORFs). In most cases (31/32) the last sORF (typically sORF VI) is in frame with the *asp* ORF. Where present, sORF VI_{alt} typically (17/25) displays an adequate sequence context, 8/25 of the sequences examined display weak sequence context.

The codon usage frequencies of the codons within the sequences of each sORF were also examined, as large proportions of less frequent codons may slow the translating ribosome and therefore have potential to regulate the speed and level of translation. Codon usage has previously been shown to limit expression of the HIV-1 envelope glycoprotein (Haas *et al*., 1996). The antiviral role of human schlafen 11 (SLFN11) and its ability to reduce production of HIV proteins by targeting tRNAs associated with rare codons has also been demonstrated (Li *et al*., 2012).

As depicted in Table 3.9, sORF I displayed the highest percentage of rare codons (frequency between 0-10), none of the codons within this sORF were within the pool of common codons (frequencies >21). Interestingly the proportion of rare codon use gradually decreases across the sORF sequences; sORF II displays an almost even spread of codon frequencies, while sORFs III-VI display higher proportions of the more common codons (frequencies >21). The *asp* coding sequence displayed a high abundance of common codons (frequencies >21), with only 26% of the coding sequence displaying rare codons (frequencies <10).

Table 3.9 Comparison of codon usage frequencies within sORF and *asp* sequences. Figures represent the percentages of each codon frequency (x100) in highly expressed human genes (frequencies reported from Haas *et al*., 1996).

sORF	**Codon Usage Frequencies**		
	0-10	**11-20**	**>21**
	Rare		**Common**
I	80%	20%	-
II	29%	36%	35%
III	11%	52%	37%
IV	-	20%	80%
V	18%	31%	51%
VI	9%	39%	52%
VI_{alt}	8%	42%	50%
asp	26%	34%	40%

3.5 Discussion

While the conservation of the *asp* ORF in HIV-1 has been reported previously in the literature (Miller, 1988; Briquet and Vaquero, 2002), the main focus of this work describes the presence and high level conservation of a series of sORFs upstream of the *asp* ORF across all groups/ subtypes of HIV-1. Upstream sORFs are typically associated with translational regulation of the downstream gene product; examples include many genes for which tight regulation of expression is critical (Davuluri *et al.*, 2000; Morris and Geballe, 2000; Suzuki *et al.*, 2000). The data presented here has demonstrated that the sORF region upstream of HIV-1 *asp* is highly conserved amongst the various clades of HIV-1 and has the potential to control gene expression.

The re-examination of the putative HIV-1 ASP amino acid sequence indicates a high level of conservation across all groups and subtypes, except for group O. The strong conservation of ASP is significant, considering its position directly opposite the gp120/gp41 junction of *env*, the gene that displays the greatest sequence diversity in HIV-1 (Martins *et al*., 1991) suggesting that conservation of ASP is functionally important. Many studies have successfully detected and partially characterized the *asp* transcript and translation products *in vitro* (Bukrinsky *et al*., 1990; Michael *et al*., 1994b; Briquet *et al*., 2002; Landry *et al*., 2007; Kobayashi-Ishihara *et al.*, 2012). The absence of an ASP ORF in Group O viruses (due to the lack of a suitable AUG codon) may reflect their more distant evolutionary relationship from the major M group HIV-1 subtypes. Further to this, Group O viruses are most likely derived from the wild gorilla and not from a chimpanzee reservoir as are Group M viruses (Van Heuverswyn *et al*., 2006).

The absence of an *asp* ORF in HIV-2 sequences is consistent with the work of Tagieva and Vaquero (1997), which showed that endogenously expressed *asp* RNA inhibits the replication of HIV-1 but not HIV-2. A selection of the SIV sequences of African Green Monkey, Sooty Mangabey and Mandrill, the close ancestors of HIV-2, and the SIV Chimpanzee strains, the closer relatives of HIV-1, was also examined. 3/7 of the SIV_{CPZ} sequences displayed an *asp*-like ORF; none of the African Green

Monkey, Sooty Mangabey or Mandrill sequences displayed an *asp* ORF. These data are consistent with primate ancestry of HIV-1 and HIV-2.

Examination of the putative ASP amino acid sequences revealed the high conservation of two previously reported motifs (Clerc *et al*., 2011); the cysteine triplets and the PxxP repeat motifs. These motifs were detected within the most highly conserved regions of the protein across strains/ groups of the virus examined. Similar cys-rich motifs have also been observed in Oct-4, where they form a transcription repressor/ activator binding site, regulating transcription of the gene (Nordhoff *et al*., 2001). Cys-rich regions may also be involved in aggregation of proteins, as observed with the extracellular portions of type 1 and 2 (p55 and p75) tumor necrosis factor receptors (Marsters *et al*., 1992) and the snake venom hemorrhagic metalloproteinase toxins (Jia *et al*., 1997). Similar PxxP motifs have previously been reported in HIV-1 Nef (Picard *et al*., 2002) and ORF-3 in Hepatitis E virus (Ray *et al*., 1992). In both of these proteins the PxxP region has been shown to interact with cellular protein kinases; perhaps this may also occur in ASP. No function has yet been ascribed to these cys-rich and PxxP motifs in ASP.

The data presented re-confirms the highly hydrophobic nature of the putative ASP, as evidenced by the high conservation of at least one transmembrane domain within the sequences examined here. Briquet and Vaquero (2002) reported three transmembrane helices at amino acid positions 10-26, 59-75 and 152-168 of HIV-1_{BRU} ASP. This study confirmed that the central helix at 152-168 is 100% conserved amongst the sequences examined. The sub-cellular localisation of the ASP protein still remains controversial. Briquet and Vaquero (2002) conducted immunoelectron microscopy experiments that indicated ASP localization to within the vicinity of the mitochondria as well as the virions released from infected cells. More recently *in vitro* studies suggested the localisation of ASP to the plasma membrane (Clerc *et al*., 2011) and the potential association with autophagosomes, suggesting a role in autophagy (Torresilla *et al*., 2013).

Examination of the upstream region of the *asp* ORF revealed high conservation of a series of sORFs. This series of (typically) six sORFs was more strongly conserved amongst the A, B, C and D clades of HIV-1; with sORFs I, V, VI and VI_{alt} being highly

conserved amongst all the clades examined. The conservation of the initiation context of the sORFs is also much stronger for sORF I, V, VI and VI_{alt}, typically displaying an adequate sequence context. This data is consistent with studies of sORF regulating systems, where the majority of the sORFs display a weak or adequate sequence context (Hinnebusch, 1997).

In addition, the sequences of each sORF were examined for the proportion of common and rare codon use. This analysis revealed that the later sORFs; V, VI and VI_{alt}; displayed higher abundance of more common codons. The sequences of sORFs I, II and III contain a mixture of codon frequencies with rare to moderately abundant codons. In particular, sORF I contains a higher percentage of rare codons (frequencies between <10). Regions with an abundance of rare codons may potentially slow or stall the translating ribosomal unit to reduce the levels of gene expression (Lavner and Kotlar, 2005; Plotkin and Kudla, 2011). Similar events have been reported in the Papillomavirus, whereby less frequent codon compositions minimize the expression of bovine papillomavirus type 1 late genes (Zhou *et al*., 1999). Interestingly SLFN11, an interferon-induced protein that can inhibit retrovirus protein synthesis, plays a vital role in HIV-1 infection by the selective inhibition of viral protein expression according to codon bias (Li *et al*., 2012). While the exact mechanism of SLFN11 interaction with tRNAs is still unknown, the higher abundance of rare codons observed here may permit the highly controlled expression of genes via a similar mechanism, particularly in the later stages of infection where *asp* expression has not been detected (Bukrinsky and Etkin, 1990).

3.6 Conclusions

ASP is highly conserved across HIV-1 and its ancestors the SIV_{CPZ} but not HIV-2 and its SIV ancestors. Several sequence motifs, including cystine triplets, a PxxP motif and a transmembrane domain; are also highly conserved and may be keys to the function/s of the putative protein.

Along with *asp*, an upstream region presenting a series of sORFs is also highly conserved, particularly among the major infecting HIV-1 strains. sORFs I, V and VI showed the highest level of conservation and present initiation codons with either a weak or adequate sequence context, suggestive of a sORF regulating system with the potential to regulate downstream expression. The proportion of rare codons within these sORFs may also play a role in regulation. It is possible, therefore, that ASP production may be regulated by this region. This potential for sORFs to regulate downstream gene expression will be addressed in the following chapters of this thesis.

CHAPTER 4 – PRELIMINARY STUDIES OF THE sORF REGION AND THE CONTROL OF DOWNSTREAM GENE EXPRESSION

4.1 General introduction

Regulation of gene expression allows for the precise timing and control of expression. However, the exact mechanisms by which genes are turned 'on and off' in different cell type/states and under different conditions is an area of ongoing research. Gene expression may be controlled at many levels in eukaryotes; including transcription, post-transcriptional processing, messenger RNA (mRNA) stability, translation, post-translational modifications and protein turnover.

In the previous chapter, the conservation of the six sORFs present on the *asp* transcript was discussed. These sORFs are highly conserved amongst the major infecting subtypes (B, C and D) of HIV-1 and may therefore play a role in the regulation of *asp* expression. In this chapter, the effects the sORF region plays on gene expression are investigated. This work was conducted by sub-cloning the *asp* upstream sORF region into the pEGFP-N1 plasmid, enabling the effects of gene expression to be measured directly according to the output of EGFP in transfected cells. This is a useful model system to study gene expression at different levels (eg. transcriptional control and/or translational control).

4.2 Aims

The effect the *asp* sORF region has on downstream gene expression was established by its direct effect on the downstream reporter EGFP. The HIV-1 *asp* (NL-43 strain) sORF region was sub-cloned into the multiple cloning site of the pEGFP-N1 vector; specifically aiming to:

1. Test the effect that the sORF region has on downstream gene expression.
2. Determine whether the transcriptional activators PMA and TSA affect the levels of downstream gene expression.
3. Use co-transfection studies to determine whether transcripts and their translation products have the potential to affect gene expression.
4. Test the role of each individual sORF on downstream gene expression by sequential mutation of the initiation codon of each sORF.

4.3 Materials and methods

4.3.1 Reporter gene constructs

The base plasmid constructs were composed as previously described in section 2.3. The entire sORF region (I-VI) was provided by Dr Nicholas Deacon and Ms Fee Yee Wong (Macfarlane Burnet Institute, Victoria, Australia). This construct contained the sORFs upstream of HIV-1 *asp* (NL-43 strain) inserted into the multiple cloning site of the pEGFP-N1 vector (Clontech) upstream of the reporter gene enhanced green florescent protein (EGFP) with a distance of 92bp between the initiation codon of EGFP and the stop codon of sORF VI. This construct was re-engineered to correct the intercistronic distance and frame shift the stop codon of sORF VI such that it sits 63bp (in-frame) upstream of the AUG initiation codon of the reporter EGFP. Unless otherwise stated all further constructs contained this corrected 63bp intercistronic distance (Figure 4.1). In all constructs the AUG initiation context of EGFP was the same as that in *asp* ORF. The pEGFP-N1 and pE-N1 were used as positive and negative controls respectively as detailed in section 2.3, Figure 2.1.

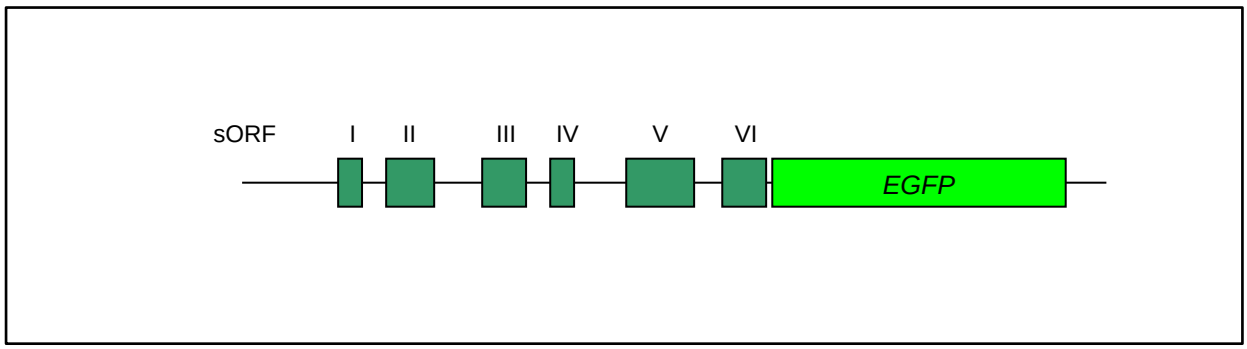


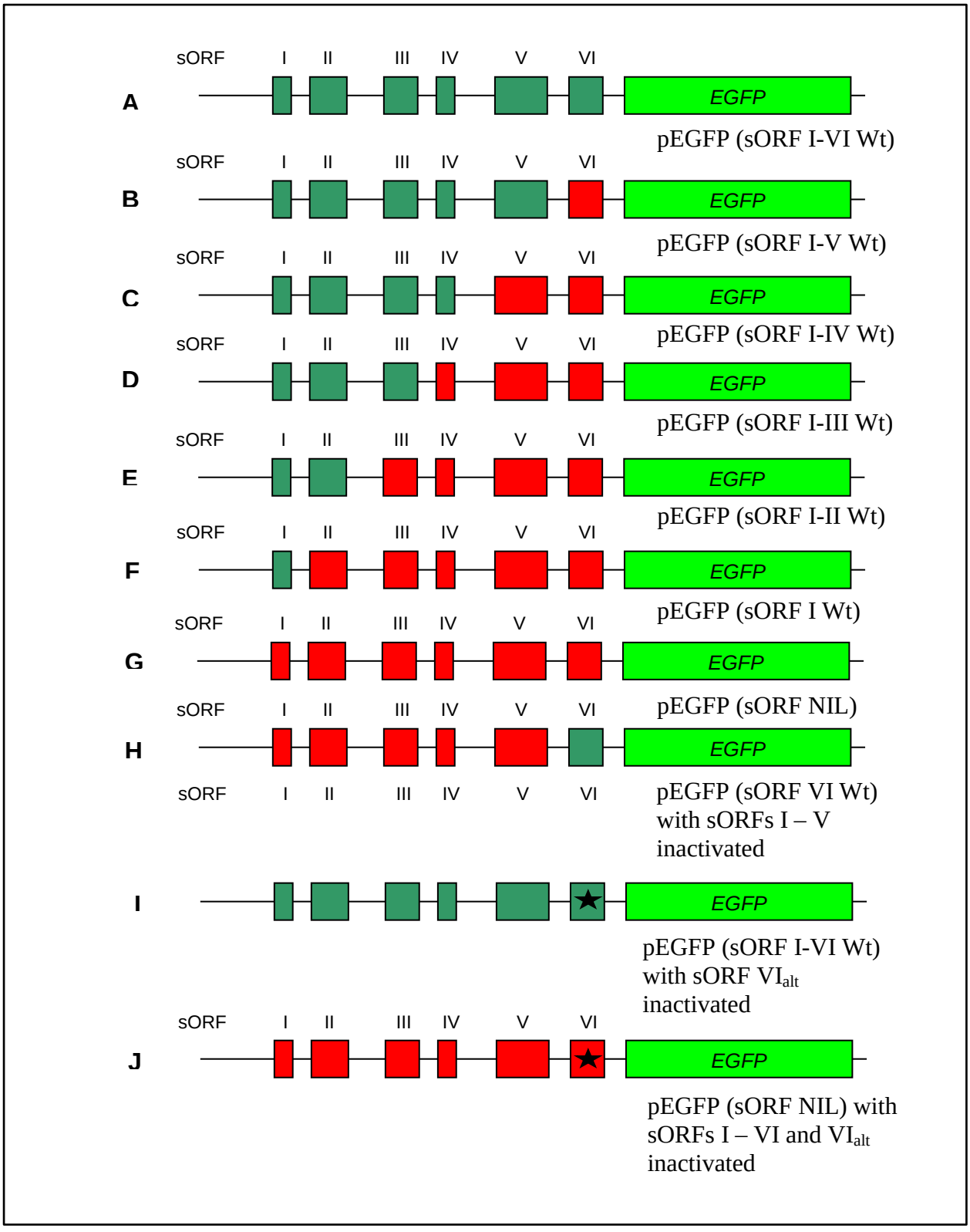
Figure 4.1 Schematic of base plasmid used throughout this study. pEGFP (sORF I-VI Wt) plasmid with sORFs I – VI sub-cloned in the MCS with an intercistronic spacing of 63bp between sORF VI and EGFP.

Site directed mutagenesis was used to knock out the initiation codons of each sORF, sequentially, starting from sORF VI. Mutagenesis was carried out using the QuikChange® II Site-Directed Mutagenesis Kit (Stratagene) according to the manufacturer's instructions using the primers depicted in Table 4.1 below. Following PCR amplification of the mutatant, the plasmid DNA was digested with 10units Dpn I and transformed into competent *E.coli* as detailed in section 2.4.2. Plasmids were amplified as described in 2.4.3, and then purified as detailed in sections 2.4.4 to 2.5.3 followed by sequencing to confirm the presence of the mutation as detailed in 2.5.4.

Table 4.1 Site directed mutagenesis primers for the knock-out of each sORF initiation codon. The following primers (and their complementary primer) were used to mutate the ATG initiation codons of each sORF with mutated sequence in bold.

sORF	**Sequence (5' - 3')**	**T_M (°C)**
I	GCTTATAGAGCTATTCGCCA**CCT**ACCTAGAAGAATAAGACAGG	67
II	GCTCAATGCCACAGC**CCT**AGCAGTAGCTGAGGGG	70
III	GGAGAGAGAGACAGGAGAACAGATC**CCT**TCGATTAGTGAACGG	70
IV	GGCAGGGATATTCAC**CCT**TATCGTTTCAGACCC	66
V	GGCTGTGGTATATAAAATTATT**CCT**AATGATAGTAGGAGGC	63
VI	GGAACAGATTTGGAATAA**CCT**GACCTGGATGGAGTGG	65
VI_{alt}	GCTCTGGAAAACT**CCT**TTGCACCACTGC	63

Figure 4.2 Schematic of plasmids used throughout this study to investigate the effect of upstream sORFs on gene expression. Red shading represents sORFs with mutated initiation codons, green represents non-mutated sORFs. (A) pEGFP (sORF I-VI Wt) plasmid with sORFs I – VI sub-cloned in the MCS, (B) pEGFP (sORF I-V Wt) plasmid with sORF VI inactive, (C) pEGFP (sORF I-IV Wt) plasmid with sORFs V-VI inactivated, (D) pEGFP (sORF I-III Wt) plasmid with sORFs IV – VI inactivated, (E) pEGFP (sORF I-II Wt) plasmid with sORFs III – VI inactivated, (F) pEGFP (sORF I Wt) plasmid with sORFs II – VI inactivated, (G) pEGFP (sORF NIL) plasmid with sORFs I – VI inactivated, (H) pEGFP (sORF VI Wt) plasmid with sORFs I – V inactivated, (I) pEGFP (sORF I-VI Wt) plasmid with sORF VI_{alt} inactivated, indicated by a star, and (J) pEGFP (sORF NIL) plasmid with sORFs I – VI and VI_{alt} inactivated.


sORF
I
II
III
IV
V
VI
EGFP
A
pEGFP (sORF I-VI Wt)
B
pEGFP (sORF I-V Wt)
C
pEGFP (sORF I-IV Wt)
D
pEGFP (sORF I-III Wt)
E
pEGFP (sORF I-II Wt)
F
pEGFP (sORF I Wt)
G
pEGFP (sORF NIL)
H
pEGFP (sORF VI Wt) with sORFs I – V inactivated
I
pEGFP (sORF I-VI Wt) with sORF VIalt inactivated
J
pEGFP (sORF NIL) with sORFs I – VI and VIalt inactivated

4.3.2 Investigation of sORF effects on reporter expression

HEK293 cells were prepared as detailed in section 2.6.1 and transfected with 1µg plasmid (as detailed in Figure 4.1) DNA as detailed in section 2.6.2. The samples were assayed as described in section 2.8 and the data analyzed as detailed in 2.9.

4.3.3 Investigation of transcriptional activators on reporter expression

HEK293 cells were prepared as detailed in section 2.6.1 and transfected with 1µg of the pEGFP (sORF I-VI Wt) plasmid along with the positive control pEGFP-N1 and negative control pE-N1 plasmids as detailed in section 2.6.2. Cells were treated with the activators NaBut or TSA 24 hours post transfection as detailed in 2.6.3. 24 hours post activation, the samples were assayed according to section 2.8 and the data analysed as detailed in 2.9.

4.3.4 Co-transfection studies with pDsRed-N1

HEK293 cells were prepared as detailed in section 2.6.1 and transfected with equal mass of pEGFP-N1 plasmid and pDsRed-N1 (Clontech) plasmid at a ratio of 1:1 (w/w) DNA and transfected as detailed in 2.6.2. 48 hours post transfection the samples were assayed as described in section 2.8. The fluorescence assay for pDsRed-N1 was measured in the cell lysates for intensity of red fluorescence as a direct measure of the amount of DsRed expression. Fluorescence was measured using the FLUOstar OPTIMA micro-plate reader (BMG Labtech) in black 96-well plates at an excitation of 560nm and emission of 580nm and the data analysed as detailed in 2.9. Total RNA was extracted and cDNA synthesized for RT-PCR analysis as detailed in 2.9. This was followed with northern blotting as per section 2.10 with the addition of a red florescent protein (RFP) transcript probe. RFP specific primers (Table 4.2) were used to amplify a small region from the pDsRed-N1 plasmid with the PCR DIG Probe Synthesis Kit (Roche) according to the manufacturer's instructions as detailed in section 2.10.1.

Table 4.2 PCR primers used to generate DIG-Labelled RFP probe for Northern blotting.

Probe	Sequence (5′ - 3′)	T_M(°C)
RFP For	GCGCTCCTCCAAGAACGTCATCAAGG	63
RFP Rev	GGAGGTGATGTCCAGCTTGGAGTCCACG	66

4.3.5 Investigation of active sORF initiation codons on reporter expression

HEK293 cells were prepared as detailed in section 2.6.1 and transfected with 1µg plasmid (as detailed in Figure 4.2) DNA as detailed in section 2.6.2. The samples were assayed according to section 2.8 and the data analyzed as detailed in 2.9. Total RNA was extracted and cDNA synthesized for RT-PCR analysis as detailed in 2.9. This was followed with northern blotting as per section 2.10.

4.4 Results

4.4.1 Investigation of sORF inhibition

The region nt 8798 to 7980 from pNL4-3, containing the six sORFs (sORF I-VI Wt), was cloned upstream of the reporter gene EGFP in the plasmid pEGFP-N1, and transfected into HEK293 cells. EGFP expression in cells transfected with the pEGFP(sORF I-VI Wt) construct was compared to cells transfected with the base plasmid, pEGFP-N1, and pE-N1 (lacking the EGFP gene). The presence of the upstream sORF region reduced EGFP expression by at least 95% ($p = <0.01$, $n=6$) (Figure 4.3). This confirms that this sORF region regulates downstream gene expression, consistent with the literature describing the effect sORFs may play on downstream expression (reviewed by Morris and Geballe, 2000).

The initial constructs contained an intercistronic distance between the stop codon of sORF VI and start codon of EGFP of 92bp. In HIV-1 NL4-3 this distance is much shorter at only 48bp and in-frame. Thus this distance was subsequently shortened to 63bp. The stop codon of sORF VI and start codon of EGFP were also fused in-frame with one another, to better reflect the orientation of sORF VI and *asp* in HIV-1 NL4-3. Cloning difficulties and the locations of restriction digest locations prevented the intercistronic distance being reduced further in this construct. The intercistronic distance did not affect downstream gene expression when all the sORFs were included in the construct (Figure 4.3) ($p=0.167$, $n=6$). All further construct manipulations were completed with this shorter intercistronic distance.

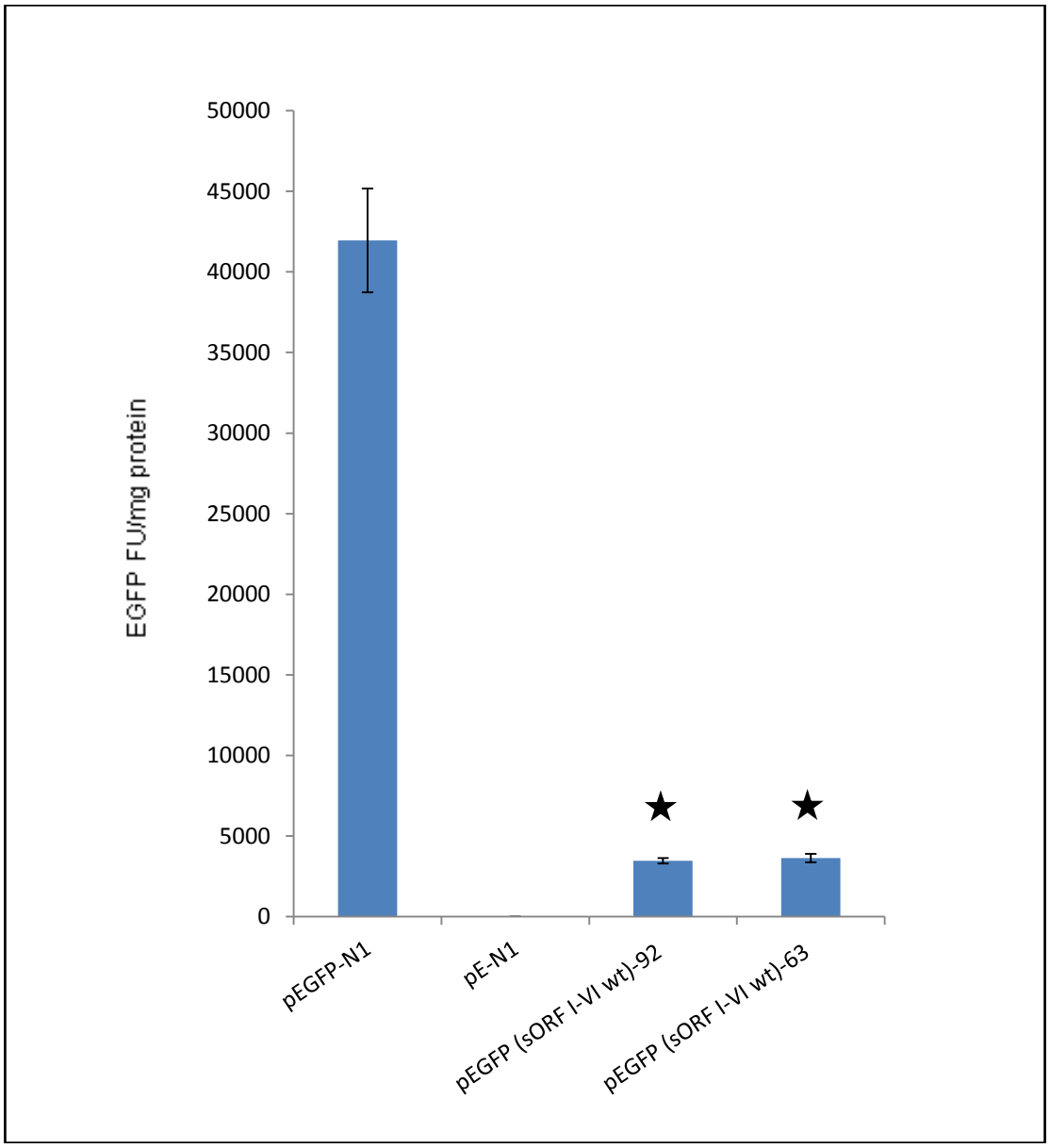


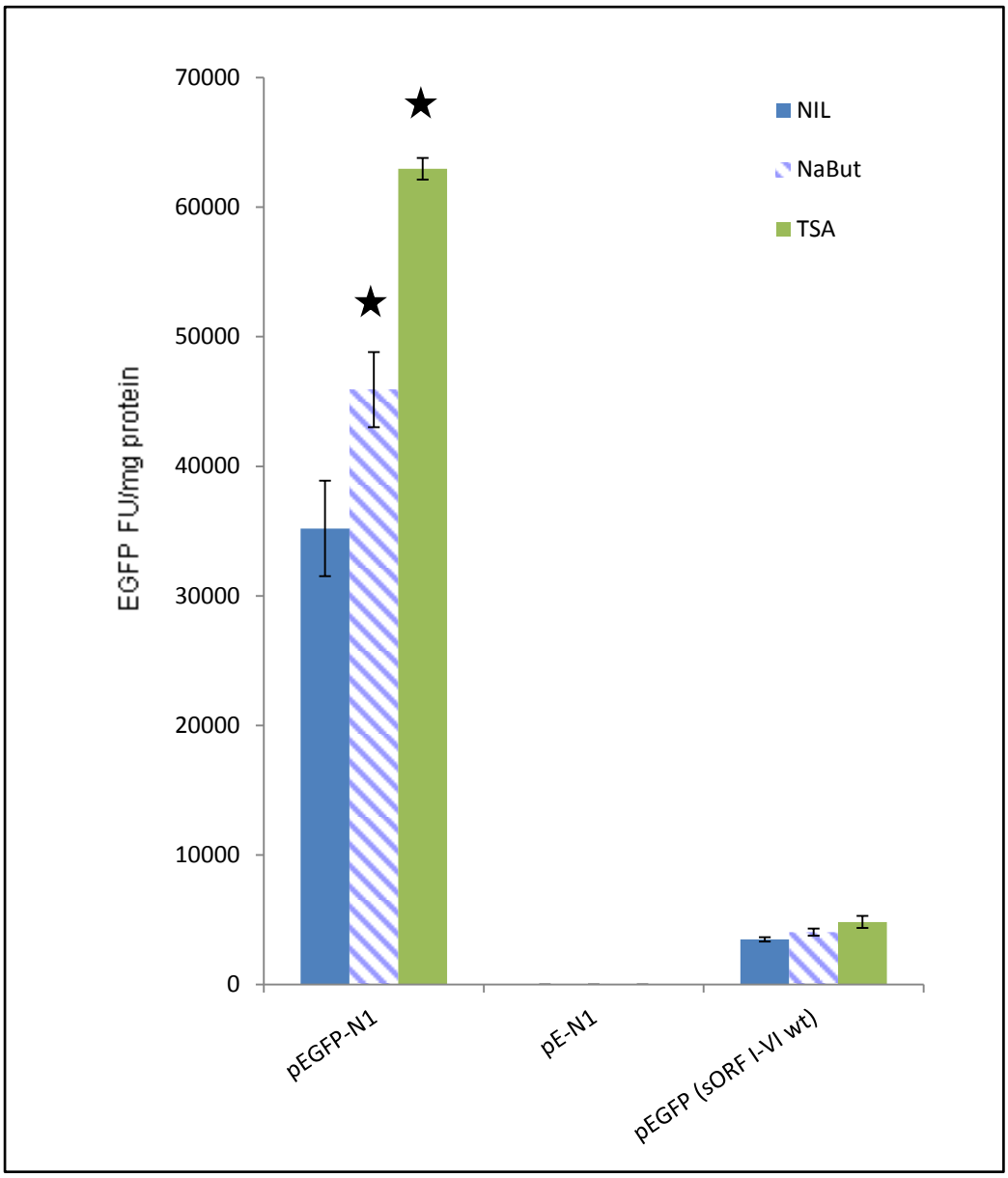
Figure 4.3 Effect of HIV-1 *asp* upstream sORFs I-VI on gene expression with shorter and longer intercistronic distances. EGFP assay of HEK293 cells transfected with constructs containing sORFs I-VI upstream of the reporter EGFP. Constructs contained the sORFs I-VI upstream of the EGFP reporter with an intercistronic distance of either 92bp or 63bp. Plasmids pEGFP-N1 and pE-N1 were used as positive and negative controls, respectively. The presence of the upstream sORF region (I-VI Wt) reduced EGFP expression by 95% in comparison to the parental plasmid, with intercistronic distance having no effect on downstream expression. All data represent the mean ±SE of six individual experiments. Significant differences (p= <0.01) in EGFP expression compared to the pEGFP-N1 control noted with a star.

4.4.2 Investigation of sORF inhibition in the presence of the transcriptional activators TSA and NaBut.

These experiments tested the effect of the transcriptional activators TSA and NaBut, on gene expression. Both the transcriptional activators TSA and NaBut act by inhibiting the function of the histone deacetylase, making DNA more accessible, thus promoting transcription (Dean *et al*., 1987; Struhl, 1998). As the pEGFP-N1 vector includes the SV40 promoter, expression in HEK293 cells is expected to increase protein production.

HEK293 cells transfected with pEGFP-N1 showed a 23% or 44% increase in the levels of EGFP expression when treated with TSA ($p = <0.01$, $n=6$) or NaBut ($p = <0.01$, $n=6$) compared to the untreated cells respectively Figure 4.4. A similar pattern was observed in cells transfected with the construct containing the sORFs I-VI upstream of the EGFP reporter, pEGFP (sORF I-VI Wt) however these differences were statistically not significant ($p=0.014$, $n=6$ and $p=0.201$, $n=6$ for both TSA and NaBut respectively compared to un-activated cells).

Figure 4.4 The effect of cell activators NaBut and TSA on EGFP expression in the presence of HIV-1 *asp* upstream sORFs I-VI. EGFP assay of HEK293 cells transfected with constructs containing sORFs I-VI upstream of the reporter EGFP. Constructs contained the sORFs I-VI upstream of the EGFP reporter. Plasmids pEGFP-N1 and pE-N1 were used as positive and negative controls, respectively. Cells were left untreated (blue), or treated with NaBut (Striped) or TSA (green). Positive control shows an increase in EGFP expression with treatment with NaBut or TSA compared to HEK293 cells left un-treated. The presence of the upstream sORF region (I-VI Wt) reduced EGFP expression by 95% in comparison to the parental plasmid as previously shown, with the cell activators having little effect on downstream expression. All data represent the mean ±SE of six individual experiments. Significant differences (p= <0.01) in EGFP expression compared to the untreated controls noted with a star.


70000
60000
50000
40000
30000
20000
10000
0
EGFP FU/mg protein
NIL
NaBut
TSA
pEGFP-N1
pE-N1
pEGFP (sORF I-VI wt)

4.4.3 Investigation of transcript abundance

To assess any potential *trans* effects that sORF transcripts, or their translation products, may have on gene expression, reporter constructs were co-transfected with an equal mass of a second vector, pDsRed-N1, which contains the RFP gene driven by the CMV promoter. This allowed the generation of red (RFP) and green (EGFP) florescent signals which could be simultaneously assayed in the one sample. Co-transfection combinations, pE-N1:pEGFP-N1 and pE-N1: pEGFP-(sORF I-VI Wt) produced low levels of red fluorescence, which reflects cross over fluorescent signal generated from EGFP as has been previously reported (Ellenberg *et al*., 1999; Miller *et al*., 1999). The results of this experiment (Figure 4.5) showed that all of the combinations co-transfected with pDsRed-N1 expressed high levels of RFP, regardless of the accompanying EGFP construct. There was no significant difference between the co-transfections carried out with pDsRed-N1 combined with pEGFP-N1 ($p= 0.862$, $n=6$) or the pEGFP-(sORF I-VI Wt) construct ($p= 0.161$, $n=6$) compared to the pDsRed-N1:pE-N1 control.

The levels of both EGFP and RFP transcripts were also determined. The results of both the EGFP (Figure 4.6) and RFP (Figure 4.7) transcript abundance assays showed that the respective levels of each transcript were consistent across all samples. There was no significant difference in the levels of EGFP transcript across the co-transfection experiments compared to the pEGFP-N1 control nor in the levels of RFP transcript compared to the pDsRed-N1 control.

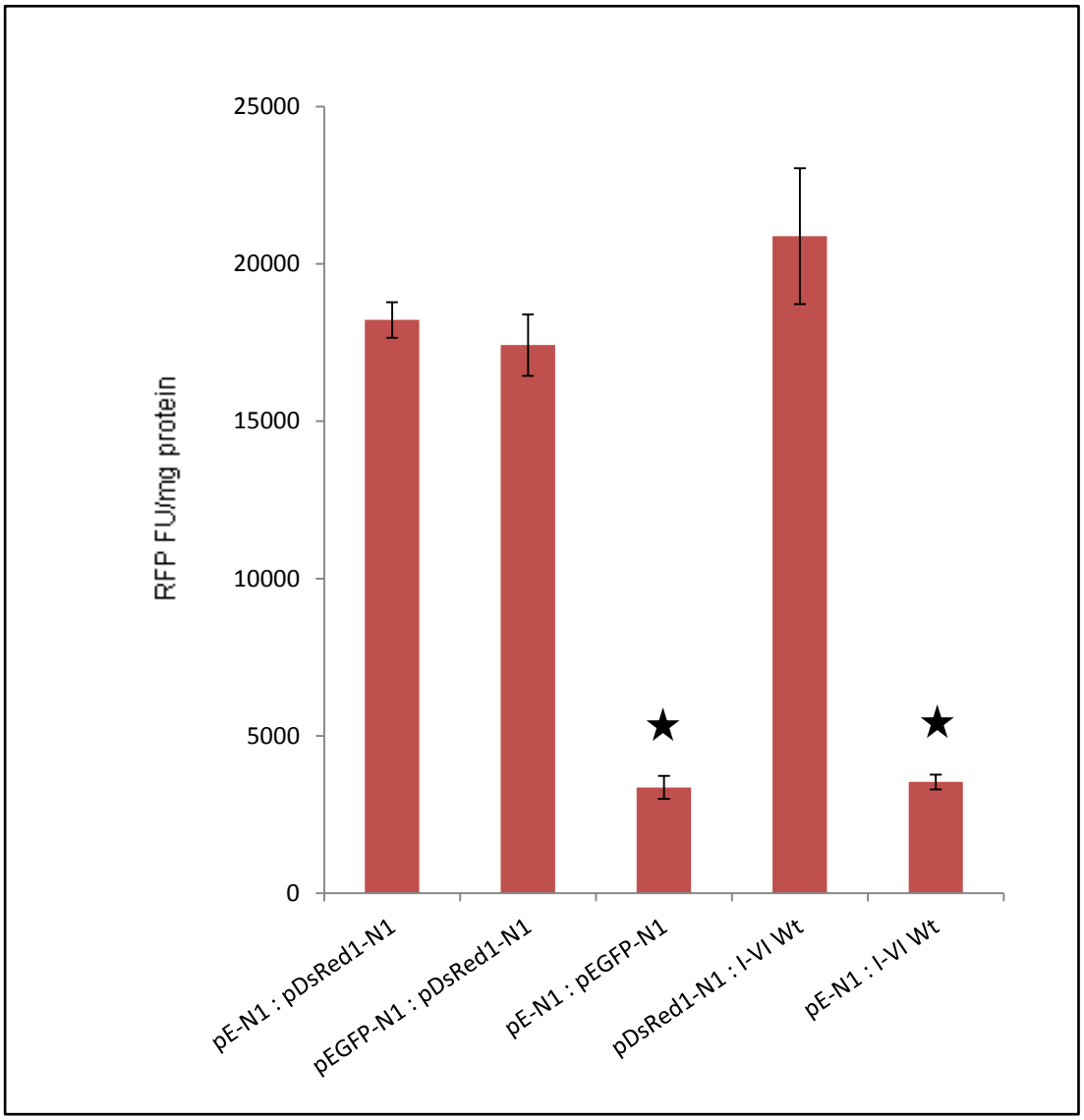


Figure 4.5 Effect of HIV-1 *asp* upstream sORFs I-VI on pDsRed expression in *trans* by co-transfection of reporter constructs with an equal mass of pDsRed-N1. RFP assay of HEK293 cells co-transfected with constructs containing sORFs I-VI upstream of the reporter EGFP and pDsRed-N1. Plasmids pEGFP-N1 and pE-N1 were used as positive and negative controls, respectively. Positive controls transfected with pDsRed-N1 show high levels of RFP expression; no significant difference was observed after co-transfection with construct containing sORFs I-VI. Experiments lacking pDsRed-N1, used as negative controls show a low RFP expression in comparison to the positive controls, due to EGFP and RFP spectra crossover. All data represent the mean ±SE of six individual experiments. Significant differences (p= <0.01) in RFP expression compared to respective controls noted with a star.

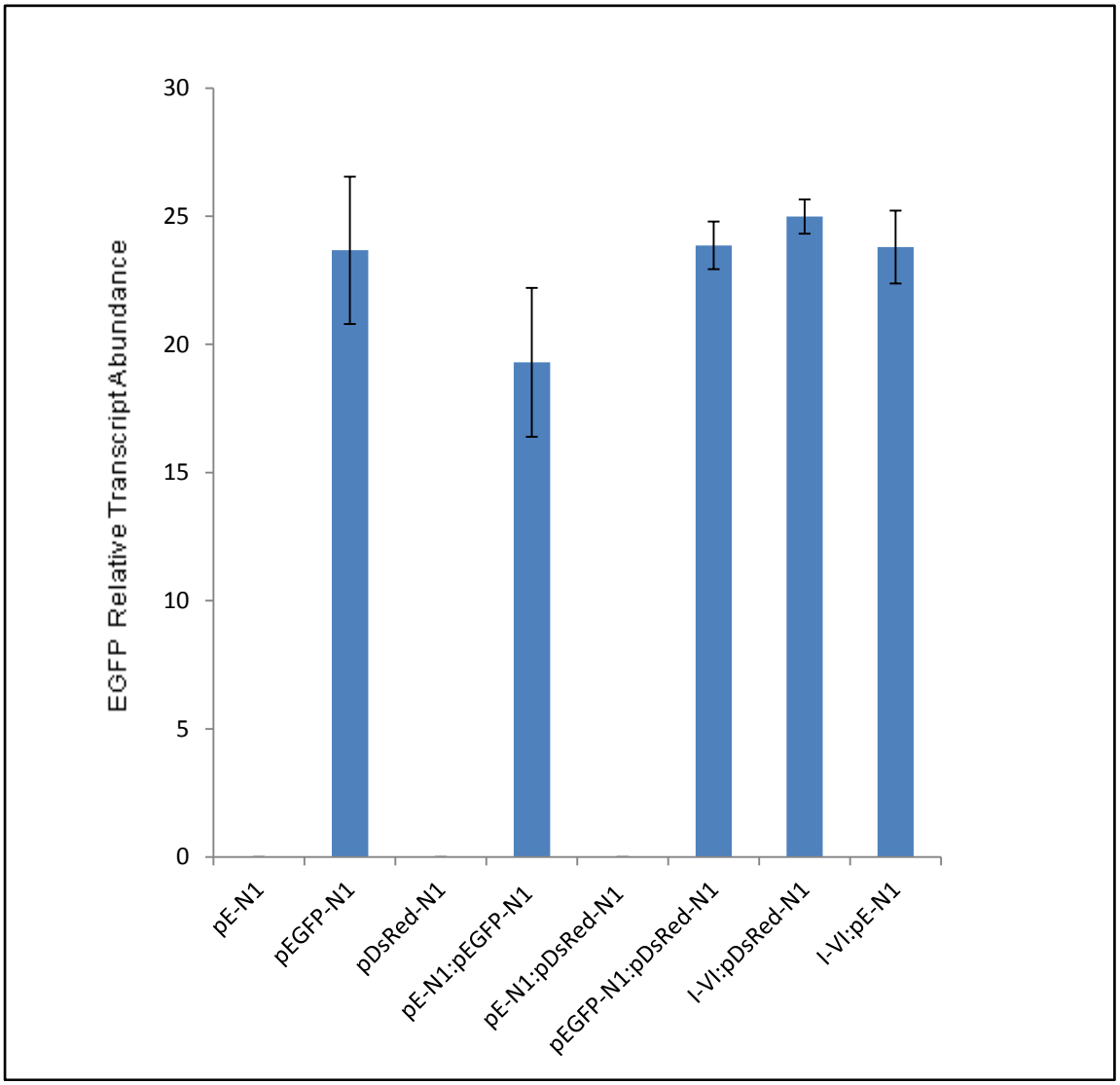


Figure 4.6 Effect of HIV-1 *asp* upstream sORFs I-VI and pDsRed expression on EGFP transcript abundance by co-transfection of reporter constructs with an equal mass of pDsRed-N1. EGFP transcript abundance of HEK293 cells co-transfected with constructs containing sORFs I-VI upstream of the reporter EGFP and pDsRed-N1. Plasmids pEGFP-N1 and pE-N1 were used as positive and negative controls, respectively. No significant difference between-transfected samples were observed. All data represent the mean ±SE of three individual experiments.

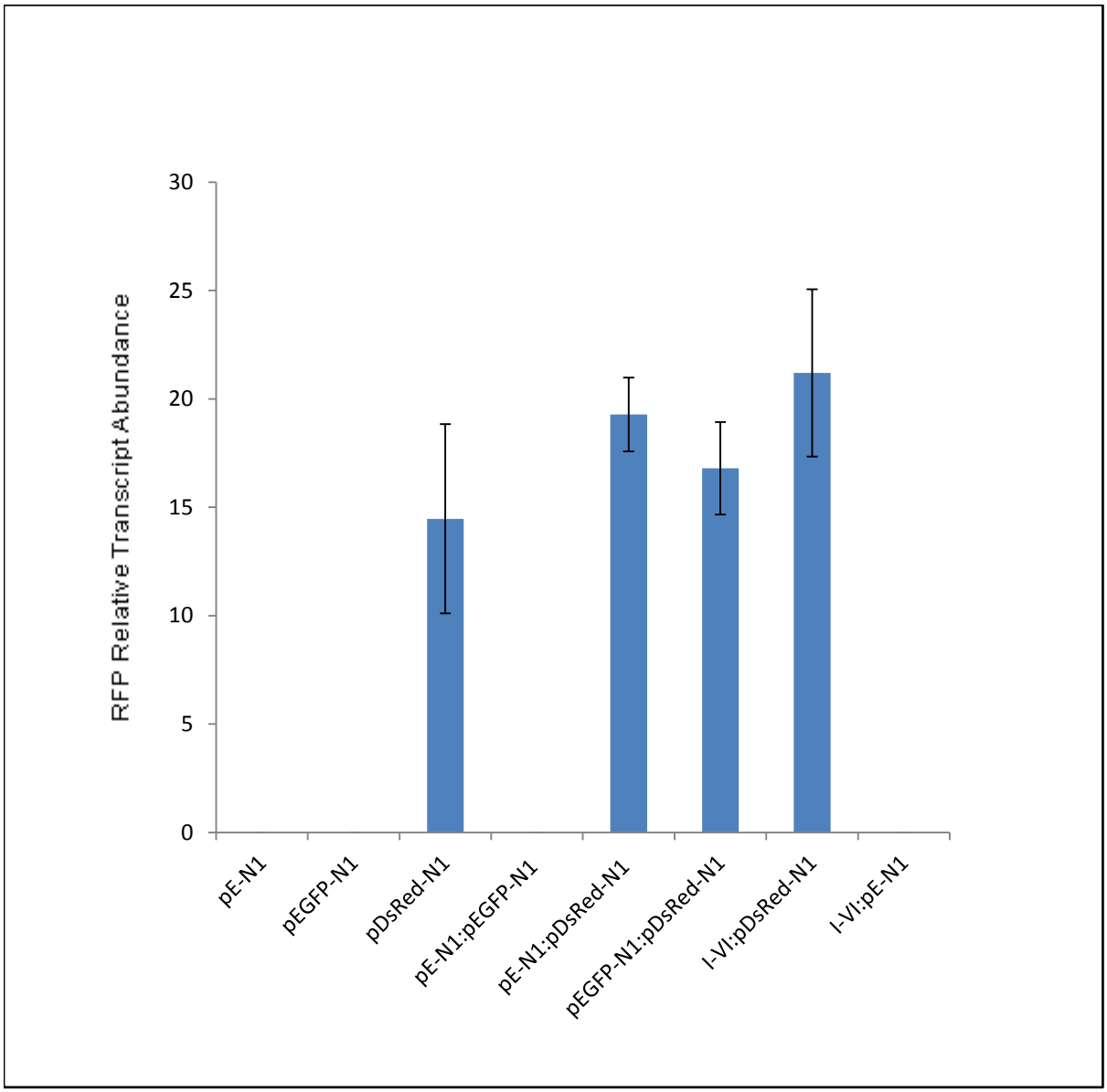


Figure 4.7 Effect of HIV-1 *asp* upstream sORFs I-VI and EGFP expression on RFP transcript abundance by co-transfection of reporter constructs with an equal mass of pDsRed-N1. RFP transcript abundance of HEK293 cells co-transfected with constructs containing sORFs I-VI upstream of the reporter EGFP and pDsRed-N1. Plasmids pEGFP-N1 and pE-N1 were used as positive and negative controls, respectively. No significant difference between co-transfected samples was observed. All data represent the mean ±SE of three individual experiments.

4.4.4 Mutational analysis of sORF initiation codons on downstream expression

The role of each sORF on downstream gene expression was examined by sequential mutation of the initiation codon of each sORF, starting at sORF VI (Figure 4.8). The inactivation of sORF VI produced a 38% increase in the level of EGFP expression in comparison to the I-VI Wt (p= <0.01, n=6). This increase in gene expression was not statistically different after sORF V had also been inactivated (p=0.330, n=6). There was no statistically significant difference when sORFs IV (p=0.494, n=6), III (p=0.336, n=6), II (0=0.043, n=6) and I (p=0.036, n=6) had also been collectively inactivated, in comparison to the sORF VI inactivation alone. The inhibitory effect of sORF VI was restored once all of the sORFs I-V were inactivated leaving the sORF VI initiation codon intact. There was no significant difference between this construct and the I-VI Wt (p=0.263, n=6). The effect of the internal AUG codon (codon 24 of sORF VI), VI_{alt}, on downstream gene expression was also tested by mutation of this AUG in constructs where sORFs I-VI were fully functional, or non-functional. A 2-fold increase in EGFP gene expression when all the sORF initiation codons had been mutated compared to the wild-type was observed. Levels of expression increased when VI_{alt} was mutated in comparison to the I-VI Wt (p= <0.01, n=6). There was little difference between constructs in which VI_{alt} had been mutated in the presence or absence of all sORFs I-VI (p=0.045, n=6).

EGFP transcript abundance was examined within these experiments to determine whether changes in EGFP levels are associated with changes at the level of transcription. This data (Figure 4.9) showed that the levels of EGFP transcript were again consistent across all the constructs used in these experiments. Importantly there was no difference between the positive control, pEGFP-N1 and the remaining constructs.

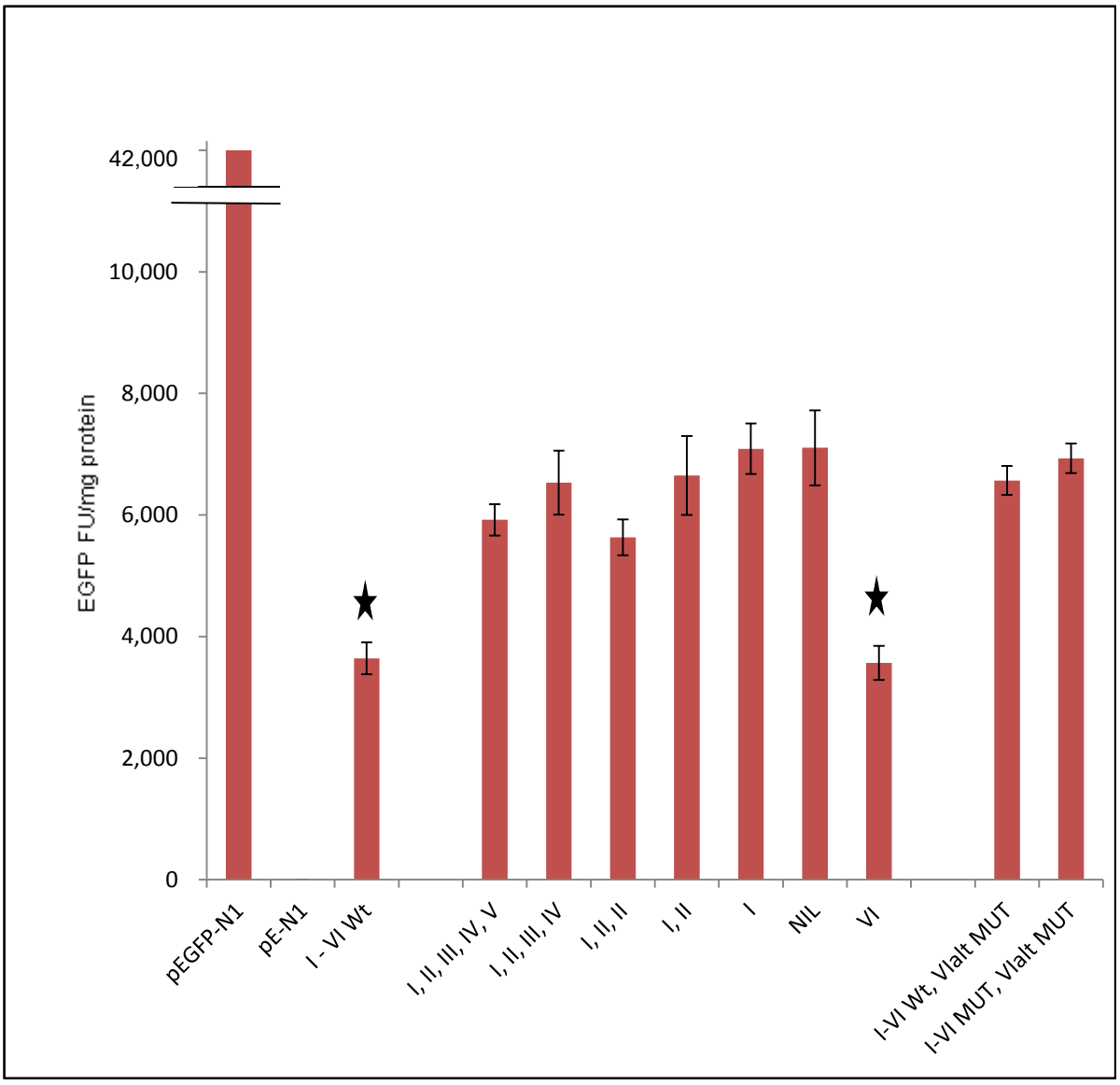


Figure 4.8 Effect of HIV-1 *asp* upstream sORFs I-VI and mutation of sORF initiation codons on the reporter EGFP. EGFP assay of HEK293 cells transfected with constructs containing sORFs I-VI upstream of the reporter EGFP. Constructs contained the sORFs I-VI upstream of the EGFP reporter, with each respective sORF initiation codon mutated. Labels depict functional sORF initiation codons. Plasmids pEGFP-N1 and pE-N1 were used as positive and negative controls, respectively. The presence of the upstream sORF region (I-VI Wt) reduced EGFP expression by 95% in comparison to the parental plasmid, with mutation of sORF VI initiation codon relieving some inhibition. There was no significant difference between the construct having only sORF VI functional and the I-VI Wt (both indicated with a star) on downstream expression. All data represent the mean ±SE of six individual experiments.

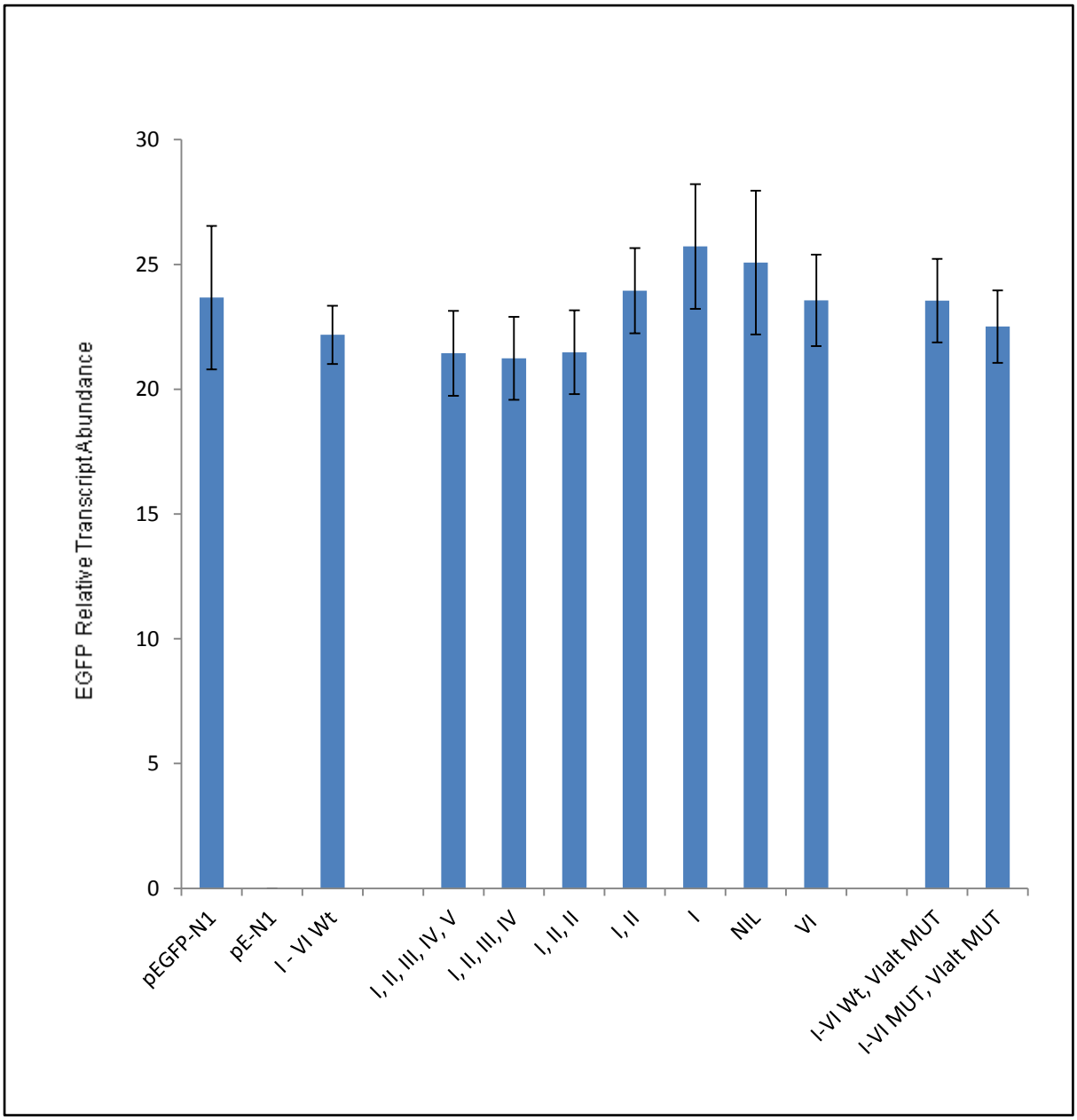


Figure 4.9 Effect of HIV-1 *asp* upstream sORFs I-VI and mutation of sORF initiation codons on EGFP transcript abundance. EGFP transcript abundance of HEK293 cells transfected with constructs containing sORFs I-VI upstream of the reporter EGFP. Constructs contained the sORFs I-VI upstream of the EGFP reporter, with each respective sORF initiation codon mutated. Labels depict functional sORF initiation codons. Plasmids pEGFP-N1 and pE-N1 were used as positive and negative controls, respectively. All samples showed consistent level of transcript, differences were not significant. All data represent the mean ±SE of three individual experiments.

4.5 Discussion

The role of the HIV-1 NL4-3 *asp* sORF region on downstream gene expression has not previously been addressed in the literature. The study of this negative sense transcript and the gene itself have been hampered by issues associated with the difficulty of working in HIV infected cells, and the overpowering expression of the positive sense transcript. This study aimed to investigate the sORF region and its potential to control expression of downstream genes.

Sub-cloning of the HIV-1 NL4-3 *asp* sORF region into the reporter vector, pEGFP-N1, showed that this sORF region inhibits downstream expression. This result was not unexpected given the array of mechanisms by which sORFs may regulate gene expression, typically by slowing/stalling translational processes (reviewed in Morris and Geballe, 2000; Calvo *et al*., 2009). This finding was further evidenced in these experiments, where knocking out each sORF initiation codon relieved some of this inhibition. Similar effects have been observed with sORFs in repression of *Lc* expression, a transcriptional activator in maize. In this example, mutation of the three sORF initiation codons upstream of the *Lc* gene produced a 2-fold increase in the level of gene expression compared to the wild-type (Damiani and Wessler, 1993). These data match this finding with a 2-fold increase in EGFP gene expression when all the sORF initiation codons had been mutated compared to the wild-type.

Previous work conducted by Kozak had reported the effects that intercistronic length plays on the efficiency of reinitiation by the scanning ribosome (Kozak, 1987c). In her studies, Kozak reported that, as the intercistronic distance between the stop codon of the sORF and the next initiation codon was widened to 79 nucleotides, the efficiency of reinitiation at the downstream AUG improved. This was also observed in the translational control of *GCN4,* where reinitiation between four sORFs in the mRNA leader sequence was shown to diminish as the distance between sORFs was reduced (Grant *et al*., 1994). This report also concluded that the shorter intercistronic distances reduced eIF2 activity, a vital initiation factor required for efficient translational initiation, and that this was the rate-limiting step. Similar results have also been reported with the mechanism of *tat* translation in HIV-1 (Luukkonen *et al*., 1995). To rule out any possible effects intercistronic spacing may have on

downstream gene expression the spacing within the constructs was shortened from longer than the critical 79 nucleotides as suggested by Kozak (1987c). No significant difference in downstream gene expression between shorter or longer intercistronic spacing was observed.

The effects of the transcriptional activators TSA and NaBut have been reported in HIV-1 infected cells (reviewed by Struhl, 1998; Quivy and Van Lint, 2002). TSA has been reported to hyperacetylate cellular histones, resulting in the activation of the HIV-1 promoter and leading to an increase in virus production (Van Lint *et al*., 1996). Further work examined the mechanism by which TSA/NaBut, along with expression of p50/p65 and tumour necrosis factor alpha/SF2 (TNF), induce NF-κB to activate the HIV-1 LTR, thus increasing viral transcription (Quivy *et al*., 2002). In line with these observations, the possibility of transcriptional activation affecting gene expression levels was investigated. While the positive control, pEGFP-N1, showed a higher level of EGFP expression in the presence of TSA/NaBut, there was no significant difference between the TSA and NaBut treated compared to the untreated cells in the presence of the sORF I-VI region. These data imply that transcriptional activation does not affect downstream gene expression and suggests that this effect of the upstream sORF region occurs via a translational mechanism.

The alteration of gene expression by *trans* acting factors has also been observed in transfection experiments. In one such example, cells expressing Renilla luciferase have shown a 2 to 8 fold increase in expression when transfected with plasmids containing the GATA-4 or GATA-6 transcription factors (Ho and Strauss, 2004). In one well studied example, single sORF of the *N.crassa arg-2* mRNA encodes the arginine attenuator peptide, AAP (Wang and Sachs, 1997). As the ribosome initiates at the upstream sORF and begins to translate the AAP, the ribosome will stall in direct response to the peptide, preventing translation of *arg-2*, the small subunit of arginine-specific carbamoyl phosphate synthetase (Wang *et al*., 1999; Fang *et al*., 2000; Fang *et al*., 2004). The experimental sets established here were designed to confirm that the sORF region does not act in *trans* to downregulate the level of EGFP gene expression. In these co-transfection experiments, there was no effect on the levels of RFP production when co-transfected with pEGFP-N1 and associated

constructs. The consistent levels of mRNA transcript abundance for both RFP and EGFP also indicate that the effects the sORF region has on EGFP expression are more likely to involve the translational, and not the transcriptional levels of gene expression.

The role of each sORF on reporter gene expression was investigated. These findings suggested that sORF VI plays the greatest role in inhibiting gene expression, with a 60% increase observed when the sORF VI initiation codon alone was mutated. Little effect was observed when the remaining sORFs (I-V) were mutated, suggesting their role in translational inhibition is minimal in comparison to sORF VI. This was also confirmed when all initiation codons except that of sORF VI were mutated, restoring the level of gene expression back to the wild-type. There was also no difference between constructs where the initiation codon VI_{alt} had been mutated.

4.6 Conclusions

In this chapter the repression of the reporter EGFP gene expression by the HIV-1 NL4-3 *asp* sORFs I-VI was confirmed. These experiments showed that the transcriptional activators TSA and NaBut did not significantly affect downstream gene expression, suggesting that the sORF region does not respond to transcriptional activation. This suggests that the downregulation of gene expression is not associated with the level of transcript produced and that control occurs at the level of translation. The possibility for the sORF region to affect global gene expression in *trans* was also examined. Co-transfection data showed constant levels of reporter gene expression and constant levels of mRNA transcript abundance indicating that the sORF series controls downstream genes only, most likely via a translational mechanism. Mutation studies of the sORF initiation codons further suggest that sORF VI plays the greatest role in the inhibition of downstream gene expression.

CHAPTER 5 – CHARACTERISATION OF THE sORF TRANSCRIPT AND THE ROLE OF SPLICING IN GENE EXPRESSION

5.1 General introduction

Transcription of *asp* is controlled by long terminal repeat (LTR) sequences within the 3′LTR and occurs early in infection, at the same time as regulatory protein gene transcription (Peeters *et al*., 1996; Bentley *et al*., 2004). Early studies by Bukrinsky and Etkin (1990) detected three polyadenylated negative sense RNA transcripts (1.6, 1.1 and 1.0kb) in acutely infected H9 cells; these transcripts were present early in infection (day 3) but were not detected later (on days 5 or 7) however sequence data was not published. A study conducted by Michael *et al*. (1994b) detected a negative sense transcript in PMBCs isolated from infected patients; sequence analysis predicted a larger full-length transcript of 2.3kb. A comprehensive study conducted by Landry *et al*. (2007) provided further evidence for negative sense transcription in HIV-1, identifying an alternative poly(A) signal that would potentially produce a 4.1kb transcript, however this transcript was not detected by northern blot analysis. The most recent study by Kobayashi-Ishihara *et al*. (2012) confirmed the presence of a 2.6kb *asp* transcript in acutely and chronically infected cells, transcribed from the U3 region of the 3′LTR.

This chapter tests the hypothesis that splicing may account for the effects of the upstream sORF region on translation observed in Chapter 4. In this chapter the potential role of splicing and/or alternate initiation sites that may explain the variable transcript sizes reported are investigated.

5.2 Aims

The sORF region is further characterised by a detailed analysis of the transcript and its splice products and their potential effects on downstream gene expression in the reporter construct system pEGFP-N1, as follows:

1. Examination of the sORF transcript, as cloned into the reporter gene construct pEGFP-N1, by reverse transcriptase PCR and sequence analysis.
2. Investigation of the conservation of vital splicing acceptor and donor motifs.
3. Examination of the effect of splicing on downstream gene expression.
4. Investigation of the role of splicing in downstream gene expression by mutation of vital splicing acceptor and donor motifs in the reporter gene construct system.
5. Examination of the relative levels of each spliced product by real-time PCR.
6. Re-examination of the role of each sORF by sequential mutation along the full length unspliced transcript.
7. Examination of the role of each sORF within the most abundant spliced product by sequential mutation.

5.3 Materials and methods

5.3.1 Reporter gene constructs

The pEGFP (sORF I-VI Wt) construct (Figure 4.1) was derived from pNL4-3 DNA by subcloning the sORF segment (position 8798 to 7980 comprising the start codon of sORF I to the stop codon of sORF VI respectively) into pGEM®-T Easy vector system (Promega) with the addition of an Age I site as detailed in Chapter 2, section 2.3. The parental plasmids, pEGFP-N1 and pE-N were used as positive control and negative controls in all transfection experiments as detailed in section 2.3.

Site directed mutagenesis was used to knock out the ATG codon of each sORF sequentially, starting at sORF VI. Mutagenesis was carried out using the QuikChange® II Site-Directed Mutagenesis Kit (Stratagene) according to the manufacturer's instructions as detailed in Chapter 4, section 4.3.1 using the primers listed in Chapter 4, Table 4.1, with the additional splicing motif mutations as detailed in Table 5.1 below. Various combinations of mutations were used to create the sets of constructs containing mutations in the splicing donor and acceptor sites (Figure 5.1) and the sORF AUG codons (Figures 5.2 and 5.3). Spliced Variants 1, 2 and 3 were cloned individually into the parental plasmid, pEGFP-N1 using the In-Fusion™ Advantage PCR Cloning Kit (Clontech) according to the manufacturer's instructions using the sets of primers detailed in Table 5.2 below. Individual reactions were established to linearize the parental plasmid pEGFP-N1 for each Spliced Variant. Individual reactions were also established to amplify the unique sequences of Spliced Variants 1, 2 and 3 and then in-fused with the parental plasmid, pEGFP-N1 with overlapping sequences for Variants 1 and 2 in-fused with primers as detailed in Table 5.2. All plasmid constructs were sequenced to check integrity.

Table 5.1 Site directed mutagenesis primers for the knock-out of splicing donor 1 and splicing acceptor 1. The following primers (and their complementary primer) were used to mutate splice site motifs with mutated sequence in bold.

Motif	Sequence (5' - 3')	T_M(°C)
SA 1	GAGCTATTCGCCACATA**T**CTAGAAGAATAAGACAGG	63
SD 1	GAAATAACATGAC**T**TGGATGGAGTGGGACAGG	64

Table 5.2 In-Fusion™ primers used to generate the various Spliced Variant constructs.

Variant	Sequence (5' - 3')		T_M (°C)
1&2 Primer 1	GATCCACCGGTCGCC	amplify parental pEGFP-N1 plasmid	50
1&2 Primer 2	AATTCGAAGCTTGAGCTCG		49
3 Primer 1	GCTCGAGATGTGAGTCCGGTAGC		56
3 Primer 2	CCATGGTGAGCAAGGGCG		55
1 Primer 1	CTCAAGCTTCGAATTCCACCCATCTTATAGCAAAATCC	amplify Spliced Variant	64
1 Primer 2	TTGGAATAACATGACACCTAGAAGAATAAGACAGGCTTGG		64
2 Primer 1	CTCAAGCTTCGAATTCCACCCATCTTATAGCAAAATCC		64
2 Primer 2	TTGGAATAACATGACCTCTTGATTGTAACGAGGATTGTG		63
3 Primer 1	GCTACCGGACTCAGATCTCGAGC		63
3 Primer 2	CGCCCTTGCTCACCATGG		55
1 Primer 1	CTTATTCTTCTAGGTGTCATGTTATTCCAAATCTGTTCCAG	sequence common to Variants 1 & 2	63
1 Primer 2	GGCGACCGGTGGATCGGATCAACAGCTCCTGGGG		73
2 Primer 1	CGTTACAATCAAGAGGTCATGTTATTCCAAATCTGTTCCAG		64
2 Primer 2	GGCGACCGGTGGATCGGATCAACAGCTCCTGGGG		73

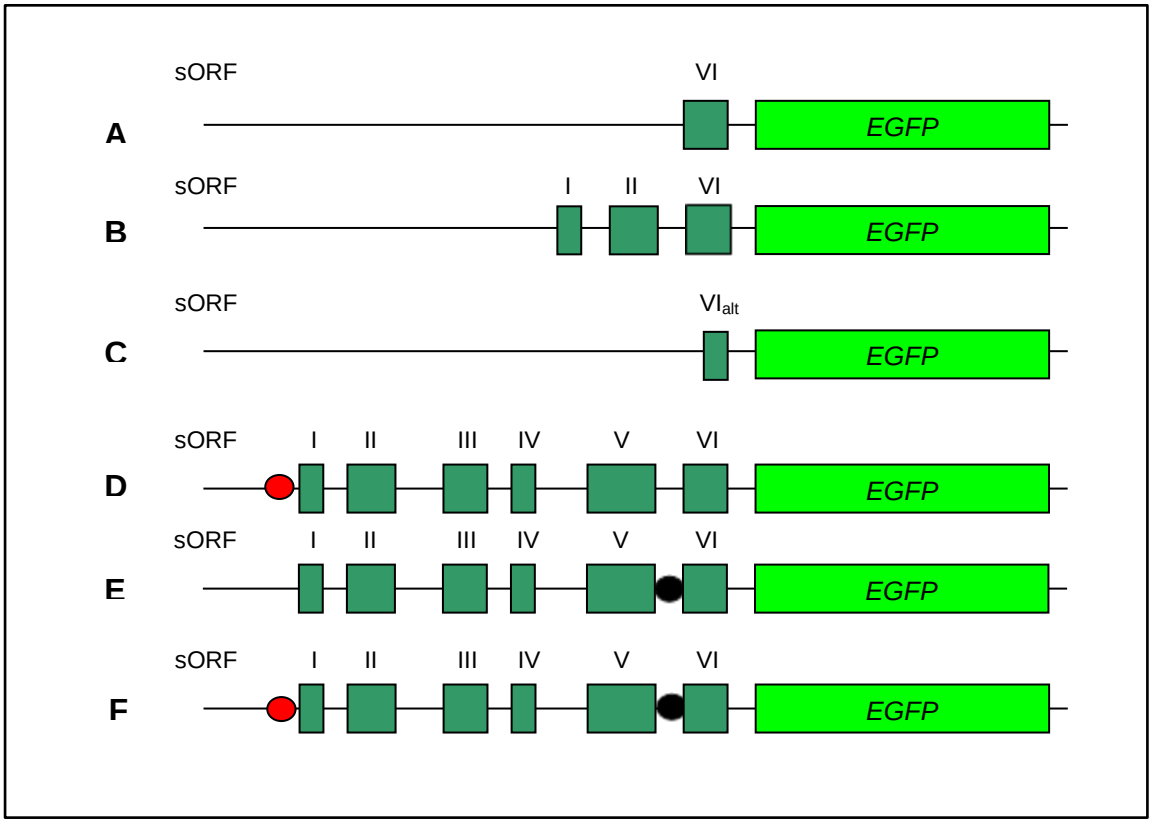


Figure 5.1 Schematic of base plasmids used throughout this study to investigate the effect of each sORF on gene expression. (A) pEGFP (sORF VI)-Variant 1 plasmid with sORF VI sub-cloned in the MCS, (B) pEGFP (sORFs I, II and VI)-Variant 2 plasmid, (C) pEGFP (sORF VI_{alt})-Variant 3 plasmid, (D) pEGFP (sORFs I-VI Wt) plasmid with Splice Donor 1 (SD 1) mutated, (E) pEGFP (sORFs I-VI Wt) plasmid with Splice Acceptor 1 (SA 1) mutated, and (F) pEGFP (sORFs I-VI Wt) plasmid with SA 1 and SD 1 mutated. SD and SA motifs indicated with a red or black circle, respectively.

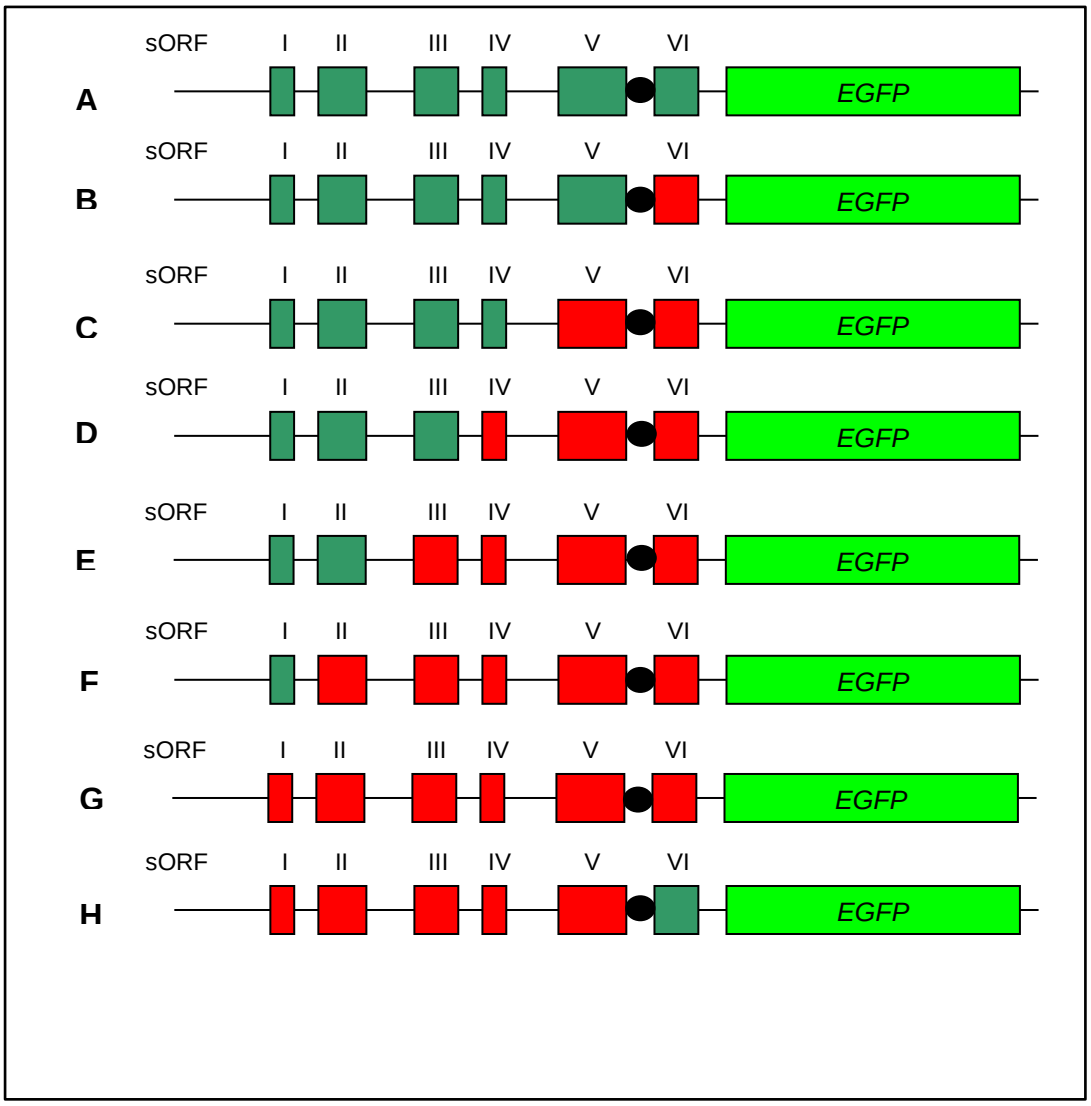


Figure 5.2 Schematic of base plasmids used throughout this study to investigate the effect of each sORF on gene expression. Red shading represents sORFs with mutated ATG initiation codons, green represent non-mutated initiation codons. (A) pEGFP (sORF I-VI Wt) plasmid with sORFs I – VI sub-cloned in the MCS, (B) pEGFP (sORF I-V Wt) plasmid with sORF VI inactivated, (C) pEGFP (sORF I-IV Wt) plasmid with sORFs V-VI inactivated, (D) pEGFP (sORF I-III Wt) plasmid with sORFs IV – VI inactivated, (E) pEGFP (sORF I-II Wt) plasmid with sORFs III – VI inactivated, (F) pEGFP (sORF I Wt) plasmid with sORFs II – VI inactivated, (G) pEGFP (sORF NIL) plasmid with sORFs I – VI inactivated and (H) pEGFP (sORF VI Wt) plasmid with sORFs I – V inactivated. All constructs contained a mutation at SA 1, indicated by the black circle to inhibit splicing activity.

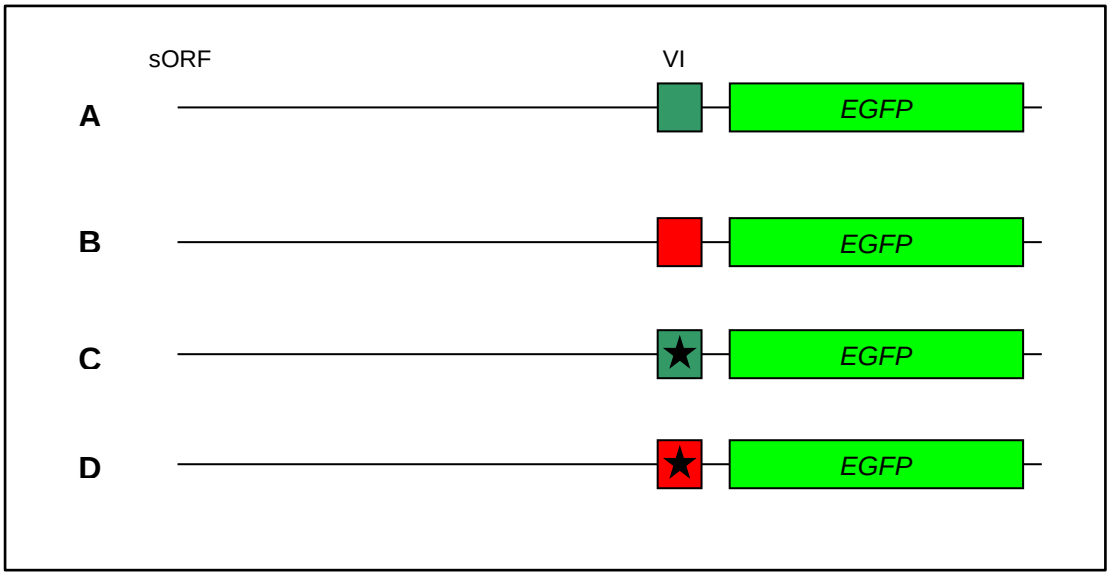


Figure 5.3 Schematic of base plasmids used throughout this study to investigate the effect of each sORF on gene expression within Variant 3. Red shading represent sORFs with mutated ATG codons, green represent non-mutated. (A) pEGFP (sORF VI Wt) plasmid, (B) pEGFP (sORF VI) plasmid with sORF VI inactivated, (C) pEGFP (sORF VI Wt) plasmid with only sORF VI_{alt} inactivated and (D) pEGFP (sORF VI) plasmid with sORFs VI and VI_{alt} inactivated. Mutations of sORF VI_{alt} indicated with a black star.

5.3.2 Transient transfections and reporter gene assays

HEK293 cells were transfected with 1µg DNA using the calcium phosphate ProFection® transfection system (Promega) according to the manufacturer's instructions. In all transfection experiments, pEGFP-N1 and pE-N1 were used as positive and negative controls, respectively as detailed in Chapter 2, section 2.6. Transfected cells were harvested 48 hours post transfection by lysis with 1x Reporter lysis buffer (Promega) and EGFP fluorescence measured with the FLUOStar OPTIMA microplate reader (BMG Labtech) and normalized as detailed in Chapter 2, sections 2.7. Each result represents the calculated mean ± SE of six transfected samples.

5.3.3 Transcript analysis

Transfected HEK293 cells were harvested for total RNA at 48 hours post transfection as detailed in Chapter 2, section 2.9. Samples of the cDNA were then PCR amplified as detailed in section 2.9.4 using the primers detailed in Chapter 2, Table 2.1, except for Spliced Variants 1, 2 and 3 where Reverse primer was replaced with 5′-CGCCCTTGCTCACCATGG-3′ (T_M 55). Amplified products were cloned in pGEM®-T Easy (Promega) prior to sequencing as detailed in Chapter 2, sections 2.9.5, 2.9.6 and 2.5.4.

Northern analysis was conducted using 1µg total RNA, blotted onto nylon membrane and probed with DIG-labelled EGFP and GAPDH cDNA, then detected by luminescence assay as detailed in Chapter 2, section 2.10. EGFP message abundance is presented as transcript relative to GAPDH message and represents the calculated mean ± SE of three samples.

5.3.4 Quantitative real-time PCR analysis

TaqMan® RT-PCR assays (Applied Biosystems) were carried out to determine the relative amounts of each spliced variant within the single mRNA pool. TaqMan®

probes were designed to span exon junctions thus targeting each species as detailed in Table 5.3 below.

Table 5.3 TaqMan® probes used for individual Variant TaqMan® quantitative real-time PCR assays.

Variant	Sequence (5' - 3')	T_M(°C)
Unspliced FOR	TCTCTCCACCTTCTTCTTCTATTCCTT	59
Unspliced REV	CAGACCCACCTCCCAATCC	59
UnSpliced PROBE	CTGTCGGGTCCCCTC	69
Variant 1 FOR	GCGTCCCAGAAGTTCCACAAT	60
Variant 1 REV	GCCTTGGAATGCTAGTTGGAGTAAT	60
Variant 1 PROBE	CAAGAGTCATGTTATTCC	69
Variant 2 FOR	CCCTGTCTTATTCTTCTAGGTCATGTT	58
Variant 2 REV	TGCCTTGGAATGCTAGTTGGA	59
Variant 2 PROBE	TCTGTTCCAGAGATTTATTA	69
Variant 3 FOR	AAATCCTTTCCAAGCCCTGTCT	59
Variant 3 REV	CCACTGCTGTGCCTTGGAAT	60
Variant 3 PROBE	TCTAGAGATTTATTACTCCAACTAG	69

Individual reactions were established according to the TaqMan® Fast Universal PCR Master Mix, without AmpErase UNG (Applied Biosystems), along with the appropriate primer/probe set and cDNA template (as described previously) in a total volume of 20µL. Isolated Spliced Variants, cloned into pGEM®-T Easy (as described previously) were used as standards and to check primer/probe specificity along with a –RT sample to ensure no DNA contamination. The concentrations of starting material were measured using the Quant-iT™ dsDNA HS assay with the Qubit™ fluorometer (Life Technologies) and copy number calculated. The thermal cycling protocol consisted of an initial melt at 95ºC for 20 sec followed by 40 cycles of 95ºC for 3 sec and 60ºC for 30 sec using the ABI 7500 Fast Real Time PCR System (Applied Biosystems). The Ct values were calculated using the software SDS2.4 in order to generate calibration curves and calculate copy numbers of target transcripts.

5.4 Results

5.4.1 Alternative splicing of the sORF region and conservation of splice sites

In Chapter 4, EGFP expression in HEK293 cells transfected with the pEGFP-(sORF I-VI Wt) construct was compared to cells transfected with the base plasmid, pEGFP-N1, and pE-N1 (lacking the EGFP gene). The presence of the upstream sORF region reduced EGFP expression by at least 95% ($p = <0.01$, $n=6$) and northern blot experiments indicated no significant difference in abundance of the reporter EGFP transcript (normalised to GAPDH transcript) in pEGFP-N1 and pEGFP-(sORFs I-IV Wt) transfected cells ($p = 0.345$, $n = 3$). Experiments in which the initiator ATG codon of each sORF was mutated, producing a construct without active sORFs (construct NIL) resulted in a 45% increase in expression compared to the wild type sORF construct ($p<0.01$, $n=6$) In this chapter we examine the potential for splicing within the upstream sORF region to regulate downstream expression.

The pEGFP-(sORFs I-IV Wt) construct was transfected into HEK293 cells, as were the positive and negative control plasmids, pEGFP-N1 and pE-N1 respectively. Primers situated immediately up and downstream of the sORF region (see Figure 5.7 B) were used to detect the sORF region within the transcript pool. The lack of non-specific amplification was confirmed by the absence of product from cells transfected with the pEGFP-N1 and pE-N1 plasmids. An amplification product corresponding to the full-length construct transcript (890bp) was amplified from the cDNA pool (Figure 5.4, fourth lane), however, it was not the most abundant product. At least two shorter amplimers were also detected in the same cDNA pool: an abundant product (240bp) and a less abundant product (480bp). All amplified products were recovered, cloned and sequenced.

Alignment of the sequences of the two shorter amplimers with the full-length transcript revealed that splicing had occurred within the transcript (Figure 5.6). The two spliced sequences used the same splice acceptor SA 1 (immediately before sORF I, NL4-3 nt. position 8095), but alternative splice donor sites, SD 1 and SD 2

(NL4-3 nt. positions 8745 and 8539, respectively) as depicted in Figure 5.7 C. The 240bp splice product (Spliced Variant 1) resulted from the removal of sORFs I to V inclusive (using SD 1 and SA 1) while the 480bp splice product (Spliced Variant 2) resulted from the removal of sORFs III to V inclusive (using SD 2 and SA 1).

Spliced Variants 1 and 2 were cloned separately into the plasmid pEGFP-N1 and these constructs were transfected into HEK293 cells. Transcript RT-PCR analysis (Figure 5.5 A) showed that the predominant species produced in Spliced Variant 1 transfected cells was the expected 240bp spliced product, containing sORF VI alone (Figure 5.5 A, lane 2). However, analysis of Spliced Variant 2 transfected cells (Figure 5.5 A, lane 5) revealed two amplified products; the expected 480bp Spliced Variant 2, containing sORFs I, II and VI, and a smaller, less abundant product (214bp). All amplified products were again recovered, cloned and sequenced as described previously.

Alignment of the sequences of the three products (Figure 5.6) revealed that splicing within the transcript of Spliced Variant 2 involving SD 1 and an alternative splice acceptor, SA 2 (NL4-3 nt. position 8069; 26bp downstream of SA 1), had produced a third product (Spliced Variant 3, Figure 5.5 B) of similar size to Spliced Variant 1. This splicing event results in the complete removal of sORFs I and II and the ATG initiation codon of sORF VI, presenting an alternative, in frame sORF ATG codon (weak Kozak sequence context) 69bp further downstream (NL4-3 nt. Position 8022); this alternative sORF is designated sORF VI_{alt}. Spliced Variant 3 was not detected in any of the six colonies screened in the initial RT-PCR and cloning experiments involving Spliced Variant 1, suggesting that Spliced Variant 3 is less abundant than Spliced Variant 1. The stronger polypyrimidine tract sequence of SA 1 (9/11 Py) may also explain the predominant utilisation of this splice site. The polypyrimidine tract of SA 2 is weaker (Py 7/11) which could affect splicing activity (Figure 5.7 C).

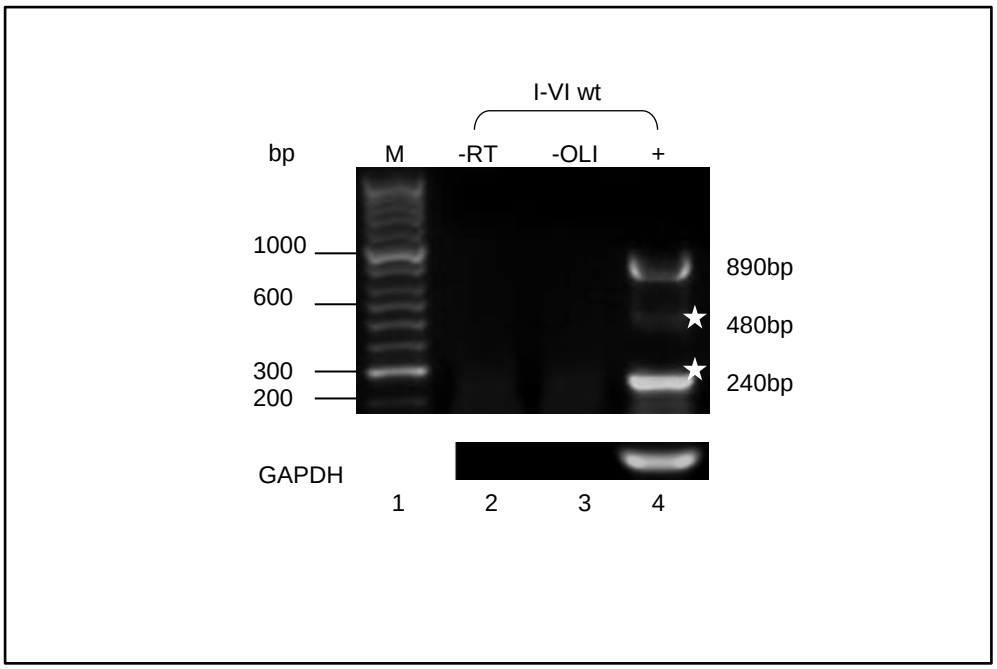


Figure 5.4 RT-PCR analysis HIV-1 *asp* upstream sORFs I-VI. I-VI Wt (spliced products indicated with a star). The samples were assessed for DNA contamination (-RT) and RNA priming (-OLI); no contamination was detected. Lane marked + depicts the cDNA sample. The cDNA sample shows the amplification products 240bp and 480bp amplimer (Spliced Variants), indicated with a star, and full length unspliced product at 890bp. GAPDH controls shown below, indicating integrity of the transcript pool.

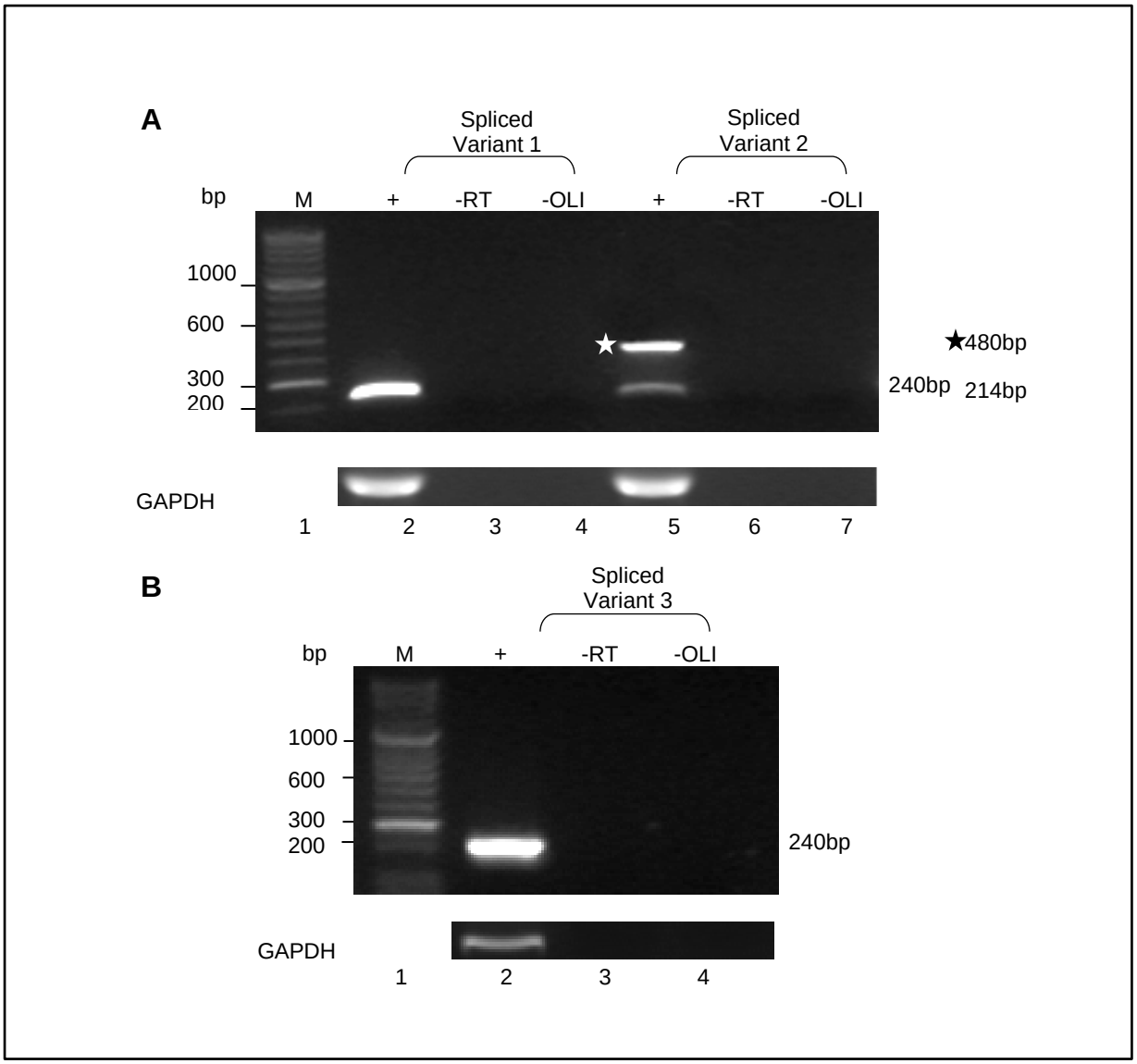


Figure 5.5 RT-PCR analysis HIV-1 *asp* upstream sORFs I-VI Spliced Variants. (A) Analysis of constructs containing Spliced Variant 1 (sORF VI alone), Spliced Variant 2 (sORFs I, II and VI, parent transcript noted with a star), (B) Spliced Variant 3 (sORF VI_{alt}). Samples were assessed for DNA contamination (-RT) and RNA priming (-OLI); no contamination was detected. Lanes marked + depict the cDNA samples. The cDNA samples show the amplification products from a sole product (240bp amplimer) in Spliced Variant 1. Spliced Variant 2 depicts two products: the major Spliced Variant 2 (480bp amplimer), indicated with a star, and a sub-spliced product, Spliced Variant 3 (240bp amplimer). GAPDH controls shown below, indicating integrity of the transcript pool.

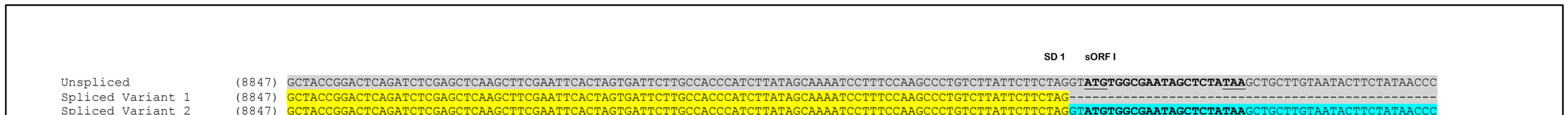
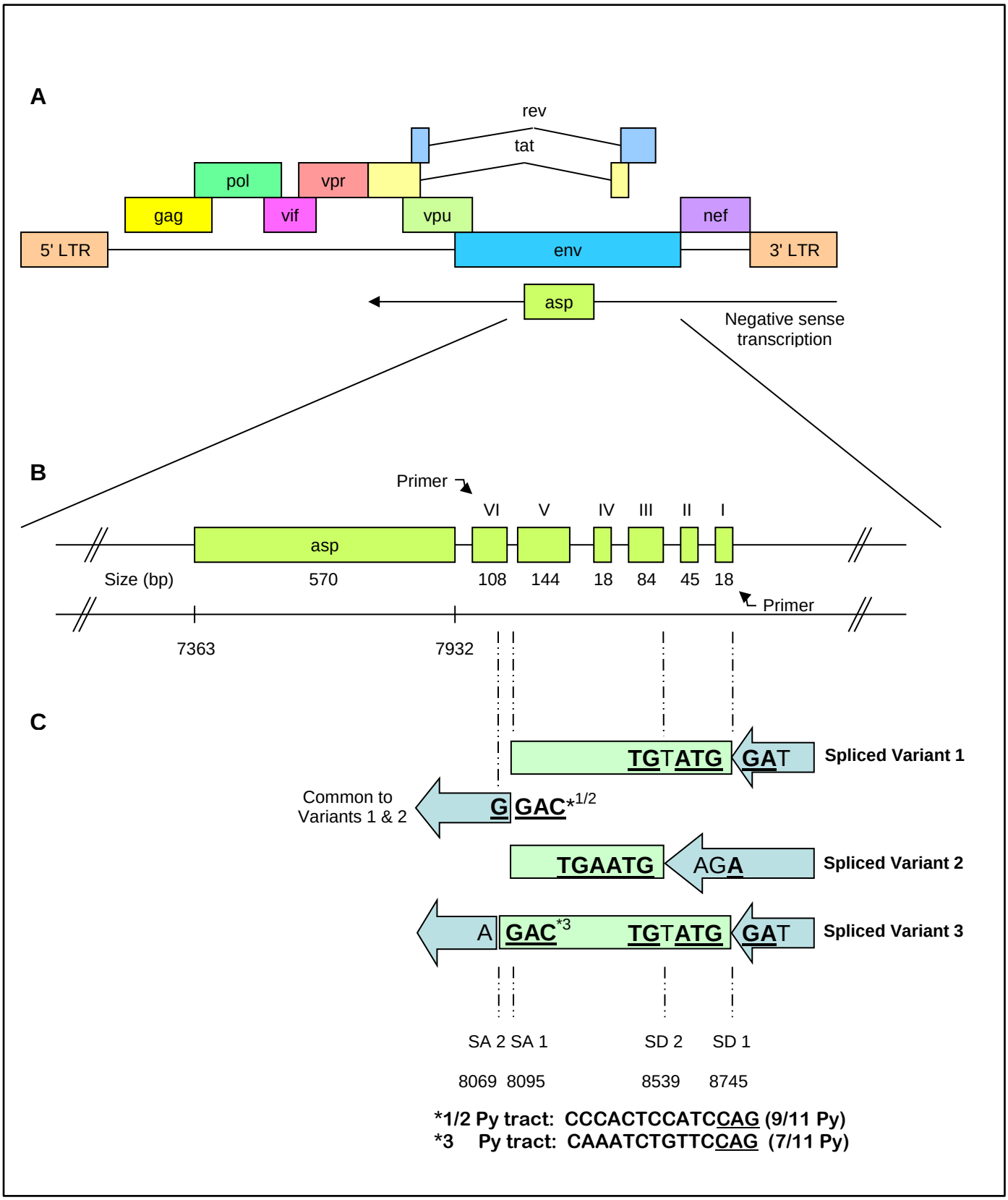


Figure 5.6 Multiple sequence alignment of Unspliced HIV-1 *asp* upstream sORF region and Spliced Variants 1-3. The sORF sequences are shown in bold with the sORF ATG initation codons (and respective sORF termination codons) underlined. The splice acceptor sites 1 and 2 (SA 1 and 2) and respective splice donor sites (SD 1 and 2) shown. Sequences highlighted in yellow include sequence common to all Spliced Variants. Sequences highlighted in blue represent sequences common only to Spliced Variants 1 or 2. The polypyrimidine tracts for both SA 1 and 2 are italicised and underlined. Nucleotide position numbers of HIV-1 NL4-3 shown to the left of the sequence.

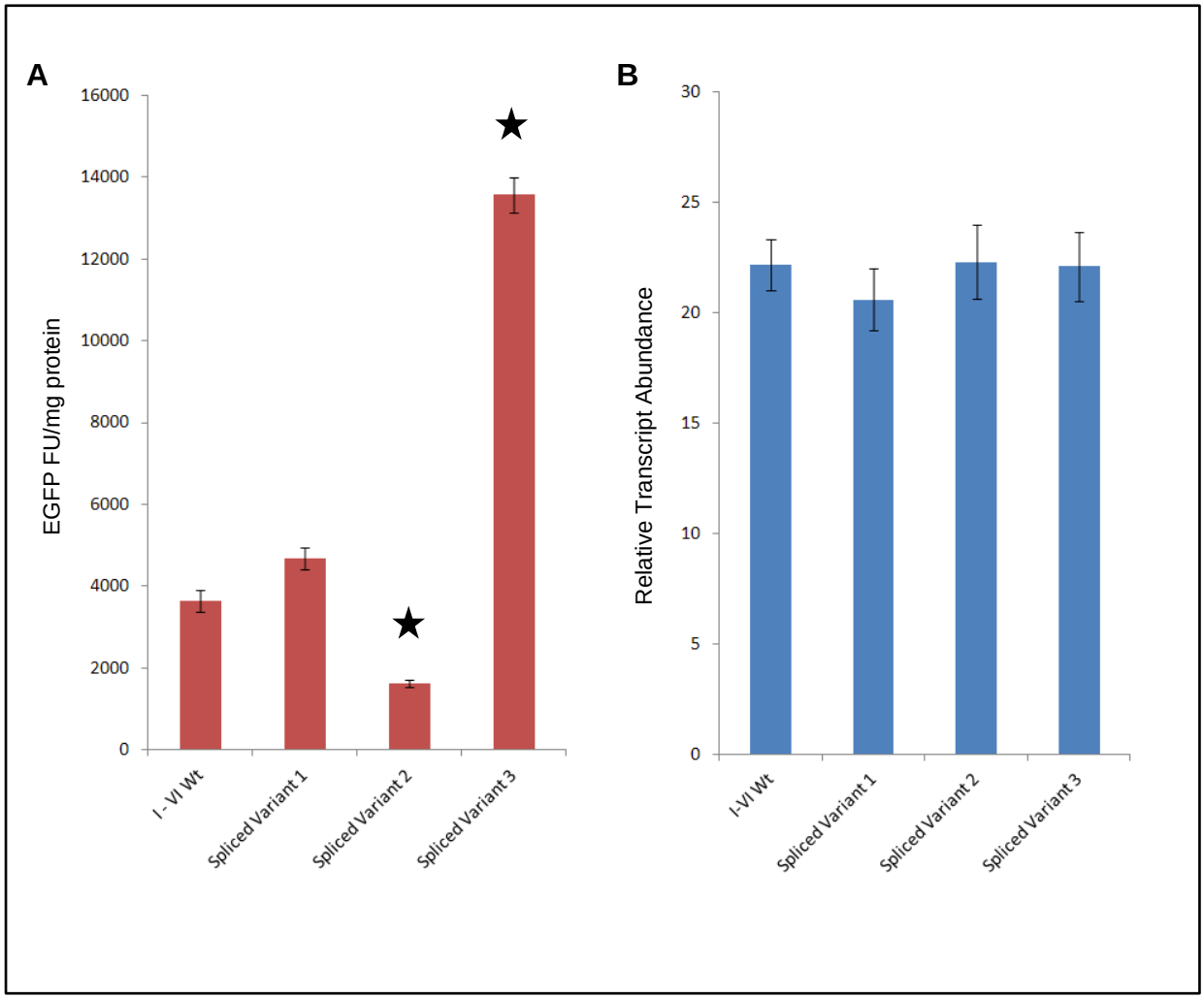
Figure 5.7 Summary of splicing events within the sORF region. (A) The HIV-1 genome with the negative sense gene, *asp* and (B) its associated sORFs, indicated in the negative sense orientation with nucleotide position numbers indicated below from NL4-3. Locations of primers used to amplify the sORF region are indicated. (C) Four alternative transcripts were detected for the sORF region; unspliced, Spliced Variant 1 (containing sORF VI only), Spliced Variant 2 (containing sORFs I, II and VI), and Spliced Variant 3 (containing part of sORF VI). Splice donor and acceptor motifs shown with 100% matches to consensus underlined. Nucleotide position numbers of SD and SA sites noted below. The polypirimidine sequences of SA 1, utilised in Spliced Variants 1 and 2 (*1/2) and SA 2, utilised in Spliced Variant 3 (*3) also shown below.


A
rev
tat
pol
vpr
gag
vif
vpu
nef
5' LTR
env
3' LTR
asp
Negative sense transcription
B
Primer
VI
V
IV
III
II
I
asp
Size (bp)
570
108
144
18
84
45
18
Primer
7363
7932
C
TGTATG
GAT
Spliced Variant 1
Common to Variants 1 & 2
GGAC*1/2
TGAATG
AGA
Spliced Variant 2
A
GAC*3
TGTATG
GAT
Spliced Variant 3
SA 2
SA 1
SD 2
SD 1
8069
8095
8539
8745
*1/2 Py tract: CCCACTCCATCCAG (9/11 Py)
*3 Py tract: CAAATCTGTTCCAG (7/11 Py)

Spliced Variants 1 and 2 were cloned separately into the plasmid pEGFP-N1 and these constructs were transfected into HEK293 cells. EGFP expression, transcript abundance and transcript integrity were analysed as described earlier. Spliced Variant 1, which contains only sORF VI, produced a slight increase (22%) in EGFP expression compared to the wild type ($p = 0.021$, $n=6$) (Figure 5.8 A); but transcript abundance did not change (Figure 5.8 B). In contrast Spliced Variant 2, which contains sORFs I, II and VI, produced a 55% decrease in EGFP expression compared to the wild type ($p= <0.01$, $n=6$), with no change in transcript abundance.

Analysis of expression showed that Spliced Variant 3, which contains only the sORF VIalt ATG initiation codon, produced a 4-fold increase in EGFP expression compared to the wild type ($p= <0.01$, $n=6$); with no change in transcript abundance (Figure 5.8 A and B, respectively). Levels of downstream gene expression observed with Spliced Variant 3 were approximately 3-fold lower than the positive control pEGFP-N1 (pEGFP-N1 gene expression approximately 42,000 FU/mg protein, Figure 4.3). Transcript RT-PCR analysis (Figure 5.5 B, lane 2) confirmed that the predominant species produced in Variant 3 transfected cells was of the size expected for the product containing sORF VIalt.

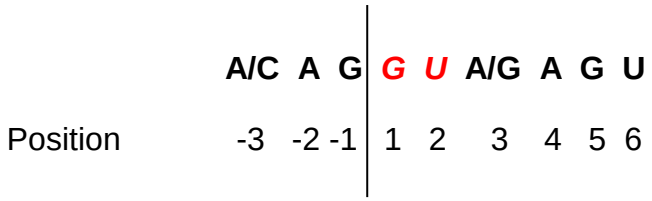
Figure 5.8 Effect of HIV-1 *asp* upstream sORFs I-VI Spliced Variants on the reporter EGFP gene expression, transcript abundance and analysis of Spliced Variants by RT-PCR. Spliced Variant 1 increased EGFP expression by 22% ($p = 0.021$) in comparison to the wild type, while Spliced Variant 2 decreased EGFP expression by 55% ($p<0.01$) and Spliced Variant 3 produced a 4-fold increase in EGFP expression ($p<0.01$). Transcript abundance was consistent for all constructs. RT-PCR analysis of Spliced Variants 1 and 2 reveals full transcript within Spliced Variant 1 and suggests that Spliced Variant 2 is further processed to produce Spliced Variant 3. (A) EGFP assay of HEK293 cells transfected with constructs containing sORFs I-VI upstream of the reporter EGFP and constructs containing Spliced Variant 1 (sORF VI), Spliced Variant 2 (sORFs I, II and VI) and Spliced Variant 3 (sORF VI_{alt}). All data represent the mean ±SE of six individual experiments. Significant differences ($p= <0.01$) in EGFP expression compared to the I-VI Wt noted with a star. (B) Transcript abundance; consistent across all constructs. All data represent the mean ±SE of three individual experiments.


A
EGFP FU/mg protein
16000
14000
12000
10000
8000
6000
4000
2000
0
I - VI Wt
Spliced Variant 1
Spliced Variant 2
Spliced Variant 3
B
Relative Transcript Abundance
30
25
20
15
10
5
0
I-VI Wt
Spliced Variant 1
Spliced Variant 2
Spliced Variant 3

Analysis of the HIV-1_{NL4-3} sequence revealed typical splice donor (at nt 8745 and 8539) and splice acceptor (at nt 8095 and 8069) sequences. SD 1 (NL4-3 nt. position 8745) was 100% conserved across 27 HIV-1 subtype B sequences examined (see Appendix 1), while SD 2 (NL4-3 nt. position 8539) of Variant 2 was slightly less well conserved (86 and 93% conservation of the two critical nucleotides, respectively) (Figure 5.9 A). SA 1 was 100% conserved across all 27 HIV-1 B clade sequences examined (Figure 5.9 B). Analysis of the SA 2 site showed that conservation of the vital A and G nucleotides at positions 13 and 14 was poor (30 and 65% respectively). There is potential for production of a fourth spliced product, from the moderately conserved SD 2 and the poorly conserved SA 2, but this product was never detected in these experiments.

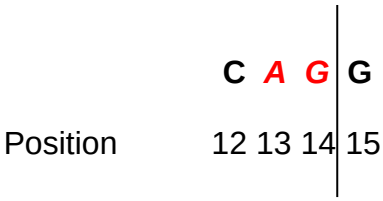
Figure 5.9 Conservation of splice donor and acceptor motifs in subtype B HIV sequences. Essential sequence components of Splice Donor 1 and Splice Acceptor 1 are 100% conserved across 27 HIV-1 subtype B sequences. (A) The splice donor motif with position numbers (with respect to the splice site) indicated below. Splice site indicated by the vertical line and essential components of the splice site indicated in italics. The conservation of each residue for both SD 1 and 2 are expressed as a % of a total of 27 HIV-1 subtype B sequences randomly selected and examined for motif conservation. The SD 1 motif is more highly conserved than SD 2 in the sequences examined. (B) The splice acceptor motif with position numbers (NL4-3) indicated below. Splice site indicated by the vertical line and essential components of the splice site indicated in italics. The critical residues of the splice acceptor (highlighted) were both 100% conserved across the 27 HIV-1 subtype B sequences examined for SA 1, utilised by Variants 1 and 2. SA 2, utilised by Variant 3, exhibits a lower degree of conservation.

A



Position	Frequency	Conservation SD 1	Conservation SD 2
-3	70	10	100
-2	60	100	0
-1	80	100	0
1	100	100	86
2	100	100	93
3	95	100	100
4	70	0	100
5	80	97	100
6	45	100	100

B



Position	Frequency	Conservation SA 1	Conservation SA 2
12	80	93	56
13	100	100	30
14	100	100	65
15	60	100	7

5.4.2 Analysis of Spliced Variants by real-time PCR

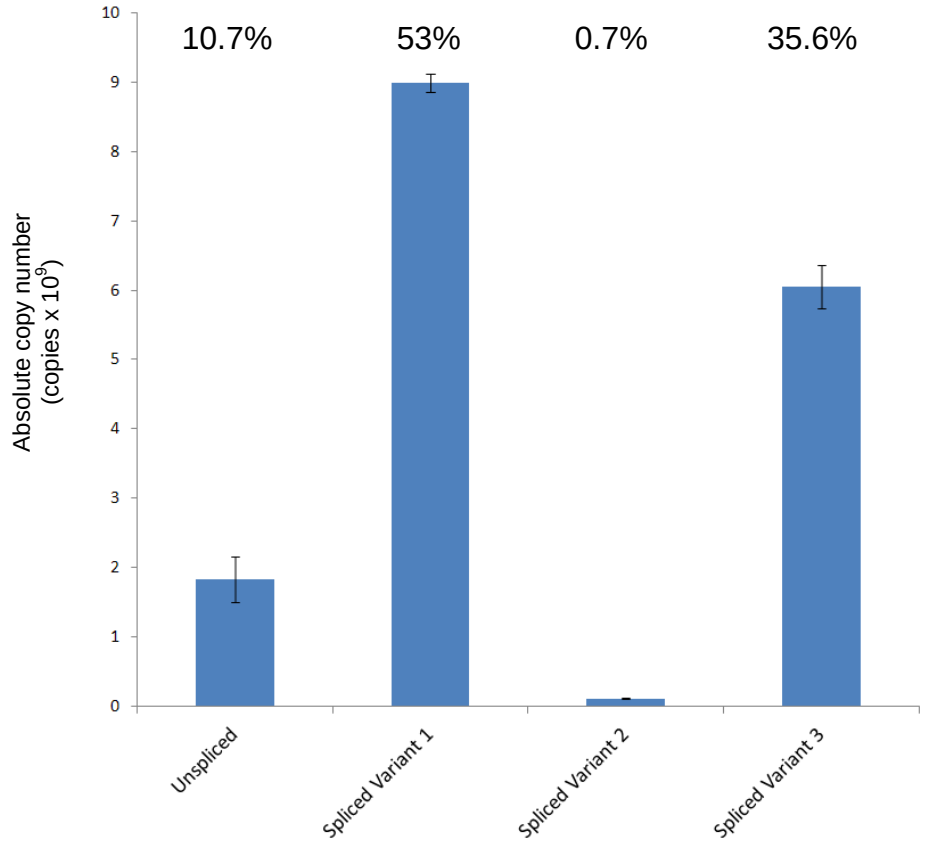
Real-time PCR was used to determine the relative levels of alternatively spliced, and unspliced, transcripts in the cDNA pool obtained from HEK293 cells transfected with the pEGFP- (sORF I-VI Wt) construct. TaqMan® probes specific to each species (unspliced product and Spliced Variants 1, 2, and 3) were used. All four probes were efficient (Figure 5.10 A); probe specificity was confirmed using positive control DNA, absence of amplification in –RT indicated samples were not contaminated with DNA. Copy numbers indicate that Variants 1 and 3 are the most abundant products (Figure 5.10 B) present in 3:2 ratio ($p<0.01$, $n=3$). In combination, these two transcripts are 8-fold more abundant than the unspliced transcript, consistent with the RT-PCR experiments (Figure 5.4, Lane 4). Spliced Variant 2 was the least abundant of all the transcripts, equivalent to 6% of the unspliced transcript ($p<0.01$, $n=3$).

Figure 5.10 Abundance of HIV-1 *asp* upstream sORFs I-VI Spliced Variants. Quantitative real-time PCR analysis of transcript abundance confirms that Spliced Variants 1 and 3 are the most abundant products present at a ratio of 3:2 ($p<0.01$) and, in combination, are 8-fold more abundant than the unspliced transcript. Spliced Variant 2 was the least abundant of all the transcripts. (A) The efficiencies of each primer/probe, with all four sets depicting efficiencies between 90 and 110%. (B) Copy numbers of each transcript (Unspliced, Spliced Variants 1, 2 and 3) within the same cDNA sample. Samples were assessed for DNA contamination (-RT) and primer specificity; no contamination was detected. Copy numbers obtained for each transcript were significantly different ($p < 0.01$). Percentage abundance of each Spliced Variant within the pool indicated above each respective bar. All data represent the mean ±SE of triplicate samples.

A

Probe	Slope	Efficiency	R^2
Unspliced	-3.10	110%	0.98
Spliced Variant 1	-3.43	96%	0.99
Spliced Variant 2	-3.26	103%	0.99
Spliced Variant 3	-3.22	104%	0.97

B



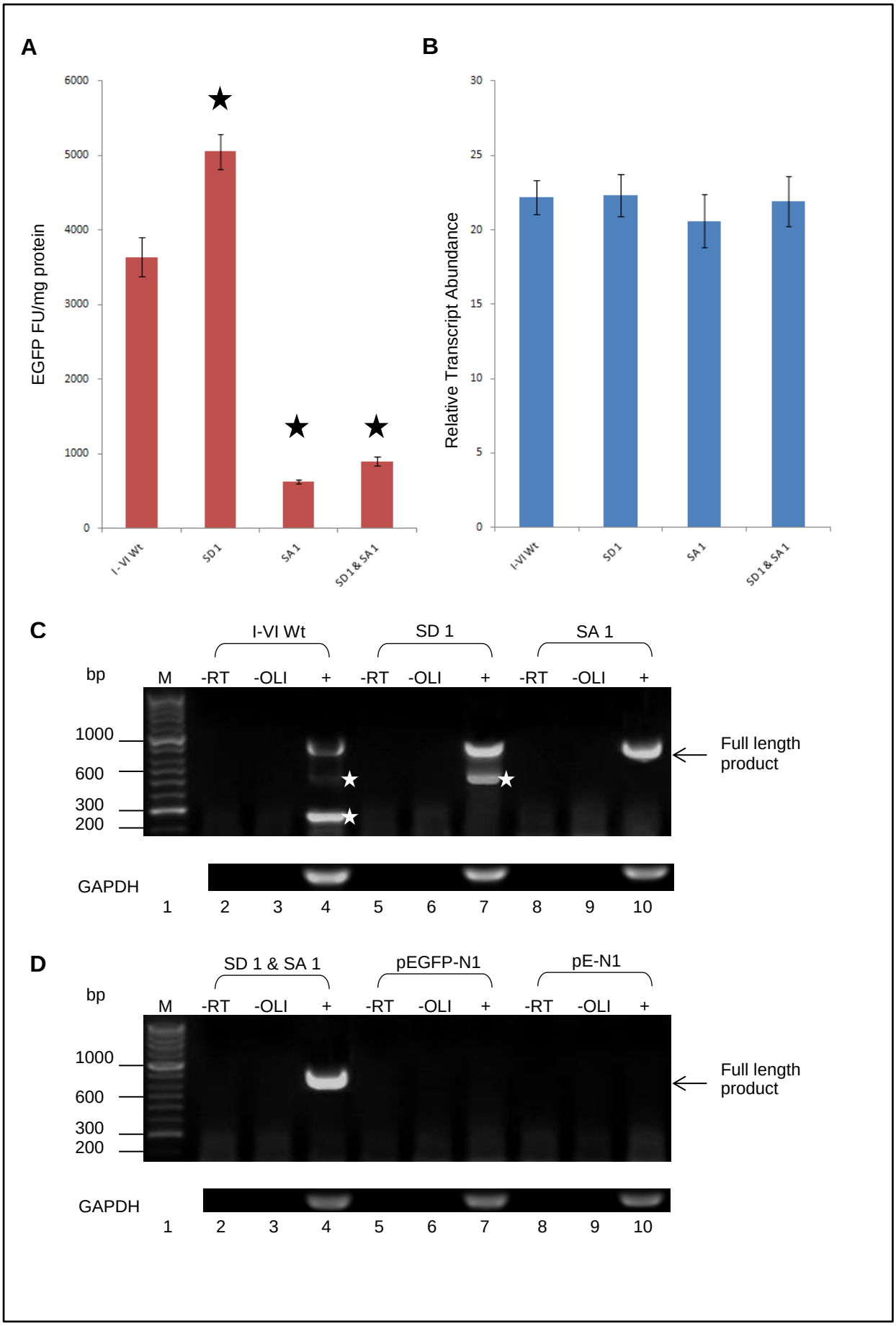
5.4.3 Effects of mutating the SA 1 and SD 1 motifs

Constructs in which the SD 1 and/or SA 1 sites had been mutated were transfected into HEK293 cells alongside the pEGFP-(sORF I-VI Wt) construct and controls (pEGFP-N1, pE-N1); analysis of reporter gene expression, transcript abundance and transcript integrity was performed as previously described.

Mutation of the conserved splice donor (SD 1) produced a 28% increase in EGFP expression compared to the wild type ($p = 0.003$, $n=6$) (Figure 5.11 A); transcript abundance did not change (Figure 5.11 B). RT-PCR analysis (Figure 5.11 C, seventh lane) showed both the full-length, unspliced transcript and Spliced Variant 2 (produced using the alternative splice donor site, SD 2, at nt 8539) were present (confirmed by PCR product sequence analysis, Figure 5.11 C, Lane 7, starred). Note that Spliced Variant 3 was not detected due to the absence of SD 1.

Similar results were obtained when SA 1 was disrupted in either the presence or absence of SD1. Production of EGFP was reduced to approximately 17% of wild type levels (Figure 5.11 A) and transcript abundance was, again, unaffected (Figure 5.11 B). The predominant cDNA/ mRNA species in each instance was the full-length, unspliced transcript; no spliced products were detected (Figure 5.11 C, Lanes 4 and 10).

Figure 5.11 Effect of mutating the major splice donor and acceptor motifs within the HIV-1 *asp* upstream sORFs I-VI region on the reporter EGFP gene expression, transcript abundance and analysis of Spliced Variants by RT-PCR. Mutation of Splice Donor 1 increased EGFP expression by 28% ($p = 0.003$) in comparison to the wild type. Mutation of Splice Acceptor 1 (with or without disruption of Splice Donor 1) decreased EGFP expression to approximately 17% of wild type levels. Transcript abundance is consistent for all constructs. RT-PCR analysis of Splice Donor 1 mutant reveals splicing to produce Spliced Variant 2, while mutation of Splice Acceptor 1, with or without Splice Donor 1 abolishes splicing activity. (A) EGFP assay of HEK293 cells transfected with constructs containing sORFs I-VI upstream of the reporter EGFP with mutation of the SD 1 and SA 1 motifs. All data represent the mean ±SE of six individual experiments. Significant differences ($p = <0.01$) in EGFP expression compared to the I-VI Wt noted with a star. (B) Transcript abundance; consistent across all constructs. All data represent the mean ±SE of three individual experiments. RT-PCR analysis performed on RNA samples from HEK293 cells. Analysis of constructs (C) with and without mutation of the SD 1 and SA 1 motifs and (D) construct containing mutation of both SD 1 and SA 1 along with pEGFP-N1 and pE-N1 parental plasmid controls. Samples were assessed for DNA contamination (-RT) and RNA priming (-OLI); no contamination was detected. Lanes marked + depict the cDNA samples. Lanes marked I-VI Wt show the amplification products from the three alternative transcripts detected: the major product, Spliced Variants 1 and 3 (240bp amplimer) and less abundant product, Spliced Variant 2 (480bp amplimer) transcripts are indicated with a star, the unspliced transcript is indicated by the 890bp amplimer. Lane marked SD 1, construct where the SD 1 motif was mutated, displays increased use of SD 2. Lanes marked SA 1, construct where the SA 1 motif was mutated and SD 1 & SA 1, where both SD 1 and SA 1 motifs have been mutated, only contain the 840bp amplification product of the full-length transcript, indicating loss of splicing activity. GAPDH controls, shown below, indicate integrity of the transcript pool.


A
EGFP FU/mg protein
6000
5000
4000
3000
2000
1000
0
I-VI Wt
SD 1
SA 1
SD 1 & SA 1
B
Relative Transcript Abundance
30
25
20
15
10
5
0
I-VI Wt
SD 1
SA 1
SD 1 & SA 1
C
I-VI Wt
SD 1
SA 1
bp
M
-RT
-OLI
+
1000
600
300
200
Full length product
GAPDH
1 2 3 4 5 6 7 8 9 10
D
SD 1 & SA 1
pEGFP-N1
pE-N1
bp
M
-RT
-OLI
+
1000
600
300
200
Full length product
GAPDH
1 2 3 4 5 6 7 8 9 10

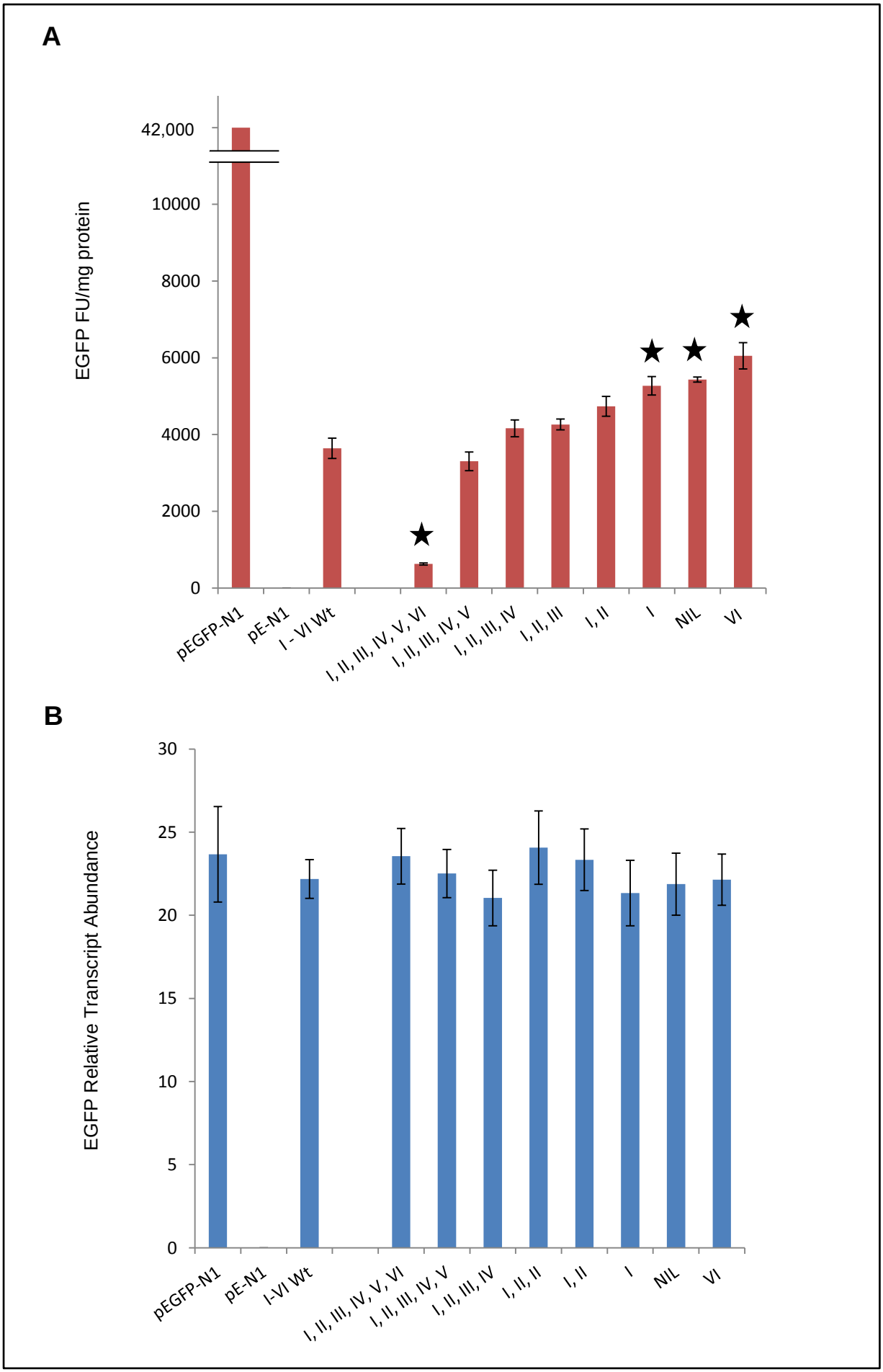
5.4.5 Re-examination of sORF effects by mutation of initiation codons

In Chapter 4 (section 4.4.4) the role of each sORF was examined by sequentially mutating each sORF initiation codon. These data showed that mutation of sORF VI AUG initiation codon produced a significant increase in the level of EGFP expression in comparison to the I-VI Wt. The inhibitory effect of sORF VI was restored once all of the sORFs I-V were mutated leaving sORF VI AUG intact. This could now be explained by the predominant Spliced Variant (Spliced Variant 1), containing sORF VI, producing the dominant expression. Thus in order to re-examine the effects of each sORF, the experiment was repeated with the additional mutation in the major splice acceptor motif, thus inhibiting splicing of the transcript.

Mutation of the major splice acceptor produced an 83% reduction in the level of EGFP reporter expression (Figure 5.12 A) in comparison to the Wt splice acceptor motif ($p = <0.01$, $n=6$). This dramatic decrease in expression can be attributed to the inhibitory nature of the sORFs. In the absence of SA 1, only unspliced full length transcript is produced, in comparison to the Wt where the predominant Spliced Variant 1 (containing sORF VI alone) is produced. RT-PCR analysis confirmed the absence of splicing activity with all constructs containing the splicing acceptor motif mutation as seen in Figure 5.11 C, lane 10 (data not shown).

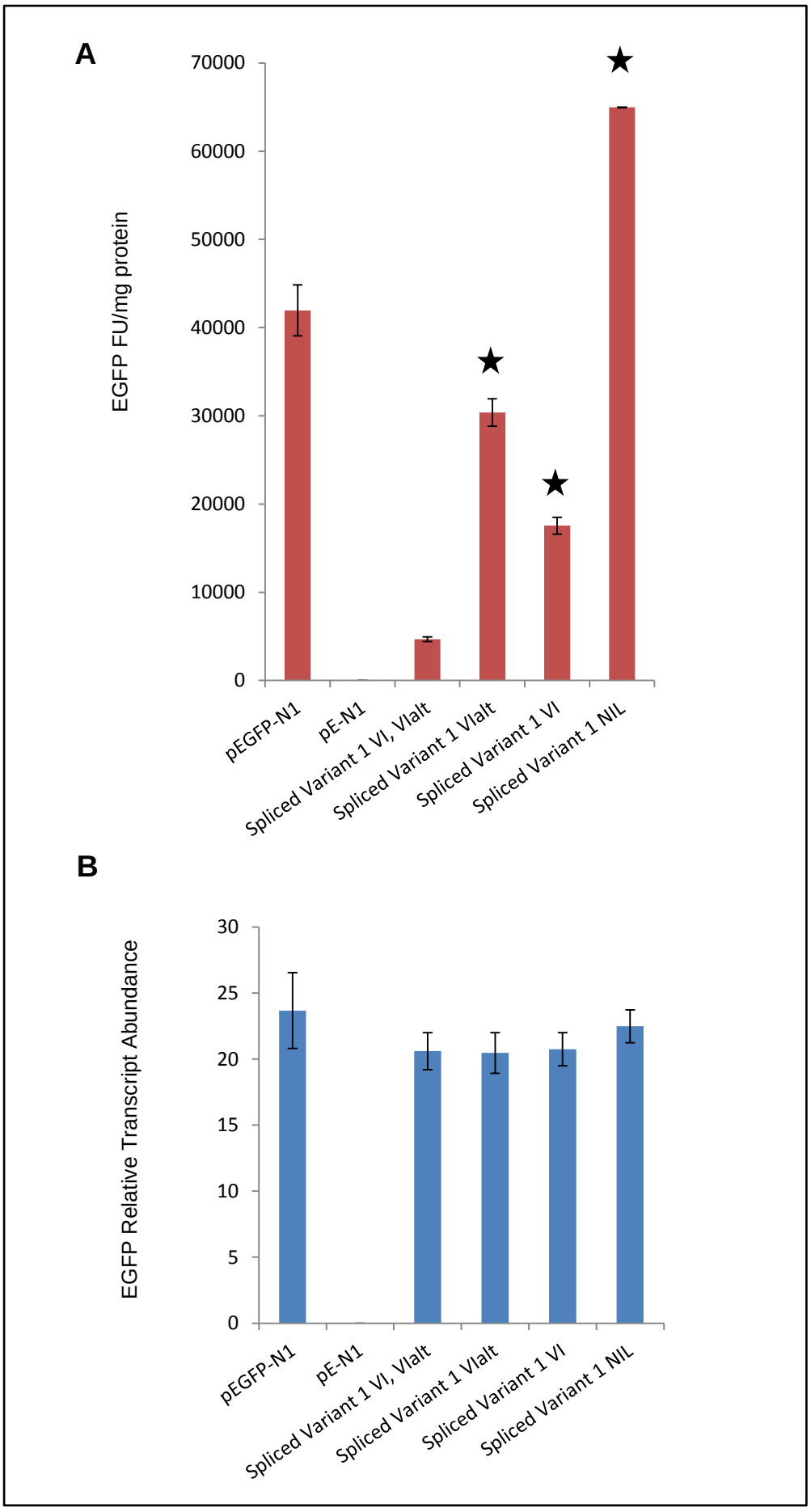
Mutation of the splicing acceptor motif in conjunction with sequential mutations in the sORF initiation codons revealed a dramatic increase in the level of reporter gene expression. Mutation of sORF VI alone, leaving sORFs I-V active, alleviated the 83% inhibition observed when sORF VI is left intact ($p = <0.01$, $n=6$), suggesting a key role for sORF VI in inhibition of downstream gene expresssion. The level of expression increases as each sORF initiation codon is sequentially mutated along the transcript. Interestingly, when all sORF initiation codons except sORF VI had been deactivated, a moderately high level of reporter gene expression ($p = <0.01$, $n=6$) resulted (40% increase compared to the I-VI Wt), implying that sORFs I to V may facilitate inhibition by sORF VI. Transcript abundance was consistent across all experiments (Figure 5.12 B).

Figure 5.12 Re-examination of the effect of unspliced HIV-1 *asp* upstream sORFs I-VI by mutation of sORF initiation codons on the reporter EGFP and transcript abundance. Knock-out of the major splice acceptor motif results in an 83% reduction in reporter gene expression in comparison to the Wt, where splicing events occur. Expression can be substantively restored by mutation of the sORF VI initiation codon, with additional mutations in each sORF initiation codon further alleviating this inhibition of expression. (A) EGFP assay of HEK293 cells transfected with constructs containing sORFs I-VI upstream of the reporter EGFP with mutation of the SA 1 motifs. I-VI Wt construct does not contain any mutations in the sORF AUG codons nor the SA motif, labels indicate active sORF AUG codons intact. Construct NIL contains no active sORF initation codons. All data represent the mean ±SE of six individual experiments. Significant differences ($p = <0.01$) in EGFP expression compared to the I-VI Wt noted with a star. (B) Transcript abundance; consistent across all constructs. All data represent the mean ±SE of three individual experiments.


A
EGFP FU/mg protein
42,000
10000
8000
6000
4000
2000
0
pEGFP-N1
pE-N1
I - VI Wt
I, II, III, IV, V, VI
I, II, III, IV, V
I, II, III, IV
I, II, III
I, II
I
NIL
VI
B
EGFP Relative Transcript Abundance
30
25
20
15
10
5
0
pEGFP-N1
pE-N1
I-VI Wt
I, II, III, IV, V, VI
I, II, III, IV, V
I, II, III, IV
I, II, II
I, II
I
NIL
VI

To better understand the potential roles of sORF VI and VI_{alt} in the regulation of gene expression, mutations in the sORF initiation codons of both sORF VI and VI_{alt} in Spliced Variant 1, the predominant product, were introduced and their respective effects on reporter gene expression noted (Figure 5.13 A). Mutation of sORF VI resulted in a 7-fold increase in the level of reporter gene expression ($p = <0.01$, $n=6$) in comparison to the non-mutated Spliced Variant 1. In contrast there was only a 4-fold increase in the level of reporter gene expression when the sORF VI_{alt} AUG initiation codon was mutated ($p = <0.01$, $n=6$). When mutations of both sORF VI and VI_{alt} were combined, the highest level of reporter gene expression (14-fold increase in comparison to the non-mutated Spliced Variant 1) was observed ($p = <0.01$, $n=6$). Confirmation of the absence of splicing was obtained by RT-PCR analysis with all constructs producing the single transcript as seen in Figure 5.5 A, lane 2 (data not shown). Transcript abundance was consistent across all experiments (Figure 5.13 B).

Figure 5.13 Examination of the effect of sORFs I-VI Spliced Variant 1, the predominant HIV-1 *asp* upstream sORF configuration, by mutation of sORF initiation codons on the reporter EGFP and transcript abundance. A) EGFP assay of HEK293 cells transfected with constructs cloned with Spliced Variant 1, containing sORFs VI and VI_{alt} upstream of the reporter EGFP. Labels indicate active sORFs with initiation codons intact. All data represent the mean ±SE of six individual experiments. (B) Transcript abundance; consistent across all constructs. All data represent the mean ±SE of three individual experiments. Significant differences (p= <0.01) in EGFP expression compared to Spliced Variant 1 with sORFs VI, VI_{alt} active noted with a star.


A
EGFP FU/mg protein
70000
60000
50000
40000
30000
20000
10000
0
pEGFP-N1
pE-N1
Spliced Variant 1 VI, VIalt
Spliced Variant 1 VIalt
Spliced Variant 1 VI
Spliced Variant 1 NIL
B
EGFP Relative Transcript Abundance
30
25
20
15
10
5
0
pEGFP-N1
pE-N1
Spliced Variant 1 VI, VIalt
Spliced Variant 1 VIalt
Spliced Variant 1 VI
Spliced Variant 1 NIL

5.5 Discussion

The conservation of the *asp* ORF in HIV-1 has been well established (Miller, 1988), and previous chapters have described the presence and high level conservation of a series of sORFs upstream of the *asp* ORF across all subtypes of HIV-1. These sORFs are typically associated with translational regulation of the downstream gene product; examples include many genes for which tight regulation of expression is critical (Davuluri *et al.*, 2000; Morris and Geballe, 2000; Suzuki *et al*., 2000). Data presented in Chapter 4 demonstrated that the sORF region upstream of HIV-1 *asp* regulates the expression of a downstream reporter gene without any change in transcript abundance. Regulation was abolished by mutation of the sORF AUG initiation codons, suggesting a translational mechanism of regulation. Experiments discussed in this chapter further examined the ability of the sORF region to regulate downstream gene expression using a splicing mechanism.

Results have shown that the *asp* sORF region undergoes alternative splicing *in vitro* producing at least three different Spliced Variants in addition to the unspliced transcript. The most abundant of these retains only the sORF immediately preceding *asp*, sORF VI. To date several transcript sizes for *asp* have been identified: 1.6kb, 1.1kb. 1.0kb (Bukrinsky and Etkin, 1990), 2.3kb (Michael *et al*., 1994b), 4.1kb (Landry *et al*., 2007) and 2.6kb (Kobayashi-Ishihara *et al*., 2012). The transcript sizes reported by Michael *et al*. (1994b) and Bukrinsky and Etkin (1990) of 2.3kb and 1.6kb respectively, can potentially be explained as the unspliced and predominant spliced forms (Spliced Variants 1 and 3), differing by the size of the 0.65kb intron that was detected here.

The importance of SA 1 and SD 1 is supported by the high conservation of these typical splice acceptor and donor sequences. While conservation of SD1 could be explained by high level conservation of the positive sense strand, as required to maintain Pro (NL4-3 Env Pro814) in this position within Env gp41, this is not the case for SA1. The inclusion of any nucleotide base at position X of the ACX nucleotide sequence would maintain Thr at this position within gp41 (NL4-3 Env Thr597), suggesting that another selective pressure, selection of the SA1 sequence, is responsible for the high level conservation observed here. SD 2 is only moderately

conserved and SA 2 is poorly conserved, so these sites may be of lesser importance. The polypyrimidine tract sequence of SA 1 is also stronger (9/11 Py) than that of SA 2 (Py 7/11) and may also explain the predominant utilisation of SA 1.

These experiments have also shown that mutation of the predominantly utilised splice acceptor, SA 1, abolishes splicing activity of the sORF region, drastically reducing reporter expression. Splicing may have a role in releasing the severe inhibitory effect of the presence of the sORFs, in a similar manner to regulation of expression of the Estrogen Receptor α by alternative splicing of upstream sORFs reported by Kos (2002).

Spliced Variant 1 (containing sORF VI alone) produces a higher level of expression in comparison to Spliced Variant 2 (containing sORFs I, II and VI), which could be related to the number of upstream sORFs. However Spliced Variant 3 (sORF VI_{alt}) produces much higher levels of expression than either Spliced Variants 1 or 2, suggesting that Spliced Variant 3 may permit downstream ORF expression and that inhibition is specifically associated with sORF VI. The mechanism by which ribosomes translate these transcripts is the subject of further work presented in the following chapter.

Given the splicing activity occurring within the sORF region, the respective role of each sORF was re-examined in constructs where splicing was inactivated. These data indicated that the sORF region severely inhibits the translational expression of the downstream reporter gene. Mutation of sORF VI alone relieves this inhibition, as does mutation of the initiation codons of the entire sORF region. Knock-out produces a step-wise increase in the level of reporter gene expression as sORFs are gradually inactivated. This suggests an additive role for each sORF in the inhibition of translation. Inactivation of sORF I produces no significant change in reporter gene expression compared to construct NIL ($p= 0.538$, $n= 6$). The largest drop in expression is observed when all sORFs are active, compared to the construct in which all sORFs except sORF VI are active ($p= <0.01$, $n=6$). This might suggest that some sORFs play lesser roles in the inhibition of downstream gene expression compared to others within the series.

The investigation of the role of sORFs VI and VI_{alt} in the regulation of gene expression within Spliced Variant 1, the predominant product, confirmed that sORF VI plays a more important role in the inhibition of gene expression than VI_{alt}. Combined together, mutations of both initiation codons in this Spliced Variant permits the highest level of gene expression, releasing the inhibitory nature of the sORF. This suggests that Spliced Variant 3, containing sORF VI_{alt} may permit the expression of *asp*.

5.6 Conclusions

Work presented in this chapter identifies four possible transcripts for the sORF transcript within the reporter gene system used. These included the full length unspliced transcript; a predominant Variant 1 produced using highly conserved splicing donor and acceptor motifs and containing sORF VI and the alternative initiation codon, VI_{alt}; Variant 2 containing sORFs I, II and VI; and Variant 3, a sub-variant of Spliced Variant 2, containing the alternative initiation codon, VI_{alt}, alone. The levels of each transcript variant within the cDNA pool were confirmed by quantitative real-time PCR and highlight the abundance of Spliced Variants 1 and 3 compared to full length transcript within each experiment. Mutational analysis further showed that knock-out of the major splice motif inhibits this splicing activity and results in the use of full length transcript, strongly inhibiting translational expression. Further examination of the role of each sORF, within both the full length unspliced transcript and the predominant Spliced Variant 1, indicates that sORF VI plays a major role in inhibition of gene expression and that sORF VI_{alt} permits expression of the main ORF.

CHAPTER 6 – POTENTIAL FOR THE TRANSLATIONAL CONTROL OF GENE EXPRESSION IN THE HIV-1 *asp* sORF REGION

6.1 General introduction

Gene expression may be controlled at several levels, including the level of transcription, stability of the transcript and/or translational efficiency. These processes may, in combination, tightly regulate expression. In particular, the efficiency with which the translational machinery reads the mRNA transcript by is of growing interest. The presence of multiple features, such as upstream sORFs and/or secondary structures within the transcript, are well known to affect translational efficiency by a complex mechanism, involving assembly and movement of the translational machinery upon the transcript (Kozak, 1990).

Previous chapters have considered the role of splicing in the regulation of HIV-1 *asp* expression. Consistent levels of transcript have been observed, suggesting the sORFs may play a role at the level of translation. Two likely mechanisms may control ribosome progression, stalling or shunting at the sORF initiation codon. Stalling of the ribosome may result from mRNA folding to form secondary structures that may slow ribosomal progression, and may permit the recognition of sub-optimal context initiation codons. This process typically occurs when a stem-loop structure is located 14 nucleotides upstream from the initiation codon (the approximate distance of the leading edge of the ribosome and its initiation AUG recognition centre) (Kozak, 1990). Secondary structures with a Gibbs Free Energy higher than $\Delta G = -30$ kcal/mol are considered strong enough to stall translating ribosomes (Svitkin *et al*., 2001; Kos *et al*., 2002). The ribosomal shunt model would predict that the scanning ribosome, upon reaching a hairpin loop structure, would 'shunt' across the structure instead of unwinding it (Jackson, 1996). The features vital for shunting activity include that the primary sORF must be between 2 to 15 codons in length and the distance between the sORF termination codon and the base of the secondary structure should be between 5 to 10 nucleotides (Dominguez *et al*., 2000; Hemmings-Mieszczak *et al*., 2000). Perhaps one of the most well studied examples of the shunt model is the cauliflower mosaic virus RNA leader, where the scanning ribosome having translated sORF A, is able to shunt across the transcript to the ORF VII AUG initiation codon

bypassing a stem-loop structure located at nt 70-550 (Dominguez *et al*., 1998; Gale *et al*., 2000; Pooggin *et al*., 2000). Similarly, the minimal sORF, implicated in the regulation of Env expression is consistent with the shunting model (Krummheuer et al., 2007). Work presented in this chapter further explores the potential role of sORFs in control of expression of HIV-1 *asp*. The mechanisms by which ribosomes may translate the sORF-containing transcript are investigated

The toeprinting technique (Figure 6.1) was developed as a means of examining the efficiency of initiation and assembly of the ribosomal complexes in protein synthesis within eukaryotes (Kozak, 1998). Kozak (1998) developed the technique to study ribosomal scanning over quite long distances and proved that the introduction of secondary structures upstream of the AUG initiation codon may block this scanning altogether. This technique comprises three parts:

1. The ribosomal complexes are bound to the mRNA transcript and allowed to move along the transcript.
2. A radio (or fluorescent) labelled primer is bound to the mRNA transcript, upstream of the sORF AUG initiation codon and reverse transcription from the primer is completed, synthesising cDNA. This reverse transcription process will be blocked by ribosomes stalled at the initiation codon (or elsewhere on the transcript).
3. The products are resolved by electrophoresis, alongside products of sequencing reactions prepared using the same primer. As ribosomes protect a region of 15-16 nts the resulting labelled reverse transcripts are short by this distance from the AUG initiation codon.

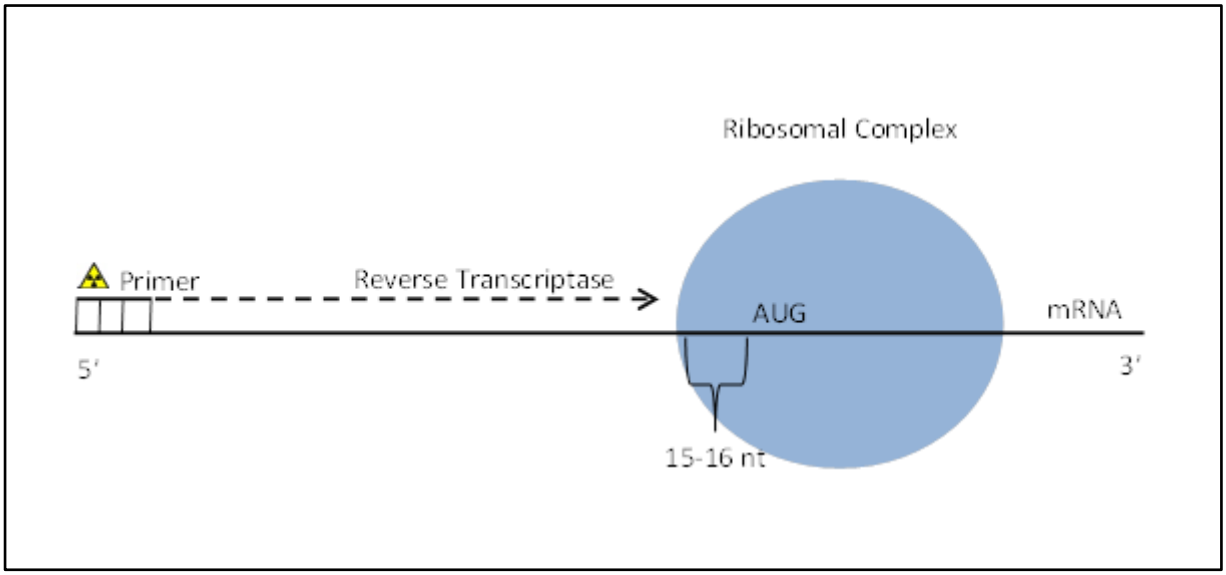


Figure 6.1 Schematic of the toeprinting technique. Initially the ribosomal complexes are allowed to form and move along the mRNA transcript. Secondly a radio (or fluorescently) labelled probe is bound upstream of the sORF AUG initation codon. The reverse transcriptase reaction then synthesizes the cDNA, until it reaches the ribosomal boundary (located 15-15 nts from the A position of the sORF AUG initation codon). The final stage of the toeprinting technique involves the separation of the cDNA products by electrophoresis.

The toeprinting assay is reliant upon the use of an inhibitor to block the elongation phase of translation and hold the 80S ribosomal complex at the AUG initiation codon, which subsequently blocks the primer extension phase of the assay. In the development of the toeprinting assay Kozak (1998) tested a series of different inhibitors (sparsomycin, anisomycin and cycloheximide). The inhibitors sparsomycin and cycloheximide act by binding the aminoacyl-tRNA to the P site, rather than the A site, (Monro *et al*., 1969; Oen *et al*., 1974) while anisomycin inhibits the binding of the aminoacyl-tRNA to the A-site (Carrasco *et al*., 1973). Kozak (1998) reported a much cleaner toeprint result when cycloheximide was used alone at a concentration of 900µg/mL, or used at a lower concentration (90µg/mL) in combination with sparsomycin (200µM). The efficiency of the toeprinting assay is also affected by a range of environmental factors. For example Kozak (1998) reported that longer incubation times typically decrease yield of extension products, possibly due to degradation, while magnesium concentrations as high as 2.5mM may be required for optimal ribosome complex formation (Snyder and Edwards, 1991).

Since the development of the toeprinting assay (Kozak, 1998), a host of studies have utilised this technique to understand the role sORFs play in a range of different settings. *S.cereviciae* extracts have been used to study the *GCN4* mRNA 5'-leader which contains a series of four upstream sORFs (Gaba *et al*., 2001). Since the availability of commercially available cell-free translation systems the technique has been used to study the translation of 15-lipoxygenase (Ostareck *et al*., 1997; Ostareck *et al*., 2001), the estrogen receptor α gene (Kos *et al*., 2002), the *her-2* oncogene (Spevak *et al*., 2006) and the mu opioid receptor (Song *et al*., 2007). The work presented in this chapter uses the toeprinting assay to similarly examine mechanisms of translational control associated with the HIV-1 NL4-3 *asp* sORF transcript and the potential role of each sORF in the translation of *asp*.

6.2 Aims

The aim of this study is to investigate the position/s at which ribosomes stall across the sORF region upstream of the HIV-1 NL4-3 *asp* gene. More specifically:

1. Secondary structure predictions, using MFOLD, were investigated to determine potential sites associated with ribosomal stalling within the upstream regions of each sORF mRNA sequence.
2. The toeprinting assay was established and optimised, then used to determine sites associated with ribosomal stalling.
3. A translational model for the ribosomal scanning of the sORF transcript was developed.

6.3 Materials and methods

6.3.1 RNA secondary structure prediction by MFOLD

Secondary structures that may form on the messenger RNA were predicted using the MFOLD server provided by the Burnet Institute, Melbourne, Victoria, Australia (http://mfold.burnet.edu.au) by feeding in the first 50nt upstream of each sORF. ΔG values are presented in kcal/mol where 1 kcal is equal to 4.184 kJ (Zuker, 2003).

6.3.2 Plasmid constructs

The plasmid constructs, pEGFP (sORF I-VI Wt) and pEGFP (sORF I-VI NIL) used in this investigation are described in Chapter 2, section 2.3. These constructs contain either the entire sORF region (sORFs I – VI) with all the AUG initiation codons intact or the entire sORF region with all the AUG initiation codons mutated as discussed in Chapter 4, section 4.3.1.

6.3.3 Probes

Deoxyoligonucleotide probes labelled in the 5´ position were designed for each respective sORF as shown in Table 6.1 below. These probes served to prime the final reverse transcription stage of the toeprinting assay and the dideoxy cycle sequencing reactions used to map the toeprint positions. Each probe was positioned a minimum of 10nt downstream from the termination codon of each respective sORF, but not more than 200nt from the sORF initiating AUG codon. To increase labelling efficiency all probes contained an A or T residue on the 5′ end.

Table 6.1 Deoxyoligonucleotide toeprinting probes.

Probe	Sequence (5′ - 3′)	T_M(°C)
sORF I and II	TGTGGAACTTCTGGGACGCAG	66
sORF III	TAGAGTTAGGCAGGGATATTCAC	66
sORF IV	TAGGAGGCTTGGTAGGTTTAAG	64
sORF V	ATGACCTGGATGGAGTGGGAC	66
sORF VI	AATCCTGGCTGTGGAAAGATAC	64

6.3.4 *In vitro* transcription

The addition of the T7 promoter used for *in vitro* transcription was made by PCR amplification. Reactions were established with 100ng plasmid DNA, 10µL 2x AmpliTaq Gold® Fast PCR Master Mix, Universal PCR (Applied Biosystems) and 10µM of the following primers: Forward 5′- CGTGACGTGCGGCATCC -3′ and Reverse 5′- CTAATACGACTCACTATAGGGAGACCGGACTCAGATCTCTCGAGC -3′ with T7 promoter sequence underlined. Reactions were amplified after an initial denaturation at 95˚C for 10 min followed by 35 cycles of 96˚C for 3 sec, 50˚C for 3 sec and 68˚C for 20 sec and a final chase cycle of 72˚C for 10 sec. The PCR products were purified using the PureLink™ PCR Micro Kit (Invitrogen) and quantitated as described in 2.5.2. PCR products were diluted to a final DNA concentration of 100ng/µL.

In vitro transcription reactions were established at room temperature with 0.5µg of template PCR product, 10mM DTT, 40units RNase OUT, 10mM rATP, 10mM rCTP, 10mM rUTP, 1mM rGTP, 10mM m^7G (5′)ppp(5′)G RNA Cap Analog (NEB) and 10µL 5x T7 Buffer (0.2M Tris-HCl (pH 8.0), 40mM $MgCl_2$, 10mM spermidine-$(HCl)_3$, 125mM NaCl). 50units of T7 RNA Polymerase (Invitrogen) were added and the volume adjusted to 50µL with RNase-free water. The reaction mixtures were incubated at 37˚C (air jacket) overnight before DNase digestion and purification using the SV Total RNA Isolation System (Promega) and quantitated as described in 2.9.1 and 2.9.2.2. The 0.9 kb RNA transcript was analysed by PAGE to ensure transcript purity and integrity.

6.3.5 Probe labelling with [γ^{32}P]ATP

Deoxyoligonucleotide toeprinting probes were labelled with [γ^{32}P]ATP in 50µL reactions. Each reaction was established with 50pmol primer and 25µL of fresh [γ^{32}P]ATP 3000Ci/mmol (10mCi/mL) along with 10µL 5x Forward Reaction Buffer (350mM Tris-HCl (pH 7.6), 50mM $MgCl_2$, 500mM KCl, 5mM 2-mercaptoethanol) and 10units T4 Polynucleotide Kinase (Invitrogen). Reaction mixtures were incubated at 37˚C for 1hr followed by inactivation at 65˚C for 10 min. Labelled probes were purified of excess [γ^{32}P]ATP using the mini Quick Spin Oligo Columns (Roche)

according to the manufacturer's instructions. Purified probes were eluted in a total volume of 50µL.

6.3.6. Toeprinting assay

6.3.6.1 Probe hybridisation and *in vitro* translations

To avoid detachment of ribosomal complexes at excessive temperatures, the labelled probes were annealed to *in vitro* transcribed mRNA transcripts prior to the translation and primer extension stages of the assay. Approximately 3pmol of purified probe was mixed with 400ng of *in vitro* transcribed mRNA template along with 20units RNase OUT, 50mM Tris (pH 7.5) in a total volume of 10µL. The mix was heated to 70˚C for 5 min followed by gradual cooling (1˚C per minute), held at 37˚C for 10 min, and then held on wet-ice.

The mix was placed directly into the *in vitro* translation reaction. The 50µL reactions were assembled on ice to contain 10µM Complete Amino Acid Mix (Promega), 40units RNase OUT, 35µL (50% total volume) Nuclease-treated Rabbit Reticulocyte Lysate (Promega) and, where required, 1µL Cycloheximide (100mg/mL in DMSO). The reactions were incubated at 25˚C for 3 min prior to the addition of the 10µL Probe-mRNA mix and then further incubated for 15 min at 25˚C before snap chilling on wet-ice.

6.3.6.2 Primer extension

Primer extension reactions were established with 4µL of the *in vitro* translation mix from section 6.3.6.1. They were immediately diluted with an equal volume of ice-cold 5x Extension Buffer (50mM Tris-HCl (pH 7.5) with 40mM KCl, 6mM $MgCl_2$), 10mM DTT, 40units RNase OUT, 750µM dNTPs and 1µL Cycloheximide (0.7µg/µL in DMSO) in a total volume of 20µL. Reverse transcription was initiated by the addition of 100units SuperScript™ II Reverse Transcriptase (Invitrogen) and incubated at 30˚C for 30 min. The reactions were stopped by extraction with an equal volume of Phenol:Chloroform (1:1). The extracted product was mixed with an equal volume of 2x STR Loading Solution (95% Formamide, 10mM NaOH, 0.05% Bromophenol blue

and 0.05% Xylene cyanol) (Promega). A 10uL sample was heated to 100˚C for 2 min and cooled on ice before layering onto a sequencing gel for electrophoresis.

6.3.7 Dideoxy-cycle sequencing reactions

Sequencing reactions using the same probes were used to map exact positions of toeprints upon the transcripts. Four separate 20µL reactions comprised 100fmol of plasmid DNA, 1.5pmol labelled probe, 2µL 10x AmpliTaq® Gold 360 Buffer, 2.5mM MgCl, 1.25units of AmpliTaq® Gold 360 DNA Polymerase (Applied Biosystems) and 4µL of one of the four ddNTP mixes: ddCTP mix (20mM of each dNTP and 400mM ddCTP), ddTTP mix (20mM of each dNTP and 800mM ddTTP), ddATP mix (20mM of each dNTP and 600mM ddATP) and ddGTP mix (20mM of each dNTP and 200mM ddGTP). The products were denatured by incubation at 95˚C for 10 min followed by 25 cycles of 95˚C for 30 sec, 50˚C for 30 sec and 72˚C for 1 min. This was followed by 10 cycles of 95˚C for 30 sec, 72˚C for 2 min. Products were extracted with an equal volume of Phenol:Chloroform (1:1), then mixed with an equal volume of 2x STR Loading Solution (95% Formamide, 10mM NaOH, 0.05% Bromophenol blue and 0.05% Xylene cyanol) (Promega). A 2uL sample was heated to 100˚C for 2 min and cooled on ice before layering onto a sequencing gel for electrophoresis.

6.3.8 Electrophoresis and autoradiography

All samples were resolved on 8% polyacrylamide sequencing gels. The samples were layered onto 0.4mm thick sequencing gels (20cm by 50cm) equilibrated in 0.5x TBE buffer (45mM Tris-borate, 1mM EDTA, pH 8.3). The gels were run at ~2,500V and 65W to maintain a constant gel temperature of 45˚C. The gel was run for ~1.5hrs and was monitored by the addition of loading dye (0.25% (w/v) bromophenol blue and 0.25% (w/v) xylene cyanol FF to the DNA samples. Completed gels were exposed to BioMax MS film (Kodak) against an intensifying screen for ~16hrs at -70˚C before being developed.

6.4 Results

6.4.1 Prediction of mRNA secondary structures

The MFOLD tool was used to predict potential structures that may cause ribosomes to stall or slow within the immediate vicinity of the sORF initiation codons. In order to do this, structures within a 50nt window upstream of each sORF initiation codon were predicted (Table 6.2). This is based upon the accepted sliding window of 30nt around the initiation codon as has been previously reported (Keller *et al*., 2011; Zhou and Wilke, 2011).

The predictions made with MFOLD for the 50nt region upstream of the initiation codon of sORF I showed a very small stem-loop structure immediately prior to the AUG initiation codon (Appendix 2, Figure A.1). The initiation codon was positioned only 2nts from the structure predicted and with the low combined ∆G value of -1.70 kcal/mol it would be considered too weak a structure to affect the rate at which the ribosome scans the mRNA transcript. This is based upon reports that a Gibbs Free Energy higher than ∆G = -30 kcal/mol is required to stall translating ribosomes (Kos *et al*., 2002). This was also the case with sORFs II, III and V where highly unstable structures with a ∆G of -4.03, -9.80 and -0.6 kcal/mol, respectively, were observed (Table 6.2, see also Appendix 2, Figure A.1).

The predicted structure for sORF IV (Figure 6.2) produced the highest Gibbs Free Energy at -18.90 kcal/mol (Table 6.2) and is positioned 7nts from the AUG initiation codon. While the Gibbs Free Energy is not greater than -30 kcal/mol it is considered moderately strong (Kozak, 1990) (∆G > -19.00 kcal/mol) and may, therefore, have an effect on slowing the scanning 40S ribosomal subunit to initiate at this AUG codon (strong sequence context).

The upstream sequences of sORF VI and the internal AUG of sORF VI_{alt} were also examined. All of the structures predicted for sORF VI and sORF VI_{alt} (Appendix 2, Figure A.2) had very low ∆G values ranging from -1.70 to -9.20 kcal/mol (Table 6.2) far less than the ∆G needed to significantly stall translating ribosomes. The predictions also showed than none of the stem-loop structures predicted for sORF VI

are within the 14 nucleotide distance from the AUG initiation codon. In most cases the small stem-loop structures are only 3-5 nucleotides from the AUG codon, or in the case of sORF VI structure 1 is >20 nucleotides from the AUG codon. Therefore, it is also unlikely that any of these structures would inhibit the scanning ribosome and/or slow the translation process.

Table 6.2 Comparison of Gibbs Free Energy between predicted secondary structures.

sORF	**ΔG (kcal/mol)**	**ΔG (kcal/mol)/nucleotide**	**Distance of structure from AUG (nt)**
sORF I	-1.70	-0.03	2
sORF II	-4.03	-0.08	1
sORF III	-9.80	-0.18	0
sORF IV	-18.90	-0.36	7
sORF V	-0.60	-0.01	7
sORF VI Structure 1	-2.70	-0.05	26
sORF VI Structure 2	-2.40	-0.04	3
sORF VI Structure 3	-2.20	-0.04	3
sORF VI Structure 4	-1.70	-0.03	5
sORF VI_{alt}	-9.20	-0.17	3

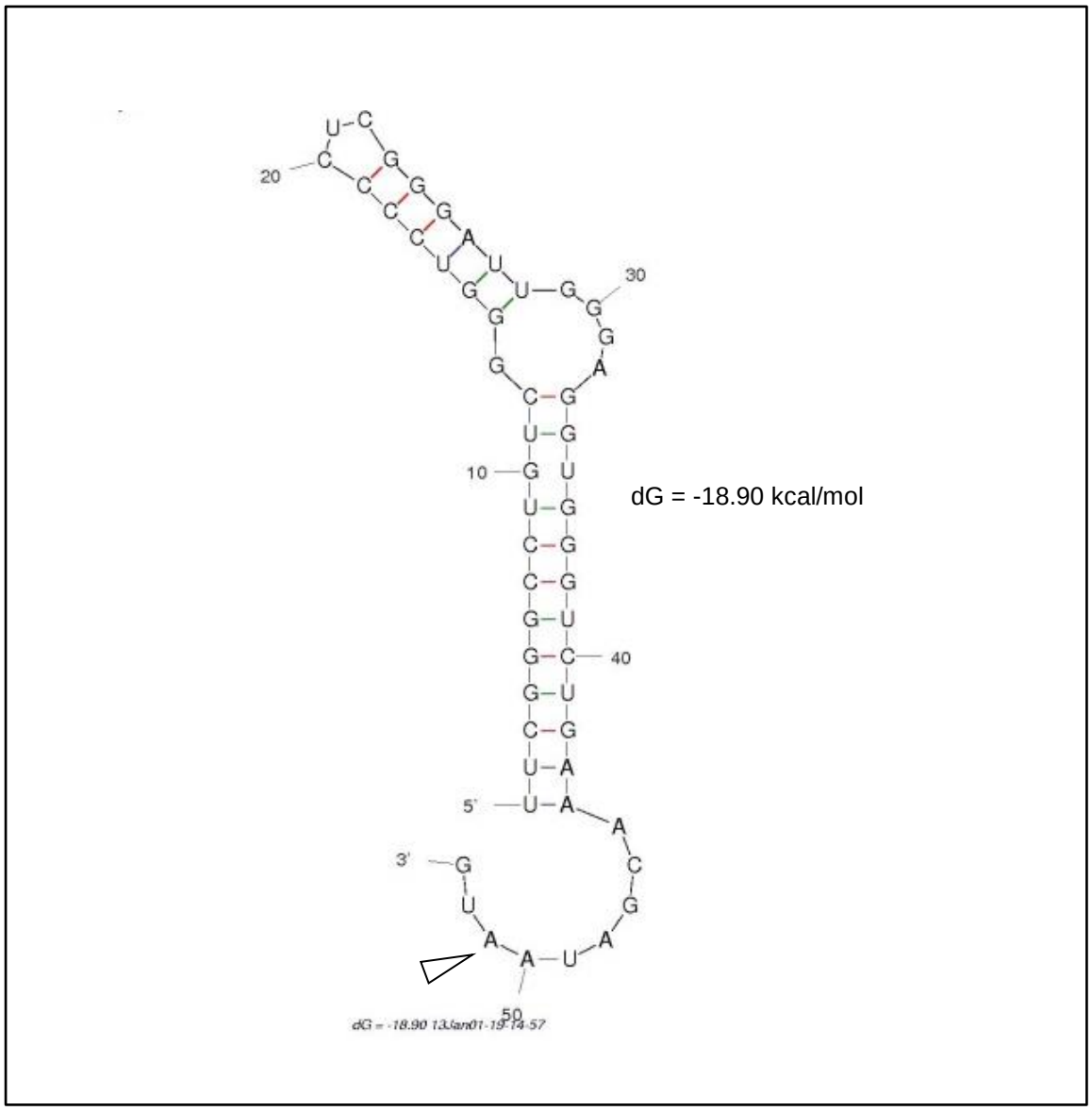


Figure 6.2 Secondary structure prediction for sORF IV. The first 50nt upstream from the sORF AUG initiation codon was folded using MFOLD. The sORF AUG initiation codon is noted with an arrow.

6.4.2 Optimisation of primer extension and ribosomal binding conditions

The toeprinting assay relies upon the successful binding of ribosomes to the mRNA transcript, and then reverse transcription from the bound radiolabelled primer to the edge of the ribosomal unit. To determine optimal conditions, two different ribosomal binding temperatures for the reticulocyte lysate were assayed, 25°C and 30°C. Simultaneously the optimal reverse transcription time was determined by allowing reverse transcription to proceed for either 10, 30 or 45 minutes. As depicted in Figure 6.3, the maximum signal was obtained when ribosomal complex formation was carried out at 25°C and the reverse transcriptase was allowed to proceed for no more than 30 minutes. Complete reverse transcript extension to the 5'-end of the mRNA template is stronger when the ribosomal complexes are allowed to form at 30°C. Of the reverse transcription times tested 30 minutes provided the strongest banding of all the toeprints observed. This is in agreement with results reported by Kozak (1998), where longer incubation times typically decrease yield of extension products, possibly due to degradation.

In order to further optimise the ribosomal binding conditions the effect of varying concentrations of cycloheximide and the effect of Mg^{2+} concentration in the ribosomal complex formation reactions was investigated. The region of the sORF transcript which encompasses sORFs I and II was chosen to test these conditions. Two concentrations of cycloheximide (1000 and 500µg/mL) were tested (Figure 6.4). This experiment showed that stalling ribosomes were held on the transcript more effectively with a higher concentration of cycloheximide, consistent with Kozak's initial experiments (Kozak, 1998).

The commercial rabbit reticulocyte lysate used in the toeprinting experiments is provided with an endogenous magnesium concentration of 0.5mM, however concentrations as high as 2.5mM may be required for optimal complex formation (Snyder and Edwards, 1991). The effects of no additional magnesium (0.5mM endogenous concentration) or an increase in magnesium concentrations to 1.9mM were tested. As depicted in Figure 6.4, the higher magnesium concentration dramatically reduced the toeprinting signal at sORFs I and II compared to

endogenous 0.5mM Mg^{2+}. Thus, in all subsequent toeprinting experiments magnesium concentration was maintained at 0.5mM. In summary the optimized toeprinting reaction parameters included ribosomal binding temperature of 25°C with primer extension time of 30 min. Cycloheximide concentration was optimized to 1000µg/mL with no additional magnesium included in the assay reaction.

Figure 6.3 Comparison of ribosomal binding temperatures and primer extension times in toeprinting reactions. In preparation for primer extension, the ribosome binding reactions were either carried out at 25°C or 30°C. Primer extension was terminated after 10, 30 or 45 min as indicated. A black (filled) arrow marks the full length product where extension to the 5' has occurred. The optimal time and temperature of 30min at 25°C was chosen as the banding intensity was maximal, without a loss of complete primer extension to the 5'-end. A negative control containing no RNA (-) shows no signal. Dideoxy cycle sequencing reactions carried out with the same toeprinting primer are depicted to the left.

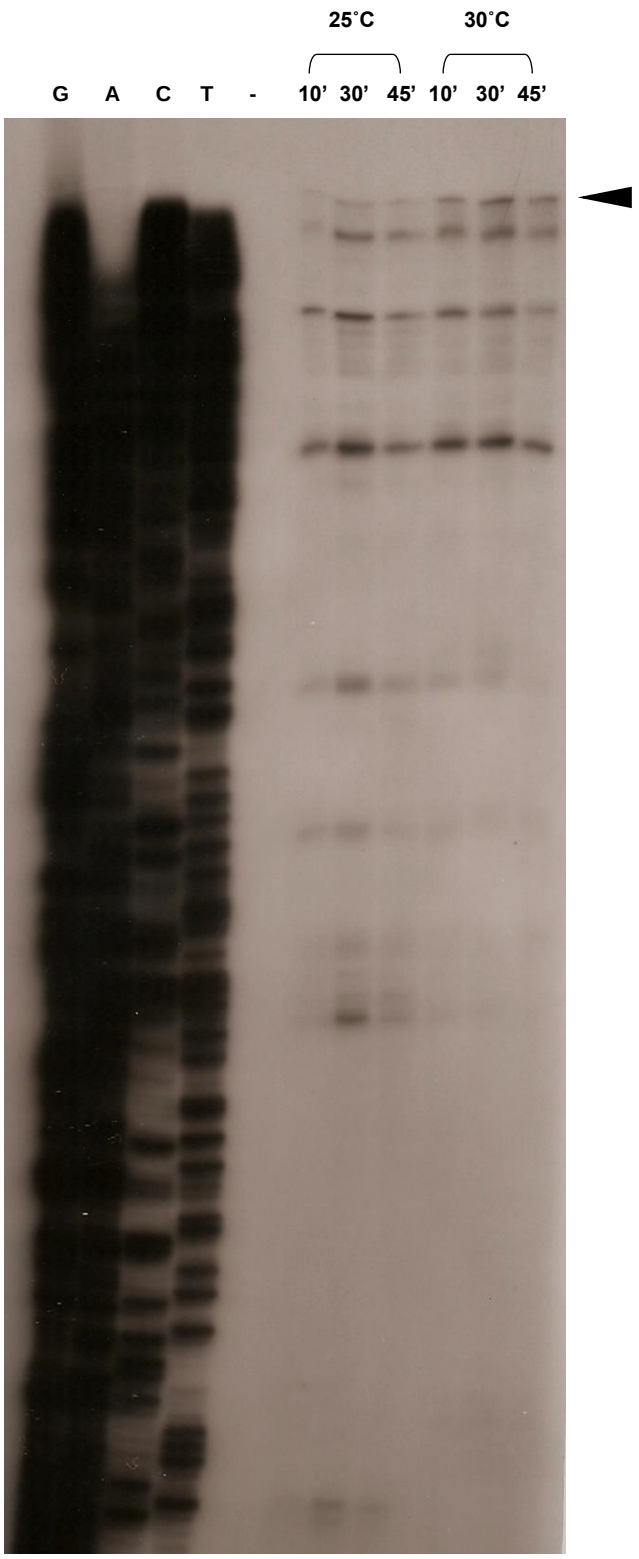
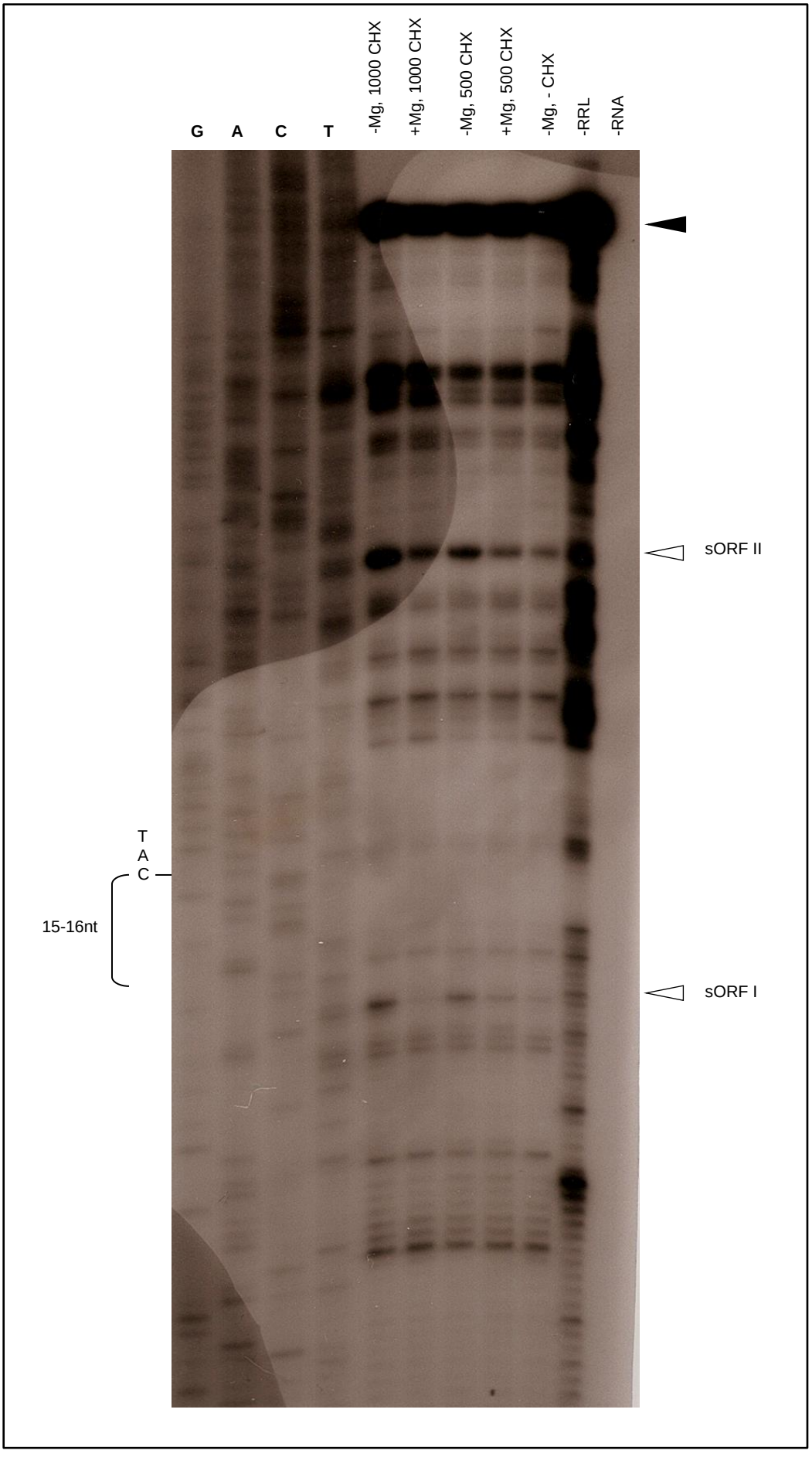

25˚C
30˚C
G A C T - 10' 30' 45' 10' 30' 45'

Figure 6.4 Optimisation of ribosomal binding conditions in toeprinting reactions. In preparation for primer extension, the ribosome binding reactions were carried out at 25°C in the presence of varinag concentrations of cycloheximide (CHX), 1000µg/mL or 500µg/mL, and Mg^{2+} (0.5mM or 1.9mM). A black (filled) arrow marks the full length product where extension to the 5' has occurred. Arrows (unfilled) mark the locations of both toeprints for sORF I and II with the AUG initiation codon for sORF I depicted 15-16nt downstream (denoted as the reverse complement CAT). Dideoxy cycle sequencing reactions carried out with the same toeprinting primer are depicted to the left. The optimal toeprints were observed in the absence of excess Mg^{2+} and 1000µg/mL CHX. A negative control containing no RNA (-) shows no toeprinting signal, while the negative control without lysate (-RRL) depicts maximal primer extension to the 5' end. A control containing no cycloheximide (-CHX) confirms ribosomal stalling at the sORF AUG initiation codon.


G
A
C
T
-Mg, 1000 CHX
+Mg, 1000 CHX
-Mg, 500 CHX
+Mg, 500 CHX
-Mg, - CHX
-RRL
-RNA
sORF II
sORF I
T
A
C
15-16nt

6.4.3 Toeprint analysis of ribosomal stalling at sORF I, II, III, IV, V, VI and VI_{alt} initiation codons

Once the toeprinting assay had been optimised, the possible locations of toeprints along the sORF transcript were examined. Because the length of the sequence exceeded 800bp, the toeprinting assay was broken down into five assays encompassing regions of 200-300bp of readable dideoxy cycle sequencing data. The toeprinting assay for sORFs I and II (Figure 6.5) showed that ribosomes stall at both AUG initiation codons upon the transcript. This is confirmed by the absence of the individual toeprints in the constructs in which the AUG initiation codons of each sORF have been mutated. Interestingly there is a stronger toeprint signal observed at sORF II than sORF I upon the same transcript. This suggests a preference for sORF II, consistent with the strong Kozak context of sORF II, while sORF I presents only an adequate context.

The toeprints for sORF III were investigated using a primer positioned in close proximity to the sORF AUG initiation codon. The toeprinting assay revealed that only a weak toeprint is observed at this sORF initiation codon. Two weak bands are seen (in both the presence and absence of cycloheximide) at the sORF AUG position, both of the same intensity, which are subsequently not visible once the AUG codon has been mutated (Figure 6.6).

The toeprinting assay was used to investigate the stalling of ribosomes around sORF IV initiation codon, a sORF with a strong Kozak context and in which a stem-loop structure with the potential to stall the ribosome was predicted (Figure 6.2). As shown in Figure 6.7, the toeprinting assay revealed a strong toeprint signal for sORF IV. Given that this sORF has a strong Kozak context it is expected that ribosomes would initiate at this AUG. This is confirmed by the absence of any signal at the AUG location in the assays where the initiator AUG codon has been mutated (Figure 6.7, lane denoted NIL).

Figure 6.5 Toeprinting analysis of initiating AUG codons of sORFs I and II. A black (filled) arrow marks the full length product where extension to the 5' has occurred. Unfilled arrows mark the locations of both positive toeprints for sORF I and II with the AUG initiation codon for sORFs I and II depicted 15-16nt downstream (denoted as the reverse complement CAT). Dideoxy cycle sequencing reactions carried out with the same toeprinting primer are depicted to the left. The sORF initiation codons mutated in each respective experiment are shown, I and II wild-type (I, II Wt) with toeprints at both sORFs I and II, sORF I mutated (II Wt) with toeprint only at sORF II and both sORFs I and II mutated (NIL) where no toeprints are observed. A negative control containing no RNA (-) depicts no toeprint signal, while the negative control without lysate (-RRL) depicts maximal primer extension to the 5' end. Controls containing no cycloheximide (-) confirm toeprint locations against samples with cyclohexmide (+).

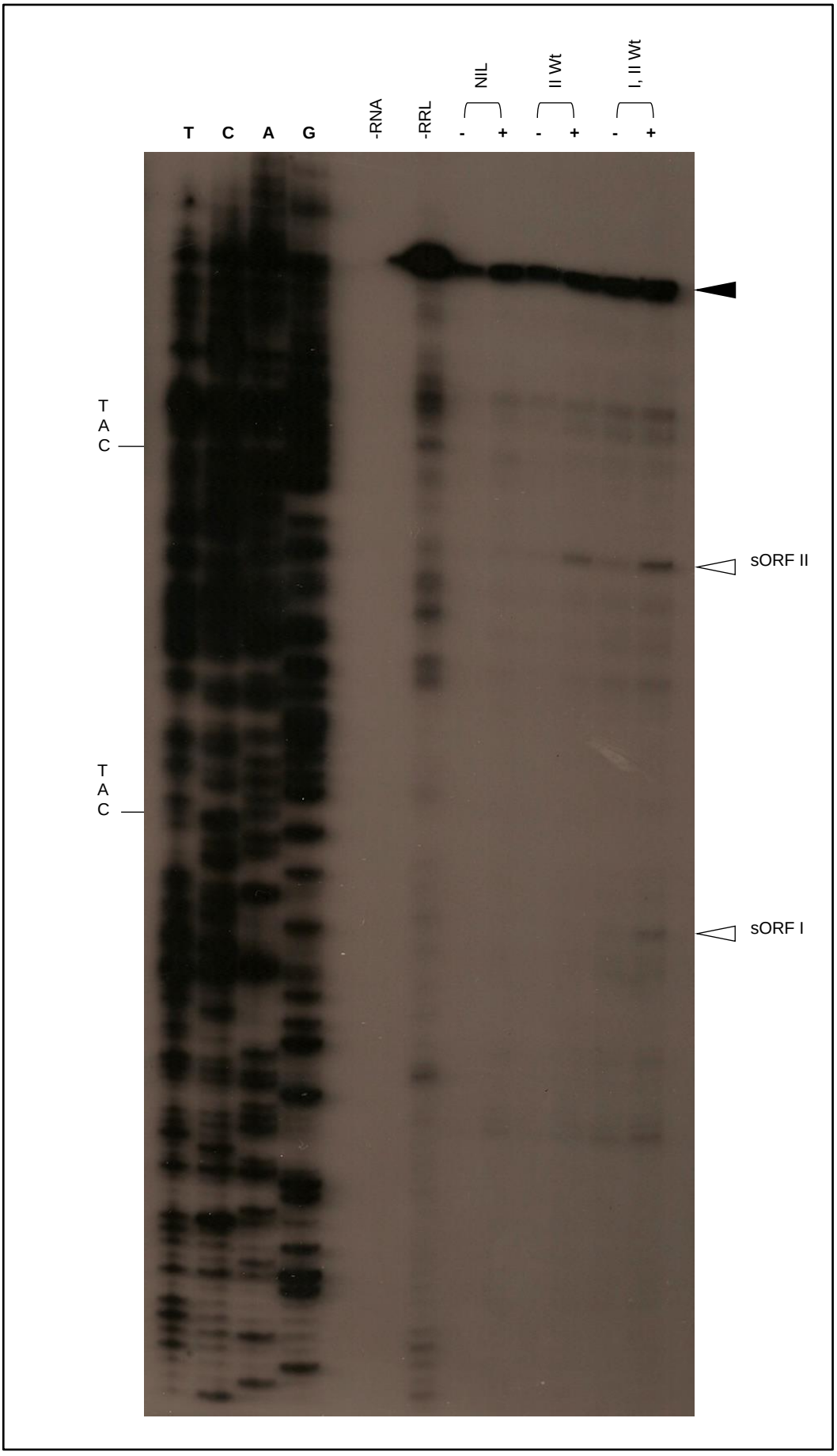

T C A G
-RNA
-RRL
NIL
II Wt
I, II Wt
- + - + - +
T
A
C
T
A
C
sORF II
sORF I

Figure 6.6 Toeprinting analysis of initiating AUG of sORF III. A black (filled) arrow marks the full length product where extension to the 5' has occurred. An unfilled arrow marks the location of a weak toeprint for sORF III with the AUG initiation codon depicted 15-16nt downstream (denoted as the reverse complement CAT). Dideoxy cycle sequencing reactions carried out with the same toeprinting primer are depicted to the left. The sORF AUG codons mutated in each respective experiment are shown, III wild-type (III Wt) with no toeprint observed and sORFs III mutated (NIL) also no toeprint observed. A negative control containing no RNA (-) depicts no toeprint signal, while the negative control without lysate (-RRL) depicts maximal primer extension to the 5' end. Controls containing no cycloheximide (-) confirm toeprint locations against samples with cyclohexmide (+).

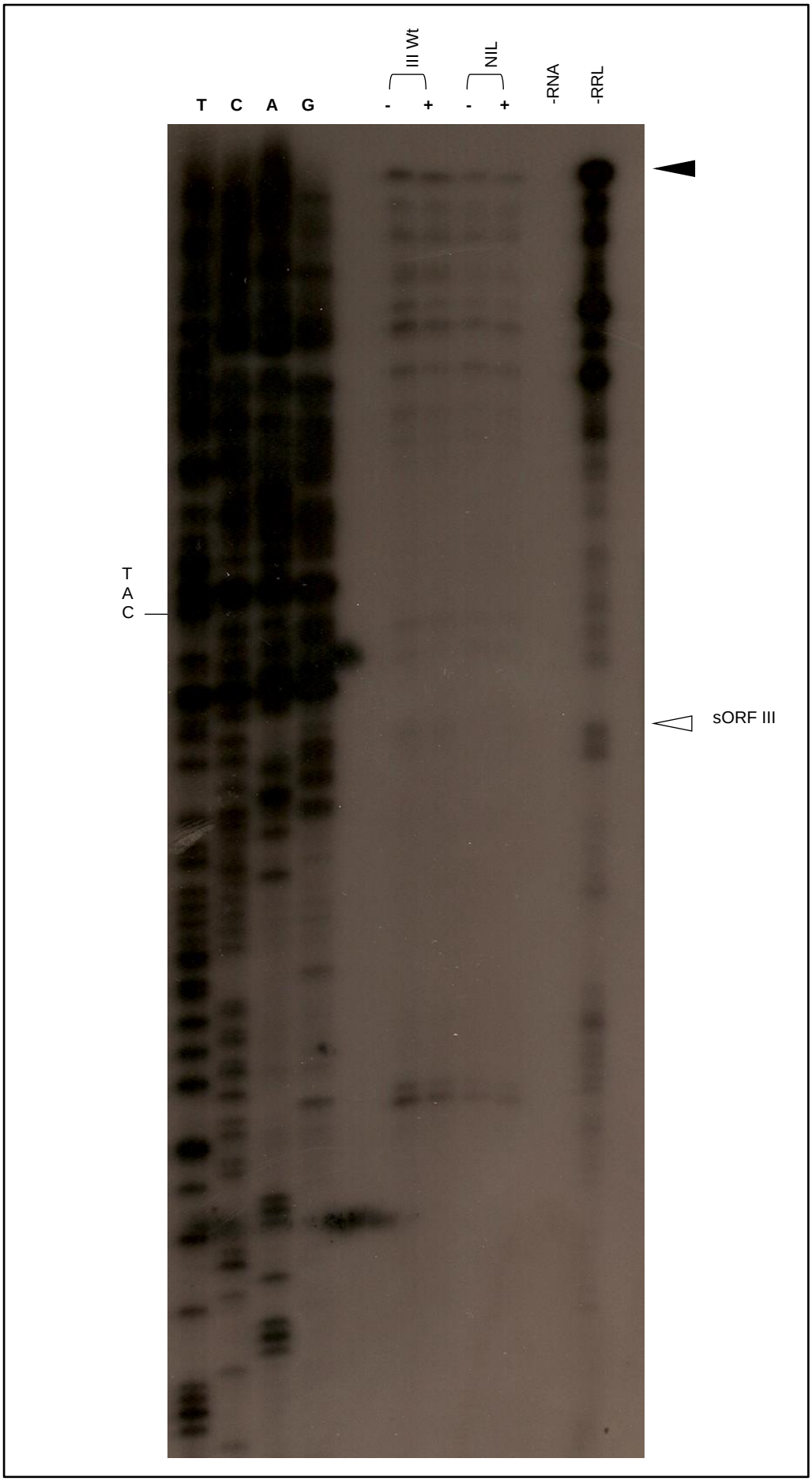
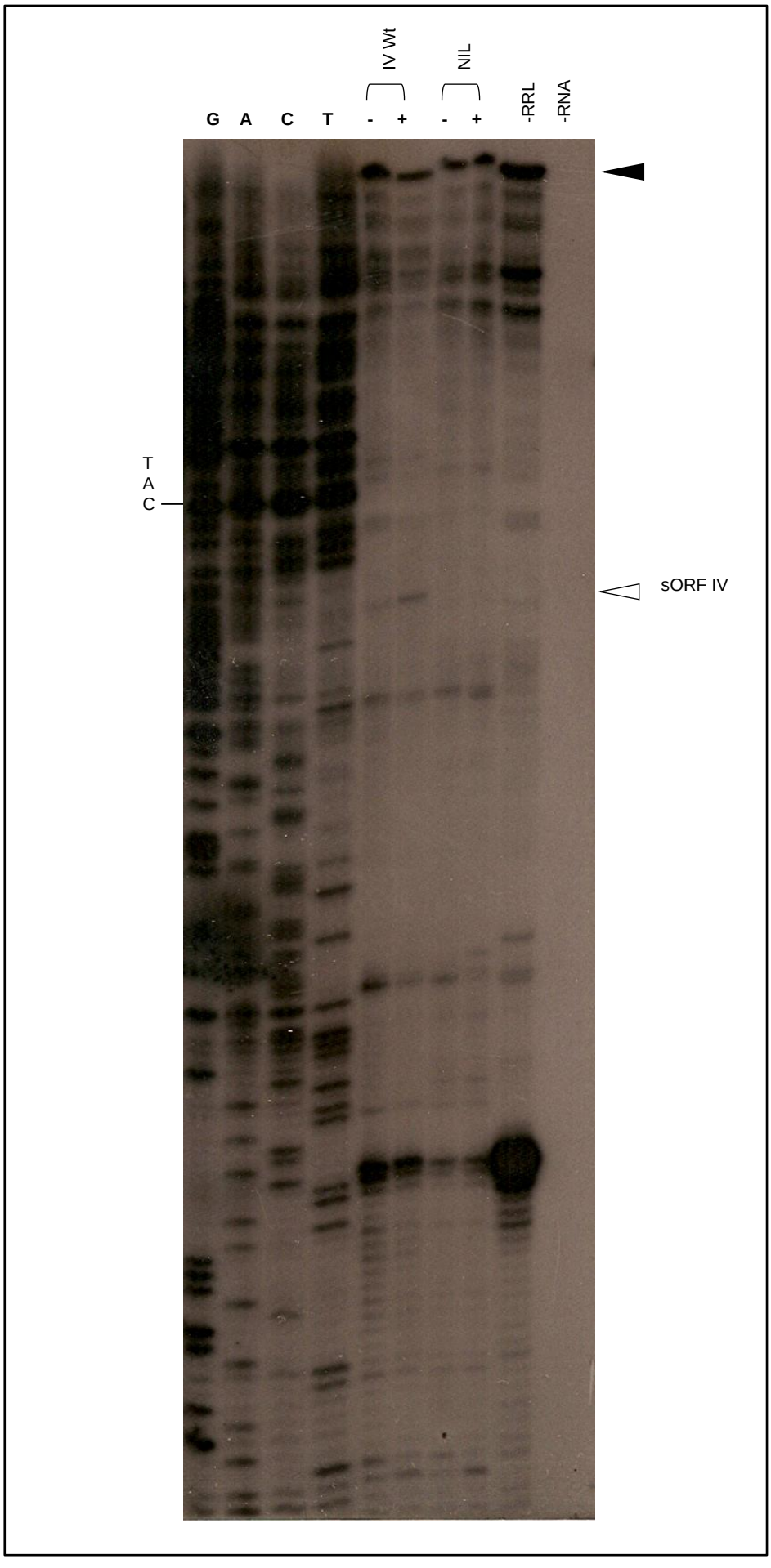

III Wt
NIL
T C A G
- + - +
-RNA
-RRL
T
A
C
sORF III

Figure 6.7 Toeprinting analysis of initiating AUG of sORF IV. A black (filled) arrow marks the full length product where extension to the 5' has occurred. An unfilled arrow marks the location of positive toeprint for sORF IV with the AUG initiation codon depicted 15-16nt downstream (denoted as the reverse complement CAT). Dideoxy cycle sequencing reactions carried out with the same toeprinting primer are depicted to the left. The sORF AUG codons mutated in each respective experiment are shown, IV wild-type (IV Wt) with positive toeprint observed and sORF IV mutated (NIL) with no toeprint observed. A negative control containing no RNA (-) depicts no toeprint signal, while the negative control without lysate (-RRL) depicts maximal primer extension to the 5' end. Controls containing no cycloheximide (-) confirm toeprint locations against samples with cyclohexmide (+).


IV Wt
NIL
G A C T - + - + -RRL -RNA
T
A
C
sORF IV

Similar to sORF III, the assay of sORF V revealed a weak toeprint at this AUG initiation codon. Two weak bands of the same intensity are observed in the presence and absence of cycloheximide at the sORF AUG position, which are subsequently not visible once the AUG codon had been mutated (Figure 6.8). This result may again be related to several factors such as the distance between the previous sORF (IV) and the initiator AUG codon of sORF V (65 bp), the Kozak consensus surrounding the initiator AUG codon (adequate) and the length of the sORF.

Lastly, a toeprinting assay for sORF VI covered a region spaning sORF VI and VI_{alt}. A toeprint was obtained at the AUG inititation codon of sORF VI, and a second toeprint, was also observed at the potential AUG initiation codon of sORF VI_{alt}, the alternative AUG codon located (in-frame) 69 bp further downstream (Figure 6.9). The AUG codon of sORF VI_{alt} is in a weak sequence context in comparison to the adequate sequence context to sORF VI, which does not correlate with the strength of the toeprints observed in this experiment (Table 6.4). The toeprints were confirmed by the absence of any signal at the AUG location in the assays where the initiator AUG codons have been mutated (NIL).

The toeprint assay for the sORF VI region also indicated a toeprint at a site located 12 bp downstream from the initiation codon of sORF VI. This particular location would suggest the potential to initiate at a CUG codon (adequate sequence context), a non-AUG triplet considered a highly recognised triplet in reticulocyte translation experiments (Peabody, 1989; Kozak, 1997). As summarised in Table 6.3, this cryptic CUG codon is conserved in only 9 of the 31 sequences examined in Chapter 3. Of these sequences, 3 (all from subtype B) display the CUG codon 12 bp downstream from the initiation codon of sORF VI, while four sequences present a CUG codon further downstream at 18 bp. In all but 2 sequences (VI850 and YBF100) the CUG codon has an adequate sequence context.

Figure 6.8 Toeprinting analysis of initiating AUG of sORF V. A black (filled) arrow marks the full length product where extension to the 5' has occurred. An unfilled arrow marks the location of a weak toeprint for sORF V with the AUG initiation codon depicted 15-16nt downstream (denoted as the reverse complement CAT). Dideoxy cycle sequencing reactions carried out with the same toeprinting primer are depicted to the left. The sORF AUG codons mutated in each respective experiment are shown, sORF V wild-type (V Wt) with no toeprint observed and sORF V mutated (NIL) with a complete loss of the toeprint observed. A negative control containing no RNA (-) depicts no toeprint signal, while the negative control without lysate (-RRL) depicts maximal primer extension to the 5' end. Controls containing no cycloheximide (-) confirm toeprint locations against samples with cyclohexmide (+).

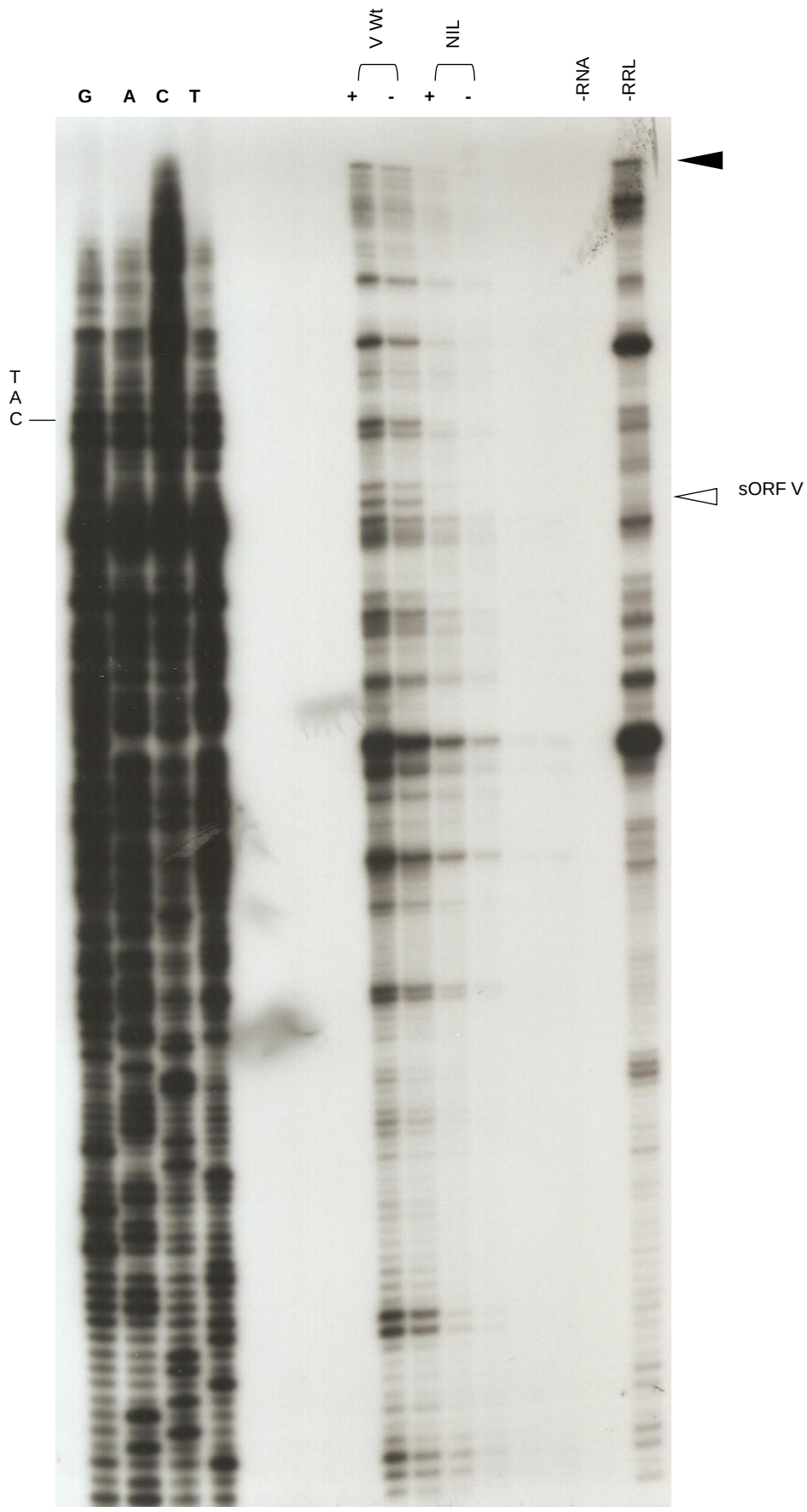

G A C T
V Wt
NIL
+ - + -
-RNA
-RRL
T
A
C
sORF V

Figure 6.9 Toeprinting analysis of initiating AUGs of sORFs VI and VI_{alt}. A black (filled) arrow marks the full length product where extension to the 5' has occurred. Unfilled arrows mark the locations of positive toeprints for sORFs VI and VI_{alt} with the AUG initiation codons depicted 15-16nt downstream (denoted as the reverse complement CAT) and a toeprint occurring at a cryptic CUG codon 12nt downstream from the initiation codon of sORF VI noted. Dideoxy cycle sequencing reactions carried out with the same toeprinting primer are depicted to the left. The sORF AUG codons mutated in each respective experiment are shown, VI and VI_{alt} wild-type (VI Wt) with toeprints observed at both AUGs, and sORFs VI and VI_{alt} mutated (NIL) with a loss of toeprint observed. A negative control containing no RNA (-) depicts no toeprint signal, while the negative control without lysate (-RRL) depicts maximal primer extension to the 5' end. Controls containing no cycloheximide (-) confirm toeprint locations against samples with cyclohexmide (+).

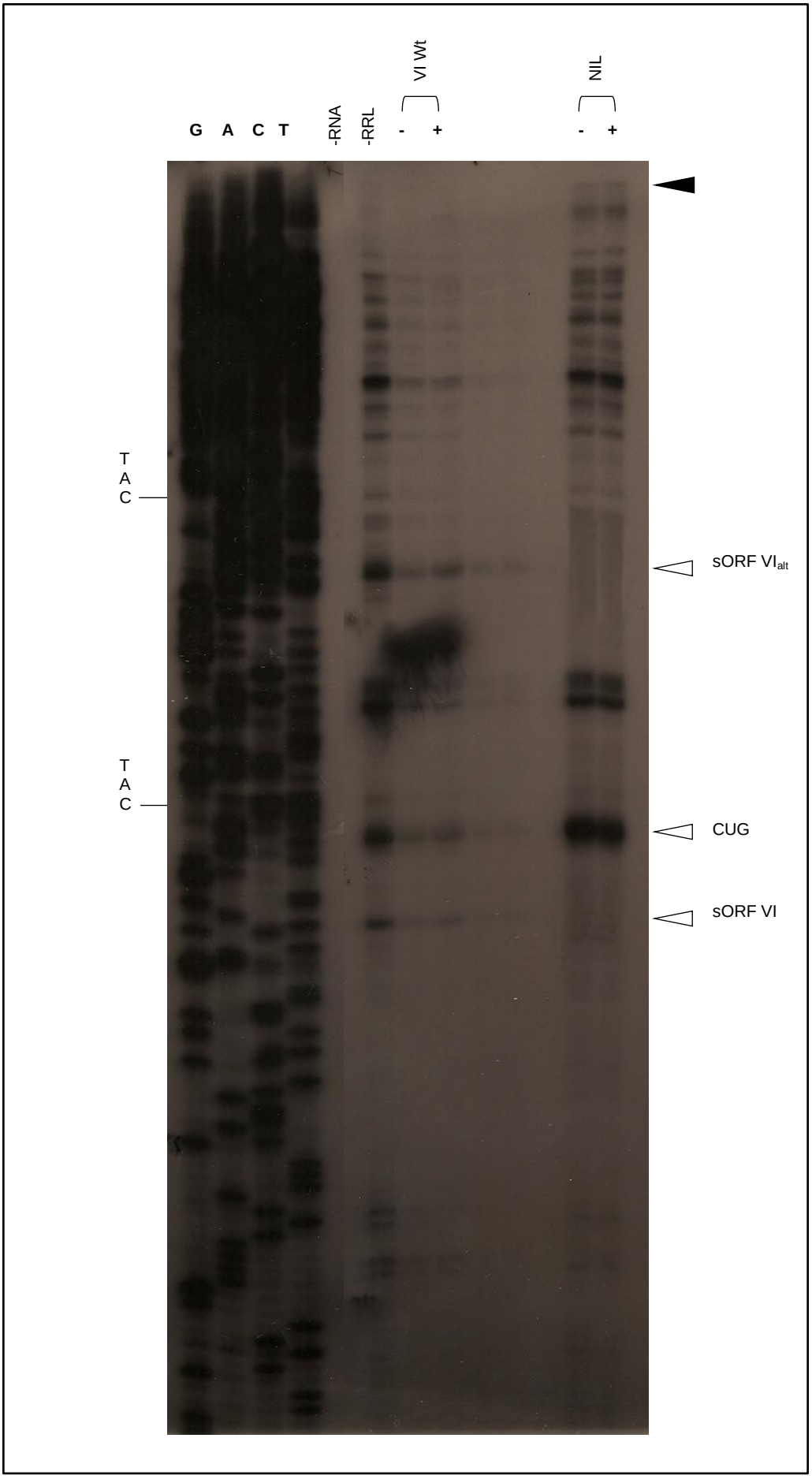

G A C T
-RNA
-RRL
VI Wt
- +
NIL
- +
T
A
C
T
A
C
sORF VI$_{alt}$
CUG
sORF VI

Table 6.3 HIV-1 conservation of cryptic CUG codon located downstream of the sORF VI AUG initiation codon. Locations indicate distances (bp) downstream from the sORF VI initiation codon.

Subtype	Strain	Kozak Context	Location (bp)
A	SE6594	Adequate	18
B	NL43	Adequate	12
B	BRU	Adequate	12
B	ACH1	Adequate	12
D	NDK	Adequate	3
F	FIN9363	Adequate	18
F	VI850	Weak	18
F	CM53657	Adequate	18
Group N	YBF100	Weak	29

The toeprinting analysis of the entire sORF region provides an overall picture of where the ribosomes initiate along the transcript, as summarised in Table 6.4. Ribosomes were able to assemble and initiate at sORFs I, II, IV, VI and VI_{alt} but not at sORFs III and V. Interestingly, the two sORF AUG initiation codons at which no toeprints were observed (sORF III and V) both have an adequate sequence context, with a much stronger AUG initiation codon presented by the preceding sORF. These prior sORFs (II and IV) are also positioned at lengthy distances from the subsequent sORF initiation codon (172 bp between sORF II termination and sORF III initiation codons; 65 bp between sORF IV termination and sORF V initiation codons), which may affect the ability of the ribosome to recognise the downstream AUG codon during scanning from the previous sORF (Figure 6.9).

Table 6.4 Summary of toeprinting assay data with Kozak context of each sORF. Sequences surrounding sORF initiation codons (underlined) with sequences key to strength in Kozak context depicted in bold. Toeprints marked (#) depict stronger toeprint signals. Stronger toeprint intensities are observed at sORFs II and VI_{alt} compared to sORFs I and VI respectively within the same toeprinting experiment.

sORF	Reading Frame	Sequence	Kozak Context	Toeprint
I	-1	CTA**G**GTATGT	Adequate	Yes
II	-1	ACT**G**CTATG**G**	Strong	Yes#
III	-1	AATCGAATG**G**	Adequate	Weak
IV	-3	ACG**A**TAATG**G**	Strong	Yes
V	-1	ATC**A**TTATGA	Adequate	Weak
VI	-1	CAG**G**TCATGT	Adequate	Yes
VI_{alt}	-1	GTGCAAATGA	Weak	Yes#

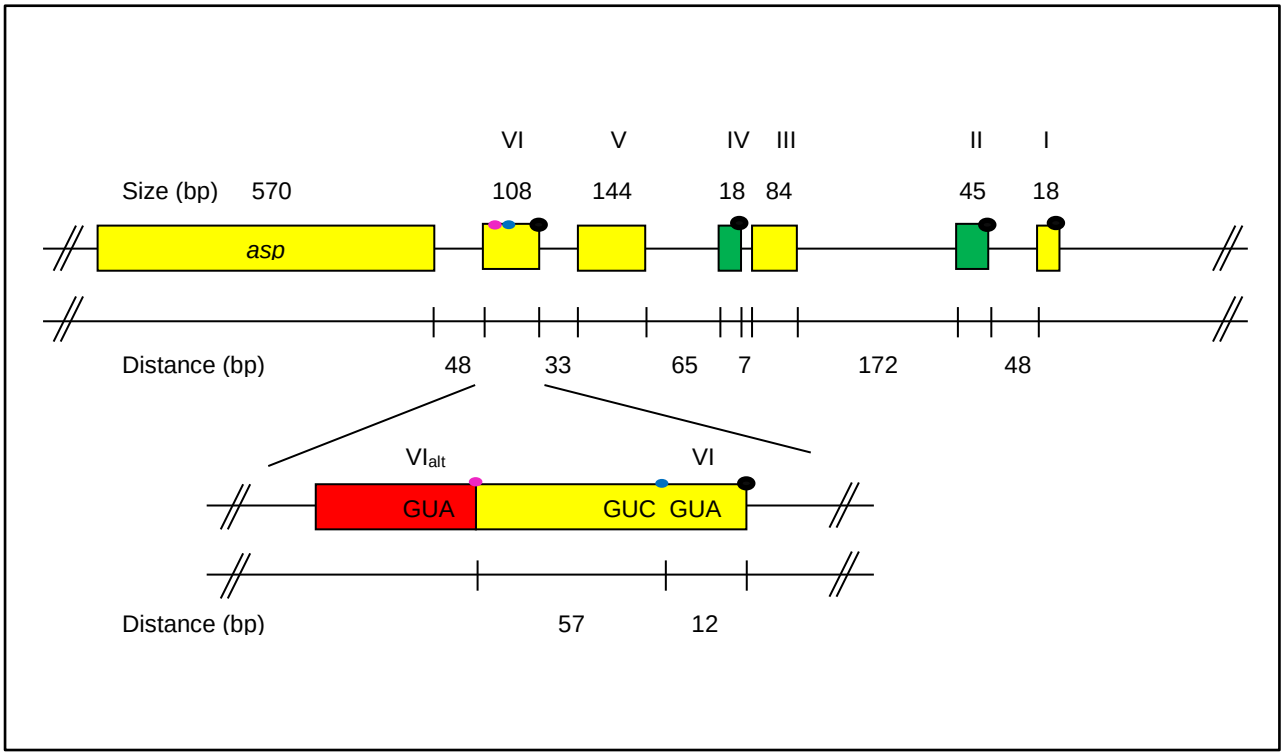


Figure 6.10 Summary of sequence features and translational characteristics of the sORF region of HIV-1NL4-3 used in the toeprinting assays. The HIV-1 *asp* sORFs are indicated in the negative sense orientation with their sizes (bp) indicated above. Colour of sORFs indicates strength of initiation context, green: strong Kozak context, yellow: adequate Kozak context and red: weak Kozak context. Black circles represent AUG codons at which ribosomal initiation was detected, small pink circle denotes initiation of translation at VI_{alt} and small blue circle represents initiation of translation at cryptic CUG codon within sORF VI coding sequence. Distances between termination and initiation codons of sORFs are indicated below (bp).

6.5 Discussion

A series of MFOLD predictions showed that sORFs I, II, III, V, VI and VI_{alt} did not present any significantly stable stem-loop structures ($\Delta G < -9.80$) that could inhibit or slow the translating ribosomal unit. MFOLD analyses did reveal one large stem-loop structure positioned 7 nt upstream from the AUG codon of sORF IV that presented a Gibbs Free Energy of -18.90 kcal/mol. While this energy is considered by most as not strong enough to impair translation, its strength is moderate (Kozak, 1990; Kozak, 1989a) and may slow the translating ribosome enough to allow recognition and initiation. Combined with the proximity of the stem-loop to the AUG codon, this structure may have an effect on slowing the scanning 40S ribosomal subunit and allowing it to initiate at this strong AUG initiation codon, a location at which toeprinting assays show that ribosomes indeed stall. Toeprinting assays established the locations of ribosomal arrest during the initiation phase of translation *in vitro*. These data, for the first time, prove that ribosomal complexes formed and initiated at sORFs I, II, IV, VI and VI_{alt} (as summarised in Table 6.4). These data are therefore consistent with the leaky scanning and the termination reinitiation mechanisms of translational regulation.

The translational model of leaky scanning predicts that the 40S ribosomal subunit attaches to the m^7G cap and begins scanning the mRNA transcript. Once the 40S subunit reaches a suitable AUG codon the complete ribosomal complex is formed and translational initialises. If, however, the 40S ribosomal unit does not recognise the AUG codon (due to either weak Kozak context or secondary structures blocking access) the 40S unit may pass this AUG codon and continue scanning until it reaches a downstream AUG codon (Kozak, 2005). Reinitiation at a downstream AUG codon is primarily dependent upon the distance between the end of the previous sORF and the new downstream AUG codon (Kozak, 1987c). In this case a distance of less than 79 nucleotides is considered too short, as the scanning ribosome is less likely to have recruited the necessary initiation factors required to initiate at the next AUG codon (Poyry *et al*., 2004). Similarly, reinitiation is less likely when the upstream sORF is greater than 35 codons long (Rajkowitsch *et al*., 2004). These models both account for the results here. In this model the first two sORFs of

the *asp* transcript are positioned far upstream and thus may be involved in ribosomal recruitment for downstream translation (Figure 6.9).

The experiments presented here did not show toeprints for both sORFs III and V, suggesting that the ribosomes miss these AUG codons and do not initiate at these two sORFs or that initiation is an infrequent event. Nevertheless a number of factors may explain why these sORFs are missed. In the case of sORF III the distance between sORF II and III is lengthy at 172 bp (Figure 6.10), and would enable the scanning ribosome to reinitiate as this distance is greater than the 79 nucleotides required. However, the data presented here imply that this does not occur and suggests that the ribosome may either dissociate from the transcript after termination at sORF II or may miss sORF III by other means such as a ribosomal shunt, this remains unclear. While the MFOLD analysis presented here does not support a shunting model, structures suggestive of a shunting mechanism may be present in other regions not analysed here. In contrast, the inability of ribosomes to initiate at sORF V can be explained by the short distance (65 bp) between the termination of sORF IV and the AUG codon of sORF V, which may be insufficient to allow the scanning 40S ribosome to gather the necessary initiation factors to allow it to recognise and begin translation at sORF V (adequate context). Interestingly, if the ribosome were to initiate at sORF V (the longest of all the sORFs at 144 bp) this could prevent subsequent initiation at sORF VI due to the length of the preceding sORF (greater than 35 codons) and the very short (33 bp) distance between the sORFs. This could favour initiation at VI_{alt} located 102 bp from the termination codon of sORF V.

The role of the termination codon in translation experiments must not be overlooked. Should a ribosome form a block at the termination codon a signal should be observed in assay samples incubated in the absence of cycloheximide with a loss of signal from the initiation codon (Sachs *et al*., 2002). In this instance, should a ribosome stall at the termination codon, the scanning ribosomes are prevented from reaching the downstream AUG codons (Cao and Geballe, 1996). No ribosomal arrest at termination codons was observed in any of the experiments conducted.

Finally, the effect of mutation of sORF AUG initiation codons as investigated in Chapter 5 (section 5.4.5) may now be more clearly explained. Mutation of sORFs I, II and III had little, if any effect on downstream expression and this may be in part due to their lengthy distance from the main ORF (EGFP, in the expression experiments in Chapter 5). The mutation of sORF IV decreased expression slightly while the largest effect was observed with the mutation of sORF VI. This can now be explained by the ability of the ribosome to initiate at either the initiation codon of sORF VI or VI_{alt} which in turn inhibits the ability of the scanning ribosomal complex to initiate at the main downstream ORF. These data, together, imply that sORF VI (or sORF VI_{alt}) inhibits downstream translation however this inhibition is predominant when sORFs I, II, and IV (where ribosomes are observed to initate translation) engage the ribosome. Importantly this reflects only 11.4% of the transcript pool (Unspliced and Spliced Variant 2). The remaining transcripts comprising Spliced Variant 1 (53%) and Spliced Variant 3 (35.6%) do not contain any of sORFs I, II or IV, contain either sORF VI (and VI_{alt}) or VI_{alt} alone. It is clear that sORF VI can negatively affect translation further downstream probably by bypassing the in-frame initiation codon of the downstream *asp* ORF or by other processes such as stalling or complete dissociation from the transcript. This mechanism could be similar to the four sORFs of the *GCN4* mRNA, which inhibit *GCN4* expression in a termination-reinitiation manner (Mueller and Hinnebusch, 1986; Hinnebusch, 1997).

6.6 Conclusions

The data presented in this chapter provide evidence that sORFs I and II may allow the scanning ribosome to initiate and then continue scanning, bypassing sORF III, reinitiating at sORF IV, in strong Kozak context, bypassing sORF V (due to the short distance between termination and reinitiation) to then reinitiate at sORF VI, VI_{alt} or the cryptic CUG codon. This may allow leaky-scanning of the *asp* ORF. These experiments clearly demonstrate for the first time that sORFs I, II, IV, VI and VI_{alt} are able to initiate translation by the scanning 40S ribosome complex, allowing the transcript to be read in a leaky-scanning and termination reinitiation mode. Whether the possible secondary structure identified here for sORF IV, or another mechanism such as internal ribosome entry, plays a role in the translation of the transcript remains unknown.

CHAPTER 7 – GENERAL DISCUSSION

7.1 Summary of findings

This study suggests that a series of sORFs may act to control the negative sense *asp* gene, by a post-transcriptional mechanism. Work undertaken specifically examined (1) the conservation of the sORF series upstream of the *asp* gene, (2) the effect of the sORF region on downstream gene expression in a reporter construct system, (3) the potential for splicing to modulate levels of gene expression and (4) the role of each active sORF initiation codon on translational initiation

This work confirmed and extended *in silico* analysis of *asp*/ASP to show that it is highly conserved across the HIV-1 strains. This analysis also described the presence and high level conservation of a series of sORFs upstream of the *asp* ORF across all subtypes of HIV-1 (Chapter 3). In particular a multiple sequence alignment of the region upstream of the *asp* ORF (nts 8742 to 7932 in HIV-1 NL4-3) in 31 HIV-1 sequences, revealed a nucleotide sequence conservation >80% with reference to the consensus sequence. This analysis also revealed that the major infecting HIV-1 subtypes (B, C and D) display 5-7 sORFs (typically 6 sORFs), while fewer sORFs (3-5) are observed in the rarer subtypes (H, K and N). The translation initiation context of each sORF was typically adequate, this included the sequences of sORFs I (26/31 sequences), V (21/31) and VI (25/31); consistent with other sORF regulating systems, such as that of yeast *GCN4* (Hinnebusch, 1997) and Protein kinase C (Raveh-Amit *et al*., 2009), suggesting that these sORFs have a role in translational regulation of *asp* gene expression. The findings presented here also suggest gene expression has the potential to be modulated by the use of rare codons, particularly in sORF I where their abundance is high (80%).

The sORF region (cloned from HIV-1 NL4-3) consistently downregulated gene expression in the EGFP reporter construct system. These data were supported by experiments involving sequential mutation of the sORF initiation codons in which, to varying levels, each mutation alleviated the inhibition initially observed. The differences observed in the levels of reporter gene expression were not a

consequence of experimental variation in transfection efficiency, or a *trans* affect. These differences were also not a consequence of varying transcriptional activity, as EGFP transcripts levels were consistent across the experiments, supporting a translational mechanism for control of *asp* expression.

Further characterisation of the sORF region identified four alternative transcripts resulting from various splicing signals. The most abundant product, Spliced Variant 1, contained sORF VI alone and is produced by splicing using the highly conserved splice donor and acceptor sites. Spliced Variant 2, containing sORFs I, II and VI, utilised the splice acceptor common to Variant 1 and a less well conserved splice donor. Cloning of Spliced Variant 2 enabled the detection of a third product, Spliced Variant 3, presenting the alternate AUG initiation codon VI_{alt}. RT-PCR analysis demonstrated that SV2 makes up less than 1% of the transcript pool, while SV3 comprises approximately one third of the pool, indicating that sub-splicing of Spliced Variant 2 is a frequent event under these experimental conditions. While Spliced Variants 1 and 2 inhibited downstream expression to varying extents, Spliced Variant 3 permitted expression of the main ORF.

Toeprinting analysis of the sORF region revealed the potential for ribosomes to initiate at sORFs I, II, IV, VI and VI_{alt}, with only weak toeprints observed for sORF III and sORF V, suggesting that the leaky scanning and/or termination reinitiation mechanisms of translation account for the mode of translation across the sORF transcript (a summary of these findings is presented in Figure 7.1). Together, these findings suggest a complex mechanism comprised of splicing and translational control may regulate *asp* gene expression in the HIV-1 infection cycle.

7.1.1 A potential mechanism for regulation of *asp* expression by upstream sORFs

In the unspliced transcript:

> sORFs I and II engage the scanning ribosome, which then continues scanning, bypassing sORF III, reinitiating at sORF IV, in strong Kozak context, bypassing sORF V (due to the short distance between termination and reinitiation) and reinitiating at sORF VI. Ribosomes that reinitiate at sORF

VI do not reinitiate at the main ORF initiation due to the short distance between the termination codon of sORF VI and the main ORF (48 bases). Leaky scanning permits some ribosomes to bypass sORF VI and reinitiate at the main ORF. Hence, the main ORF is expressed, but at low levels.

In Spliced Variant 1:

A single sORF, sORF VI, is present upstream of the main ORF. Leaky scanning permits some ribosomes to initiate at sORF VI (unlikely to reinitiate at the main ORF); however other ribosomes scan through sORF VI and proceed to initiate at the main ORF. Expression of the main ORF is moderately reduced.

In Spliced Variant 2:

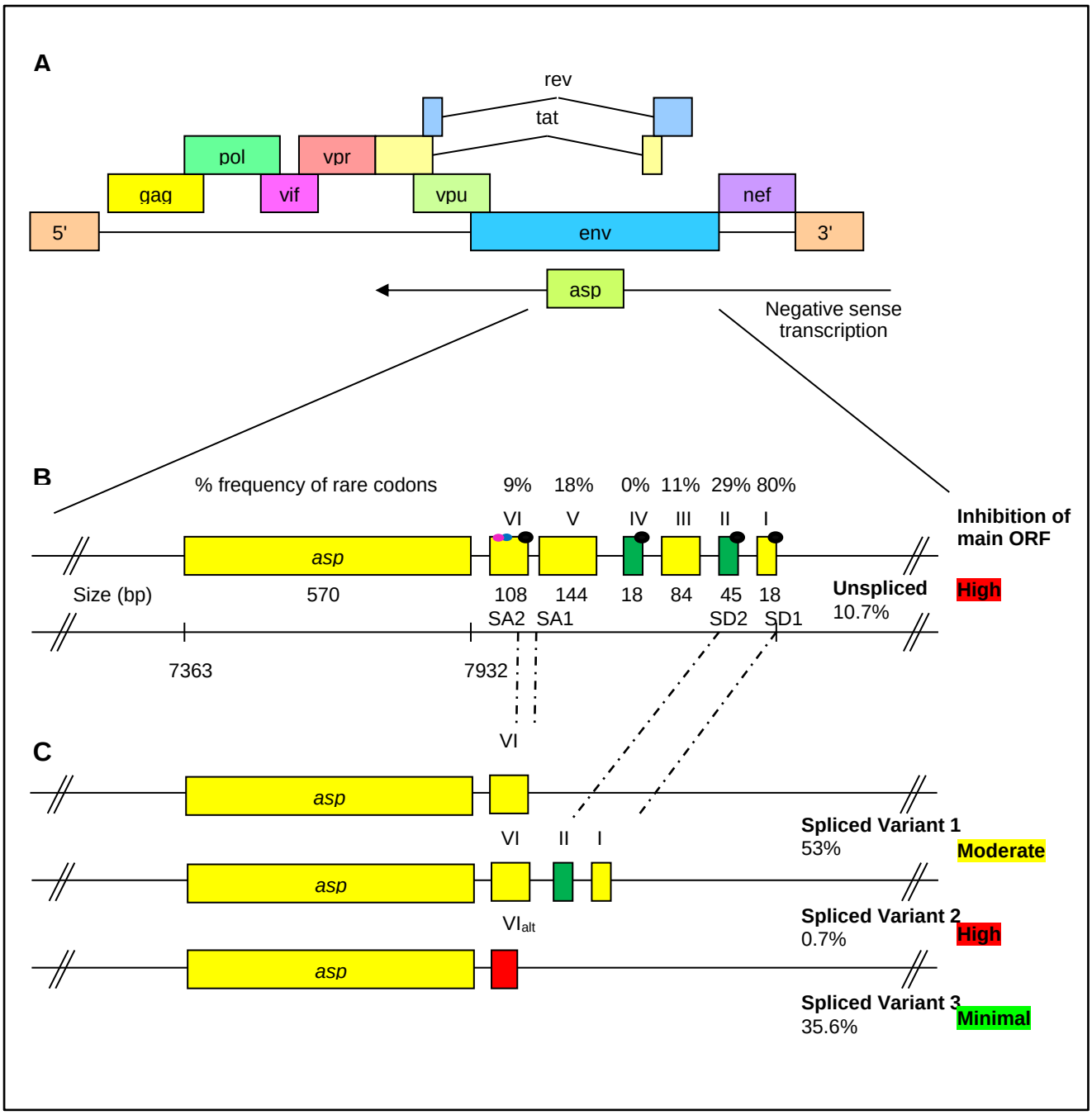
sORFs I and II engage the scanning ribosome, which is able to reinitiate at sORF VI, as distance between the termination codon of sORF II and the VI initiation codon is sufficient (90 bases). The ribosome fails to reinitiate at the main ORF so expression of the downstream ORF is strongly inhibited.

In Spliced Variant 3:

A single sORF, sORF $VI_{alt,}$ is present upstream of the main ORF. The VI_{alt} initiation codon has a weak Kozak context, so few ribosomes initiate at VI_{alt}; most ribosomes scan through to initiate and express the main ORF.

Therefore, expression of the main ORF can be finely modulated by control of splicing events, to provide varying proportions of the Spliced Variants in the transcript pool.

Figure 7.1 Summary of the characteristics of the sORF region, its associated spliced products and translational events in HIV-1 NL4-3. (A) The HIV-1 genome with the negative sense gene, *asp,* and (B) its associated upstream sORFs, indicated in the negative sense orientation with nucleotide position numbers indicated below from NL4-3. Splicing donor and acceptor sites located underneath, with the % of low frequency codons for each sORF depicted above. Shading within each sORF represents the strength in initiation codon context: green, represents strong Kozak context, yellow, represents adequate Kozak context, and red, represents weak Kozak context (C) Four alternative transcripts were detected for the sORF region; unspliced, Spliced Variant 1 (containing sORF VI only), Spliced Variant 2 (containing sORFs I, II and VI), and Spliced Variant 3 (containing part of sORF VI, presenting the alternate AUG initiation codon, VI_{alt}). Percentage abundance of each spliced product indicated below and the level of inhibition of the main ORF depicted. Black circles represent AUG codons at which ribosomal initiation was detected, small pink circle denotes initiation of translation at VI_{alt} and small blue circle represents initiation of translation at a cryptic CUG codon within the sORF VI coding sequence.


A
rev
tat
pol
vpr
gag
vif
vpu
nef
5'
env
3'
asp
Negative sense transcription
B
% frequency of rare codons
9% 18% 0% 11% 29% 80%
VI V IV III II I
asp
Inhibition of main ORF
Size (bp)
570
108 144 18 84 45 18
Unspliced
10.7%
High
SA2 SA1 SD2 SD1
7363
7932
VI
C
asp
Spliced Variant 1
53%
Moderate
VI II I
asp
Spliced Variant 2
0.7%
High
VIalt
asp
Spliced Variant 3
35.6%
Minimal

7.2 General discussion

This thesis has characterised the splicing events within the sORF region and examined the potential role such events play on gene expression. These data indicated that such events have potential to modulate the level of downstream gene expression with fine control. This is evidenced by the distinct differences in the levels of gene expression given by each spliced product, and confirmed by mutational studies (Chapter 4). These observations are similar to those of splicing events of the *elk-1* mRNA, in which two spliced isoforms enable the tight translational regulation of ELK-1 in different tissues and conditions (Araud *et al*., 2007). More recently the transcript variation of the human microsomal epoxide hydrolase (EPHX1) transcript has confirmed that sORFs regulate the post-transcriptional expression of microsomal epoxide hydrolase (mEH) (Nguyen *at al*., 2013). Similar events have also been reported for the regulation of the plant M flax rust resistance gene (Schmidt *et al*., 2007) and the human Disc large 1 oncosuppressor (Cavatorta *et al*., 2011). This work indicates that expression of *asp* may be tightly regulated; the conditions under which gene expression is permitted may be cell specific.

The present study utilised a series of sORF constructs with sORF initiation codons mutated in the VI to I direction. Mutation of sORF initiation codons in the opposite direction would allow further analysis of the termination-reinitiation and leaky scanning mechanisms of translation and the potential use of the cryptic CUG initiation codon (Chapter 6). The use of non-AUG-initiated reading frames is not uncommon and has been shown to rely on the use of elongator leucine-bound transfer RNAs to initiate translation and to be dependent upon the expression of eIF2A (Starck *et al*., 2012). Mutational analysis and toeprinting assays would facilitate further examination of the importance of this codon as a non-AUG initiator.

As reported by Kozak (1987c) the ability of a ribosome to reinitiate at a downstream AUG codon is dependent upon the distance between the end of the previous sORF and the downstream initiation codon. Distances less than 79 nucleotides are too

short to allow the ribosome to have recruited the necessary initiation factors required for initiation at the next initiation codon (Poyry *et al*., 2004). In the context of the HIV-1 *asp* sORF region, the length of sequence between sORF II and III is greater than 170 nucleotides; this distance should enable the scanning ribosome to reinitiate as this distance is greater than 79 nucleotides. However, the data presented here implies that only a weak toeprint is observed at sORF III suggesting that the ribosome may either dissociate from the transcript after termination of sORF II, or may miss sORF III (adequate sequence context). In contrast, the inability of ribosomes to initiate at sORF V can be explained by the short distance (64 nucleotides) between the termination of sORF IV and the AUG initiation codon of sORF V. Analysis of the coding sequences for sORFs I-VI also revealed the high abundance of rare codons, particularly within sORF I. Taken together with the toeprinting data this might suggest an additional method of control. In this instance the high percentage of rare codons within sORF I might significantly slow the translating ribosome to enable recognition of downstream sORFs, particularly sORF II whose initiation codon is in close proximity to the termination codon of sORF I (47nt). The mutation of rare codons to more common codons for the same amino acid may alter the level of gene expression. The ability to adjust translation efficiency by the use of such rare codons should be further examined. However, it must be noted that any mutations in the sequence of the sORF series will be restricted by the need to conserve the complementary Env coding sequence.

It is likely that sORFs VI and VI_{alt} are used by the translational machinery to initiate and engage the ribosomal complex and translate *asp* by a leaky scanning mechanism, possibly in a similar mode described for the translation of *env* and *vpu* (Schwartz *et al*., 1990; Anderson *et al.*, 2007). In both cases, alternative splicing of the HIV-1 positive sense transcript enables multiple AUG codons to be utilised to control gene expression. Similar events have also been reported for the translational regulation of Rev and Env in SIVmac239 in which up to five sORFs and mRNA splicing modulate expression (van der Velden *et al*., 2012) and the translational regulation of two Dot1 isoforms, a histone methyltransferase in *S.cerevisiae* by leaky scanning (Frederiks *et al*., 2009).

The conditions under which *asp* expression is permitted need further clarification. Toeprinting assays for each of the three Spliced Variants (1, 2 and 3) would build a complete picture of the effect these sORFs play when different leader sequences are produced via the splicing activity observed here. Similarly further examination via MFOLD of each of these different leader sequences may uncover new structures that may be implicated in either ribosome stalling and/or shunting mechanisms. Importantly, the abundance of these spliced transcripts in the context of HIV-1 infection is required to ascertain their relative importance with respect to *asp* expression. Differences in splicing activity and/or levels, particularly amongst T and monocyte cell linages, may offer insights into the role of *asp*. This is of particular interest, given the role of autophagy as suggested by Torresilla *et al.* (2013) and the implications this may have on HIV-1 replication in both T and monocytic cell lines (Espert *at al*., 2009; Kyei *et al*., 2009).

To date, only a handful of reports have been published on regulation of HIV-1 *asp* expression and have focused on the role of the LTR (promoter) and of Tat on negative sense transcription (Bukrinsky and Etkin, 1990; Michael *et al*., 1994a; Vanhee-Brossollet *et al*., 1995; Briquet and Vaquero, 2002; Bentley *et al*., 2004; Landry *et al*., 2007; Kobayashi-Ishihara *et al*., 2012; Laverdure *et al*., 2012). Most studies on *asp* and its expression, including this work, have used gene or specific region (eg upstream region, LTR) constructs. The present studies have relied on the reporter construct pEGFP-N1, into which the sORF region was cloned to allow a preliminary study of their potential role on downstream expression. Their relevance to the biology of HIV now must be extended by examining these events in HIV-1 infected cell systems, which would require physical containment facilities not available within our laboratory. Previous reports have established the difficulties in detection of this negative sense transcript (Bukrinsky and Etkin, 1990; Michael *et al*., 1994b; Landry *et al*., 2007; Kobayashi-Ishihara *et al*., 2012). Given the difficulties in detection of the negative sense transcript, presumably due to the transient expression of *asp* regulated by transcription from the 3'-LTR, the potential for alternative splicing of the transcript and sORF inhibition of translation, future studies of this transcript will be challenging.

Transcription of the negative sense gene is restricted to the early phase of infection by T-cell transcription factors acting on the negative sense promoter (Michael *et al*., 1994a; Bentley *et al*., 2004). The present investigation has shown that translation may also be strongly inhibited by the presence of multiple sORFs upstream of *asp*; inhibition may be modulated by alternative splicing of the upstream transcript. Such multi-modal regulation would permit a combination of regulatory signal inputs to precisely control *asp* expression, limiting ASP production to a particular phase in the virus replication cycle, and/or in response to host signalling events. The use of alternative splicing as a means of controlling gene expression in HIV-1 has the potential to extend to a global control of viral gene expression. In this study at least three intron sequences were identified. These sequences have the potential to control global viral gene expression by non-coding RNA (ncRNA) interactions in RNAi pathways. In recent years the identification of ncRNA (miRNA, piRNA and/or siRNA) sequences in HIV-1 has increased, with much unknown about the interaction of these sequences (Klase *et al*., 2012). The possible use of these intronic sequences as precursors for RNAi processes should not be ignored and may reveal novel mechanisms of gene expression in HIV-1. While not examined here, the potential for intron sequences to control expression of proviral and host cell genes via siRNA or miRNA pathways, activities that are becoming increasingly apparent (Yeung *et al*., 2005; Klase *et al*., 2009; Schopman *et al*., 2012), must also be considered.

7.3 Concluding remarks

In summary this thesis has established the potential for a series of six sORFs present upstream of the HIV-1 *asp* to regulate gene expression with fine control. While the exact function of *asp*/ASP still remains unclear, its reported association with autophagosomes (Torresilla *et al*., 2013), together with the potential for highly controlled expression described here, suggests the protein has a vital role in HIV-1 infection. Given *env* expression in infected cells leads to the induction of autophagy in uninfected CD4 T cells (Brass *et al*., 2008; Espert *et al*., 2008; Espert and Biard-Piechaczyk, 2009), the association of ASP to the autophagosome suggests it may be involved in the regulation of cell death. Further research into the autophagic

pathway in the context of HIV-1 infection will be of great interest, and may be of benefit in the development of highly active antiretroviral therapy (HAART) where HIV protease inhibitors have been shown to induce autophagy in cancer cells (Gills *et al*., 2007).

Since the discovery of HIV-1 and the advances made in understanding its mechanisms of infectivity a number of innovations have been made in the treatment of HIV-1, most of which only slow the replication of the virus in the host and the progression of the disease into AIDS. The majority of treatments available to date rely on cocktails of reverse transcriptase, protease and integrase blocking drugs, thus slowing the rate of viral incorporation into the cell, and thus spread of the virus from cell to cell (Marchand *et al*., 2006). More recently however, RNA based strategies including: ribozymes, anti-sense RNA, monoclonal antibodies, siRNA, oligonucleotides and shRNA/miRNA have been shown to inhibit the various stages of the viral life cycle (Nielsen *et al*., 2005). The present study acts as a foundation for further research into the regulatory role of sORFs for the expression of *asp*/ASP and may lead to further characterisation of the negative sense transcript and the potential role of *asp*/ASP in the life cycle of HIV-1; therefore opening the way to ASP activity and *asp* expression becoming an additional target for anti-HIV therapy.

REFERENCES

Acland, P., M. Dixon, G. Peters and C. Dickson (1990). "Subcellular fate of the int-2 oncoprotein is determined by choice of initiation codon." Nature 343(6259): 662-665.

Alderete, J. P., S. J. Child and A. P. Geballe (2001). "Abundant early expression of gpUL4 from a human cytomegalovirus mutant lacking a repressive upstream open reading frame." J Virol 75(15): 7188-7192.

Ali, I. K., L. McKendrick, S. J. Morley and R. J. Jackson (2001). "Activity of the hepatitis A virus IRES requires association between the cap-binding translation initiation factor (eIF4E) and eIF4G." J Virol 75(17): 7854-7863.

Anderson, J. L., A. T. Johnson, J. L. Howard and D. F. Purcell (2007). "Both linear and discontinuous ribosome scanning are used for translation initiation from bicistronic human immunodeficiency virus type 1 env mRNAs." J Virol 81(9): 4664-4676.

Araud, T., R. Genolet, P. Jaquier-Gubler and J. Curran (2007). "Alternatively spliced isoforms of the human elk-1 mRNA within the 5' UTR: implications for ELK-1 expression." Nucleic Acids Res 35(14): 4649-4663.

Aravin, A. and T. Tuschl (2005). "Identification and characterization of small RNAs involved in RNA silencing." FEBS Lett 579(26): 5830-5840.

Arnold, J., B. Yamamoto, M. Li, A. J. Phipps, I. Younis, M. D. Lairmore and P. L. Green (2006). "Enhancement of infectivity and persistence in vivo by HBZ, a natural antisense coded protein of HTLV-1." Blood 107(10): 3976-3982.

Atkins, JF., and Gesteland, RF. (1996) “Regulatory Recoding”. In Translational Control, Ed. Hershey, JWB., Cold Spring Harbour Laboratory Press, USA, pp. 653-684.

Augood, S. J., J. L. Ruth and P. C. Emson (1990). "A rapid method of non radioactive Northern blot analysis." Nucleic Acids Res 18(14): 4291.

Baker, K. E. and R. Parker (2004). "Nonsense-mediated mRNA decay: terminating erroneous gene expression." Curr Opin Cell Biol 16(3): 293-299.

Beltzer, J. P., S. R. Morris and G. B. Kohlhaw (1988). "Yeast LEU4 encodes mitochondrial and nonmitochondrial forms of alpha-isopropylmalate synthase." J Biol Chem 263(1): 368-374.

Benkirane, M., R. F. Chun, H. Xiao, V. V. Ogryzko, B. H. Howard, Y. Nakatani and K. T. Jeang (1998). "Activation of integrated provirus requires histone acetyltransferase. p300 and P/CAF are coactivators for HIV-1 Tat." J Biol Chem 273(38): 24898-24905.

Bennasser, Y., S. Y. Le, M. L. Yeung and K. T. Jeang (2004). "HIV-1 encoded candidate micro-RNAs and their cellular targets." Retrovirology 1: 43.

Bentley, K., N. Deacon, S. Sonza, S. Zeichner and M. Churchill (2004). "Mutational analysis of the HIV-1 LTR as a promoter of negative sense transcription." Arch Virol 149(12): 2277-2294.

Bergamini, G., M. Reschke, M. C. Battista, M. C. Boccuni, F. Campanini, A. Ripalti and M. P. Landini (1998). "The major open reading frame of the beta2.7 transcript of human cytomegalovirus: in vitro expression of a protein posttranscriptionally regulated by the 5' region." J Virol 72(10): 8425-8429.

Berthelot, K., M. Muldoon, L. Rajkowitsch, J. Hughes and J. E. McCarthy (2004). "Dynamics and processivity of 40S ribosome scanning on mRNA in yeast." Mol Microbiol 51(4): 987-1001.

Boeck, R. and D. Kolakofsky (1994). "Positions +5 and +6 can be major determinants of the efficiency of non-AUG initiation codons for protein synthesis." Embo j 13(15): 3608-3617.

Borman, A. M., Y. M. Michel and K. M. Kean (2000). "Biochemical characterisation of cap-poly(A) synergy in rabbit reticulocyte lysates: the eIF4G-PABP interaction increases the functional affinity of eIF4E for the capped mRNA 5'-end." Nucleic Acids Res 28(21): 4068-4075.

Brass, A. L., D. M. Dykxhoorn, Y. Benita, N. Yan, A. Engelman, R. J. Xavier, J. Lieberman and S. J. Elledge (2008). "Identification of host proteins required for HIV infection through a functional genomic screen." Science 319(5865): 921-926.

Briquet, S., J. Richardson, C. Vanhee-Brossollet and C. Vaquero (2001). "Natural antisense transcripts are detected in different cell lines and tissues of cats infected with feline immunodeficiency virus." Gene 267(2): 157-164.

Briquet, S. and C. Vaquero (2002). "Immunolocalization studies of an antisense protein in HIV-1-infected cells and viral particles." Virology 292(2): 177-184.

Brosnan, C. A. and O. Voinnet (2009). "The long and the short of noncoding RNAs." Curr Opin Cell Biol 21(3): 416-425.

Buck, C. B., X. Shen, M. A. Egan, T. C. Pierson, C. M. Walker and R. F. Siliciano (2001). "The human immunodeficiency virus type 1 gag gene encodes an internal ribosome entry site." J Virol 75(1): 181-191.

Bukrinsky, M. I. and A. F. Etkin (1990). "Plus strand of the HIV provirus DNA is expressed at early stages of infection." AIDS Res Hum Retroviruses 6(4): 425-426.

Calkhoven, C. F., C. Muller and A. Leutz (2002). "Translational control of gene expression and disease." Trends Mol Med 8(12): 577-583.

Calvo, S. E., D. J. Pagliarini and V. K. Mootha (2009). "Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans." Proc Natl Acad Sci U S A 106(18): 7507-7512.

Cao, F. and J. E. Tavis (2011). "RNA elements directing translation of the duck hepatitis B Virus polymerase via ribosomal shunting." J Virol 85(13): 6343-6352.

Cao, J. and A. P. Geballe (1994). "Mutational analysis of the translational signal in the human cytomegalovirus gpUL4 (gp48) transcript leader by retroviral infection." Virology 205(1): 151-160.

Cao, J. and A. P. Geballe (1995). "Translational inhibition by a human cytomegalovirus upstream open reading frame despite inefficient utilization of its AUG codon." J Virol 69(2): 1030-1036.

Cao, J. and A. P. Geballe (1996). "Coding sequence-dependent ribosomal arrest at termination of translation." Mol Cell Biol 16(2): 603-608.

Carlson, M., R. Taussig, S. Kustu and D. Botstein (1983). "The secreted form of invertase in Saccharomyces cerevisiae is synthesized from mRNA encoding a signal sequence." Mol Cell Biol 3(3): 439-447.

Carrasco, L., M. Barbacid and D. Vazquez (1973). "The trichodermin group of antibiotics, inhibitors of peptide bond formation by eukaryotic ribosomes." Biochim Biophys Acta 312(2): 368-376.

Cavanagh, M. H., S. Landry, B. Audet, C. Arpin-Andre, P. Hivin, M. E. Pare, J. Thete, E. Wattel, S. J. Marriott, J. M. Mesnard and B. Barbeau (2006). "HTLV-I antisense transcripts initiating in the 3'LTR are alternatively spliced and polyadenylated." Retrovirology 3: 15.

Cavatorta, A. L., F. Facciuto, M. B. Valdano, F. Marziali, A. A. Giri, L. Banks and D. Gardiol (2011). "Regulation of translational efficiency by different splice variants of the Disc large 1 oncosuppressor 5'-UTR." Febs j 278(14): 2596-2608.

Celano, P., C. M. Berchtold, D. L. Kizer, A. Weeraratna, B. D. Nelkin, S. B. Baylin and R. A. Casero, Jr. (1992). "Characterization of an endogenous RNA transcript with homology to the antisense strand of the human c-myc gene." J Biol Chem 267(21): 15092-15096.

Chappell, S. A., G. M. Edelman and V. P. Mauro (2000). "A 9-nt segment of a cellular mRNA can function as an internal ribosome entry site (IRES) and when present in linked multiple copies greatly enhances IRES activity." Proc Natl Acad Sci U S A 97(4): 1536-1541.

Chappell, S. A., G. M. Edelman and V. P. Mauro (2006a). "Ribosomal tethering and clustering as mechanisms for translation initiation." Proc Natl Acad Sci U S A 103(48): 18077-18082.

Chappell, S. A., J. Dresios, G. M. Edelman and V. P. Mauro (2006b). "Ribosomal shunting mediated by a translational enhancer element that base pairs to 18S rRNA." Proc Natl Acad Sci U S A 103(25): 9488-9493.

Cheah, M. T., A. Wachter, N. Sudarsan and R. R. Breaker (2007). "Control of alternative RNA splicing and gene expression by eukaryotic riboswitches." Nature 447(7143): 497-500.

Child, S. J., M. K. Miller and A. P. Geballe (1999a). "Translational control by an upstream open reading frame in the HER-2/neu transcript." J Biol Chem 274(34): 24335-24341.

Child, S. J., M. K. Miller and A. P. Geballe (1999b). "Cell type-dependent and -independent control of HER-2/neu translation." Int J Biochem Cell Biol 31(1): 201-213.

Clerc, I., S. Laverdure, C. Torresilla, S. Landry, S. Borel, A. Vargas, C. Arpin-Andre, B. Gay, L. Briant, A. Gross, B. Barbeau and J. M. Mesnard (2011). "Polarized expression of the membrane ASP protein derived from HIV-1 antisense transcription in T cells." Retrovirology 8: 74.

Cochrane, A. W., K. S. Jones, S. Beidas, P. J. Dillon, A. M. Skalka and C. A. Rosen (1991). "Identification and characterization of intragenic sequences which repress human immunodeficiency virus structural gene expression." J Virol 65(10): 5305-5313.

Conti, E. and E. Izaurralde (2005). "Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species." Curr Opin Cell Biol 17(3): 316-325.

Coolidge, C. J., R. J. Seely and J. G. Patton (1997). "Functional analysis of the polypyrimidine tract in pre-mRNA splicing." Nucleic Acids Res 25(4): 888-896.

Damiani, R. D., Jr. and S. R. Wessler (1993). "An upstream open reading frame represses expression of Lc, a member of the R/B family of maize transcriptional activators." Proc Natl Acad Sci U S A 90(17): 8244-8248.

Danthinne, X., J. Seurinck, F. Meulewaeter, M. Van Montagu and M. Cornelissen (1993). "The 3' untranslated region of satellite tobacco necrosis virus RNA stimulates translation in vitro." Mol Cell Biol 13(6): 3340-3349.

Das, F., N. Ghosh-Choudhury, A. Bera, B. S. Kasinath and G. G. Choudhury (2013). "TGFbeta-induced PI 3 kinase-dependent Mnk-1 activation is necessary for Ser-209 phosphorylation of eIF4E and mesangial cell hypertrophy." J Cell Physiol 228(7): 1617-1626.

Davuluri, R. V., Y. Suzuki, S. Sugano and M. Q. Zhang (2000). "CART classification of human 5' UTR sequences." Genome Res 10(11): 1807-1816.

de Breyne, S., V. Simonet, T. Pelet and J. Curran (2003). "Identification of a cis-acting element required for shunt-mediated translational initiation of the Sendai virus Y proteins." Nucleic Acids Res 31(2): 608-618.

de la Fuente, C., F. Santiago, L. Deng, C. Eadie, I. Zilberman, K. Kehn, A. Maddukuri, S. Baylor, K. Wu, C. G. Lee, A. Pumfery and F. Kashanchi (2002). "Gene expression profile of HIV-1 Tat expressing cells: a close interplay between proliferative and differentiation signals." BMC Biochem 3: 14.

Dean, F. B., P. Bullock, Y. Murakami, C. R. Wobbe, L. Weissbach and J. Hurwitz (1987). "Simian virus 40 (SV40) DNA replication: SV40 large T antigen unwinds DNA containing the SV40 origin of replication." Proc Natl Acad Sci U S A 84(1): 16-20.

Degnin, C. R., M. R. Schleiss, J. Cao and A. P. Geballe (1993). "Translational inhibition mediated by a short upstream open reading frame in the human cytomegalovirus gpUL4 (gp48) transcript." J Virol 67(9): 5514-5521.

Delbecq, P., M. Werner, A. Feller, R. K. Filipkowski, F. Messenguy and A. Pierard (1994). "A segment of mRNA encoding the leader peptide of the CPA1 gene confers repression by arginine on a heterologous yeast gene transcript." Mol Cell Biol 14(4): 2378-2390.

Derry, M. C., A. Yanagiya, Y. Martineau and N. Sonenberg (2006). "Regulation of poly(A)-binding protein through PABP-interacting proteins." Cold Spring Harb Symp Quant Biol 71: 537-543.

Dever, T. E. (1999). "Translation initiation: adept at adapting." Trends Biochem Sci 24(10): 398-403.

Dominguez, D. I., L. A. Ryabova, M. M. Pooggin, W. Schmidt-Puchta, J. Futterer and T. Hohn (1998). "Ribosome shunting in cauliflower mosaic virus. Identification of an essential and sufficient structural element." J Biol Chem 273(6): 3669-3678.

Donze, O., P. Damay and P. F. Spahr (1995). "The first and third uORFs in RSV leader RNA are efficiently translated: implications for translational regulation and viral RNA packaging." Nucleic Acids Res 23(5): 861-868.

El Kharroubi, A., G. Piras, R. Zensen and M. A. Martin (1998). "Transcriptional activation of the integrated chromatin-associated human immunodeficiency virus type 1 promoter." Mol Cell Biol 18(5): 2535-2544.

Ellenberg, J., J. Lippincott-Schwartz and J. F. Presley (1999). "Dual-colour imaging with GFP variants." Trends Cell Biol 9(2): 52-56.

Espert, L. and M. Biard-Piechaczyk (2009). "Autophagy in HIV-induced T cell death." Curr Top Microbiol Immunol 335: 307-321.

Espert, L., P. Codogno and M. Biard-Piechaczyk (2008). "What is the role of autophagy in HIV-1 infection?" Autophagy 4(3): 273-275.

Evanko, D. S., C. E. Ellis, V. Venkatachalam and T. Frielle (1998). "Preliminary analysis of the transcriptional regulation of the human beta 1-adrenergic receptor gene." Biochem Biophys Res Commun 244(2): 395-402.

Fang, P., C. C. Spevak, C. Wu and M. S. Sachs (2004). "A nascent polypeptide domain that can regulate translation elongation." Proc Natl Acad Sci U S A 101(12): 4059-4064.

Fang, P., Z. Wang and M. S. Sachs (2000). "Evolutionarily conserved features of the arginine attenuator peptide provide the necessary requirements for its function in translational regulation." J Biol Chem 275(35): 26710-26719.

Filipowicz, W., L. Jaskiewicz, F. A. Kolb and R. S. Pillai (2005). "Post-transcriptional gene silencing by siRNAs and miRNAs." Curr Opin Struct Biol 15(3): 331-341.

Frankel, A. D., D. S. Bredt and C. O. Pabo (1988). "Tat protein from human immunodeficiency virus forms a metal-linked dimer." Science 240(4848): 70-73.

Fraser, C. S. and J. A. Doudna (2007). "Structural and mechanistic insights into hepatitis C viral translation initiation." Nat Rev Microbiol 5(1): 29-38.

Frederiks, F., G. J. Heynen, S. J. van Deventer, H. Janssen and F. van Leeuwen (2009). "Two Dot1 isoforms in Saccharomyces cerevisiae as a result of leaky scanning by the ribosome." Nucleic Acids Res 37(21): 7047-7058.

Friebe, P. and E. Harris (2010). "Interplay of RNA elements in the dengue virus 5' and 3' ends required for viral RNA replication." J Virol 84(12): 6103-6118.

Futterer, J., I. Potrykus, Y. Bao, L. Li, T. M. Burns, R. Hull and T. Hohn (1996). "Position-dependent ATT initiation during plant pararetrovirus rice tungro bacilliform virus translation." J Virol 70(5): 2999-3010.

Futterer, J., H. M. Rothnie, T. Hohn and I. Potrykus (1997). "Rice tungro bacilliform virus open reading frames II and III are translated from polycistronic pregenomic RNA by leaky scanning." J Virol 71(10): 7984-7989.

Gaba, A., Z. Wang, T. Krishnamoorthy, A. G. Hinnebusch and M. S. Sachs (2001). "Physical evidence for distinct mechanisms of translational control by upstream open reading frames." Embo j 20(22): 6453-6463.

Gale, M., Jr., S. L. Tan and M. G. Katze (2000). "Translational control of viral gene expression in eukaryotes." Microbiol Mol Biol Rev 64(2): 239-280.

Gaudray, G., F. Gachon, J. Basbous, M. Biard-Piechaczyk, C. Devaux and J. M. Mesnard (2002). "The complementary strand of the human T-cell leukemia virus type 1 RNA genome encodes a bZIP transcription factor that down-regulates viral transcription." J Virol 76(24): 12813-12822.

Geballe, AP. (1996) “Translational Control Mediated by Upstream AUG Codons”. In Translational Control, Ed. Hershey, JWB., Cold Spring Harbour Laboratory Press, USA, pp. 173-197.

Gills, J. J., J. Lopiccolo, J. Tsurutani, R. H. Shoemaker, C. J. Best, M. S. Abu-Asab, J. Borojerdi, N. A. Warfel, E. R. Gardner, M. Danish, M. C. Hollander, S. Kawabata, M. Tsokos, W. D. Figg, P. S. Steeg and P. A. Dennis (2007). "Nelfinavir, A lead HIV protease inhibitor, is a broad-spectrum, anticancer agent that induces endoplasmic reticulum stress, autophagy, and apoptosis in vitro and in vivo." Clin Cancer Res 13(17): 5183-5194.

Goergen, D. and M. Niepmann (2012). "Stimulation of Hepatitis C Virus RNA translation by microRNA-122 occurs under different conditions in vivo and in vitro." Virus Res 167(2): 343-352.

Graham, F. L. and A. J. van der Eb (1973). "A new technique for the assay of infectivity of human adenovirus 5 DNA." Virology 52(2): 456-467.

Grant, C. M. and A. G. Hinnebusch (1994). "Effect of sequence context at stop codons on efficiency of reinitiation in GCN4 translational control." Mol Cell Biol 14(1): 606-618.

Grant, C. M., P. F. Miller and A. G. Hinnebusch (1994). "Requirements for intercistronic distance and level of eukaryotic initiation factor 2 activity in reinitiation on GCN4 mRNA vary with the downstream cistron." Mol Cell Biol 14(4): 2616-2628.

Grant, C. M., P. F. Miller and A. G. Hinnebusch (1995). "Sequences 5' of the first upstream open reading frame in GCN4 mRNA are required for efficient translational reinitiation." Nucleic Acids Res 23(19): 3980-3988.

Greenway, A. L., G. Holloway, D. A. McPhee, P. Ellis, A. Cornall and M. Lidman (2003). "HIV-1 Nef control of cell signalling molecules: multiple strategies to promote virus replication." J Biosci 28(3): 323-335.

Grunert, S. and R. J. Jackson (1994). "The immediate downstream codon strongly influences the efficiency of utilization of eukaryotic translation initiation codons." Embo j 13(15): 3618-3630.

Haas, J., E. C. Park and B. Seed (1996). "Codon usage limitation in the expression of HIV-1 envelope glycoprotein." Curr Biol 6(3): 315-324.

Halin, M., E. Douceron, I. Clerc, C. Journo, N. L. Ko, S. Landry, E. L. Murphy, A. Gessain, I. Lemasson, J. M. Mesnard, B. Barbeau and R. Mahieux (2009). "Human T-cell leukemia virus type 2 produces a spliced antisense transcript encoding a protein that lacks a classic bZIP domain but still inhibits Tax2-mediated transcription." Blood 114(12): 2427-2438.

Han, F. and X. Zhang (2006). "Internal initiation of mRNA translation in insect cell mediated by an internal ribosome entry site (IRES) from shrimp white spot syndrome virus (WSSV)." Biochem Biophys Res Commun 344(3): 893-899.

Harding, H. P., I. Novoa, Y. Zhang, H. Zeng, R. Wek, M. Schapira and D. Ron (2000). "Regulated translation initiation controls stress-induced gene expression in mammalian cells." Mol Cell 6(5): 1099-1108.

Harigai, M., T. Miyashita, M. Hanada and J. C. Reed (1996). "A cis-acting element in the BCL-2 gene controls expression through translational mechanisms." Oncogene 12(6): 1369-1374.

Hemmings-Mieszczak, M., T. Hohn and T. Preiss (2000). "Termination and peptide release at the upstream open reading frame are required for downstream translation on synthetic shunt-competent mRNA leaders." Mol Cell Biol 20(17): 6212-6223.

Henke, J. I., D. Goergen, J. Zheng, Y. Song, C. G. Schuttler, C. Fehr, C. Junemann and M. Niepmann (2008). "microRNA-122 stimulates translation of hepatitis C virus RNA." Embo j 27(24): 3300-3310.

Hill, J. R. and D. R. Morris (1993). "Cell-specific translational regulation of S-adenosylmethionine decarboxylase mRNA. Dependence on translation and coding capacity of the cis-acting upstream open reading frame." J Biol Chem 268(1): 726-731.

Hinnebusch, A. G. (1997). "Translational regulation of yeast GCN4. A window on factors that control initiator-trna binding to the ribosome." J Biol Chem 272(35): 21661-21664.

Ho, C. K. and J. F. Strauss, 3rd (2004). "Activation of the control reporter plasmids pRL-TK and pRL-SV40 by multiple GATA transcription factors can lead to aberrant normalization of transfection efficiency." BMC Biotechnol 4: 10.

Holzmann, K., I. Ambrosch, L. Elbling, M. Micksche and W. Berger (2001). "A small upstream open reading frame causes inhibition of human major vault protein expression from a ubiquitous mRNA splice variant." FEBS Lett 494(1-2): 99-104.

Huang, H. K., H. Yoon, E. M. Hannig and T. F. Donahue (1997). "GTP hydrolysis controls stringent selection of the AUG start codon during translation initiation in Saccharomyces cerevisiae." Genes Dev 11(18): 2396-2413.

Iakova, P., G. L. Wang, L. Timchenko, M. Michalak, O. M. Pereira-Smith, J. R. Smith and N. A. Timchenko (2004). "Competition of CUGBP1 and calreticulin for the regulation of p21 translation determines cell fate." Embo j 23(2): 406-417.

Imataka, H., A. Gradi and N. Sonenberg (1998). "A newly identified N-terminal amino acid sequence of human eIF4G binds poly(A)-binding protein and functions in poly(A)-dependent translation." Embo j 17(24): 7480-7489.

Ito, T. and M. M. Lai (1999). "An internal polypyrimidine-tract-binding protein-binding site in the hepatitis C virus RNA attenuates translation, which is relieved by the 3'-untranslated sequence." Virology 254(2): 288-296.

Jackson, RJ. (1996) “A Comparative View of Initiation Site Selection Mechanisms”. In Translational Control, Ed. Hershey, JWB., Cold Spring Harbor Laboratory Press, USA, pp. 71-112.

Jackson, R. J. and M. Wickens (1997). "Translational controls impinging on the 5'-untranslated region and initiation factor proteins." Curr Opin Genet Dev 7(2): 233-241.

Janzen, D. M., L. Frolova and A. P. Geballe (2002). "Inhibition of translation termination mediated by an interaction of eukaryotic release factor 1 with a nascent peptidyl-tRNA." Mol Cell Biol 22(24): 8562-8570.

Jeang, K. T., H. Xiao and E. A. Rich (1999). "Multifaceted activities of the HIV-1 transactivator of transcription, Tat." J Biol Chem 274(41): 28837-28840.

Jia, L. G., X. M. Wang, J. D. Shannon, J. B. Bjarnason and J. W. Fox (1997). "Function of disintegrin-like/cysteine-rich domains of atrolysin A. Inhibition of platelet aggregation by recombinant protein and peptide antagonists." J Biol Chem 272(20): 13094-13102.

Johansen, H., D. Schumperli and M. Rosenberg (1984). "Affecting gene expression by altering the length and sequence of the 5' leader." Proc Natl Acad Sci U S A 81(24): 7698-7702.

Jousse, C., A. Bruhat, V. Carraro, F. Urano, M. Ferrara, D. Ron and P. Fafournoux (2001). "Inhibition of CHOP translation by a peptide encoded by an open reading frame localized in the chop 5'UTR." Nucleic Acids Res 29(21): 4341-4351.

Kahvejian, A., Y. V. Svitkin, R. Sukarieh, M. N. M'Boutchou and N. Sonenberg (2005). "Mammalian poly(A)-binding protein is a eukaryotic translation initiation factor, which acts via multiple mechanisms." Genes Dev 19(1): 104-113.

Keller, T. E., S. D. Mis, K. E. Jia and C. O. Wilke (2012). "Reduced mRNA secondary-structure stability near the start codon indicates functional genes in prokaryotes." Genome Biol Evol 4(2): 80-88.

Kim, K. M., H. Cho and Y. K. Kim (2012). "The upstream open reading frame of cyclin-dependent kinase inhibitor 1A mRNA negatively regulates translation of the downstream main open reading frame." Biochem Biophys Res Commun 424(3): 469-475.

Kiss-Laszlo, Z., S. Blanc and T. Hohn (1995). "Splicing of cauliflower mosaic virus 35S RNA is essential for viral infectivity." Embo j 14(14): 3552-3562.

Klase, Z., L. Houzet and K. T. Jeang (2012). "MicroRNAs and HIV-1: complex interactions." J Biol Chem 287(49): 40884-40890.

Kobayashi-Ishihara, M., M. Yamagishi, T. Hara, Y. Matsuda, R. Takahashi, A. Miyake, K. Nakano, T. Yamochi, T. Ishida and T. Watanabe (2012). "HIV-1-encoded antisense RNA suppresses viral replication for a prolonged period." Retrovirology 9: 38.

Korkaya, H., S. Jameel, D. Gupta, S. Tyagi, R. Kumar, M. Zafrullah, M. Mazumdar, S. K. Lal, L. Xiaofang, D. Sehgal, S. R. Das and D. Sahal (2001). "The ORF3 protein of hepatitis E virus binds to Src homology 3 domains and activates MAPK." J Biol Chem 276(45): 42389-42400.

Kos, M., S. Denger, G. Reid and F. Gannon (2002). "Upstream open reading frames regulate the translation of the multiple mRNA variants of the estrogen receptor alpha." J Biol Chem 277(40): 37131-37138.

Kozak, M. (1978). "How do eucaryotic ribosomes select initiation regions in messenger RNA?" Cell 15(4): 1109-1123.

Kozak, M. (1984a). "Point mutations close to the AUG initiator codon affect the efficiency of translation of rat preproinsulin in vivo." Nature 308(5956): 241-246.

Kozak, M. (1984b). "Selection of initiation sites by eucaryotic ribosomes: effect of inserting AUG triplets upstream from the coding sequence for preproinsulin." Nucleic Acids Res 12(9): 3873-3893.

Kozak, M. (1986). "Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes." Cell 44(2): 283-292.

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

Kozak, M. (1987b). "At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells." J Mol Biol 196(4): 947-950.

Kozak, M. (1987c). "Effects of intercistronic length on the efficiency of reinitiation by eucaryotic ribosomes." Mol Cell Biol 7(10): 3438-3445.

Kozak, M. (1989a). "Circumstances and mechanisms of inhibition of translation by secondary structure in eucaryotic mRNAs." Mol Cell Biol 9(11): 5134-5142.

Kozak, M. (1989b). "The scanning model for translation: an update." J Cell Biol 108(2): 229-241.

Kozak, M. (1990). "Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes." Proc Natl Acad Sci U S A 87(21): 8301-8305.

Kozak, M. (1991a). "An analysis of vertebrate mRNA sequences: intimations of translational control." J Cell Biol 115(4): 887-903.

Kozak, M. (1991b). "Structural features in eukaryotic mRNAs that modulate the initiation of translation." J Biol Chem 266(30): 19867-19870.

Kozak, M. (1991c). "A short leader sequence impairs the fidelity of initiation by eukaryotic ribosomes." Gene Expr 1(2): 111-115.

Kozak, M. (1995). "Adherence to the first-AUG rule when a second AUG codon follows closely upon the first." Proc Natl Acad Sci U S A 92(7): 2662-2666.

Kozak, M. (1997). "Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6." Embo j 16(9): 2482-2492.

Kozak, M. (1998). "Primer extension analysis of eukaryotic ribosome-mRNA complexes." Nucleic Acids Res 26(21): 4853-4859.

Kozak, M. (1999). "Initiation of translation in prokaryotes and eukaryotes." Gene 234(2): 187-208.

Kozak, M. (2001a). "Constraints on reinitiation of translation in mammals." Nucleic Acids Res 29(24): 5226-5232.

Kozak, M. (2001b). "New ways of initiating translation in eukaryotes?" Mol Cell Biol 21(6): 1899-1907.

Kozak, M. (2002). "Pushing the limits of the scanning mechanism for initiation of translation." Gene 299(1-2): 1-34.

Kozak, M. (2003). "Alternative ways to think about mRNA sequences and proteins that appear to promote internal initiation of translation." Gene 318: 1-23.

Kozak, M. (2005). "Regulation of translation via mRNA structure in prokaryotes and eukaryotes." Gene 361: 13-37.

Kronstad, L. M., K. F. Brulois, J. U. Jung and B. A. Glaunsinger (2013). "Dual short upstream open reading frames control translation of a herpesviral polycistronic mRNA." PLoS Pathog 9(1): e1003156.

Krummheuer, J., A. T. Johnson, I. Hauber, S. Kammler, J. L. Anderson, J. Hauber, D. F. Purcell and H. Schaal (2007). "A minimal uORF within the HIV-1 vpu leader allows efficient translation initiation at the downstream env AUG." Virology 363(2): 261-271.

Kumar, M. and G. G. Carmichael (1998). "Antisense RNA: function and fate of duplex RNA in cells of higher eukaryotes." Microbiol Mol Biol Rev 62(4): 1415-1434.

Kyei, G. B., C. Dinkins, A. S. Davis, E. Roberts, S. B. Singh, C. Dong, L. Wu, E. Kominami, T. Ueno, A. Yamamoto, M. Federico, A. Panganiban, I. Vergne and V. Deretic (2009). "Autophagy pathway intersects with HIV-1 biosynthesis and regulates viral yields in macrophages." J Cell Biol 186(2): 255-268.

Lagunoff, M. and B. Roizman (1994). "Expression of a herpes simplex virus 1 open reading frame antisense to the gamma(1)34.5 gene and transcribed by an RNA 3' coterminal with the unspliced latency-associated transcript." J Virol 68(9): 6021-6028.

Landry, S., M. Halin, S. Lefort, B. Audet, C. Vaquero, J. M. Mesnard and B. Barbeau (2007). "Detection, characterization and regulation of antisense transcripts in HIV-1." Retrovirology 4: 71.

Larocca, D., L. A. Chao, M. H. Seto and T. K. Brunck (1989). "Human T-cell leukemia virus minus strand transcription in infected T-cells." Biochem Biophys Res Commun 163(2): 1006-1013.

Latorre, P., D. Kolakofsky and J. Curran (1998). "Sendai virus Y proteins are initiated by a ribosomal shunt." Mol Cell Biol 18(9): 5021-5031.

Laverdure, S., A. Gross, C. Arpin-Andre, I. Clerc, B. Beaumelle, B. Barbeau and J. M. Mesnard (2012). "HIV-1 antisense transcription is preferentially activated in primary monocyte-derived cells." J Virol 86(24): 13785-13789.

Lavner, Y. and D. Kotlar (2005). "Codon bias as a factor in regulating expression via translation rate in the human genome." Gene 345(1): 127-138.

Lazar, M. A., R. A. Hodin, G. Cardona and W. W. Chin (1990). "Gene expression from the c-erbA alpha/Rev-ErbA alpha genomic locus. Potential regulation of alternative splicing by opposite strand transcription." J Biol Chem 265(22): 12859-12863.

Lazar, M. A., R. A. Hodin, D. S. Darling and W. W. Chin (1989). "A novel member of the thyroid/steroid hormone receptor family is encoded by the opposite strand of the rat c-erbA alpha transcriptional unit." Mol Cell Biol 9(3): 1128-1136.

Lazarowitz, S. G. and H. D. Robertson (1977). "Initiator regions from the small size class of reovirus messenger RNA protected by rabbit reticulocyte ribosomes." J Biol Chem 252(21): 7842-7849.

Lee, J., E. H. Park, G. Couture, I. Harvey, P. Garneau and J. Pelletier (2002). "An upstream open reading frame impedes translation of the huntingtin gene." Nucleic Acids Res 30(23): 5110-5119.

Lee, N. S. and J. J. Rossi (2004). "Control of HIV-1 replication by RNA interference." Virus Res 102(1): 53-58.

Lemaire, P., C. Vesque, J. Schmitt, H. Stunnenberg, R. Frank and P. Charnay (1990). "The serum-inducible mouse gene Krox-24 encodes a sequence-specific transcriptional activator." Mol Cell Biol 10(7): 3456-3467.

Lemasson, I., M. R. Lewis, N. Polakowski, P. Hivin, M. H. Cavanagh, S. Thebault, B. Barbeau, J. K. Nyborg and J. M. Mesnard (2007). "Human T-cell leukemia virus type 1 (HTLV-1) bZIP protein interacts with the cellular transcription factor CREB to inhibit HTLV-1 transcription." J Virol 81(4): 1543-1553.

Li, A. W., G. Seyoum, R. P. Shiu and P. R. Murphy (1996). "Expression of the rat BFGF antisense RNA transcript is tissue-specific and developmentally regulated." Mol Cell Endocrinol 118(1-2): 113-123.

Li, H., W. M. Havens, M. L. Nibert and S. A. Ghabrial (2011). "RNA sequence determinants of a coupled termination-reinitiation strategy for downstream open reading frame translation in Helminthosporium victoriae virus 190S and other victoriviruses (Family Totiviridae)." J Virol 85(14): 7343-7352.

Li, M., E. Kao, X. Gao, H. Sandig, K. Limmer, M. Pavon-Eternod, T. E. Jones, S. Landry, T. Pan, M. D. Weitzman and M. David (2012). "Codon-usage-based inhibition of HIV protein synthesis by human schlafen 11." Nature 491(7422): 125-128.

Lincoln, A. J., Y. Monczak, S. C. Williams and P. F. Johnson (1998). "Inhibition of CCAAT/enhancer-binding protein alpha and beta translation by upstream open reading frames." J Biol Chem 273(16): 9552-9560.

Linz, B., N. Koloteva, S. Vasilescu and J. E. McCarthy (1997). "Disruption of ribosomal scanning on the 5'-untranslated region, and not restriction of translational initiation per se, modulates the stability of nonaberrant mRNAs in the yeast Saccharomyces cerevisiae." J Biol Chem 272(14): 9131-9140.

Liu, C. C., C. C. Simonsen and A. D. Levinson (1984). "Initiation of translation at internal AUG codons in mammalian cells." Nature 309(5963): 82-85.
Lovett, P. S. and E. J. Rogers (1996). "Ribosome regulation by the nascent peptide." Microbiol Rev 60(2): 366-385.

Ludwig, L. B., J. L. Ambrus, Jr., K. A. Krawczyk, S. Sharma, S. Brooks, C. B. Hsiao and S. A. Schwartz (2006). "Human Immunodeficiency Virus-Type 1 LTR DNA contains an intrinsic gene producing antisense RNA and protein products." Retrovirology 3: 80.

Luttermann, C. and G. Meyers (2009). "The importance of inter- and intramolecular base pairing for translation reinitiation on a eukaryotic bicistronic mRNA." Genes Dev 23(3): 331-344.

Luukkonen, B. G., W. Tan and S. Schwartz (1995). "Efficiency of reinitiation of translation on human immunodeficiency virus type 1 mRNAs is determined by the length of the upstream open reading frame and by intercistronic distance." J Virol 69(7): 4086-4094.

Maggiolini, M., O. Donze and D. Picard (1999). "A non-radioactive method for inexpensive quantitative RT-PCR." Biol Chem 380(6): 695-697.

Mandel, M. and A. Higa (1970). "Calcium-dependent bacteriophage DNA infection." J Mol Biol 53(1): 159-162.

Mangus, D. A., M. C. Evans and A. Jacobson (2003). "Poly(A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression." Genome Biol 4(7): 223.

Marchand, C., A. A. Johnson, E. Semenova and Y. Pommier (2006). "Mechanisms and inhibition of HIV integration." Drug Discov Today Dis Mech 3(2): 253-260.

Marsters, S. A., A. D. Frutkin, N. J. Simpson, B. M. Fendly and A. Ashkenazi (1992). "Identification of cysteine-rich domains of the type 1 tumor necrosis factor receptor involved in ligand binding." J Biol Chem 267(9): 5747-5750.

Marth, J. D., R. W. Overell, K. E. Meier, E. G. Krebs and R. M. Perlmutter (1988). "Translational activation of the lck proto-oncogene." Nature 332(6160): 171-173.

Martin, KJ., and Green, MR. (1992) “Transcriptional Activation by Viral Immediate-Early Proteins: Variations on a Common Theme”. In Transcriptional Regulation, Eds. McKnight, SL., and Yama, KR. Cold Spring Harbour Laboratory Press, New York, pp. 695-725.

Martins, L. P., N. Chenciner, B. Asjo, A. Meyerhans and S. Wain-Hobson (1991). "Independent fluctuation of human immunodeficiency virus type 1 rev and gp41 quasispecies in vivo." J Virol 65(8): 4502-4507.

Marzio, G., M. Vink, K. Verhoef, A. de Ronde and B. Berkhout (2002). "Efficient human immunodeficiency virus replication requires a fine-tuned level of transcription." J Virol 76(6): 3084-3088.

Mata, J., S. Marguerat and J. Bahler (2005). "Post-transcriptional control of gene expression: a genome-wide perspective." Trends Biochem Sci 30(9): 506-514.

Mathews, MB. (1996) “Interactions between Viruses and the Cellular Machinary for Protein Synthesis”. In Translational Control, Ed. Hershey, JWB., Cold Spring Harbor Laboratory Press, USA, pp. 505-548.

Matsuda, D. and T. W. Dreher (2006). "Close spacing of AUG initiation codons confers dicistronic character on a eukaryotic mRNA." Rna 12(7): 1338-1349.

Mattick, J. S. (2005). "The functional genomics of noncoding RNA." Science 309(5740): 1527-1528.

Medenbach, J., M. Seiler and M. W. Hentze (2011). "Translational control via protein-regulated upstream open reading frames." Cell 145(6): 902-913.

Meijer, H. A. and A. A. Thomas (2002). "Control of eukaryotic protein synthesis by upstream open reading frames in the 5'-untranslated region of an mRNA." Biochem J 367(Pt 1): 1-11.

Michael, N. L., L. D'Arcy, P. K. Ehrenberg and R. R. Redfield (1994a). "Naturally occurring genotypes of the human immunodeficiency virus type 1 long terminal repeat display a wide range of basal and Tat-induced transcriptional activities." J Virol 68(5): 3163-3174.

Michael, N. L., M. T. Vahey, L. d'Arcy, P. K. Ehrenberg, J. D. Mosca, J. Rappaport and R. R. Redfield (1994b). "Negative-strand RNA transcripts are produced in human immunodeficiency virus type 1-infected cells and patients by a novel promoter downregulated by Tat." J Virol 68(2): 979-987.

Miller, D. M., 3rd, N. S. Desai, D. C. Hardin, D. W. Piston, G. H. Patterson, J. Fleenor, S. Xu and A. Fire (1999). "Two-color GFP expression system for C. elegans." Biotechniques 26(5): 914-918, 920-911.

Miller, P. F. and A. G. Hinnebusch (1989). "Sequences that surround the stop codons of upstream open reading frames in GCN4 mRNA determine their distinct functions in translational control." Genes Dev 3(8): 1217-1225.

Miller, R. H. (1988). "Human immunodeficiency virus may encode a novel protein on the genomic DNA plus strand." Science 239(4846): 1420-1422.

Mize, G. J., H. Ruan, J. J. Low and D. R. Morris (1998). "The inhibitory upstream open reading frame from mammalian S-adenosylmethionine decarboxylase mRNA has a strict sequence specificity in critical positions." J Biol Chem 273(49): 32500-32505.

Monro, R. E., T. Staehelin, M. L. Celma and D. Vazquez (1969). "The peptidyl transferase activity of ribosomes." Cold Spring Harb Symp Quant Biol 34: 357-368.

Morris, D. R. and A. P. Geballe (2000). "Upstream open reading frames as regulators of mRNA translation." Mol Cell Biol 20(23): 8635-8642.

Moustakas, A., T. S. Sonstegard and P. B. Hackett (1993). "Alterations of the three short open reading frames in the Rous sarcoma virus leader RNA modulate viral replication and gene expression." J Virol 67(7): 4337-4349.

Moustakas, A., T. S. Sonstegard and P. B. Hackett (1993). "Alterations of the three short open reading frames in the Rous sarcoma virus leader RNA modulate viral replication and gene expression." J Virol 67(7): 4337-4349.

Mueller, P. P. and A. G. Hinnebusch (1986). "Multiple upstream AUG codons mediate translational control of GCN4." Cell 45(2): 201-207.

Mueller, P. P. and A. G. Hinnebusch (1986). "Multiple upstream AUG codons mediate translational control of GCN4." Cell 45(2): 201-207.

Munzarova, V., J. Panek, S. Gunisova, I. Danyi, B. Szamecz and L. S. Valasek (2011). "Translation reinitiation relies on the interaction between eIF3a/TIF32 and progressively folded cis-acting mRNA elements preceding short uORFs." PLoS Genet 7(7): e1002137.

Murata, T., M. Hijikata and K. Shimotohno (2005). "Enhancement of internal ribosome entry site-mediated translation and replication of hepatitis C virus by PD98059." Virology 340(1): 105-115.

Myasnikov, A. G., A. Simonetti, S. Marzi and B. P. Klaholz (2009). "Structure-function insights into prokaryotic and eukaryotic translation initiation." Curr Opin Struct Biol 19(3): 300-309.

Napthine, S., R. A. Lever, M. L. Powell, R. J. Jackson, T. D. Brown and I. Brierley (2009). "Expression of the VP2 protein of murine norovirus by a translation termination-reinitiation strategy." PLoS One 4(12): e8390.

Nguyen, H. L., X. Yang and C. J. Omiecinski (2013). "Expression of a novel mRNA transcript for human microsomal epoxide hydrolase (EPHX1) is regulated by short open reading frames within its 5'-untranslated region." Rna 19(6): 752-766.

Nguyen, L. S., L. Jolly, C. Shoubridge, W. K. Chan, L. Huang, F. Laumonnier, M. Raynaud, A. Hackett, M. Field, J. Rodriguez, A. K. Srivastava, Y. Lee, R. Long, A. M. Addington, J. L. Rapoport, S. Suren, C. N. Hahn, J. Gamble, M. F. Wilkinson, M. A. Corbett and J. Gecz (2012). "Transcriptome profiling of UPF3B/NMD-deficient lymphoblastoid cells from patients with various forms of intellectual disability." Mol Psychiatry 17(11): 1103-1115.

Nielsen, M. H., F. S. Pedersen and J. Kjems (2005). "Molecular strategies to inhibit HIV-1 replication." Retrovirology 2: 10.

Nomura, A., Y. Iwasaki, M. Saito, Y. Aoki, E. Yamamori, N. Ozaki, K. Tachikawa, N. Mutsuga, M. Morishita, M. Yoshida, M. Asai, Y. Oiso and H. Saito (2001). "Involvement of upstream open reading frames in regulation of rat V(1b) vasopressin receptor expression." Am J Physiol Endocrinol Metab 280(5): E780-787.

Nordhoff, V., K. Hubner, A. Bauer, I. Orlova, A. Malapetsa and H. R. Scholer (2001). "Comparative analysis of human, bovine, and murine Oct-4 upstream promoter sequences." Mamm Genome 12(4): 309-317.

Oliveira, C. C. and J. E. McCarthy (1995). "The relationship between eukaryotic translation and mRNA stability. A short upstream open reading frame strongly inhibits translational initiation and greatly accelerates mRNA degradation in the yeast Saccharomyces cerevisiae." J Biol Chem 270(15): 8936-8943.

Ostareck, D. H., A. Ostareck-Lederer, I. N. Shatsky and M. W. Hentze (2001). "Lipoxygenase mRNA silencing in erythroid differentiation: The 3'UTR regulatory complex controls 60S ribosomal subunit joining." Cell 104(2): 281-290.

Ostareck, D. H., A. Ostareck-Lederer, M. Wilm, B. J. Thiele, M. Mann and M. W. Hentze (1997). "mRNA silencing in erythroid differentiation: hnRNP K and hnRNP E1 regulate 15-lipoxygenase translation from the 3' end." Cell 89(4): 597-606.

Peabody, D. S. (1989). "Translation initiation at non-AUG triplets in mammalian cells." J Biol Chem 264(9): 5031-5035.

Peabody, D. S. and P. Berg (1986). "Termination-reinitiation occurs in the translation of mammalian cell mRNAs." Mol Cell Biol 6(7): 2695-2703.

Peeters, A., P. F. Lambert and N. J. Deacon (1996). "A fourth Sp1 site in the human immunodeficiency virus type 1 long terminal repeat is essential for negative-sense transcription." J Virol 70(10): 6665-6672.

Pellegrini, M., H. Oen, D. Eilat and C. R. Cantor (1974). "The mechanism of covalent reaction of bromoacetyl-phenylalanyl-transfer RNA with the peptidyl-transfer RNA binding site of the Escherichia coli ribosome." J Mol Biol 88(4): 809-829.

Pelletier, J. and N. Sonenberg (1988). "Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA." Nature 334(6180): 320-325.

Pereira, L. A., K. Bentley, A. Peeters, M. J. Churchill and N. J. Deacon (2000). "A compilation of cellular transcription factor interactions with the HIV-1 LTR promoter." Nucleic Acids Res 28(3): 663-668.

Persing, D. H., H. E. Varmus and D. Ganem (1985). "A frameshift mutation in the pre-S region of the human hepatitis B virus genome allows production of surface antigen particles but eliminates binding to polymerized albumin." Proc Natl Acad Sci U S A 82(10): 3440-3444.

Pesole, G., C. Gissi, G. Grillo, F. Licciulli, S. Liuni and C. Saccone (2000). "Analysis of oligonucleotide AUG start codon context in eukariotic mRNAs." Gene 261(1): 85-91.

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

Picard, C., A. Greenway, G. Holloway, D. Olive and Y. Collette (2002). "Interaction with simian Hck tyrosine kinase reveals convergent evolution of the Nef protein from simian and human immunodeficiency viruses despite differential molecular surface usage." Virology 295(2): 320-327.

Pisarev, A. V., N. E. Shirokikh and C. U. Hellen (2005). "Translation initiation by factor-independent binding of eukaryotic ribosomes to internal ribosomal entry sites." C R Biol 328(7): 589-605.

Plotkin, J. B. and G. Kudla (2011). "Synonymous but not the same: the causes and consequences of codon bias." Nat Rev Genet 12(1): 32-42.

Polacek, C., P. Friebe and E. Harris (2009). "Poly(A)-binding protein binds to the non-polyadenylated 3' untranslated region of dengue virus and modulates translation efficiency." J Gen Virol 90(Pt 3): 687-692.

Pooggin, M. M., T. Hohn and J. Futterer (2000). "Role of a short open reading frame in ribosome shunt on the cauliflower mosaic virus RNA leader." J Biol Chem 275(23): 17288-17296.

Pooggin, M. M., R. Rajeswaran, M. V. Schepetilnikov and L. A. Ryabova (2012). "Short ORF-dependent ribosome shunting operates in an RNA picorna-like virus and a DNA pararetrovirus that cause rice tungro disease." PLoS Pathog 8(3): e1002568.

Powell, M. L., K. E. Leigh, T. A. Poyry, R. J. Jackson, T. D. Brown and I. Brierley (2011). "Further characterisation of the translational termination-reinitiation signal of the influenza B virus segment 7 RNA." PLoS One 6(2): e16822.

Poyry, T. A., A. Kaminski, E. J. Connell, C. S. Fraser and R. J. Jackson (2007). "The mechanism of an exceptional case of reinitiation after translation of a long ORF reveals why such events do not generally occur in mammalian mRNA translation." Genes Dev 21(23): 3149-3162.

Poyry, T. A., A. Kaminski and R. J. Jackson (2004). "What determines whether mammalian ribosomes resume scanning after translation of a short upstream open reading frame?" Genes Dev 18(1): 62-75.

Prang, N., H. Wolf and F. Schwarzmann (1995). "Epstein-Barr virus lytic replication is controlled by posttranscriptional negative regulation of BZLF1." J Virol 69(4): 2644-2648.

Prats, A. C., G. De Billy, P. Wang and J. L. Darlix (1989). "CUG initiation codon used for the synthesis of a cell surface antigen coded by the murine leukemia virus." J Mol Biol 205(2): 363-372.

Preiss, T. and W. H. M (2003). "Starting the protein synthesis machine: eukaryotic translation initiation." Bioessays 25(12): 1201-1211.

Provost, P., C. Barat, I. Plante and M. J. Tremblay (2006). "HIV-l and the microRNA-guided silencing pathway: an intricate and multifaceted encounter." Virus Res 121(2): 107-115.

Purcell, D. F. and M. A. Martin (1993). "Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity." J Virol 67(11): 6365-6378.

Quivy, V., E. Adam, Y. Collette, D. Demonte, A. Chariot, C. Vanhulle, B. Berkhout, R. Castellano, Y. de Launoit, A. Burny, J. Piette, V. Bours and C. Van Lint (2002). "Synergistic activation of human immunodeficiency virus type 1 promoter activity by NF-kappaB and inhibitors of deacetylases: potential perspectives for the development of therapeutic strategies." J Virol 76(21): 11091-11103.

Quivy, V. and C. Van Lint (2002). "Diversity of acetylation targets and roles in transcriptional regulation: the human immunodeficiency virus type 1 promoter as a model system." Biochem Pharmacol 64(5-6): 925-934.

Racine, T. and R. Duncan (2010). "Facilitated leaky scanning and atypical ribosome shunting direct downstream translation initiation on the tricistronic S1 mRNA of avian reovirus." Nucleic Acids Res 38(20): 7260-7272.

Rajkowitsch, L., C. Vilela, K. Berthelot, C. V. Ramirez and J. E. McCarthy (2004). "Reinitiation and recycling are distinct processes occurring downstream of translation termination in yeast." J Mol Biol 335(1): 71-85.

Raney, A., G. L. Law, G. J. Mize and D. R. Morris (2002). "Regulated translation termination at the upstream open reading frame in s-adenosylmethionine decarboxylase mRNA." J Biol Chem 277(8): 5988-5994.

Rao, C. D., M. Pech, K. C. Robbins and S. A. Aaronson (1988). "The 5' untranslated sequence of the c-sis/platelet-derived growth factor 2 transcript is a potent translational inhibitor." Mol Cell Biol 8(1): 284-292.

Raveh-Amit, H., A. Maissel, J. Poller, L. Marom, O. Elroy-Stein, M. Shapira and E. Livneh (2009). "Translational control of protein kinase Ceta by two upstream open reading frames." Mol Cell Biol 29(22): 6140-6148.

Ray, R., S. Jameel, V. Manivel and R. Ray (1992). "Indian hepatitis E virus shows a major deletion in the small open reading frame." Virology 189(1): 359-362.

Redondo, N., M. A. Sanz, J. Steinberger, T. Skern, Y. Kusov and L. Carrasco (2012). "Translation directed by hepatitis A virus IRES in the absence of active eIF4F complex and eIF2." PLoS One 7(12): e52065.

Remm, M., A. Remm and M. Ustav (1999). "Human papillomavirus type 18 E1 protein is translated from polycistronic mRNA by a discontinuous scanning mechanism." J Virol 73(4): 3062-3070.

Reynolds, K., A. M. Zimmer and A. Zimmer (1996). "Regulation of RAR beta 2 mRNA expression: evidence for an inhibitory peptide encoded in the 5'-untranslated region." J Cell Biol 134(4): 827-835.

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

Rohwedel, J., S. Kugler, T. Engebrecht, W. Purschke, P. K. Muller and C. Kruse (2003). "Evidence for posttranscriptional regulation of the multi K homology domain protein vigilin by a small peptide encoded in the 5' leader sequence." Cell Mol Life Sci 60(8): 1705-1715.

Roy, B., J. N. Vaughn, B. H. Kim, F. Zhou, M. A. Gilchrist and A. G. Von Arnim (2010). "The h subunit of eIF3 promotes reinitiation competence during translation of mRNAs harboring upstream open reading frames." Rna 16(4): 748-761.

Ruiz-Echevarria, M. J., C. I. Gonzalez and S. W. Peltz (1998). "Identifying the right stop: determining how the surveillance complex recognizes and degrades an aberrant mRNA." Embo j 17(2): 575-589.

Ruiz-Echevarria, M. J. and S. W. Peltz (2000). "The RNA binding protein Pub1 modulates the stability of transcripts containing upstream open reading frames." Cell 101(7): 741-751.

Ryabova, L. A. and T. Hohn (2000). "Ribosome shunting in the cauliflower mosaic virus 35S RNA leader is a special case of reinitiation of translation functioning in plant and animal systems." Genes Dev 14(7): 817-829.

Sachs, M. S., Z. Wang, A. Gaba, P. Fang, J. Belk, R. Ganesan, N. Amrani and A. Jacobson (2002). "Toeprint analysis of the positioning of translation apparatus components at initiation and termination codons of fungal mRNAs." Methods 26(2): 105-114.

Scheper, G. C. and C. G. Proud (2002). "Does phosphorylation of the cap-binding protein eIF4E play a role in translation initiation?" Eur J Biochem 269(22): 5350-5359.

Schepetilnikov, M., K. Kobayashi, A. Geldreich, C. Caranta, C. Robaglia, M. Keller and L. A. Ryabova (2011). "Viral factor TAV recruits TOR/S6K1 signalling to activate reinitiation after long ORF translation." Embo j 30(7): 1343-1356.

Schleiss, M. R., C. R. Degnin and A. P. Geballe (1991). "Translational control of human cytomegalovirus gp48 expression." J Virol 65(12): 6782-6789.

Schluter, G., D. Boinska and S. C. Nieman-Seyde (2000). "Evidence for translational repression of the SOCS-1 major open reading frame by an upstream open reading frame." Biochem Biophys Res Commun 268(2): 255-261.

Schmidt, S., M. Lombardi, D. M. Gardiner, M. Ayliffe and P. A. Anderson (2007). "The M flax rust resistance pre-mRNA is alternatively spliced and contains a complex upstream untranslated region." Theor Appl Genet 115(3): 373-382.

Schopman, N. C., M. Willemsen, Y. P. Liu, T. Bradley, A. van Kampen, F. Baas, B. Berkhout and J. Haasnoot (2012). "Deep sequencing of virus-infected cells reveals HIV-encoded small RNAs." Nucleic Acids Res 40(1): 414-427.

Schwartz, S., B. K. Felber, E. M. Fenyo and G. N. Pavlakis (1990). "Env and Vpu proteins of human immunodeficiency virus type 1 are produced from multiple bicistronic mRNAs." J Virol 64(11): 5448-5456.

Schwartz, S., B. K. Felber and G. N. Pavlakis (1992). "Mechanism of translation of monocistronic and multicistronic human immunodeficiency virus type 1 mRNAs." Mol Cell Biol 12(1): 207-219.

Sedman, S. A., G. W. Gelembiuk and J. E. Mertz (1990). "Translation initiation at a downstream AUG occurs with increased efficiency when the upstream AUG is located very close to the 5' cap." J Virol 64(1): 453-457.

Seino, A., Y. Yanagida, M. Aizawa and E. Kobatake (2005). "Translational control by internal ribosome entry site in Saccharomyces cerevisiae." Biochim Biophys Acta 1681(2-3): 166-174.

Shabalina, S. A., A. Y. Ogurtsov, I. B. Rogozin, E. V. Koonin and D. J. Lipman (2004). "Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals." Nucleic Acids Res 32(5): 1774-1782.

Smith, C. W., J. G. Patton and B. Nadal-Ginard (1989). "Alternative splicing in the control of gene expression." Annu Rev Genet 23: 527-577.

Smith, C. W. and J. Valcarcel (2000). "Alternative pre-mRNA splicing: the logic of combinatorial control." Trends Biochem Sci 25(8): 381-388.

Snyder, R. D. and M. L. Edwards (1991). "Effects of polyamine analogs on the extent and fidelity of in vitro polypeptide synthesis." Biochem Biophys Res Commun 176(3): 1383-1392.

Song, K. Y., C. K. Hwang, C. S. Kim, H. S. Choi, P. Y. Law, L. N. Wei and H. H. Loh (2007). "Translational repression of mouse mu opioid receptor expression via leaky scanning." Nucleic Acids Res 35(5): 1501-1513.

Spevak, C. C., E. H. Park, A. P. Geballe, J. Pelletier and M. S. Sachs (2006). "her-2 upstream open reading frame effects on the use of downstream initiation codons." Biochem Biophys Res Commun 350(4): 834-841.

Starck, S. R., V. Jiang, M. Pavon-Eternod, S. Prasad, B. McCarthy, T. Pan and N. Shastri (2012). "Leucine-tRNA initiates at CUG start codons for protein synthesis and presentation by MHC class I." Science 336(6089): 1719-1723.

Steffy, K. and F. Wong-Staal (1991). "Genetic regulation of human immunodeficiency virus." Microbiol Rev 55(2): 193-205.

Stoltzfus, C. M. (2009). "Chapter 1. Regulation of HIV-1 alternative RNA splicing and its role in virus replication." Adv Virus Res 74: 1-40.

Stripecke, R., C. C. Oliveira, J. E. McCarthy and M. W. Hentze (1994). "Proteins binding to 5' untranslated region sites: a general mechanism for translational regulation of mRNAs in human and yeast cells." Mol Cell Biol 14(9): 5898-5909.

Strudwick, S. and K. L. Borden (2002). "The emerging roles of translation factor eIF4E in the nucleus." Differentiation 70(1): 10-22.

Struhl, K. (1998). "Histone acetylation and transcriptional regulatory mechanisms." Genes Dev 12(5): 599-606.

Sullivan, C. S. and D. Ganem (2005). "MicroRNAs and viral infection." Mol Cell 20(1): 3-7.

Suzuki, Y., D. Ishihara, M. Sasaki, H. Nakagawa, H. Hata, T. Tsunoda, M. Watanabe, T. Komatsu, T. Ota, T. Isogai, A. Suyama and S. Sugano (2000). "Statistical analysis of the 5' untranslated region of human mRNA using "Oligo-Capped" cDNA libraries." Genomics 64(3): 286-297.

Svitkin, Y. V., A. Pause, A. Haghighat, S. Pyronnet, G. Witherell, G. J. Belsham and N. Sonenberg (2001). "The requirement for eukaryotic initiation factor 4A (eIF4A) in translation is in direct proportion to the degree of mRNA 5' secondary structure." Rna 7(3): 382-394.

Sweet, T., C. Kovalak and J. Coller (2012). "The DEAD-box protein Dhh1 promotes decapping by slowing ribosome movement." PLoS Biol 10(6): e1001342.

Tagieva, N. E. and C. Vaquero (1997). "Expression of naturally occurring antisense RNA inhibits human immunodeficiency virus type 1 heterologous strain replication." J Gen Virol 78 (Pt 10): 2503-2511.

Tang, H., K. L. Kuhen and F. Wong-Staal (1999). "Lentivirus replication and regulation." Annu Rev Genet 33: 133-170.

Thiebeauld, O., M. Schepetilnikov, H. S. Park, A. Geldreich, K. Kobayashi, M. Keller, T. Hohn and L. A. Ryabova (2009). "A new plant protein interacts with eIF3 and 60S to enhance virus-activated translation re-initiation." Embo j 28(20): 3171-3184.

Timchenko, L. T., E. Salisbury, G. L. Wang, H. Nguyen, J. H. Albrecht, J. W. Hershey and N. A. Timchenko (2006). "Age-specific CUGBP1-eIF2 complex increases translation of CCAAT/enhancer-binding protein beta in old liver." J Biol Chem 281(43): 32806-32819.

Torresilla, C., E. Larocque, S. Landry, M. Halin, Y. Coulombe, J. Y. Masson, J. M. Mesnard and B. Barbeau (2013). "Detection of the HIV-1 minus-strand-encoded antisense protein and its association with autophagy." J Virol 87(9): 5089-5105.

Touriol, C., S. Bornes, S. Bonnal, S. Audigier, H. Prats, A. C. Prats and S. Vagner (2003). "Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons." Biol Cell 95(3-4): 169-178.

Touriol, C., S. Bornes, S. Bonnal, S. Audigier, H. Prats, A. C. Prats and S. Vagner (2003). "Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons." Biol Cell 95(3-4): 169-178.

Treder, K., E. L. Kneller, E. M. Allen, Z. Wang, K. S. Browning and W. A. Miller (2008). "The 3' cap-independent translation element of Barley yellow dwarf virus binds eIF4F via the eIF4G subunit to initiate translation." Rna 14(1): 134-147.

van der Velden, G. J., B. Klaver, A. T. Das and B. Berkhout (2012). "Upstream AUG codons in the simian immunodeficiency virus SIVmac239 genome regulate Rev and Env protein translation." J Virol 86(22): 12362-12371.

Van Heuverswyn, F., Y. Li, C. Neel, E. Bailes, B. F. Keele, W. Liu, S. Loul, C. Butel, F. Liegeois, Y. Bienvenue, E. M. Ngolle, P. M. Sharp, G. M. Shaw, E. Delaporte, B. H. Hahn and M. Peeters (2006). "Human immunodeficiency viruses: SIV infection in wild gorillas." Nature 444(7116): 164.

Van Lint, C., S. Emiliani, M. Ott and E. Verdin (1996). "Transcriptional activation and chromatin remodeling of the HIV-1 promoter in response to histone acetylation." Embo j 15(5): 1112-1120.

Vanhee-Brossollet, C., H. Thoreau, N. Serpente, L. D'Auriol, J. P. Levy and C. Vaquero (1995). "A natural antisense RNA derived from the HIV-1 env gene encodes a protein which is recognized by circulating antibodies of HIV+ individuals." Virology 206(1): 196-202.

Vanhee-Brossollet, C. and C. Vaquero (1998). "Do natural antisense transcripts make sense in eukaryotes?" Gene 211(1): 1-9.

Vilela, C., C. V. Ramirez, B. Linz, C. Rodrigues-Pousada and J. E. McCarthy (1999). "Post-termination ribosome interactions with the 5'UTR modulate yeast mRNA stability." Embo j 18(11): 3139-3152.

Vlcek, C., Z. Kozmik, V. Paces, S. Schirm and M. Schwyzer (1990). "Pseudorabies virus immediate-early gene overlaps with an oppositely oriented open reading frame: characterization of their promoter and enhancer regions." Virology 179(1): 365-377.

Wachter, A. (2010). "Riboswitch-mediated control of gene expression in eukaryotes." RNA Biol 7(1): 67-76.

Wachter, A., M. Tunc-Ozdemir, B. C. Grove, P. J. Green, D. K. Shintani and R. R. Breaker (2007). "Riboswitch control of gene expression in plants by splicing and alternative 3' end processing of mRNAs." Plant Cell 19(11): 3437-3450.

Wagner, E. G. and R. W. Simons (1994). "Antisense RNA control in bacteria, phages, and plasmids." Annu Rev Microbiol 48: 713-742.

Walsh, D. and I. Mohr (2004). "Phosphorylation of eIF4E by Mnk-1 enhances HSV-1 translation and replication in quiescent cells." Genes Dev 18(6): 660-672.

Wang, L. and S. R. Wessler (1998). "Inefficient reinitiation is responsible for upstream open reading frame-mediated translational repression of the maize R gene." Plant Cell 10(10): 1733-1746.

Wang, X. Q. and J. A. Rothnagel (2004). "5'-untranslated regions with multiple upstream AUG codons can support low-level translation via leaky scanning and reinitiation." Nucleic Acids Res 32(4): 1382-1391.

Wang, Z., P. Fang and M. S. Sachs (1998). "The evolutionarily conserved eukaryotic arginine attenuator peptide regulates the movement of ribosomes that have translated it." Mol Cell Biol 18(12): 7528-7536.

Wang, Z., A. Gaba and M. S. Sachs (1999). "A highly conserved mechanism of regulated ribosome stalling mediated by fungal arginine attenuator peptides that appears independent of the charging status of arginyl-tRNAs." J Biol Chem 274(53): 37565-37574.

Wang, Z. and M. S. Sachs (1997). "Ribosome stalling is responsible for arginine-specific translational attenuation in Neurospora crassa." Mol Cell Biol 17(9): 4904-4913.

Ward, A. J. and T. A. Cooper (2010). "The pathobiology of splicing." J Pathol 220(2): 152-163.

Weischenfeldt, J., J. Waage, G. Tian, J. Zhao, I. Damgaard, J. S. Jakobsen, K. Kristiansen, A. Krogh, J. Wang and B. T. Porse (2012). "Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns." Genome Biol 13(5): R35.

Werner, M., A. Feller, F. Messenguy and A. Pierard (1987). "The leader peptide of yeast gene CPA1 is essential for the translational repression of its expression." Cell 49(6): 805-813.

Wigler, M., S. Silverstein, L. S. Lee, A. Pellicer, Y. Cheng and R. Axel (1977). "Transfer of purified herpes virus thymidine kinase gene to cultured mouse cells." Cell 11(1): 223-232.

Williams, M. A. and R. A. Lamb (1989). "Effect of mutations and deletions in a bicistronic mRNA on the synthesis of influenza B virus NB and NA glycoproteins." J Virol 63(1): 28-35.

Withers, J. B. and K. L. Beemon (2010). "Structural features in the Rous sarcoma virus RNA stability element are necessary for sensing the correct termination codon." Retrovirology 7: 65.

Wortman, B., N. Darbinian, B. E. Sawaya, K. Khalili and S. Amini (2002). "Evidence for regulation of long terminal repeat transcription by Wnt transcription factor TCF-4 in human astrocytic cells." J Virol 76(21): 11159-11165.

Wu, H., C. R. Ross and F. Blecha (2002). "Characterization of an upstream open reading frame in the 5' untranslated region of PR-39, a cathelicidin antimicrobial peptide." Mol Immunol 39(1-2): 9-18.

Xi, Q., R. Cuesta and R. J. Schneider (2004). "Tethering of eIF4G to adenoviral mRNAs by viral 100k protein drives ribosome shunting." Genes Dev 18(16): 1997-2009.

Xi, Q., R. Cuesta and R. J. Schneider (2005). "Regulation of translation by ribosome shunting through phosphotyrosine-dependent coupling of adenovirus protein 100k to viral mRNAs." J Virol 79(9): 5676-5683.

Xu, G., C. Rabadan-Diehl, M. Nikodemova, P. Wynn, J. Spiess and G. Aguilera (2001). "Inhibition of corticotropin releasing hormone type-1 receptor translation by an upstream AUG triplet in the 5' untranslated region." Mol Pharmacol 59(3): 485-492.

Yamashita, R., Y. Suzuki, K. Nakai and S. Sugano (2003). "Small open reading frames in 5' untranslated regions of mRnas." C R Biol 326(10-11): 987-991.

Yap, SH., Vardarli, N., and Deacon, NJ. (2005) Unpublished Results.

Yeung, M. L., Y. Bennasser, S. Y. Le and K. T. Jeang (2005). "siRNA, miRNA and HIV: promises and challenges." Cell Res 15(11-12): 935-946.

Yoshida, M., Y. Satou, J. Yasunaga, J. Fujisawa and M. Matsuoka (2008). "Transcriptional control of spliced and unspliced human T-cell leukemia virus type 1 bZIP factor (HBZ) gene." J Virol 82(19): 9359-9368.

Yueh, A. and R. J. Schneider (2000). "Translation by ribosome shunting on adenovirus and hsp70 mRNAs facilitated by complementarity to 18S rRNA." Genes Dev 14(4): 414-421.

Zelent, A., C. Mendelsohn, P. Kastner, A. Krust, J. M. Garnier, F. Ruffenach, P. Leroy and P. Chambon (1991). "Differentially expressed isoforms of the mouse retinoic acid receptor beta generated by usage of two promoters and alternative splicing." Embo j 10(1): 71-81.

Zhou, J., W. J. Liu, S. W. Peng, X. Y. Sun and I. Frazer (1999). "Papillomavirus capsid protein expression level depends on the match between codon usage and tRNA availability." J Virol 73(6): 4972-4982.

Zhou, T. and C. O. Wilke (2011). "Reduced stability of mRNA secondary structure near the translation-initiation site in dsDNA viruses." BMC Evol Biol 11: 59.

Zimmer, A., A. M. Zimmer and K. Reynolds (1994). "Tissue specific expression of the retinoic acid receptor-beta 2: regulation by short open reading frames in the 5'-noncoding region." J Cell Biol 127(4): 1111-1119.

Zuker, M. (2003). "Mfold web server for nucleic acid folding and hybridization prediction." Nucleic Acids Res 31(13): 3406-3415.

APPENDIX 1- HIV-1 SUBTYPE B SEQUENCES

Table A.1 HIV-1 B sequences used in this study.

Name	Subtype /Group	Accession	Source	Author	Reference
MBC925	B	AF042101	Australia	Oelrichs, RB	*J Biomed Sci* **7**(2): 128-135 (2000)
1027-03	B	AY332237	USA	Bernardin, F	*J Virol* **79**(17): 11523-11528 (2005)
TWCYS	B	AF086817	Taiwan	Huang, LM	*Unpublished*
01UYTRA1179	B	AY81127	Uruguay	Vinoles, J	*Am J Trop Med Hyg* **72**(4): 495-500 (2005)
NL43	B	AF003887	U.S.A	Fang, G	*J Acquir Immune Defic Syndr Hum Retrovirol* **12**(4):352-7 (1996)
OYI	B	M26726	Gabon	Huet, T	*AIDS* **3**(11):707-15 (1989)
WC1PR	B	U69584	-	Fang, G	*Unpublished*
MBCC98	B	AF042104	Australia	Oelrichs, RB	*Unpublished*
WC10P-6	B	AY3104048	USA	Philpott, S	*J Virol* **79**(1): 353-363 (2005)
WC10C-11	B	AY314062	USA	Philpott, S	*J Virol* **79**(1): 353-363 (2005)
PCM033	B	AY561238	Colombia	Sanchez, GI	*Unpublished*
S61G7	B	AF256207	Spain	Yuste, E	*J Virol* **74**(20):9546-9552 (2000)
CANA1	B	AY779564	Canada	Cali, L	*AIDS Res Humm Retroviruses* **21**(8): 728-733 (2005)
CANB6	B	AY779556	Canada	Cali, L	*AIDS Res Humm Retroviruses* **21**(8): 728-733 (2005)
85US Ba-L	B	AY713409	USA	Brown, BK	*J Virol* **79**(10):6089-6101 (2005)
HIVMCK1	B	D86068	-	Cloyd, MW	*Viology* **174**(1):103-116 (1990)
HIV2132	B	D86069	-	Cloyd, MW	*Virology* **174**(1):103-116 (1990)
BH10	B	M15654	-	Wong-Staal, F	*Nature* **313**(6000):277-284 (1985)
1058-11	B	AY331295	USA	Bernardin, F	*J Virol* **79**(17): 11523-11528 (2005)
5157-86	B	AY835756	USA	Mikhail, M	*Unpublished*
US4	B	AY173955	USA	Hierholzer, J	*AIDS Res Humm Retroviruses* **18**(18): 1339-1350 (2002)
PMC013	B	AY561236	Colombia	Sanchez, GI	*Unpublished*
02HNsc11	B	DQ007903	China	Liu, L	*Unpublished*
05AR163052	B	DQ383755	Argentina	Pando, MA	*Retrovirology* **3**: 59 (2006)
BZ167	B	AY173956	Brazil	Hierholzer, J	*AIDS Res Humm Retroviruses* **18**(18): 1339-1350 (2002)
1001-09	B	AY331283	USA	Bernardin, F	*J Virol* **79**(17): 11523-11528 (2005)
PCM001	B	AY561236	Colombia	Sanchez, GI	*Unpublished*

APPENDIX 2- MFOLD PREDICTIONS

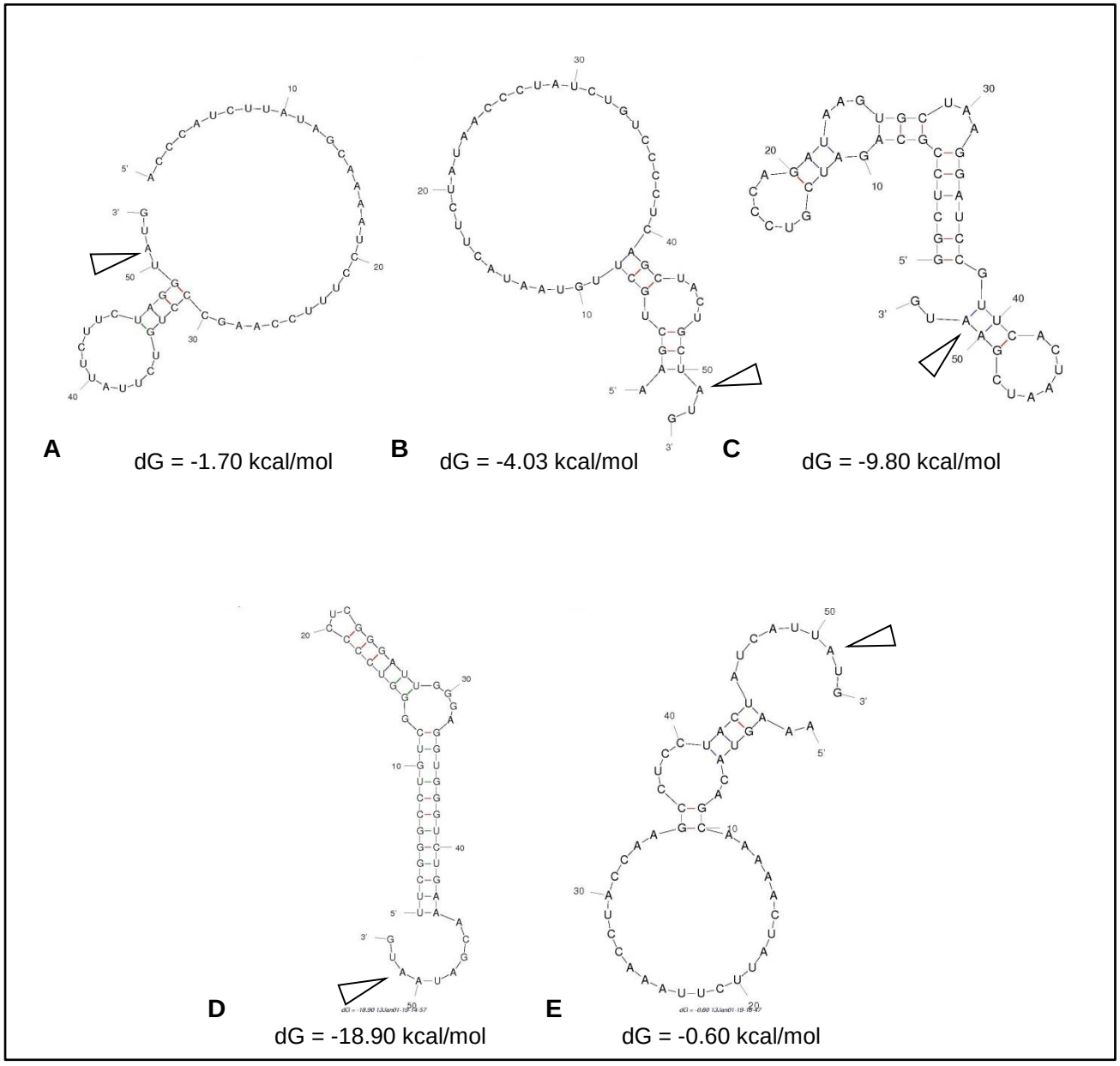


Figure A.1 Secondary structure predictions for sORFs I to V. The first 50nt upstream from each sORF AUG codon was folded using MFOLD. The structures are shown for (A) sORF I, (B) sORF II, (C) sORF III, (D) sORF IV and (E) sORF V with the sORF AUG codons noted with an arrow.

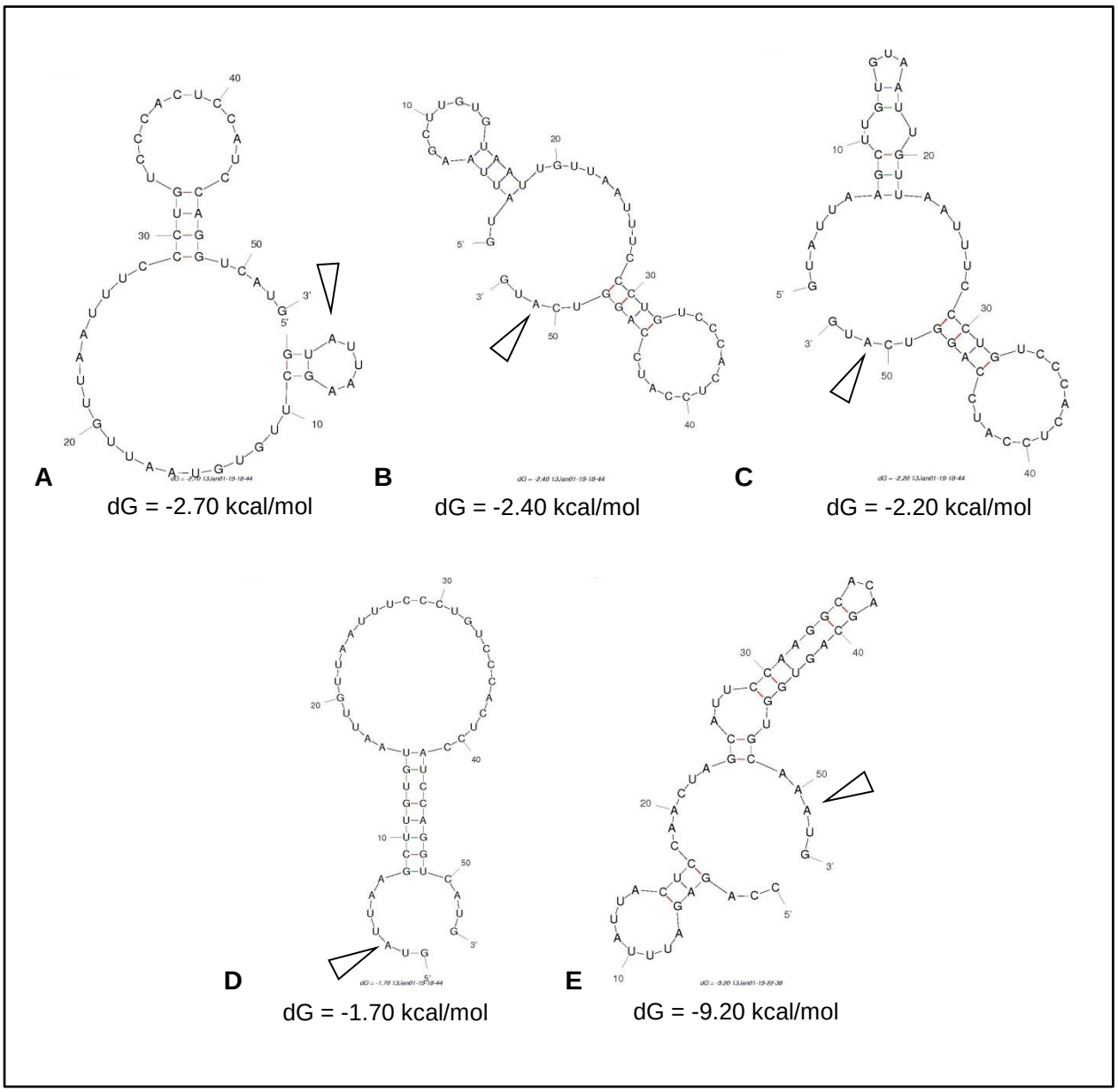


Figure A.2 Secondary structure predictions for sORFs VI and VI_{alt}. The first 50nt upstream from each sORF AUG codon was folded using MFOLD. The structures are shown for (A - D) sORF VI structures 1 to 4 respectively and, (E) sORF VI_{alt} with the sORF AUG codons noted with an arrow.

APPENDIX 3- PUBLICATIONS

Barbagallo, MS., Birch, KE., Deacon, NJ., and Mosse, JA. (2012) “Potential Control of Humman Immunodefficency Virus Type 1 *asp* Expression by Alternative Splicing in the Upstream Untranslated Region”. *DNA and Cell Biology*, **31**(7), 1303-1313.