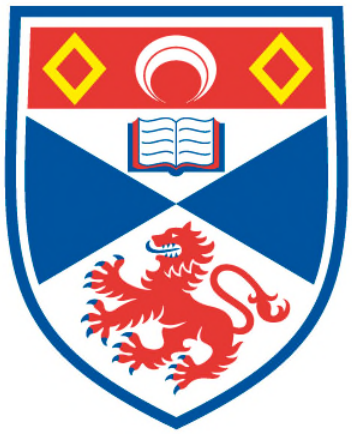
OCCURRENCE & FUNCTION OF CELLULAR 2A SEQUENCES

Claire Roulston

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews



2015

Occurrence & Function of Cellular 2A Sequences

Claire Roulston

This thesis is submitted in partial fulfilment for the degree of

PhD at the

University of St Andrews

2015

1. Candidate's declarations:

I, Claire Roulston, hereby certify that this thesis, which is approximately 60000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student in September 2010 and as a candidate for the degree of Doctor of Philosophy in Molecular Biology in September 2011; the higher study for which this is a record was carried out in the University of St Andrews between 2011 and 2015.

Date …… Signature of Candidate ……

2. Supervisor's declaration:

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of Doctor of Philosophy in Molecular Biology in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date …… Signature of Supervisor ……

3. Permission for publication:
In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. I have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

PRINTED COPY
a) Embargo on all or part of print copy for a period of one year on the following ground(s):
    - Publication would preclude future publication

Supporting statement for printed embargo request:

I intend to publish papers using the material presented in Chapters 3-8.

ELECTRONIC COPY
a) Embargo on all or part of electronic copy for a period of one year on the following ground(s):
    - Publication would preclude future publication

Supporting statement for electronic embargo request:

I intend to publish papers using the material presented in Chapters 3-8.

Date …… Signature of Candidate …… Signature of Supervisor ……

‘It is best to prove things by actual experiment; then you KNOW; whereas if you depend on guessing and supposing and conjecturing, you never get educated. Some things you CAN'T find out; but you will never know you can't by guessing and supposing: no, you have to be patient and go on experimenting until you find out that you can't find out. And it is delightful to have it that way, it makes the world so interesting.’

Eve's Diary – Mark Twain, 1905

Dedication

This thesis is dedicated to the memory of George Rae, my honorary Grandfather, for literally showing me how to appreciate the ground beneath my feet and the wonders that can be encased in a rock-pebble. For the fossils, the good advice, and the days out in the beat-up old green Volvo. Without you, I would not be a Scientist, with love always, Claire.

Acknowledgements

First and foremost I would like to take this opportunity to thank my supervisor Martin Ryan for initially granting me asylum when I was a refugee doctoral student without a lab, and then for providing unceasing support and patiently sharing his expertise over the last three years.

I'd like to thank all the members of the Ryan lab, both past and present, for all your help, encouragement and laughter. Extra special thanks are due to Dr. Garry Luke for teaching me the basics of cloning technique, John Nicolson for microscopy, and to Dr. Ekaterina Minskaia for just about everything else laboratory-based. Also to Fiona Tulloch, Ashley Pearson and Séan Wilson for all the office banter. I'll miss you all. Thank you.

I'd like to thank all the people round and about the BMSRC building who helped out with their expertise, spare reagents, free cake and general advice and support. Special thanks to Dr. Catherine Botting for mass spec analysis, and to Dr. Jens Tilsner for undertaking the plant infections. Thank you also to Dr. Lars Brunner at SAMS for gifting me the sea-urchin eggs.

To all my friends who supported me, you know who you are, and thank you. Special thanks to Derek Simpson for all the impromptu outward bound courses. To Kay and Jennifer Drysdale, for merely being yourselves, and for loaning out *Bluebell*. To all my shipmates at the JST, particularly Iain, Sean, Steve, Mo and Graham for all the adventures and for teaching me to counter adversity with grace and humour, and to never ever give up, even come hell and high water. I've seen you sail through it all with smiles on your faces and long may you continue to do so. To my little ships, *Swallow* and *Iolar*, and our crewmates: I wish you fair winds and calm waters and a safe harbour at voyage end.

Lastly, but very certainly not least, thanks to my family, particularly my parents, for all their love and support and belief in me (and taxi rides home!), and to my granddad, one hundred years young and full of life.

Abstract

This thesis describes experiments investigating the translational recoding activities and the novel dual signalling properties of eukaryotic ribosome skipping 2A sequences. Over twenty years ago, the 19 amino acid 2A region of a *Picornavirus*; namely, *Foot-and-Mouth Disease Virus* (*FMDV*) polypeptide was shown to possess apparent "self-cleaving" abilities, cutting at its own C-terminus during translation (Ryan *et al.*, 1991). Active *FMDV 2A*-like sequences were subsequently found in a number of related viruses (Luke *et al.*, 2008), with several now utilised as essential biotechnology multi-gene transfer tools (Luke *et al.*, 2010b). Then, in 2006, eukaryotic 2A-like sequences were identified from trypanosome non-LTR sequences. These were found to be functional *in vitro* (Heras *et al.*, 2006). I have been able to identify over 400 putative eukaryotic 2A-like sequences through searching the freely available online proteomic and genomic databases. Data is presented to show that these 2As were encoded in frame with non-LTRs, or metabolic, or immune function genes, from a wide range of eukaryotic organisms; but I could not discern any obvious phylogenetic distribution for 2A. I have discovered that the majority of eukaryotic 2A sequences tested can mediate ribosome skipping *in vitro*. Modelling *in silico* indicated that active 2A-like sequences possessed the propensity to form a central alpha-helical region, whereas the models suggested that inactive 2A-like sequences would be essentially unstructured. I also report that some of these eukaryotic 2A peptides constitute a novel form of dual protein targeting as they play a dual role as exocytic pathway signal peptides mediating extracellular protein trafficking. I have shown that this protein trafficking ability is evolutionarily conserved, with an echinoderm sequence able to direct protein targeting in both plant and mammalian cells. I therefore propose that these novel eukaryotic 2A sequences could potentially become extremely valuable in biotechnological engineering.

Table of Contents

Table of Contents (continued)

Table of Contents (continued)

Table of Contents (continued)

Table of Contents (continued)

Table of Contents (continued)

Table of Contents (continued)

List of Figures

List of Figures (continued)

List of Figures (continued)

List of Tables

List of Abbreviations

aa – amino acid
ank/ankyrin - ankyrin repeat protein
BFA - Brefeldin A
BHK21 - baby hamster kidney cells
bp - base pair
DAPI - diamino phenylindole (double-stranded nuclear DNA stain)
°C – Degrees Centigrade
DMEM - Dulbecco's modified eagle medium
DMSO - dimethyl sulfoxide
DNA - deoxyribonucleic acid
eEF2 - eukaryotic elongation factor 2
eGFP - enhanced green fluorescent protein
eIF4 - eukaryotic initiation factor 4
eRF1 & 3 - eukaryotic release factors 1 and 3
FCS - foetal calf serum
GUS - beta-glucuronidase
HGT - horizontal gene transfer
IPTG - isopropyl β-D-1-thiogalactopyranoside
kb - kilobase (of DNA or RNA)
LB – Lysogeny broth/agar
mCherryFP - cherry fluorescent protein
[^{35}S]-Met - radiolabelled [^{35}S]-methionine
MOI - multiplicity of infection
MW - molecular weight
NLR - Natch-like repeat/receptor protein (innate immune protein)
Non-LTR - non-long terminal repeat retrotransposon
ORF - open reading frame
PBS - phosphate buffered saline
PCR - polymerase chain reaction
PEG polyethylene glycol (used here as a transfection reagent)
PTC - peptidyltransferase centre
RNA - ribonucleic acid
rpm - revolutions per minute
SDS-PAGE - sodium dodecyl sulphate polyacrylamide gel electrophoresis
SNAT - Sodium-dependent amino acid transporter protein (membrane transporter protein)
TAE - Tris base, acetic acid and EDTA buffer
TE - transposable element
Tm - melting temperature
TnT - cell-free coupled transcription/translation reaction
UV - ultraviolet
v/v - volume for volume
w/v - weight for volume
XGal - 5-bromo-4-chloro-3-indolyl-beta-D-galactopyranoside

Chapter 1. Discovery of 2A-Like Sequences

'Every thing must have a beginning ... … and that beginning must be linked to something that went before.'

Frankenstein; Or, The Modern Prometheus - Mary Wollstonecraft Shelley, 1823 edition

1.1 **Viruses**

Viruses are obligate intracellular parasites consisting of infectious particles of genetic material (RNA or DNA) encapsulated by a protein coat which, in many cases, is surrounded by a lipid membrane (Figure 1.1). The discovery of viruses as infectious entities was the result of independent observations by a small number of researchers in the final decade of the 19th century (reviewed in Lustig and Levine, 1992; Bos, 2000; Lecoq, 2001). They concluded that a plant ailment – namely tobacco mosaic disease, was caused by an agent smaller than any known bacterium and too small to visualise using light microscopy. This agent could be filter-isolated from infected plant sap, but it could not replicate outside of its host's tissue. This agent, now known as *Tobacco Mosaic Virus* (*TMV*) was the first identified virus. The first such filterable agent from animals, *Foot-and-Mouth Disease Virus* (*FMDV*) was identified in 1898, followed shortly, in 1901, of the first isolation of a virus infecting humans, *Yellow Fever Virus*. With these discoveries the concept of viruses, infectious filterable particles as disease causing agents began to gain general acceptance.

Biochemical and electron microscopy analyses in the mid-20th century revealed that viruses were essentially nucleic acid and protein complexes, sized on the order of tens to hundreds of nanometres in diameter. Deficient in the translational machinery necessary for their replication, they are obligate parasites of cellular organisms. There have been numerous attempts to classify viruses based on genome composition, host range, biochemical characteristics or mode of replication. The Baltimore classification divided viruses into seven categories based upon genome type and mode of replication (Baltimore, 1971) and has been in widespread use from the 1970s until the present, but it is now superseded by the International Committee on Taxonomy of Viruses (ICTV, homepage http://ictvonline.org/) nomenclature system, also based on virus genome type, which further identifies viruses using a Linnaean system where viruses are placed in orders, families, and genera in a similar manner to that used to classify cellular life.

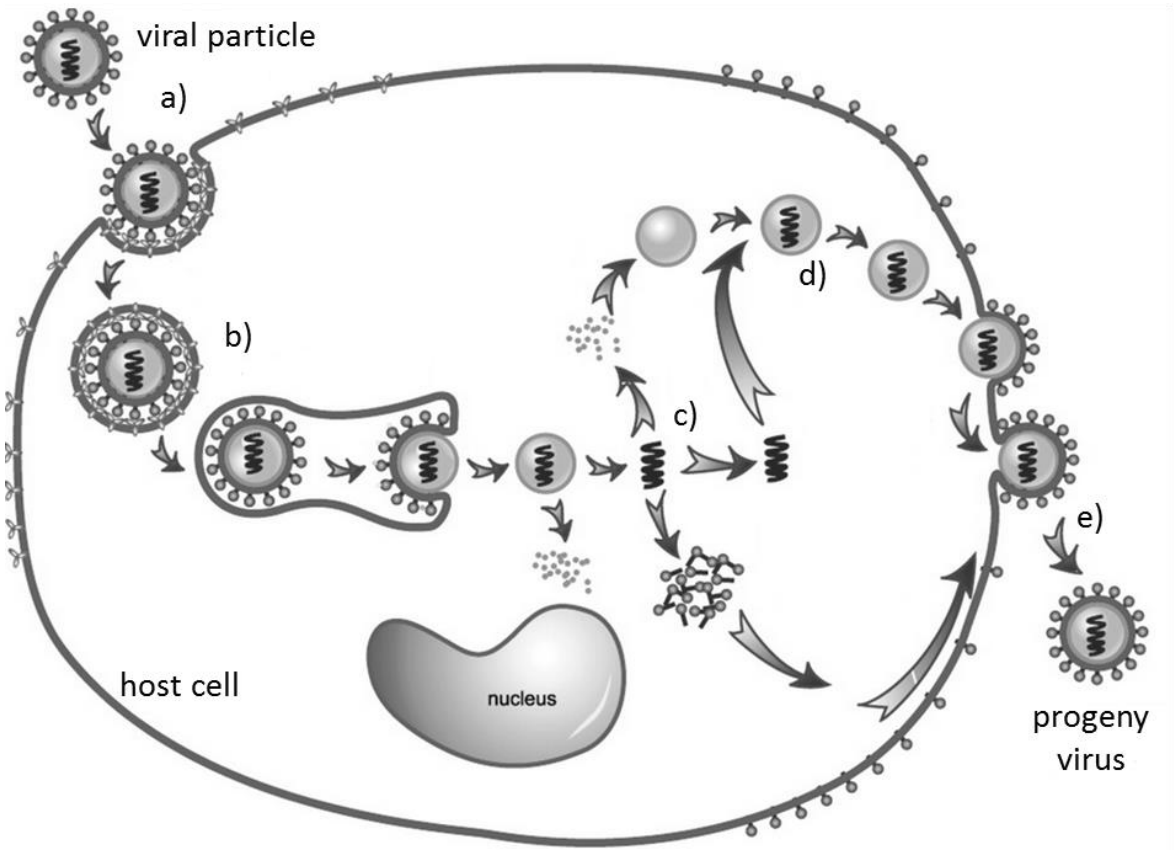


Figure 1.1 Characteristic life-cycle of a eukaryotic cell infecting virus
(a) Adsorption or docking with host cell surface receptor protein, **(b)** Entry into cell cytoplasm, **(c)** biosynthesis of viral components (proteins & nucleic acid), **(d)** assembly into new viruses, and **(e)** budding from the host cell (image modified from Gao *et al.*, 2005).

1.2 **Positive Sense Single-Strand RNA Viruses – Translational Tricks**

Viruses rely on "hi-jacking" host cell machinery to complete their replication cycle; therefore, virus-encoded proteins are translated by host cell ribosomes (see Section 1.3) in a similar manner to cellular proteins. In the case of positive sense RNA viruses their genomic RNA is treated comparably to host mRNA with cytoplasmic protein translation. All RNA viruses have extremely compact genomes, the largest being only around 30 kilobases (kb), while most are under 10 kb (Firth and Brierley, 2012). From these small genomes they must encode all the proteins necessary for their reproduction. Typically each positive sense RNA viral genome encodes one or more precursor polyproteins that are both cleaved and modified into mature structural and replicative viral proteins. However, most viral genomes encode only a single copy of each viral protein (Palmenberg, 1990), whereas the relative quantities required for a successful infection vary with infection stage (at the start, replicative proteins and immune system interacting proteins are required; later, high levels of structural proteins for packaging). Therefore, one of the major challenges for these viruses is overcoming the 5' end dependence of canonical eukaryotic

translation where typically only a single protein product can be produced from each mRNA strand (Firth and Brierley, 2012).

Whilst the central dogma of molecular biology stated that a single message or gene (DNA) encoded a single messenger (RNA), translated into a single product (protein) (Crick, 1970), it is now known that for some genes, the situation can be considerably more complex, with translational recoding mechanisms permitting a single gene to encode multiple protein products. Positive sense RNA viruses employ a number of translational recoding mechanisms (Figure 1.2) which successfully subvert the usual ribosome translation of mRNA. In non-canonical initiation the translation begins at a site other than the canonical AUG codon immediately downstream of a 5'UTR with a methylated 5' cap protein.

One method of non-canonical initiation makes use of Internal Ribosome Entry Sites (IRES, see Figure 1.2b i). First discovered in picornaviruses in the late 1980s, these are highly structured RNA regions that recruit ribosomes to internal positions on mRNAs (Jang *et al.*, 1988; Pelletier and Sonenberg, 1988). Viral IRESes are often employed to instigate translation initiation while allowing replication elements and/or packing signals to be accommodated within the 5'UTR. They can be used to access internal otherwise untranslated ORFs, and can also facilitate viral mRNA translation when host-cell translation has been inhibited, for example by viral proteases cleaving the initiation factors required for 5' cap-dependent translation (reviewed in Firth and Brierley, 2012). Viral IRESes are highly variable, both in the degree to which they depend on host cell initiation factors and in their precision of initiation site selection. IRESes are presently grouped into several classes based on their initiation mechanisms (reviewed in in Firth and Brierley, 2012). In recent years viral IRESes have been successfully used to encode multiple proteins from a single vector in biotechnological applications (for a recent review see Minskaia *et al.*, 2015): the first gene being expressed from the 5' cap and the second from an internal IRES.

IRESes instigate 5'-independent internal translation initiation whereas another form of non-canonical translation, namely ribosome shunting (Figure 2.1b ii), permits ribosomes to translate downstream ORFs in a manner that is only partially 5' dependent. Here alternate translation products are produced when secondary structure in the 5' UTR results in translation starting at a downstream AUG codon (reviewed in Firth and Brierley, 2012).

Leaky scanning is also 5' cap-dependent, here a significant proportion of the 40 ribosome subunits (see Section 1.3, to follow) that bind near the 5' cap and scan downstream on the mRNA until they reach an appropriate AUG codon, will fail to initiate translation at the first AUG codon they encounter. These ribosomes continue scanning until they reach an alternate downstream initiation codon where translation commences (reviewed by Kozak, 2002). Leaky scanning results in multiple isoforms of the translation product, some N-terminally truncated (Figure 1.2b iii).

Non-AUG initiation is similar to leaky scanning in that it results in multiple isoforms of the translation product, some N-terminally truncated (see Figure 1.2b iv). In eukaryotes, protein synthesis begins almost exclusively with methionine, encoded by AUG. The tRNA carrier Met–tRNA*i*, delivers the initial methionine. Met–tRNA*i* differs slightly from the standard Met-tRNA used during elongation. Under certain circumstances Met–tRNA*i* can be recognised by near-cognate codons such as CUG or ACG. Initiation at a non-AUG codon is enhanced by particular flanking nucleotides, and by secondary structure (typically stemloops) forming close to the exterior of the ribosome. Non-AUG initiation is typically inefficient, occurring only in in 2-30% of cases, in the remainder of instances leaky scanning mechanisms permit the ribosome to scan downstream until it encounters the next AUG codon which is treated as a start codon (reviewed in Firth and Brierley, 2012).

Another non-canonical initiation process used by positive sense RNA viruses is re-initiation (Figure 1.2b v). This depends on the ribosome 40S subunit (see following, Section 1.3) remaining associated with the mRNA after protein translation ceases (normally the subunits would disassociate after translation termination) and resuming scanning to re-initiate translation at a downstream AUG codon. In cellular genes, re-initiation most commonly occurs after a short (less than 30 codons in length) upstream ORF, but in some viruses (for example, in the *Calicivirus* and *Sapovirus* genera) re-initiation occurs after a translation of a longer upstream ORF. Re-initiation is dependent on initiation factors (see following, Section 1.3) remaining attached to the ribosome during translation (hence why it is more common after short upstream ORFs as there is less time for the factors to disassociate). After the upstream ORF translation terminates, the 40S subunit is not immediately competent to reinitiate, but becomes competent after scanning downstream along the mRNA for some distance. This time spent scanning is thought to permit the reacquisition of the necessary initiation factors and the eIF2–Met–tRNA*i*–GTP ternary complex required for translation initiation. Re-initiation is also thought to rely on complex RNA structure in the region between ORFs that can delay the scanning and permit initiation factors re-binding, or may be used to retain the bound ribosome-initiation factor complex between translation of each ORF (reviewed in Firth and Brierley, 2012; Jackson *et al.*, 2012).

In non-canonical elongation and termination, events during translation, or in the altered reading of termination signals can result in protein products other than those predicted by an in-frame reading of the mRNA sequence. Collectively such events are termed ribosome recoding.

Viral-directed ribosome recoding events include frameshifting, whereby a proportion of ribosomes are directed into an alternate reading frame by mRNA movement back or forwards by one or two nucleotides (Figure 1.2c i). Frameshifting -1 was discovered to be the mechanism by which the viral Gag-Pol polyprotein was produced from overlapping ORFs (Jacks and Varmus, 1985; Jacks

et al., 1988). Many positive-strand RNA viruses, most retroviruses, plus some *Totiviridae* dsRNA viruses use -1 frameshifting to express their RdRp or reverse transcriptase domains. Frameshifting is thought to rely both on a "slippery" mRNA sequence with repeated bases around the point of slippage, and on a strongly structured downstream RNA. It is thought that a delay in "unwrapping" the downstream RNA and preparing it for reading by the ribosome permits the one or two nucleotide slippage of the mRNA within the translating ribosome (reviewed in Firth and Brierley, 2012).

Another viral recoding mechanism is readthrough (Figure 1.2c ii), whereby a proportion of ribosomes fail to terminate at a stop codon but instead insert a standard amino acid (hence the term readthrough). Termination is generally highly efficient, but the efficiency is known to be influenced by the particular stop codon (whether UAA, UAG or UGA) and the flanking nucleotides, especially the immediately adjacent 3' nucleotide, and in some viruses secondary structure beginning eight nucleotides downstream in the mRNA sequence. During readthrough the stop codon is decoded by a near-cognate or suppressor tRNA and termination continues until the next stop codon is encountered. Readthrough mechanisms are used by some viruses to control expression of polymerase proteins or to add extensions to a subset of coat proteins (reviewed in Firth and Brierley, 2012).

Finally, another viral recoding mechanism is stop–carry on, also referred to as ribosome skipping (Figure 1.2c iii). It is the 2A peptides that instigate ribosome skipping that will form the focus of this thesis. 2A ribosome skipping will be described in detail (to follow, Chapter 1.9). 2A-mediated ribosome skipping differs from the other forms of non-canonical mRNA processing as it is believed to rely solely on the peptide sequence of the nascent protein rather than structural elements within the mRNA sequence in order to instigate ribosome recoding.

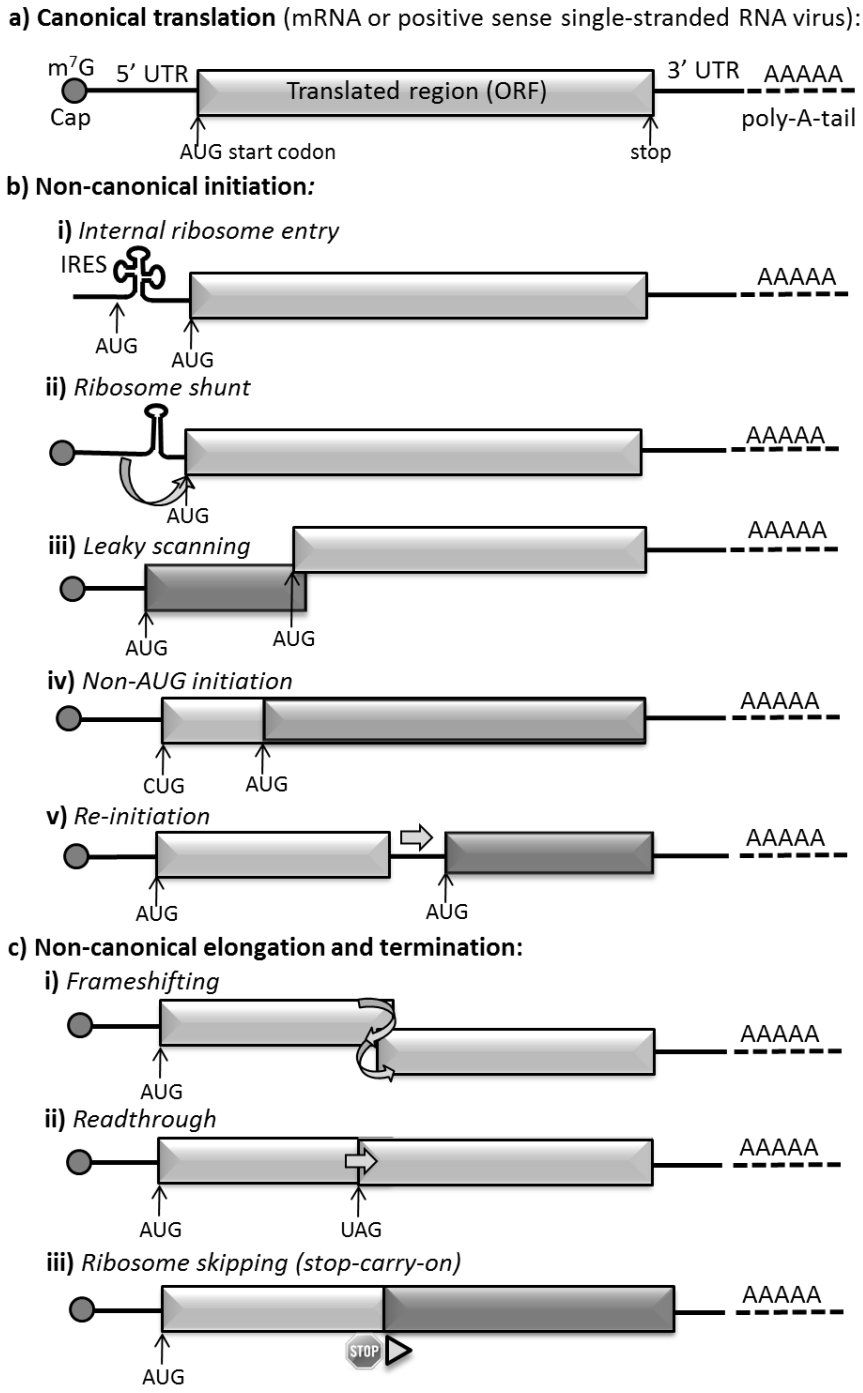


Figure 1.2 Non-canonical mRNA processing by positive sense RNA viruses
a) Canonical protein translation, (included for comparison). Examples of positive sense RNA virus non-canonical protein translation: **b)** non-canonical initiation mechanisms, **b i)** use of an internal ribosome entry site (IRES), an mRNA stem-loop structure which permits ribosome initiation complex binding without the involvement of the 5' canonical cap. **b ii)** and **b iii),** ribosome shunt and leaky scanning, respectively, in these cases translation begins downstream of the expected AUG. **b iv)** Non-AUG initiation, translation can start at non-AUG codons (typically CUG). **b v)** re-initiation of translation at a downstream AUG codon. **c)** Non-canonical elongation and termination can result from: **c i)** frameshifting in mid-translation altering the expressed reading frame, and **c ii)** read-through of stop codons and **c iii)** ribosome skipping to produce two proteins from one ORF. This thesis will focus on 2A-mediated ribosome skipping. Diagram after Firth and Brierley, 2012.

1.3 **Ribosomes**

In order to describe the process of ribosome skipping, it is first necessary to introduce briefly eukaryotic ribosomes. Ribosomes are the protein factories of the cell, they are large complexes of RNA and proteins that "read" messenger (mRNA) and manufacture new peptide chains. They are essentially ribozymes, as it is their RNA components that act to catalyse the formation of new peptide chains, whereas their protein component acts to stabilise the structure. In the last few years, high resolution cryoEM and crystallography studies (Armache *et al.*, 2010; Demeshkina *et al.*, 2010; Rabl *et al.*, 2011) have provided detailed ribosomal models (Figure 1.3).

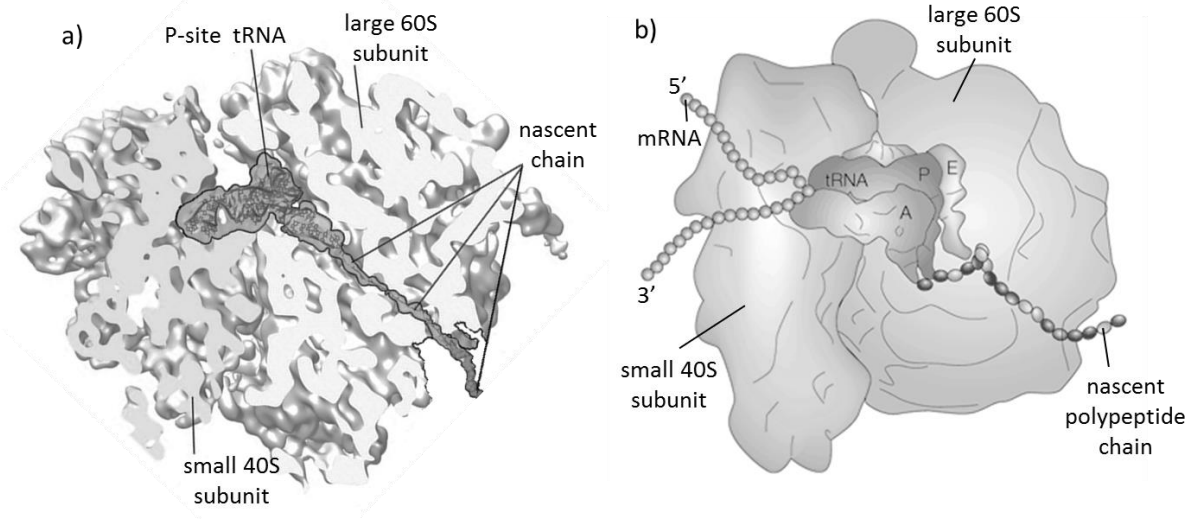


Figure 1.3 Cross-section through a translating eukaryotic ribosome
a) Structure of a translating wheat-germ (*Triticum aestivum*) ribosome illustrating the relative positions of the ribosomal subunits, the P-site tRNA in the PTC, and the nascent polypeptide (modified from Becker *et al.*, 2009), **b)** conceptual model which additionally shows the relative position of the translated mRNA and the A and E-site tRNAs (modified from Lafontaine and Tollervey, 2001).

It is now known that both eukaryotic and prokaryotic ribosomes comprise two subunits (termed large and small), each composed of ribosomal RNA (rRNA) and proteins. Prokaryotic 70S ribosomes comprise ~4500 nucleotides ribosomal RNA (rRNA) and 54 proteins, their large 50S subunit contains two RNAs (23S and 5S) and 32 proteins whilst their small subunit possesses a single 16S RNA and 22 proteins. Eukaryotic ribosomes are substantially larger than prokaryotic, for example, mammalian ribosomes are approximately 50% larger than those from *E. coli* bacteria. Eukaryotic ribosomes comprise ~5500 nucleotides rRNA and 80 ribosomal proteins with their large 60S subunits containing three RNAs (28S, 5.8S and 5S), and their small 40S subunit one 18S RNA. The increased size and complexity of eukaryotic ribosomes is due to the presence of additional rRNA expansion sequences (rRNAs that do not form part of the core ribosome structure) and eukaryotic specific proteins (Armache *et al.*, 2010). Ribosomes from different eukaryotic

lineages also differ, typically mammalian ribosomes are 12% larger than yeast ribosomes; this increase is due to additional rRNA expansion sequences (Morgan *et al.*, 2000).

However, the mechanics of protein translation at the central active site of the peptidyltransferase centre (PTC) is extremely conserved across all domains of life. First, the ribosome binds to the mRNA and translation initiates at a start (methionine, nucleotides ATG) codon. A supply of the relevant amino acids are transported to the ribosome 3' attached to transfer RNAs (tRNAs), a class of small RNA molecules. Each tRNA is specific for a single amino acid. The ribosome possesses three tRNA binding sites within the PTC. Firstly, a tRNA with its specific attached amino acid (aminoacyl-tRNA) encounters the A-site where it is presented to the mRNA being decoded. Slight changes in ribosome conformation then shunt the tRNA into the peptidyl site (P-site), where it and its amino acid burden are attached to the C-terminal end of the growing peptide chain as a peptidyl-tRNA complex. The tRNA-amino acid link is then broken and the deacylated "empty" tRNA moves into the exit site (E-site) before exiting the PTC, and the ribosome moves forward on the mRNA to read the next codon to be translated. Translation is an iterative process with one amino acid residue being added to the chain at each step, at a rate of approximately 15–20 residues per second in bacteria (reviewed in Lafontaine and Tollervey, 2001).

Unsurprisingly, given the high level of sequence and structure conservation in ribosomal RNA and proteins between organisms, ribosome synthesis is also very highly conserved. In virtually all living cells, rRNAs are generated by post-transcriptional processing from a polycistronic precursor rRNA (pre-rRNA). The rRNAs are then assembled along with the ribosomal proteins to form the ribosome complex. Events during assembly must follow a strict temporal order in order to permit the correct folding and modifications of all components. Deviations, either in component composition, modification, or assembly order will normally result in mis-folding and inhibit assembly (see Lafontaine and Tollervey, 2001). Ribosome functional impairment is normally lethal to the cell, hence the extremely conserved structure of the ribosome in all living cells.

1.4 **Ribosome Exit Tunnel**

Nascent polypeptide chains exit the ribosome through the exit tunnel. This comprises a hollow tube approximately 80-100 Å in length and 10-20Å in diameter (Figure 1.4), its walls predominately composed of rRNA (82% rRNA atoms, 18% protein atoms, respectively). Generally the tunnel walls are negatively charged and hydrophilic (Voss *et al.*, 2006; Bhushan *et al.*, 2010). At one point the "arms" of two ribosomal proteins, L4 and L17 (L22 in bacteria), form a constriction reducing the passageway width to 10Å. The tunnel dimensions preclude extensive folding of the peptide nascent chain. However, it can accommodate 30-40 peptide residues (first experiments Malkin and Rich, 1967; confirmed by Bernabeu and Lake, 1982; revisited Voss *et al.*, 2006), whereas a fully extended chain would contain only 20 residues. These observations led to the

hypothesis that the nascent chain adopts a partial α-helical confirmation within the tunnel. Recent fluorescence resonance energy transfer (FRET) (Woolhead *et al.*, 2004) and cryo-EM (Bhushan *et al.*, 2010) studies confirmed that nascent peptides do indeed form an α-helix within certain tunnel regions.

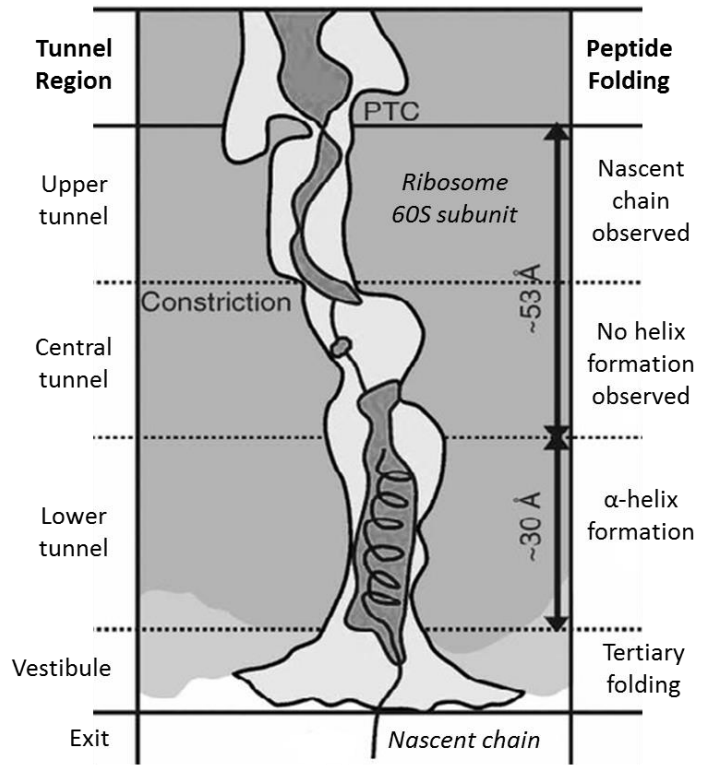


Figure 1.4 Eukaryotic ribosome tunnel – potential for nascent chain secondary structure?
Model of the wheat-germ (*Triticum aestivum*) ribosome tunnel displaying the exit pathway of the polypeptide nascent chain and the regions where some helical folding can occur (modified from Bhushan *et al.,* 2010).

1.5 **The Ribosome Tunnel & Translational Recoding – Stalling Peptides**

The constricted environs of the ribosome tunnel not only determine the permissible nascent chain secondary structures which can form within its confines, but in forming into such structures, the nascent chain can interact with the tunnel walls, and in doing so influence events in the PTC to alter or halt peptide synthesis. Such events are termed translational recoding (recently reviewed by Ito *et al.*, 2010; Cruz-Vera *et al.*, 2011; Wilson and Beckmann, 2011; for examples see Table 1.1). The specific amino acid sequence of each nascent peptide influences its rate and ease of transit through the ribosome tunnel. Peptide geometry and mass contribute, but the principle governing factor is electrostatic force, particularly nascent peptide charge and hydropathicity (Lu and Deutsch, 2008). The tunnel constriction site proteins (Figure 1.4), the vestibule exit protein L39 (bacteria L23), and rRNA residues adjacent to each of these proteins, have also been identified as essential players. These tunnel wall proteins can interact with specific residues within the nascent chain (see Table 1.1) resulting in the chain transitorily stopping within the tunnel. Chain stoppage

causes slight conformational changes in the P- and/or A-sites which block PTC activities including peptide bond formation and/or tRNA movement. Translation ceases releasing the tRNAs and the nascent peptide, then, subsequent mRNA rearrangement may permit translation re-initiation from an ORF downstream to that encoding the stalling peptide.

The majority of these stalling peptides rely on co-effector binding either with the tunnel vestibule or within the lower reaches of the tunnel to create the particular peptide conformation that causes stalling (Figure 1.5). Macrolide antibiotics provide the co-effectors for the most bacterial stalling peptides, and the downstream ORF that is later translated encodes an antibiotic resistance factor, thus providing the host with antibiotic resistance when antibiotic is present (reviewed in Tenson and Ehrenberg, 2002; Ito *et al.*, 2010). This provides an evolutionary fitness advantage to the carriers as they do not have to expend energy producing resistance factors constitutively, instead, they can divert resources to producing the resistance factors only when and if they are required. These bacterial antibiotic resistance genes tend to be carried on plasmid DNA and are responsible for the growing phenomena of drug-resistance (reviewed by Ito *et al.*, 2010). Where the stalling peptide functions to regulate cellular metabolic pathways, high intracellular levels of a constitutively expressed molecule can act as the co-effector, such as tryptophan in the case of tnaC, and arginine for the fungal arginine attenuation peptides, respectively (Fang *et al.*, 2004). Or, more rarely, the specific amino acid sequence of the stalling peptide can be sufficient to stall the ribosome without the aid of a co-effector molecule.

A number of the stalling peptides end with proline (see Table 1.1) or possess proline residue(s) close to their C-terminus. It has been suggested (Jenni and Ban, 2003) that the unique properties of proline (the only naturally occurring N-alkylamino or "imino" acid in proteins) could contribute to the ribosome stalling in these instances as there is a greater energy barrier to proline, an imino acid, forming peptide bonds than with the other naturally occurring 19 amino acids, making proline slower to bond (Pavlov *et al.*, 2009).

Each stalling peptide occurs in a distinct monophyletic (single ancestor) group of organisms, or in a single organism, suggesting that while they share a common function, namely ribosome arrest, stalling peptides are an example of convergent evolution with separate origins for each sequence (Ito *et al.*, 2010; Cruz-Vera *et al.*, 2011; Wilson and Beckmann, 2011). Additionally, every stalling peptide discovered to date can only function in either prokaryote or eukaryote ribosomes, suggesting that subtle differences in tunnel topography and characteristics may influence nascent chain–tunnel interaction.

Table 1.1 Example ribosome stalling nascent peptides
The peptide name, its host organism and the peptide sequence are given in each instance. Where specific residues are known to be essential to function these are been underlined. If the peptide requires a co-effector to function, this is also listed.

Peptide	Host organism	Active stalling peptide sequence(s)	Co-effector
Bacterial			
cat, cmlA	*Salmonella spp.* *Enterobacter spp.* *Pseudomonas spp.*	VKTD KNAD	chloramphenicol
ermC, ermCL	*Enterococcus spp.*	SFVI MxxxxIFVIs	erythromycin
tnaC **tnaC Ec** **tnaC Pv**	*Escherichia coli* & *Proteus vulgaris*	KWFNID WxxxDxxIxxxxP* WxxxDxxLxxxxPK	tryptophan
secM **secM Ms**	*Escherichia coli*	FxxxxWIxxxxGIRxGP xxxxxxxxxxHAPIRGSP	membrane translocation (SecA)
MifM Bs	*Bacilus subtilis*	RIxxWIxxxxxMNxxxxxxxxx	
Fungal/Yeast			
CPA1	*Saccharomyces cerevisiae*	NSQYTCQDYISDHIWKTS	arginine
arg	*Neurospora crassa*	PSxFTSQDYxSDHLWxAx	
Mammalian			
AdoMetDC	mammals	MAGDIS	spermidine, spermine
B2-AdRec	mammals	MKLPGVRPRPAAPRRRCTR	No co-effector
RAR- B2	mammals	MIRGWEKDQQPTCQKRGRV	
CMV UL4	mammalian cytomegaloviruses	MQPLVLSAKKLSSLLTCKYIPP	

Note: underlined residues are essential to function, x=any amino acid. *=unknown residue, (information from Tenson and Ehrenberg, 2002; Ito *et al.*, 2010)

Another class of translational recoding peptide which stalls the ribosome prior to translation re-initiation was discovered in the *Picornaviridae* virus family in the early 1990s (Figure 1.2 and Figure 1.5). Termed 2A ribosome skipping peptides, these are the sequences that form the focus of this PhD investigation. The mechanism of activity of ribosome skipping 2A will be addressed in detail in Chapter 1.8, but first, protein production in the *Picornaviridae* will now be outlined with a focus on the role of 2A in these viruses.

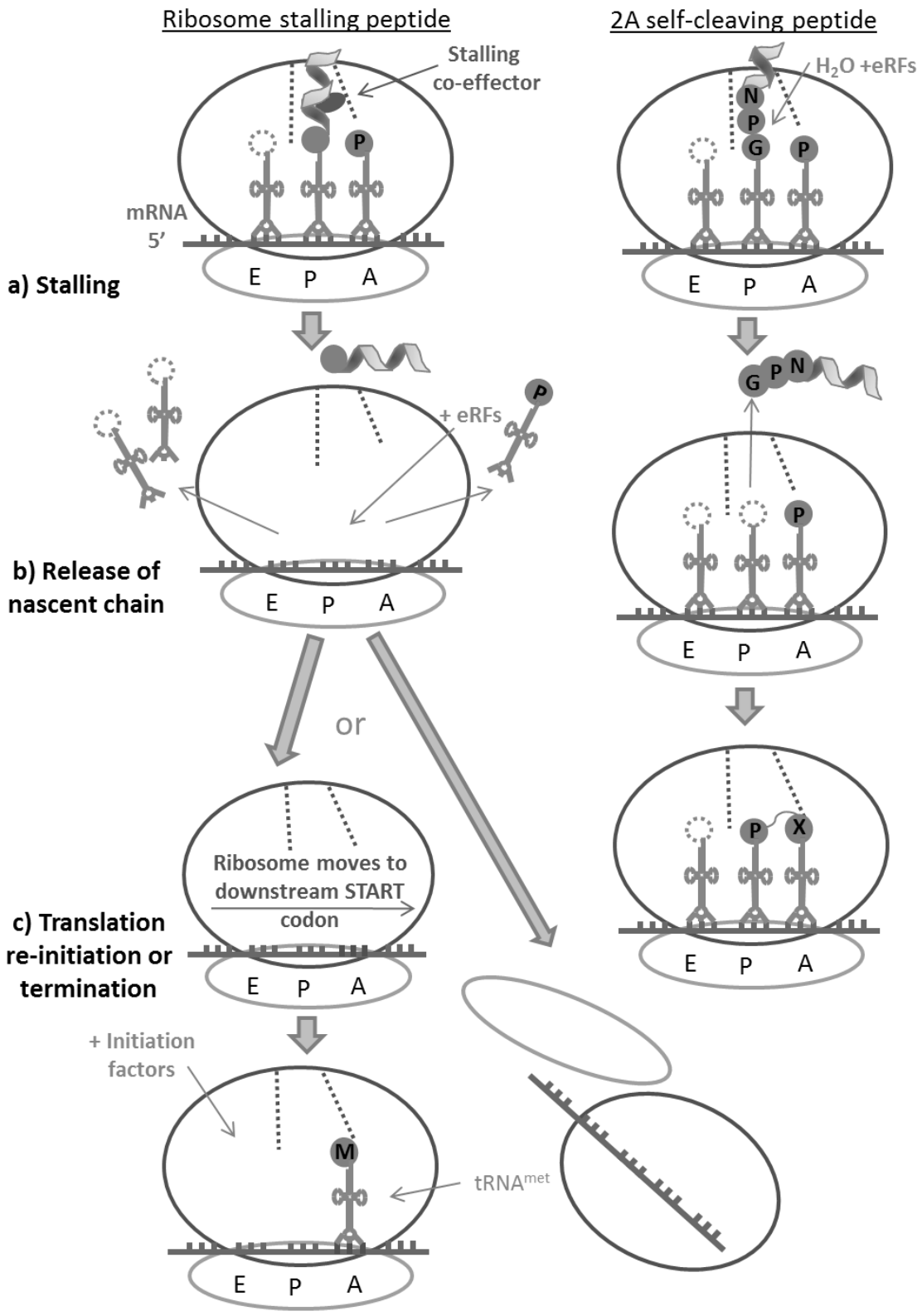


Figure 1.5 Ribosome stalling & 2A skipping peptides contrasted
a) Most stalling peptides require co-effectors (Table 1.1), 2A stalls autonomously. **b)**, Stalling peptides release the nascent chain and tRNAs. In 2A, nascent chains are released but tRNAs remain in the PTC **c)** Stalling peptides cause termination, followed by translation re-initiation from a distal start codon or ribosome subunit disassociation. In 2A translation can recommence with the proline residue as the first amino acid (Information from Ryan *et al.*, 1999; Donnelly *et al.*, 2001b; Doronina *et al.*, 2008b; Ito et al., 2010; Cruz-Vera *et al.*, 2011; Wilson & Beckmann, 2011).

1.6 **The *Picornaviridae***

Under the ICTV system the *Picornaviridae* virus family is placed in the order *Picornavirales*. The *Picornaviridae* are a large and varied family that contains many socially and economically significant human and animal pathogens including *Polio*, *Hepatitis A*, the *Common Cold Virus*, and the cattle disease *FMDV*. Their classification has recently undergone a major reorganisation due to the inclusion into the taxonomic system of numerous newly identified *Picornavirus* and *Picorna-like Viruses*, with a current total of 46 known picornavirus species grouped into 26 genera (Knowles, 2012; Adams *et al.*, 2014).

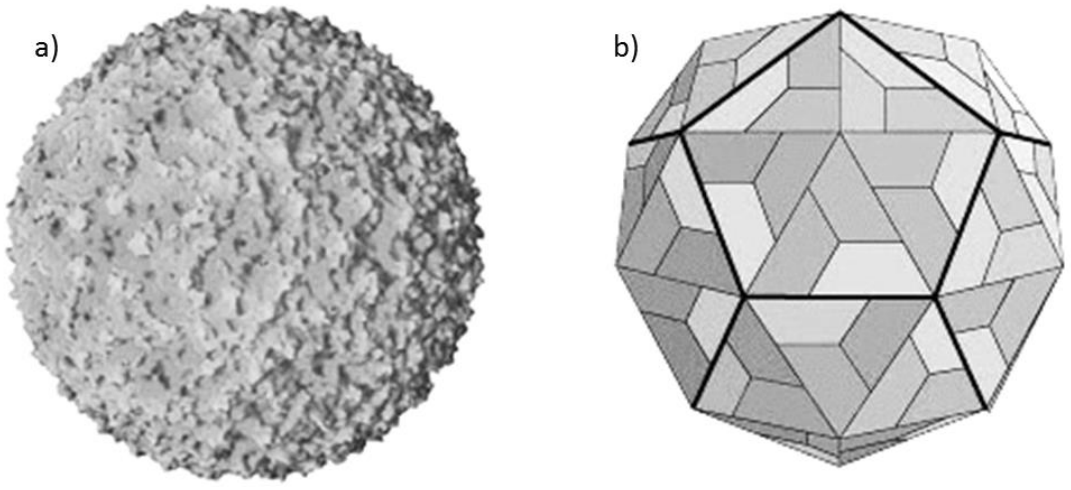


Figure 1.6 Picornavirus virion
a) Electron microscopy image of the outer view of a *FMDV* virion, **b)** schematic diagram detailing the icosahedral arrangement of the protein subunits composing the *FMDV* virion capsid (images courtesy of Prof. Martin Ryan).

As their name suggests, these viruses possess small (pico = small) RNA genomes. Their single-stranded positive sense RNA genome is typically between 7500-8000 bases in length and is organised in a single central open reading frame (ORF), flanked by highly structured 5' and 3' untranslated regions (UTRs). Picornavirus replication is thought to be entirely cytoplasmic, with the viral genomic RNA utilising the translational machinery of infected cells in a similar manner to host cell mRNAs. They encode their own RNA-dependent RNA-polymerase to enable replication of their RNA genome. Their single ORF encodes a single precursor polyprotein that is co-and post-translationally cleaved to form the structural (capsid) and replicative elements of the virus. The virion particle is composed of a single copy of the RNA genome encapsidated within an non-enveloped icosahedral capsid (Figure 1.6) composed of 60 protein subunits each consisting of 4 non-identical proteins (240 proteins in all, surrounding each RNA strand) (Palmenberg, 1990).

1.7 **Picornavirus Translation**

Picornavirus RNA is capped by an oligopeptide (Vpg or 3B) covalently attached to the 5' terminus, and the 3' end terminates in a poly(A) tract. The 5' UTR folds into a highly structured clover-leaf configuration which possesses an internal ribosome entry site (IRES). The IRES facilitates cap-independent translation and allows the virus to effectively halt host-cell protein synthesis and to henceforth "hijack" the translational apparatus to synthesise their own proteins (reviewed in Martinez-Salas, 2008; Martinez-Salas *et al.*, 2008). The ORF is translated to form a large precursor protein of approximately 250kDa, but this protein is rarely observed in cell-culture as it is co- and post-translationally cleaved to form the active structural and regulatory viral proteins. Co-translational cleavages by viral proteases 2A and 3C divide this polyprotein into three regions (P1, P2 and P3 (Figure 1.7.). P1 contains the four structural proteins (1A-1D) that form the virion capsid coat; both P2 and P3 contain the non-structural proteins, 2A-2C, and 3A-D, respectively (Palmenberg, 1990, following the nomenclature proposed by Rueckert and Wimmer, 1984). The non-structural proteins facilitate viral replication, shut-down of cellular protein synthesis, and the re-arrangement of the cell membranes.

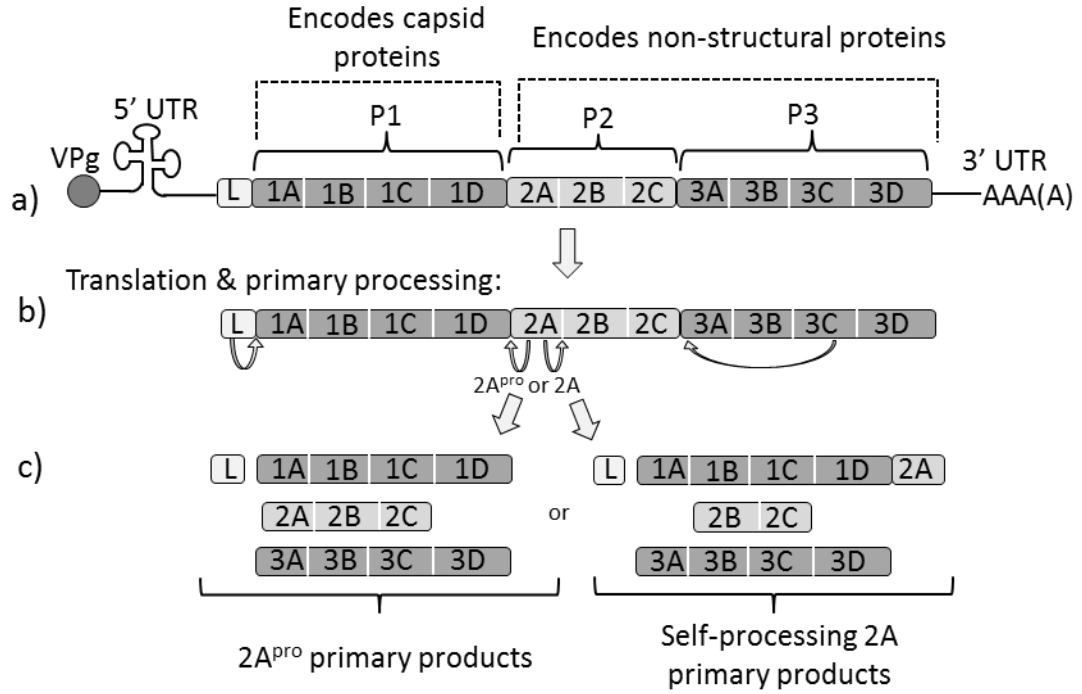


Figure 1.7 Picornavirus genome organisation, translation & primary polypeptide processing. Not to scale. **a)** RNA genome organisation, note not all picornaviruses possess a Leader (L) region **b)** Translation and primary processing of the polyprotein, **c)** note the difference in cleavage products between genera with a 2Apro or self-processing 2A region (information Palmenberg, 1990; Ryan and Flint, 1997).

1.8 **Picornavirus Translational Recoding 2As**

Virtually all picornaviruses encode a 2A protein, but the size, structure and function of 2A differs widely between genera. In the majority of genera, the 2A region apparently functions in primary polypeptide processing: for example, in *Enteroviruses* and *Rhinoviruses*, the 2A polypeptide is a thiol proteinase (termed $2A^{pro}$) which cleaves the polyprotein at the P1/P2 juncture in *cis*. Whereas, in the *Aphtoviruses* and *Cardioviruses*, the 2A region apparently self-cleaves at its own C-terminus, meaning that the 2A polypeptide remains as a C-terminal extension of the upstream polyprotein (P1) until it is removed by secondary proteinase cleavage (Ryan *et al.*, 1991; Ryan and Drew, 1994). However, in some other picornavirus genera (such as the *Parechoviruses*, *Kobuvirus* and *Megrivirus*), the 2A region has apparently no protease or protease-like activity, and instead its apparent function is to alter host cell metabolism as it possesses a high homology to cellular protein H-rev107 that regulates cell proliferation (H-box 2A) (Hughes and Stanway, 2000). The processes of recombination and the re-arrangement of genome segments have resulting in some picornaviruses possessing multiple 2A regions of various types (Table 1.2).

The mechanisms of thiol protease 2As and cell-cycle regulatory 2As were readily apparent. However, the elucidation of the functional mechanism of the non-protease "cleavage" 2As has taken over 20 years of careful step-wise investigations, primarily using *FMDV* 2A as the model system. Early investigations into the later proteolytic cleavage steps between *FMDV* 1D and 2A revealed that the 2A region in *FMDV* was only 18aa (-LLNFDLLKLAGDVESNPG-) in length (Belsham, 1993). This was considerably shorter than any than any known protease enzyme, but there was the possibility that the *FMDV* 2A-2B cleavage could have resulted from the activities of exogenous host cell proteases or of another virus encoded protease? Investigative studies conclusively demonstrated that this was not the case, as neither host cell proteases, nor the *FMDV* viral encoded proteases (namely $3C^{pro}$ or L^{pro}) cleaved at the 2A-2B boundary (Ryan *et al.*, 1989; Ryan *et al.*, 1991; Palmenberg *et al.*, 1992). It was also found that it was the amino acid not the nucleic acid sequence that was instrumental to function as synonymous mutations within the RNA sequence did not influence function (most recently revisited by Gao *et al.*, 2014).

Sequence comparison of the *Enterovirus* and *Rhinovirus* $2A^{pro}$ thiol proteases with the 2As of *Cardioviruses* and *Aphthoviruses* found that although their $2A^{pro}$ were of a similar length to the *Cardiovirus* 2A proteins, approximately 150 amino acids (aa) there was no apparent sequence similarity. However, the C-terminal region of the *Cardiovirus* 2As were found to be highly similar to the much shorter (approximately 20-30aa) 2A peptide of the *Aphthoviruses* (Donnelly *et al.*, 1997). This difference in 2A region sequence, coupled with dissimilarity to both to the *Entero*- and *Rhinovirus* $2A^{pro}$, led researchers in the late 1980s and early 1990s to further speculate that the co-

translational “cleavage” event identified at the end of the *Aphthovirus* and *Cardiovirus* 2As, occurred co-translationally through a novel non-protease mediated system (Palmenberg, 1990).

The next examinable hypothesis was that the *FMDV* 2A “cleavage” was due to the specific C-terminus amino acid sequence. The C-terminus tri-peptide from *FMDV* 2A (-NPG-) and the N-terminal residue of 2B (-P-), together formed a tetrapeptide motif (-NPGP-), which was found very rarely in natural proteins, making it almost unique to the 2A sequences. Therefore, it was suspected that this tetrapeptide might be the key to 2A function (Palmenberg, 1990). However, it was shown that the -NPGP- motif alone was ineffective at instigating 2A “cleavage” (reviewed in Luke *et al.*, 2010b).

Models of the *FMDV* 2A nascent peptide (Figure 1.8) suggest that it has the propensity to form into α-helix along most of its length with a tight reverse turn motif (-ESNPG-) turn at its C-terminus (Ryan *et al.*, 1999; Donnelly *et al.*, 2001b). Similar to ribosome stalling peptides (Table 1.1 & Figure 1.5), it was proposed that its unusual geometry could assist the *FMDV* 2A nascent chain to transiently pause (a pause was observed in puromycin incorporation experiments, see Donnelly *et al.*, 2001b) in its transverse of the ribosome exit tunnel. Indeed, the current favoured hypothesis is that 2A activity is a result of the particular geometry of the nascent 2A amino acid sequence, and its ability to interact with the ribosome exit tunnel (Ryan *et al.*, 1991; Donnelly *et al.*, 2001b; de Felipe *et al.*, 2003; Atkins *et al.*, 2007; Doronina *et al.*, 2008b). The supposition is that the N-terminus portion of 2A (the helix) might interact with the tunnel to obtain the specific stereo-chemical constraints required for the turn motif (-ESNPG-) to literally be in a position to influence events within the peptidyl transferase centre (PTC) of the ribosome. The nascent chain pausing in its exit of the ribosome tunnel halts translation and results in minute shifts in the peptidyl-tRNA^{Gly} position within the PTC, leading to a configuration whereby nucleophilic attack on the carbonyl group of the P-site peptidyl-tRNA^{Gly} by the A-site aminoacyl-tRNA^{Pro} amino group is prevented by the unfavourable energetics for peptide bond formation to the imino acid proline. The formation of a glycine-proline peptide bond is inhibited whereas the hydrolysis of the peptidyl-tRNA^{Gly} ester bond between glycine and its carrier tRNA is favoured.

It was proposed that release (termination) factors eRF1 and eRF3 contributed to the hydrolysis and release of the 2A nascent chain even although the proline codon was still occupying the ribosome A site (Doronina *et al.*, 2008a; Doronina *et al.*, 2008b). Normally, eRF1 & 3 can only form into a complex with elongating eukaryotic ribosomes whenever a stop codon enters the A-site. (Zhouravleva *et al.*, 1995), and it is this complex which permits hydrolysis of the ester bond between the final amino acid and its carrier tRNA. Therefore, in the case of 2A, the stalled ribosome-2A complex must promote eRF entry without reading the specific mRNA, as the eRFs must recognize the A-site proline codon as, in effect, the termination signal. Next, eRF complex

dissociation must occur to facilitate the entry of prolyl-tRNA to the A site of the ribosome. The aminoacyl-tRNA is then translocated from the A to the P site to become the initiating N terminus peptidyl-tRNA of the downstream nascent protein chain (Figure 1.9). However, a newly published study, using *in vitro* cell extracts, found that for *Encephalomyocarditis Virus* (*EMCV*) 2A-mediated translational recoding could occur even in the absence of eRFs (Machida *et al.*, 2014). In light of these recent findings, it will be interesting to determine whether eRFs play any role in the functioning of other viral 2As, or, if indeed, release factors are redundant for 2A processing. It is suspected that the concentration of eukaryotic elongation factors may determine whether, after hydrolysis, the ribosome dissociates or continues to translate the downstream context (Luke and Ryan, 2013).

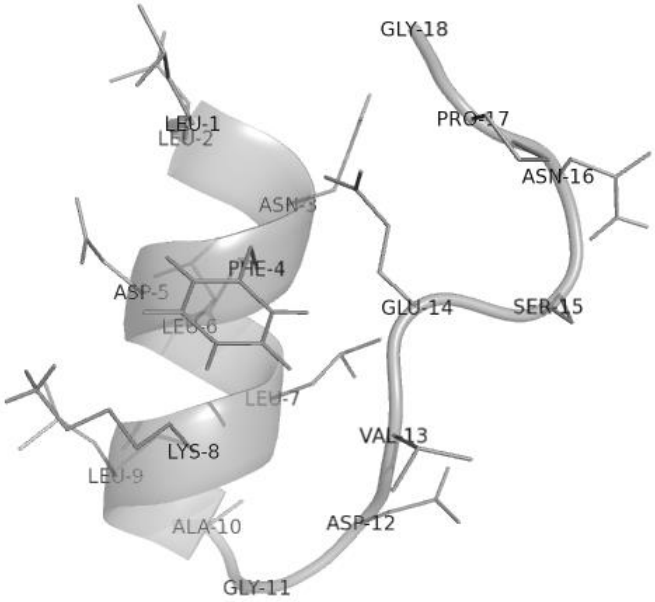


Figure 1.8 Model of *FMDV* 2A sequence.
The image shows the hypothetical configuration of *FMDV* 18aa 2A at neutral charge and pH in a non-spatially constrained hydrophilic environment, however the ribosome exit tunnel is known to be spatially constrained with negatively charged walls (model created in PyMOL through alignment of the five best-guess models from PEP-FOLD). Amino acid residues are numbered sequentially from N to C-termini. Note the N-terminal alpha-helix, and the sharp turn at the C-terminus.

Therefore, there are three possible outcomes when a eukaryotic ribosome translates 2A: either translational read-through of the 2A sequence; or, translational recoding instigated by the nascent peptide resulting in non-canonical termination at the final glycine; or, the formation of the glycine-proline peptide bond can be inhibited. In this third case, two nascent proteins are produced from a single mRNA transcript due to the "skipping" of the glycine-proline peptide bond (never formed, as opposed to formed and subsequently broken). This 2A-driven ribosome translational recoding mechanism has variously been termed "stop-go" (Atkins *et al.*, 2007), "stop carry-on" (Sharma *et al.*, 2012) and "CHYSEL" (cis-acting-hydrolase-element) (de Felipe, 2004). All four names are

used in the literature; with no one term emerging as the most popular nomenclature. In the interests of brevity and clarity this report will refer to the ribosomal recoding peptide as 2A, whereas the *enterovirus*-like 2A proteases will be referred to $2A^{pro}$.

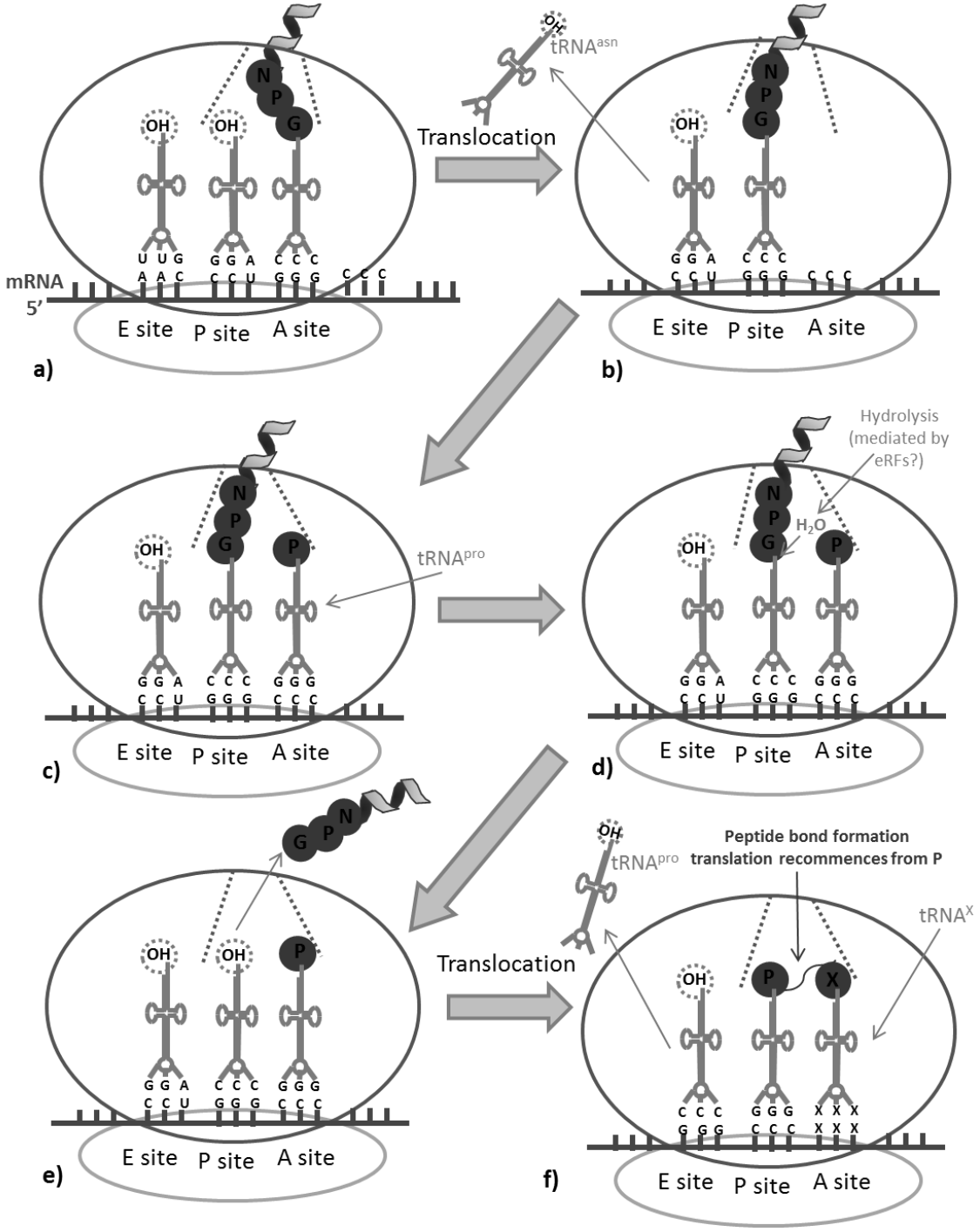


Figure 1.9 Simple schematic of 2A activity
Illustration of events in the PTC during 2A-mediated ribosome skipping (Diagram after Ryan *et al.*, 1999; Donnelly *et al.*, 2001b; Doronina *et al.*, 2008b). It was thought that eRF complex formation was essential for ester bond hydrolysis, but a recent *in vitro* study found eRFs were not required for 2A translational recoding (Machida *et al.*, 2014).

Sequence comparison of *FMDV* 2A with other *Aphthovirus* and *Cardiovirus* 2As found considerable variation within the N-terminal portion, but revealed the existence of a highly conserved C-terminus, namely -D[V/I]ExNPG- where x represents any amino acid. This was always followed by proline (P) as the first N-terminal residue of 2B. This 8 amino acid motif, -D[V/I]ExNPG$^{\downarrow}$P-, where $^{\downarrow}$=cut in nascent chain, was crucial to function as point mutations within it either severely reduced or ablated activity altogether (Luke *et al.*, 2008).

To function optimally the 2A sequence also required an appropriate upstream sequence (Donnelly *et al.*, 2001a). Hybrid 2A peptides manufactured by switching the D[V/I]ExNPG$^{\downarrow}$P motif from *FMDV* onto the 22 amino acid upstream sequence from *EMCV*, and *vice versa*, showed low or no activity (Sharma *et al.*, 2012), confirming that the entirety of 2A contributes to function. Sequence length was also important, because when, *in vitro*, *FMDV* 2A was elongated by the addition of upstream (1D C-terminus) residues, increasing chain length to up to 30 amino acids, its activity was enhanced (Donnelly *et al.*, 1997; Donnelly *et al.*, 2001b; Minskaia *et al.*, 2013). The same pattern held for the other picornavirus 2As, with nascent chain lengths of around 30 amino acids exhibiting the highest activity (Luke *et al.*, 2008). However, it has proven difficult to distinguish critical residues which interact with the ribosome tunnel from flanking "space-filling" residues. The majority of substitutions within a 2A sequence tend to reduce its self-cutting activity, which in wild-type 2As can be as high as 100%. Indeed, it was recently suggested that 2A sequences are: "fine-tuned to function as a whole" and "each 2A may then represent a specific solution for positioning the conserved C-terminus within the peptidyl-transferase centre to promote recoding" (Sharma *et al.*, 2012).

1.9 **Ribosome Skipping 2As - Viral Phylogeny**

Following the discovery of ribosome skipping 2A sequences in the *Aphthoviruses* and *Cardioviruses*, such sequences were also found in a number of other (but not all) *Picornavirus* genera (Table 1.2). In all cases the active sequence was short, being fully functional at only 30 amino acids), and ended in the conserved C-terminus motif D[V/I]ExNPG$^{\downarrow}$P. Interestingly, in a number of picornavirus genera, one or more species possessed ribosome skipping 2As whereas others within the same genera possessed 2A^{pro} (Table 1.2).

This led to the hypothesis that, for some picornaviruses, due to later recombination events an ancestral ribosome skipping 2A may have been replaced by 2A^{pro} (Luke *et al.*, 2008). Extending the search for 2A sequences beyond the picornaviruses, further online database probing using the 2A C-terminus motif, revealed the occurrence of ribosome skipping 2A sequences in a number of other viruses (Table 1.3, Figure 1.10), namely, in positive-stranded RNA viruses belonging to the *Iflaviridae*, *Tetraviridae*, and *Dicistroviridae* families (insect-infecting viruses); in the *Reoviridae* (mammalian or insect-infecting segmented double-stranded RNA viruses); and the *Totiviridae* (non-segmented double-stranded RNA viruses).

Phylogenetic analyses of the non-picornavirus 2A containing viruses (Figure 1.10) indicate that the ribosome skipping 2As have been adopted, possibly as a replacement for an ancestral protease. Viral 2A phylogenetic trees do not correspond to a reference tree of viral RNA-dependent RNA polymerase (RdRp), suggesting at least 6 independent acquisitions or evolutions of 2A (Luke *et al.*, 2008). Multiple acquisitions of 2As appear more plausible than multiple cases of independent convergent evolution, albeit given the short length of the active 2A sequence, multiple independent origins, although implausible, are not impossible. The difficulties of tracking 2A evolution are compounded by the facts that *FMDV*, the virus at the forefront of 2A research, is extremely young, originating only 430 years ago (Tully and Fares, 2008; Yoon *et al.*, 2011) and that in picornaviruses in particular, it is known that there can be a high degree of recombination between strains, and even, more rarely, between certain genera (Heath *et al.*, 2006).

Thus, the inference from their phylogenetic distribution is that ribosome skipping 2As occurred in the ancestral *Picornavirus*, as they are present in most current picornavirus genera, although they have been replaced by proteases in the *Enteroviruses*. Non-picornaviruses with ribosome skipping 2As appear to have acquired them more recently to replace proteases. For ribosome skipping 2As to have been acquired, retained and transferred between viral species, they must presumably confer an evolutionary fitness advantage on their hosts in comparison with primary protease cleavage as a method of undertaking primary polyprotein processing.

Table 1.2 2A sequence types in the *Picornaviridae*
Information from www.picornaviridae.com using the most recent classification guidelines, genera with ribosome skipping 2As are shown in in **bold**.

Picornavirus genera	2A type	Sample viruses
Aphthovirus	Ribosome skipping	*Bovine Rhinitus Viruses, Equine Rhinitus Viruses, Foot-and-Mouth-Disease Viruses*
Aquamavirus	Ribosome skipping	*Seal Aquamavirus A1*
Avihepatovirus	Ribosome skipping & 2A-H-Box/NC	*Duck Hepatitis A Virus*
Avisivirus	(x2 ribosome skipping) & 2A-H-Box	*Avisivirus A*
Cardiovirus	Ribosome skipping	*Encephalomyocarditis Virus (EMCV), Theiloviruses (Theiler's Murine Encephalomyelitis Virus (TMEV), Vilyuisk Human Encephalomyelitis Virus (VHEV), Thera Virus (TRV)*
Cosavirus	Ribosome skipping	*Cosavirus A*
Dicipivirus	Dicistronic, 2 IRESes	*Cadicivirus A*
Enterovirus	$2A^{pro}$	*Enteroviruses A-J, Rhinoviruses A-C*
Erbovirus	Ribosome skipping	*Equine Rhinitis B Virus (ERBV)*
Gallivirus	H-box	*Gallivirus A*
Hepatovirus	$2A^{pro}$	*Hepatitis A Virus*
Hunnivirus	Ribosome skipping	*Hunnivirus A*
Kobuvirus	H-box	*Aichiviruses A-C, Canine Kobuvirus 1, Feline Kobuvirus 1, Murine Kobuvirus 1*
Megrivirus	H-box	*Melegrivirus A*
Mischivirus	Ribosome skipping	*Mischivirus A*
Mosavirus	Ribosome skipping	*Mosavirus A*
Oscivirus	$2A^{pro}$	*Oscivirus A*
Parechovirus	H-box only in *Human Parechovirus*. Ribosome skipping then H-box in the others	*Human Parechovirus, Ljungan Virus, Sebokele Virus 1*
Pasivirus	Ribosome skipping	*Pasivirus A*
Passerivirus	H-box	*Passerivirus A*
Rosavirus	H-Box	*Roasvirus A*
Salivirus	$2A^{pro}$	*Salivirus A*
Sapelovirus	$2A^{pro}$	*Porcine Sapelovirus, Simian Sapelovirus, Avian Sapelovirus*
Senecavirus	Ribosome skipping	*Seneca Valley Virus*
Teschovirus	Ribosome skipping	*Porcine Teschovirus*
Tremovirus	H-box	*Avian Encephalomyelitis Virus*

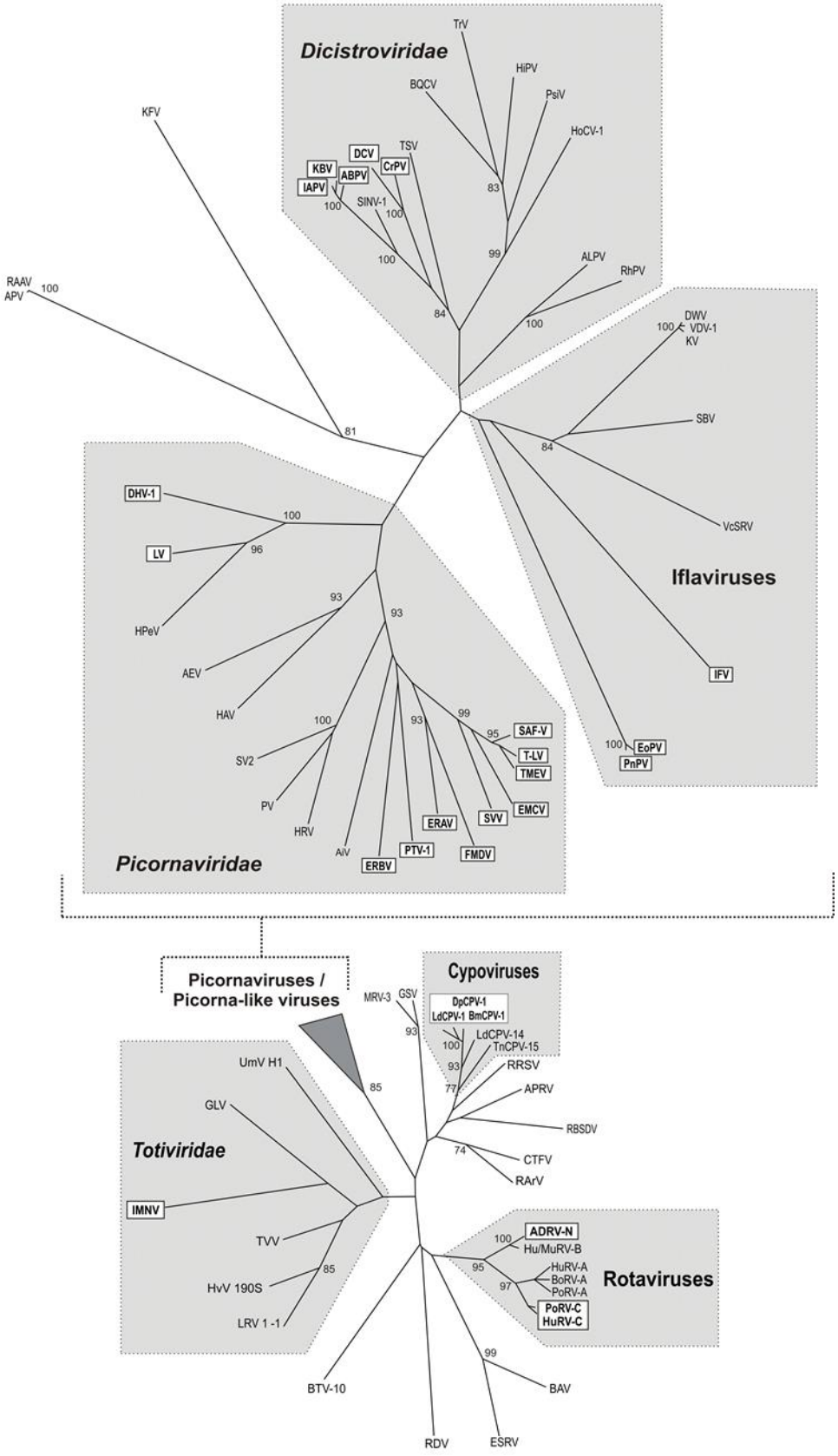


Figure 1.10 Phylogenetic analysis of viral RdRp sequences
Virus groups are indicated by grey shading, viruses with functionally active ribosome skipping 2As are indicated in white boxes (from Luke *et al.*, 2008). Key to abbreviated virus names can be found in Table 1.3 (following page).

Table 1.3 Sample viruses with active ribosome skipping 2As
A sample of active viral 2A sequences (activity data from Luke *et al.*, 2008), note the conserved C-terminus motif, but the high degree of variability in the upstream portion of the sequence (Table modified from Luke and Ryan, 2013).

Virus	Abbrev.	2A amino acid sequence
Positive single-stranded RNA viruses		
Picornaviruses (predominately mammalian viruses)		
Theiler's murine Encephalomyelitis Virus	*TMEV*	-FREFFKAVRGYHADYYKQRLIH**DVEMNPG P**-
Foot-and-Mouth Disease Virus	*FMDV*	-HKQKIVAPVKQTLNFDLLKLAG**DVESNPG P**-
Encephalomyocarditis Virus	*EMCV*	-VFGLYRIFNAHYAGYFADLLIH**DIETNPG P**-
Saffold Virus	*SAF-V*	-FTDFFKAVRDYHASYYKQRLQH**DVETNPG P**-
Equine Rhinitis B Virus	*ERBV-1*	-EATLSTILSEGATNFSLLKLAG**DVELNPG P**-
Ljungan Virus	*LV*	-YFNIMHSDEMDFAGGKFLNQCG**DVETNPG P**-
Iflaviruses (insect viruses)		
Infectious Flacherie Virus	*IFV*	-PSIGNVARTLTRAEIEDELIRA**GIESNPG P**-
Ectropis oblique Picorna-like Virus	*EoPV-2A*$_1$ *EoPV-2A*$_2$	-GQRTTEQIVTAQGWAPDLTQDG**DVESNPG P**- -TRGGLQRQNIIGGGQRDLTQDG**DIESNPG P**-
Perina nuda Picorna-like Virus	*PnPV-2A*$_1$ *PnPV-2A*$_2$	-GQRTTEQIVTAQGWVPDLTVDG**DVESNPG P**- -TRGGLRRQNIIGGGQKDLTQDG**DIESNPG P**-
Tetraviruses (insect viruses)		
Euprosterna elaeasa Virus	*EeV*	-RRLPESAQLPQGAGRGSLVTCG**DVEENPG P**-
Providence Virus	*PRV-2A*$_1$ *PRV-2A*$_2$ *PRV-2A*$_3$	-LEMKESNSGYVVGGRGSLLTCG**DVESNPG P**- -NSDDEEPEYPRGDPIEDLTDDG**DIEKNPG P**- -TIMGNIMTLAGSGGRGSLLTAG**DVEKNPG P**-
Dicistroviruses (insect viruses)		
Cricket Paralysis Virus	*CrPV*	-LVSSNDECRAFLRKRTQLLMSG**DVESNPG P**-
Acute Bee Paralysis Virus	*ABPV*	-TGFLNKLYHCGSWTDILLLLSG**DVETNPG P**-
Double-stranded RNA viruses		
Rotaviruses (mammalian viruses)		
Bovine Rotavirus C	*BoRV-C*	-GIGNPLIVANSKFQIDRILISG**DIELNPG P**-
Human Rotavirus C	*HuRV-C*	-GAGYPLIVANSKFQIDKILISG**DIELNPG P**-
New Adult Diarrhoea Virus	*ADRV-N*	-FFDSVWVYHLANSSWVRDLTRE**CIESNPG P**-
Cypoviruses (insect viruses)		
Bombyx mori Cypovirus 1	*BmCPV-1*	-RTAFDFQQDVFRSNYDLLKLCG**DIESNPG P**-
Operophtera brumata Cypovirus-18	*OpbuCPV-18*	-IHANDYQMAVFKSNYDLLKLCG**DVESNPG P**-
Totiviruses (crustacean viruses)		
Infectious Myonecrosis Virus	*IMNV-2A*$_1$ *IMNV-2A*$_2$	-WDPTYIEISDCMLPPPDLTSCG**DVESNPG P**- -RDVRYIEKPEDKEEHTDILLSG**DVESNPG P**-

1.10 **2As as Translational Regulators**

To study picornavirus 2A translational recoding activities *in vitro*, in order to prove that 2A could autonomously program ribosome skipping without requiring any viral or cellular co-factors, artificial polyprotein reporters were constructed (Figure 1.11) and used to program eukaryotic ribosomes preparations in cell-free coupled translation-transcription reactions (TnTs). Protein synthesis was measured through incorporation of radio-labelled methionine. These reporter plasmids contained two reporter proteins: in the first instance chloroamphenicol-acetyl-transferase (CAT) and β-glucuronidase (GUS) flanking the 2A [CAT-2A-GUS], later green fluorescent protein (GFP) and GUS [GFP-2A-GUS] and [GUS-2A-GFP] in a single open-reading frame run from a CMV promoter, with the methionine start codon of the second protein removed (Doronina *et al.*, 2008a).

It was discovered that although 2A-directed translational recoding could be >90% effective, typical *in vitro* activity levels ranging from around 70-95% depending on the viral 2A tested. There was, for virtually every sequence tested in this *in vitro* assay system, a band of read-through product and a slight, but measurable, molar excess of the upstream protein, and opposed to that of the downstream protein. It has been suggested that the reliance of ribosomal translation on cofactors such as elongation factor eEF2 during translation initiation and nascent peptide chain elongation, and possibly eRF1 & 3 to facilitate nascent chain detachment, could leave the 2A mechanism vulnerable to changes in cellular metabolism altering its efficiency. eEF2 in particular is suspected to play a major role. eEF2 availability is regulated by eEF2 kinase, which is in its turn, regulated by phosphorylation governed by input from a number of cell signalling pathways sensitive to the effects of stress (temperature extremes, hypoxia, nutrient deficit etc), meaning that when the cell is under stress then both initiation and elongation are less efficient due to a lack of available eEF2. Therefore, under these conditions, re-initiation downstream of 2A would be impaired, and this would lead to an increase instance of ribosome stalling at the end of 2A (Doronina *et al.*, 2008b; Brown and Ryan, 2010; Luke and Ryan, 2013).

These findings, coupled with the fact that 2A sequences in the *Picornaviridae* occur at the juncture between the upstream structural and downstream replicative proteins, led to the hypothesis that the function of 2A may be two-fold. Firstly serving in polyprocessing, cutting the nascent chain, but also (through stalling without subsequent re-initiation) down-regulating downstream translation. The picornavirus genome/polyprotein encodes a single copy of each viral protein, but, to assemble one complete virion, sixty copies of each of the four structural proteins are needed, whereas only one copy of the replicative proteases, RNA polymerases and cap proteins are required. Therefore, an excess of the upstream structural proteins over that of the downstream regulative proteins would prove beneficial to the virus, especially during the later stages of infection during virion assembly.

If, by late virus infection, when the host cell was under extreme stress as cellular protein translation was virtually shut-down and overtaken by viral translation, inability to replace cellular proteins would lead to a deficit of the release factors eRF1 and 3 and a stress-induced reduction in eEF2. Conditions that would favour a higher instance of stalling at 2A, and therefore provide an excess of the upstream capsid proteins over the levels of the downstream replicative proteins, exactly the conditions that would most benefit efficient rapid viral assembly.

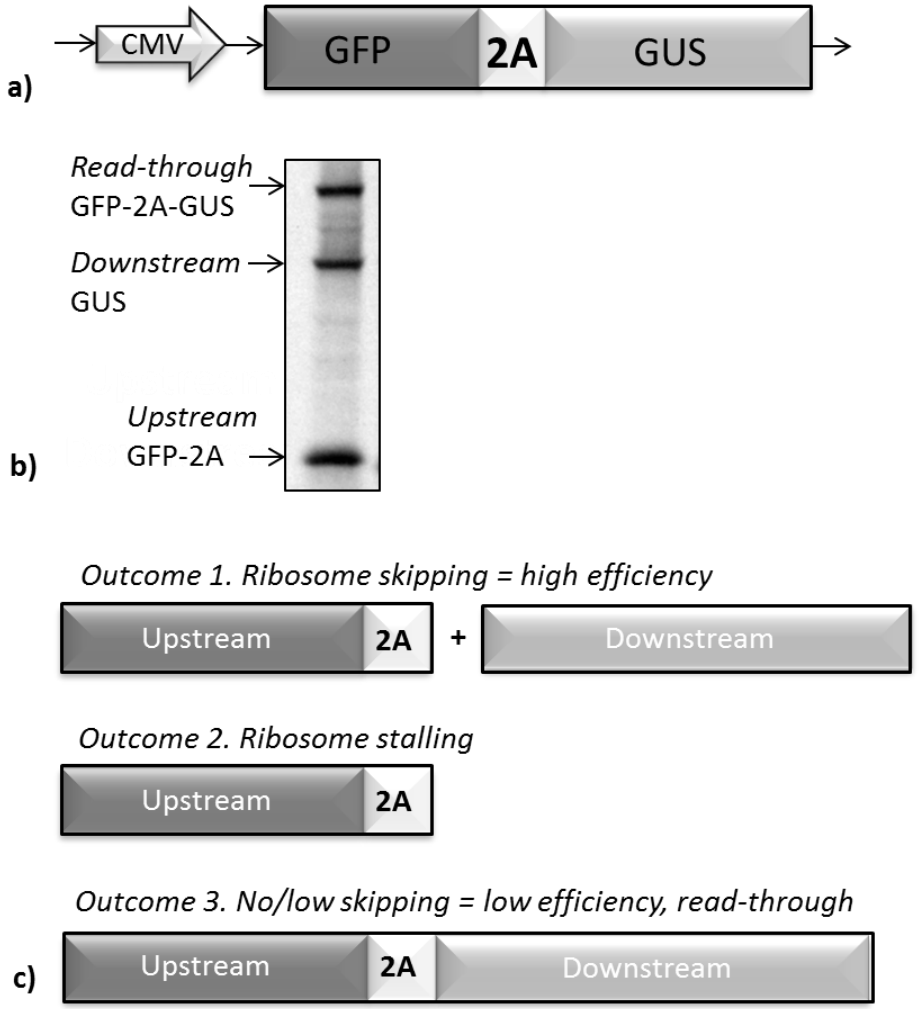


Figure 1.11 *In vitro* reporter assay system & 2A translation products
a) ORF encoding an artificial polyprotein consisting of a ribosome skipping 2A flanked by two reporter proteins. **b)** TnT products from the artificial polyprotein resolved by SDS-PAGE gel electrophoresis. **c)** The three possible outcomes of 2A translation, 1= effective ribosome skipping, 2= stalling, 3= read-through of the 2A containing polyprotein. (Not to scale).

In the insect infecting *Iflaviridae*, a 2A sequence separates the capsid and the replicative protein domains similarly to its positioning in the *Picornaviridae*, but *Perina nuda Picorna-like Virus* (*PnPV*) and *Ectropis obliqua Picorna-like virus* (*EoPV*) also possess a second 2A between their VP2 and VP4 structural proteins. In the *Dicistroviridae*, there is a highly efficient 2A positioned at the N-terminus region of the replicative protein ORF, whereas in the *Tetraviridae*, there is a 2A at

the N-terminus of the structural ORF (which has been shown to possess extremely high processing ability with ribosome skipping occurring in ~99% of instances) (Donnelly *et al.*, 2001a), and some possess second, or third, slightly less efficient 2As in non-structural protein ORFs. Here, as with the picornaviruses, the 2A are thought to control relative levels of translation.

Drosophila C Virus (*DCV*) and the closely related *Cricket Paralysis Virus* (*CrPV*) both encode a protein upstream of 2A that functions in RNA interference (RNAi) suppression. RNAi, alternately termed virus-induced gene silencing, is a key arthropod antiviral response mechanism (Li *et al.*, 2002; Ding and Voinnet, 2007; Aliyari *et al.*, 2008 ; van Rij and Berezikov, 2009). During viral infection dsRNA and microRNA (miRNA) viral replication intermediates are recognised by and processed into small interfering RNA (siRNA) duplexes by the dsRNA cellular sensor Dicer proteins (Meister and Tuschl, 2004). The siRNAs are assembled, along-with an Argonaute (Ago) protein into an RNA-induced silencing complex (RISC) and rendered inoperable (Hutvagner and Simard, 2008). In the viral "arms-race" to out-compete the cellular response, a number of viruses have evolved RNA silencing suppressors, (viral silencing repressors, VSRs) which target and block the various steps of the RNA*i* antiviral defence pathway (Ding and Voinnet, 2007; Swarts *et al.*, 2014). The VSR encoded by DCV is a dsRNA-binding protein that specifically blocks Dicer2 processing of virus dsRNA into siRNAs (van Rij *et al.*, 2006). The *CrPV* VSR interacts with Ago2 to inhibit RISC formation and activities (Nayak *et al.*, 2010).

In the double-stranded RNA-viruses: in the *Totiviridae*, *Infectious Myonecrosis Virus* (*IMNV*) the polyprotein ORF1 encodes non-structural proteins with two highly active 2As. The N-terminus of *IMNV* ORF1 consists of a double-stranded RNA (dsRNA) binding protein (dsRBP) immediately followed by a 2A. The *Reoviridae* 2As are in genome segments encoding non-structural proteins. In insect *Cypoviruses*, a highly active 2A occurs within segment 5 in *Bombyx mori Cypovirus 1* (*BmCPV-1*) and *Operophtera brumata Nucleopolyhedrovirus 18* (*OpbuNPV-18*). In the *Rotaviruses*, the human infecting *New Adult Diarrhoea Virus* (*ADRV-N*), virus segment 5 contains a highly active 2A, whereas in the *Porcine* and *Human Type C Rotaviruses* there is a lower efficiency 2A in segment 6. These 2As link the single-stranded RNA (ssRNA) binding protein NSP3 to double-stranded binding protein (dsRBP) in *Rotavirus C* and in the case of *ADRV-N*, NSP1 is linked to a dsRBP, respectively. The dsRBPs downstream of 2A sequester viral dsRNA (11–16 nucleotides, without apparent sequence specificity) from the cellular sensors of dsRNA, thus counteracting activation of the innate immune response through the antiviral cellular interferon system (Langland *et al.*, 1994). Our supposition (Odon *et al.*, 2013) is that 2A provides these viruses with the ability to acquire additional functions (here innate immune avoidance) through the "bolting-on" by 2A-linkage of extra functions (dsRBP or VSRs) into their polyprotein/genome.

1.11 **2As Beyond Viruses – Eukaryotic 2As**

The proposal that 2A acquisition may regulate gene expression and/or provide an add-on of co-expressed novel functions was furthered by the discovery of viral 2A-like sequences from the genomes of eukaryotic cellular organisms, namely trypanosomes, single-celled protozoan parasites responsible for diseases such as sleeping sickness and Chagas disease. These trypanosome 2A sequences were found to be reasonably active *in vitro* using the polyprotein reporter system [GFP-2A-GUS] developed for viral 2A analysis (Heras *et al.*, 2006).

Analysis of the flanking gene and protein structure revealed that the 2As occurred within the N-terminus of open reading frame 2 (ORF2) immediately upstream and in frame with the endonuclease and reverse transcriptase domains of the *L1Tc LINE element* – a type of non-long terminal repeat (non-LTR) retrotransposon. Non-LTRs are genetic factors that can autonomously self-replicate and spread throughout their host genome by means of a RNA intermediate. The full evolutionary role of retrotransposons in determining genetic and presumable phenotypic traits is at present unknown, but, genome sequencing studies have revealed the ubiquitous nature of these elements within our own genomes. It has been estimated that 42% of the entire human genome is composed of non-LTR retrotransposons (International-Human-Genome-Sequencing-Consortium, 2001). Chapter 3 will provide a detailed report of 2As associated with non-LTR sequences.

1.12 **Summary**

These discoveries have contributed but a little to the understanding of 2A origins and subsequent evolution. 2As represent an evolutionary paradox – how could incremental evolution create these peptides that require existence in their entirety to function? But, 2As are short, and their crucial eight amino acid C-terminal motif still shorter; therefore, they could have first arisen through chance mutations. In addition, a 2A need not have 100% functionality; a partially effective 2A that could potentially generate both the full-length protein and the breakage products could be beneficial both in increasing proteome diversity and in acting in a regulatory capacity during protein translation. Events where 2A integration resulted in a fitness disadvantage leading to extinction will regrettably not be identifiable from the genomes of extant organisms and it is possible that, in most instances, mutations leading to 2A "self-cleaving" would result in loss of protein function, except if the 2A genes occurred at the downstream end of an ORF, between gene regions, or in non-coding DNA. For example, in viruses, successful 2A integration is apparently only possible between gene regions (as in *FMDV* between the capsid and replicative protein encoding gene blocks) (Luke *et al.*, 2008).

Many unanswered questions remain. Firstly, did 2A originate as a viral or a eukaryotic cellular trait? Are there indeed no active prokaryotic 2As? How many eukaryotic 2A-like sequences are

there? How, once acquired, does the 2A sequence spread, is there horizontal gene transfer (perhaps through a viral or parasitic vector) or are there multiple independent origins of ribosome skipping 2As, evolving *de novo*?

The efficiency of 2A as a gene linker has been recognised by the biotechnology sector where it is currently the most effective tool for multiple gene expression from a single vector (de Felipe and Ryan, 2004; de Felipe *et al.*, 2006;de Felipe *et al.*, 2010). In the field of genetic engineering, the ease with which two genes linked by 2A can be expressed in series has made 2A sequences invaluable. In the last 15 years some notable successes using 2A include: synthesising antibody heavy and light chains, co-translating all subunits of the TCR-CD3 complex, gene transfer for cancer immunotherapy, multi-gene insertions of novel metabolic or pharmaceutical traits into agricultural crop plants (such as the vitamin enriched "golden rice"), and creation of pluripotent stem cells for gene therapy of genetic diseases (reviewed in de Felipe *et al.*, 2006; Luke *et al.*, 2010b; Luke and Ryan, 2013).

However, there may be hidden inherent costs of using 2A in protein biogenesis. The most obvious potential cost is that mutations in the upstream flanking protein could be doubly deleterious, resulting in loss of not one but two functional proteins if 2A ribosome skipping activity were negated, proving lethal if the 2A were linking essential structural or metabolic proteins. Another factor is that the 2A tail remains attached to the C-terminus of the upstream protein. In *FMDV*, $3C^{pro}$ removes 2A from the mature 1D protein. To date, in biotechnology applications, 2A tails have rarely proved problematic, and are useful as identification tags for antibody recognition, but they can cause loss of function in the upstream protein if C-terminus tertiary folding is essential to protein function (de Felipe *et al.*, 2010; reviewed in Luke *et al.*, 2010b; Minskaia *et al.*, 2013). Thus far there have been no reports of immune system activation by foreign 2A peptides in transgenic organisms (de Felipe *et al.*, 2006) but it would be imprudent to discount this risk factor. It is too early to evaluate the long-term multigenerational stability of gene suites inserted with 2A linkers.

There is, therefore, a pressing need to undertake fundamental research on the origins, functional ability, and evolutionary stability of non-viral 2A-like sequences in order to increase the repertoire of such sequences available for use in biotechnology, and to ascertain their safety and stability.

1.13 **Aims**

I will begin my researches by probing online genomic and proteomic databases for more instances of 2A-like sequences from eukaryotic organisms. I will then determine the translational recoding abilities of a selection of novel 2As through means of *in vitro* cell-free coupled transcription-translation assays.

I will consider the gene and protein structure surrounding any eukaryotic 2As, and undertake phylogenetic analyses to investigate their probable origins and evolutionary history.

I will use computer modelling of likely 2A peptide topology to determine if there is a relationship between theoretical 2A sequence architecture and 2A function. Site-direct mutagenesis creating hybrid 2A sequences (between active and inactive forms) will aid in further understanding 2A at the molecular level.

I will utilise standard tissue culture techniques to transfect mammalian cells with plasmids encoding 2A in-frame with fluorescent reporter proteins, to ascertain the intracellular localisation of proteins interspaced by eukaryotic 2A sequences that I have identified as potential signal sequences, in order to reveal whether these 2As might play a dual role in protein targeting.

I hope that these investigations will help to bring to attention the uniqueness and usefulness of 2A as a gene-linker, both naturally occurring and as a tool in the biotechnology toolbox.

Chapter 2. Methodology

'The first lesson… …and the last is: do what is needful! And no more.

The lessons in between, then, must consist in learning what is needful. They do.'

The Farthest Shore – Ursula K. LeGuin, 1972

2.1 Computer-Based Analyses

2.1.1 Database Searches

The main publically available DNA and protein sequence repositories of: NCBI (http://www.ncbi.nlm.nih.gov/nucleotide/ and http://www.ncbi.nlm.nih.gov/protein/), Baylor College of Medicine Purple Sea-Urchin database: https://www.hgsc.bcm.edu/other-invertebrates/sea-urchin-genome-project, Sea-Urchin database at the Max Planck Institute for Molecular Genetics http://goblet.molgen.mpg.de/cgi-bin/seaurchin-genombase.cgi, (Baylor database mirror) and Repbase http://www.girinst.org/ were probed for 2A-like sequences using the BLAST or tBLASTn algorithms against the *FMDV 2A* conserved motif D[V/I]ExNPGP. Hits were downloaded and manually screened for similarity to known 2A sequences.

2.1.2 Sequence Alignment & Phylogeny

Protein sequences were aligned using ClustalX 2.1 (http://www.clustal.org/clustal2/, Larkin *et al.*, 2007). Phylogenetic trees were drawn using Phylodraw http://pearl.cs.pusan.ac.kr/phylodraw/ Choi *et al.*, 2000) or Figtree v1.4.2 (Andrew Rambaut (2007), http://tree.bio.ed.ac.uk/software/figtree/).

2.1.3 Protein Conserved Domain Identification & Modelling

Protein sequences flanking 2As were screened using the NCBI BLAST "find conserved domain" tool, (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml), and aligned using MACAW (http://iubio.bio.indiana.edu/soft/molbio/ncbi/old/macaw/). Non-LTR/transposable elements were identified by the presence of reverse transcriptase (RT) and/or exo-endonuclease (EEP) domains. Sequences were assigned to non-LTR clade based on their RT domain using the Repbase RTclass1 web server (www.girinst.org/RTphylogeny/RTclass1, (as detailed in Odon *et al.*, 2013) using core RT domain amino acid homology). PEP-FOLD 2011 (http://bioserv.rpbs.univ-paris-diderot.fr/services/PEP-FOLD/, Maupetit *et al.*, 2009) was used to investigate the topology of 2A peptides. The program reported a set of five models representing the most frequently occurring outcomes from 100 (short form) or 500 (long form) simulations (.pdb files in Supplementary Materials). Models were visualised and aligned using PyMOL v1.3 (Schrodinger LLC Camberley, UK http://www.pymol.org/) and the model most closely corresponding to the central model in the alignment was selected to represent each 2A. Any models with β-sheet configurations were omitted from the analysis as space restrictions prevent β-sheet folding within the ribosome tunnel environs.

2.1.4 Signal Sequence Identification

Putative signal sequences were identified using SignalP 4.0 (Petersen *et al.*, 2011) http://www.cbs.dtu.dk/services/SignalP/) and/or WoLF PSORT (Horton *et al.*, 2007) an updated version of the PSORT algorithm hosted at http://www.genscript.com/psort/wolf_psort.html that unfortunately has been offline since spring 2014. When WoLF PSPORT was unavailable, the older program, PSORT II http://psort.hgc.jp/cgi-bin/runpsort.pl (Nakai and Horton, 1999) was substituted.

2.1.5 *In Silico* Plasmid Maps & Cloning Strategies

Sequence alignments were performed using DNAMAN 5.1 (Lynnon Corporation, Pointe-Claire, Quebec, Canada). Cloning strategies were determined using Snapgene Viewer 2.4.3 (GSL Biotech LLC, Chicago, Illinois, USA: http://www.snapgene.com/products/snapgene_viewer/) to generate plasmids maps and FastPCR 6.5 (PrimerDigital Ltd., Helsinki, Finland http://primerdigital.com/fastpcr.html) was used to design PCR primers and gene-blocks.

2.1.6 Amino Acid Classification Scheme

When the amino acid composition of 2A sequences was examined, the amino acids were coloured using the following scheme (Figure 2.1) based on the characteristics, in particular the hydrophobicity, of each respective amino acid.

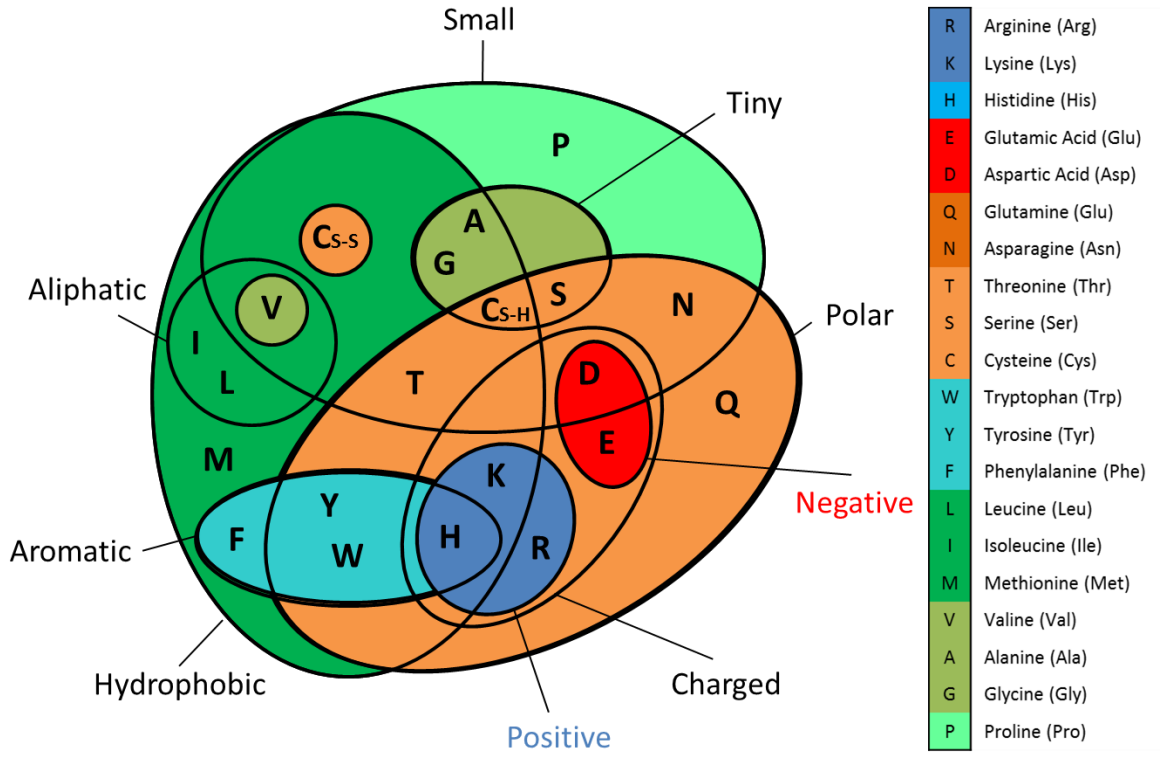


Figure 2.1 Amino acid properties
Schematic illustration of amino acid properties. This colour scheme was used in the cone graphs of 2A consensus sequences in Chapter 8.

2.2 **Laboratory Techniques**

Where reagents/consumables were purchased from external suppliers their names are stated. If no supplier name is listed then the item was purchased from university stores. All reagents were of at least analytical grade quality. Solutions were diluted in sterile deionised water, where appropriate solutions were further sterilised by autoclaving or 0.22 µm filter sterilisation. Solutions were made following standard laboratory protocols (Sambrook and Russell, 2001). Primers were manufactured by IDT (Integrated DNA Technologies, Leuven, Belgium), gene-blocks by Dundee Cell Products (Dundee, Scotland, UK). Generally all experiments were repeated 3x to ensure reproducibility of data, the gels presented in this thesis are typical of one such replicate.

2.2.1 Cloning

Putative 2A sequences were identified as detailed in Section 2.1.1 and the following Chapters. Representative 2A sequences were then cloned for *in vitro* analyses (methods to follow, sequence lists are given in each subsequent Chapter). 2A consensus sequences were determined as detailed in Chapter 8.2.1 and 8.2.2, where not previously cloned, these were cloned for analysis (methods to follow and in Chapter 8.2.2).

2.2.1.1 Plasmid Vectors

Several plasmids based on the *pcDNA*™3.1 (Invitrogen Life Technologies Ltd., Paisley, UK) vector backbone were used for routine cloning. These were obtained from Dr. Garry Luke and John Nicholson. The full sequences are provided in Appendix A.

2.2.1.2 Restriction Enzyme Digests

Restriction enzymes were purchased from Promega (Promega Ltd, Southampton, UK) or NEB (New England Biolabs, Ipswich, Massachusetts, USA) and were used following manufacturer's recommendations. Routinely for analytical digests 500 ng of DNA was digested in 1-2 U of enzyme in a final volume of 10 µL including 1 µL 10x restriction buffer. For preparative digests, up to 1 µg of DNA was digested in 2-5 U of enzyme in a final volume of 20 µL including 2 µL of 10x buffer. All reactions were incubated at the optimum temperature for their specific enzyme(s), (typically 37 °C), for 1 hour in the case of analytic digests and 2 hours for preparative digests.

Then, in the case of digests undertaken to obtain linearized vector DNA, 2 µL of calf intestinal alkaline phosphatase (CIAP (Promega)) enzyme was added and the reaction incubated at 37 °C for a further 30 minutes before heat-inactivation at 65 °C for 15 minutes. CIAP treatment removed 5' phosphate groups from linearized vector DNA to avoid self-ligation and re-circularization of linearized plasmids, the enzyme being active on 5' overhangs, 5' recessed, and blunt ends.

Digest products were separated by agarose gel electrophoresis.

2.2.1.3 Agarose Gel Electrophoresis

Agarose gels were prepared with freshly diluted 1x TAE buffer made up from 50x stock TAE (242 g Tris base, 57.1 mL glacial acetic acid, 100 mL 0.5 M EDTA) made up to 1L. Typically, 100 mL 1x TAE buffer was added to 0.8, 1, or 2 g agarose (to give 0.8, 1, or 2 % [w/v] respectively), in a small glass beaker and microwaved on full power for 6-8 minutes, stirring every 2 minutes until completely dissolved and boiling. The beaker was allowed to cool until tepid, and then 4 µl ethidium bromide (10 mg/mL, Sigma-Aldrich Company Ltd. Dorset, England) was added before pouring the gel. Once set, the gel was submerged in 1x TAE buffer in the electrophoresis tank.

DNA samples were prepared by adding 6x agarose gel loading buffer (50 % [v/v] glycerol, 0.005 % [w/v] bromophenol blue) before loading into the wells. Indicator ladders 100 bp and 1 kb (Promega) were used to determine the relative sizes of DNA bands. Routinely, agarose gel electrophoresis was performed at 100 V for 40-60 minutes using a Horizon 11.14 Horizontal Gel Electrophoresis System (Whatman Inc. Clifton, New Jersey, USA).

DNA bands were visualized by UV illumination using an UV trans-illuminator. Product bands were excised using a sterile scalpel, and subsequently purified using the Wizard SV PCR Prep Purification System (Promega) as per the manufacturer's instructions, eluting into a final volume of 20-40 µL sterile water. DNA concentration was measured using a NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific UK Ltd, Loughborough, UK)

2.2.1.4 Polymerase Chain Reaction (PCR)

PCR primers for routine cloning were designed using FastPCR 6.5 (http://primerdigital.com/). Primers for mutagenesis were designing using the online QuikChange Primer Design tool (http://www.genomics.agilent.com/primerDesignProgram.jsp, Agilent Technologies, Santa Clara, California, USA). The specific primer sequences are detailed in the relevant Chapters. Annealing temperatures were estimated by subtracting 5°C from the melting temperature of the primer pair.

2.2.1.4.1 Amplification PCR

PCR was used to amplify DNA segments from plasmids or gene-blocks, and to insert restriction sites or 2A-like sequences. PCR reactions were performed using an AB Applied Biosystems Veriti 96-well thermal cycler (Invitrogen Life Technologies). Routinely, 1 µl of 2 U/µl GoTaq DNA polymerase (Promega) in the manufacturer's 10x buffer (330 mM Tris-acetate, pH 7.9, 660 mM potassium acetate, 100 mM magnesium acetate, 5 mM DTT) was added to 2 µL each of 10 µM forward and reverse primer, 30 ng template DNA, and dNTPs at a final concentration of 250 µM made up to a final volume of 50 µL in sterile water. After an initial denaturation step of 94 °C for 2 minutes, each round of PCR synthesis consisted of denaturation (94 °C for 1 minute), annealing at the primer pair specific temperature (between 55-70 °C) for 30 seconds, then elongation at 72 °C

for 60 seconds per kb of template. Typically this three-part PCR cycle was repeated 30 times followed by a final elongation of 72 °C for 10 minutes. Samples were held at 4 °C until removal from the machine.

2.2.1.4.2 Mutagenesis PCR

Mutagenesis PCR was used to insert point mutations to change 2A coding sequences and to insert or remove novel restriction sites. Either the QuikChange II XL Site-Directed Mutagenesis Kit (Agilent Technologies, Santa Clara, California, USA) or KOD Hot Start DNA polymerase (Merck Millipore, Livingston, UK) were used as per the manufacturer's instructions. In either case reactions were set up in a total volume of 50 µL sterile water with 1 µL of polymerase, 100 ng each of forward and reverse primers, 20 ng of template DNA, 5 µL proprietary 10x buffer, and 1 uL of proprietary dNTP mix. The PCR was run for a total of 18 cycles with denaturation (94 °C for 1 minute), annealing at the primer pair specific temperature (between 55-70 °C) for 30 seconds, then elongation at 78 °C for 2 minutes per kb of template, followed by a final elongation of 72 °C for 10 minutes. Samples were held at 4 °C until removal from the machine.

The template DNA was then digested by adding 1 µL of 10 U/µL DpnI enzyme (Thermo Fisher Scientific UK Ltd, Loughborough, UK), digesting at 37 °C for one hour, followed by heating to deactivate the enzyme (65 °C for 30 minutes).

2.2.1.5 PCR Visualisation & Purification

All PCR products were run out on agarose gels (see Section 2.2.1.3), both for visualisation and for purification. Products with an expected size greater than 5 kb were run on 0.8 %, whereas products between 5 kb and 300 bp were run on 1 % and products less than 300 bp on 2 % [w/v] agarose gels respectively, all made up and run in TAE Buffer. DNA bands were UV visualised, excised, and purified (see 2.2.1.3).

2.2.1.6 Ligations

DNA Ligation reactions were undertaken in a final volume of 10 µL with 1 µL each of DNA ligase and 10x buffer (300 mM Tris-HCl, pH 7.8, 100 mM MgCl2, 100 mM DTT, 10 mM ATP) (Promega), following the manufacturer's instructions using 50-100 ng of vector DNA and an equal to three-fold concentration of insert. Reactions were incubated at 4 °C for at least 16 hours.

2.2.1.7 Transformation of Competent *E.coli*

Plasmid stocks were grown in chemically competent *E.coli* cells: strain JM109 (Promega) or DH5α (Invitrogen Life Technologies). Briefly, an aliquot of chemically competent cells was removed from storage at -80 °C, and defrosted on ice. Up to 5 µL of ligation reaction, purified mutagenesis PCR product, or plasmid DNA, typically <100 ng of DNA was added to 30-50 µL JM109 or DH5α cells in a 1.5 mL microfuge tube. The tube was incubated on ice for 30 minutes, then heat-shocked

at 42 °C for 45 seconds, chilled on ice for 2 minutes before the addition of 950 µL room temperature lysogeny broth (LB) broth. The cultures were then incubated at 37 °C with shaking for 1 hour. Then, 50 µL was spread on one half of an agar plate supplemented with the appropriate filter-sterilised antibiotic solution (typically ampicillin, to a final concentration of 100 µg/mL); the remainder was centrifuged at 3000 rpm in a Mikro 120 microcentrifuge (Andreas Hettich GmbH & Co., Tuttlingen, Germany). The cell-pellet was re-suspended in 40 µL of LB, and spread on the remaining half of the plate. Plates were air-dried then incubated at 37 °C for 16 hours.

2.2.1.8 Minprep Plasmid DNA Preparation

Individual colonies were picked from the plates, into sterile universals containing 10 mL of LB with filter-sterilised ampicillin (Promega), to a final concentration 100 µg/mL, and incubated for 16 hours at 37 °C with shaking. Pelleted by centrifugation at 3000 *g* for 10 minutes in a bench-top Rotina 420R centrifuge (Andreas Hettich GmbH & Co), then the supernatant was discarded and the pellet processed using an Omega Bio-Tek E.Z.N.A. Plasmid DNA Mini Kit II (VWR International Ltd., Leighton Buzzard, UK) following the manufacturer's instructions. DNA was eluted in 80 µL sterile water and the concentration measured by use of a Nanodrop 1000 Spectrophotometer (Thermo Fisher) before storage at -20 °C.

2.2.1.9 DNA sequencing

Plasmid DNA sequences were verified by DNA Sequencing & Services (MRCPPU, College of Life Sciences, University of Dundee, Scotland, www.dnaseq.co.uk) using Applied Biosystems Big-Dye Ver 3.1 chemistry on an Applied Biosystems model 3730 automated capillary DNA sequencer. Briefly, 500 ng of DNA was sent in 30 µL volume with 2 µL of 3.2 µM primer. Primers were either *T7* or unique primers as stated in Table 2.1. Clones were verified by sequence alignment using DNAMAN 5.1 (Lynnon).

Table 2.1 Sequencing primers
List of the routinely used sequencing primers, all primers were designed using FastPCR and manufactured by IDT except the T7 primer which was provided by DNA Sequencing & Services.

Primer	Sequence (5' to 3')	Notes
T7	TAATACGACTCACTATAGGG	Provided by DNA Sequencing and Services
GFPf	CTTACCCTTAAATTTATT	*pSTA1* sequencing primer, binds midway through eGFP
GUS seq F	CAATAATCAGGAAGTGATGGAG	*pSTA1* sequencing primer, binds at the start of GUS
GUS seq R	CTACTCAGACAATGCGATGC	*pSTA1* sequencing primer, binds end of GUS (reverse sequencing)
spCherR	CAGACAATGCGATGCAATTTCCTC	*pEMX* (reverse sequencing) binds 3' end of mCherryFP

2.2.2 Cell-Free Coupled Transcription-Translation Assays (TnTs)

Plasmids containing 2A sequences were used to programme Quick TnT transcription/translation rabbit reticulocyte lysate systems (Promega), following the manufacturer's instructions. In brief, plasmid DNA (100 ng) was used to programme the lysate master mix (10 µL) supplemented with 35S-Methionine (10 μCi). Reactions were incubated at 30 °C (90 minutes) before the addition of an equal volume of 2x SDS PAGE loading buffer (2 % [w/v] SDS, 20 % [v/v] glycerol, 2 % [v/v] β-mercaptoethanol, 0.2 % [w/v] bromophenol blue, 100mM Tris, pH 6.8) and boiling at 94 °C for 4 minutes. The TnT reaction products were then analysed by running 5 µL aliquots on SDS-PAGE gels, (see Section 2.2.3.) Gels were dried and the distribution of radiolabel determined by autoradiography using Kodak autoradiography film (Thermo Fisher) exposed to the gel for 3 hours or overnight prior to developing using a Kodak X-OMAT 1000 processor. Typically, each construct was examined by TnT analyses on three separate occasions, with the results from one replicate used to form part of a Results Figure. Preliminary investigation (see Figure 2.2) using 2As cloned in *pSTA1* revealed that the recoding activity of each construct was comparable on multiple occasions, as determined by ImageJ v.1.48 (http://imagej.nih.gov/) analyses of scanned autoradiographs (3 hour to avoid over-exposure). The relative intensity of each SDS-PAGE resolved radiolabelled protein was compared after adjusting for methionine content (eighteen methionine in [GFP-2A-GUS], six in [GFP-2A] and twelve in GUS, respectively) to determine recoding efficiency (after Donnelly *et al.*, 2001a) as calculated by:

$$([GUS^{corrM}] + [GFP\text{-}2A^{corrM}]) / ([GFP\text{-}2A\text{-}GUS^{corrM}] + [GUS^{corrM}] + [GFP\text{-}2A^{ccorrM}]) \times 100$$

The proportion of radiolabel in each band was determined. For read-through [GFP-2A-GUS]:

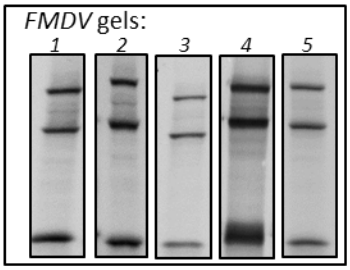
$$[GFP\text{-}2A\text{-}GUS^{corrM}] / ([GFP\text{-}2A\text{-}GUS^{corrM}] + [GUS^{corrM}] + [GFP\text{-}2A^{ccorrM}]) \times 100$$

For GUS: $GUS^{corm} / ([GFP\text{-}2A\text{-}GUS^{corrM}] + [GUS^{corrM}] + [GFP\text{-}2A^{corrM}]) \times 100$

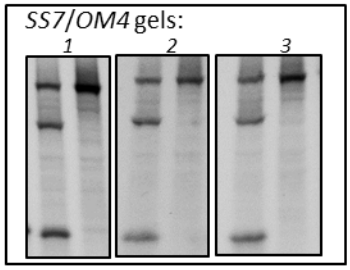
For [GFP-2A]: $[GFP\text{-}2A^{corrM}] / ([GFP\text{-}2A\text{-}GUS^{corrM}] + [GUS^{corrM}] + [GFP\text{-}2A^{ccorrM}]) \times 100$

Next, the recoding efficiency was corrected for the proportion of ribosomes where translation terminated without re-initiation at the end of [GFP-2A]. Simply, the proportion of radiolabel in the GUS band was subtracted from that in [GFP-2A]. This proportion was added to the proportion of read-through product (see above) and then subtracted from 100 to reveal the proportion of ribosomes where the products both upstream and downstream of 2A were translated. This recoding efficiency incorporating lack of re-initiation was found to be more highly variable between occasions (see Figure 2.2). Therefore, it was decided to score 2A recoding ability using a qualitative score where the intensity of the radiolabelled bands seen after SDS-PAGE resolution were evaluated by eye against those generated by *pSTA1* (*FMDV 2A*).

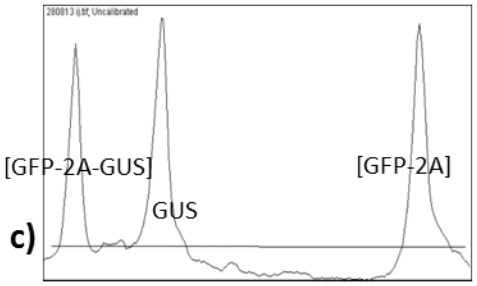
a)



b)



c)



d)

Band	Intensity	Normalised for M content
[GFP-2A-GUS]	4810.2	267.2
GUS	7163.8	597.0
[GFP-2A]	6823.6	1137.3

e)

Name / Amino Acid Sequence	Replicate/ Date	% Recoding Efficiency (after Donnelly 2001a)	% Recoding Efficiency (correcting for [GFP-2A] termination)
pSTA1 (*FMDV* 24aa) RGACQLLNFLLDKLAGDVESNPGP	*1* - 02/05/12	88.6	47.0
	2 - 24/08/12	89.2	55.5
	3 - 28/02/13	88.0	61.1
	4 - 05/07/13	87.6	56.5
	5 - 28/08/13	88.6	59.7
	Mean, SD	**88.4, ±0.62**	**56.0, ±5.5**
SS7 QRSRRPVLIAFSRTLILLLLCSSGDVEVNPGP	*1* - 02/05/12	85.5	42.1
	2 -24/08/12	86.3	50.3
	3 - 07/09/12	83.6	60.7
	Mean, SD	**85.1, ±1.39**	**51.0, ±9.32**
OM4 TRRPVILAFSCTLILLLFCSSGDVEVNPGP	*1* - 02/05/12	7.8	2.4
	2 - 24/08/12	11.8	2.6
	3 - 07/09/12	5.8	5.7
	Mean, SD	**8.5, ±3.06**	**3.57 ±1.85**

Figure 2.2 Calculation of recoding activity analyses

a) Lanes from five replicate TnTs analyses resolved by SDS-PAGE gel electrophoresis showing *pSTA1* (*FMDV 2A* see Figure 1.11) and **b)** three replicates of *SS7* and *OM4* (left to right) cloned in *pSTA1*. The date of each analysis was listed in **e)**. A separate batch of TnT reagent was used on each date. Protein synthesis *de novo* was monitored by the incorporation of S^{35}-methionine as determined by ImageJ quantification of the gel bands. **c)** Typical ImageJ intensity profile (showing the bands from *pSTA1* replicate 5 dated 28-08-13 with the background gated out (horizontal line) then the relative intensity of each peak is measured, values from *pSTA1* replicate 5 date 28-08-13 given in **d)** then normalised for methionine content, **d)**. The formulae presented in Section 2.2.2 (above) were used to calculate the relative recoding efficiency of each construct from the normalised intensity values. Data is presented in **e)**, note the decrease in value, and the increase in range for recoding efficiency of active 2As when both read-through and termination at the end of the upstream product [GFP-2A], are considered as failure to efficiently instigate recoding.

2.2.3 SDS-PAGE Protein Resolving

Proteins were resolved by SDS-PAGE separation. SDS-PAGE gels consisted of an upper layer of stacking gel (4 % [v/v] polyacrylamide) and a lower layer of 10 % or 12 % [v/v] polyacrylamide resolving gel, made following standard protocols (Sambrook *et al.*, 2001). Typically the gels were run at 100 V until the protein samples ran into the stacking gel, then at 140 V until the dye front reached the base of the resolving gel using a vertical electrophoresis 1010Y system (VWR International Ltd.).

Briefly, for Western Blot analysis, 20 µg of protein in 5 µl was added to each well of a 10 or 12 % polyacrylamide gel run in tris-glycine buffer. For TnTs, 5 µL sample per well was loaded on 10 % gels run in tris-glycine buffer or precast 4-20 % gradient 17-well gels run in tris-tricine buffer (Expedeon Ltd., Cambridgeshire, UK), and for mass spectrometry analysis precast 12-well gels with tris-glycine buffer were used (Expedeon). Protein samples were run against PageRuler Plus (Thermo Fisher) molecular marker.

2.2.4 Mammalian Cell Culture

All tissue culture was undertaken in Level 2 Containment hoods, following standard aseptic techniques. Waste was decontaminated with 1 % [w/v] Virkon disinfectant (VWR International Ltd.) prior to disposal.

2.2.4.1 Cell Lines

Mammalian cell lines: HeLa (human cervical cancer) BHK, (baby hamster kidney from the Syrian hamster, *Mesocricetus auratus*) and 293T (human embryo kidney) were kindly provided by Dr. Ekaterina Miniskaia and Fiona Tulloch (University of St Andrews). After initial trials to determine their respective suitability, HeLa cells were selected for use in all subsequent experiments.

2.2.4.2 Cell Line Maintenance

Cells were maintained in 75 mL (T75) sterile vent-capped plastic flasks (Griener Bio-One, Monroe, North Carolina, USA), in antibiotic free DMEM (Dubelco's Modified Eagle Medium) supplemented with 10 % [v/v] FCS (foetal calf serum) at 37 ºC in a 5 % CO_2 humidified incubator. Trypsin/EDTA solution (a buffered salt solution containing 0.5 % [w/v] trypsin and 0.2 % [w/v] EDTA) was used to detach adherent cell-monolayers when passaging cells. DMEM-10 % FCS and trypsin/EDTA were stored at 4 °C, and pre-warmed to 37 °C for 30 minutes prior to use.

Cell passaging was undertaken when the flasks were >80 % confluent: first, the media was aspirated, the cell-layer was washed twice with pre-warmed Trypsin/EDTA solution; then, 2 mL Trypsin/EDTA solution was added to cover the cells and the flask was returned to the incubator for 3 minutes until the adherent cells were disassociated. 8 mL of pre-warmed DMEM-10 % FCS was added to inactivate the trypsin. The cell suspension was pipetted vigorously to dislodge cells and

disassociate cell clumps. Finally, 1 mL or 2 mL cell suspension (to split the culture 1:10 or 1:5 respectively) was transferred into a T75 flask containing a final volume of 10 mL pre-warmed DMEM-10 % FCS and placed in an 37 °C incubator.

2.2.4.3 Cell-Line Long-Term Freeze Storage

Cells were grown until over-confluency in T75 flasks. Cells were washed and detached with tryspin/EDTA as detailed in Section 2.2.4.2. The resulting cell suspension from each flask was made up to 6mL with pre-warmed DMEM-10 % FCS, transferred into a sterile 15 mL conical-based centrifuge tube, and centrifuged for 5 minutes at 3000 rpm at room temperature. The media was decanted and the cell pellet re-suspended in 3 mL freeze-down mix (DMEM media supplemented with 20 % [v/v] FCS and 10 % [v/v] sterile DMSO (Sigma-Aldrich) at room-temperature). 1 mL aliquots were transferred into 1.5 mL cryovials and immediately placed on ice and transferred into a -80°C freezer. After 24 hours the vials were placed in liquid nitrogen for longer-term storage.

2.2.4.4 Reconstitution of Frozen Cell Stocks

A 1.5 mL cryovial containing 1 mL cell suspension was removed from storage at -80 °C or from liquid nitrogen and immediately placed on ice. The vial was permitted to defrost. The suspension was aliquoted into a T75 flask containing 15 mL pre-warmed DMEM with 10 % FCS and placed in the culture incubator. As soon as cell settlement and adherence was observed (typically 5-6 hours), the flask was washed 3x with pre-warmed DMEM-10 % FCS (to remove any traces of DMSO from the cryopreserving) and the media was replaced with 10 mL of pre-warmed DMEM-10 % FCS. The flask was then cultured following usual procedures.

2.2.5 Mammalian Cell Transient Transfection

Transfection is the term given to the introduction of novel genetic material into eukaryotic cells. One day prior to transfection, cells were seeded into the culture dish/tray in DMEM with 10 % FCS, to give a confluency of 50-60 % at the time of transfection. When the transfections were to be examined by Deltavision microscopy, a single sterile glass coverslip (one 12 mm round coverslip per well in a 12-well plate or one 15 mm square coverslip in a 6-well, both coverslips 1 mm thickness) was placed in each well prior to adding the cell suspension.

Transfections were undertaken using Lipofectamine® 2000 Reagent and OptiMEM reduced-serum media (Invitrogen Life Technologies) following manufacturer's protocols. To transfect 1 mL of ~50 % confluency cell monolayer 200 µL OptiMEM was gently mixed with 800 ng of plasmid DNA. 4 µL Lipofectamine was added to the OptiMEM/DNA mixture prior to incubation at room temperature for 5 minutes to allow encapsulation of DNA in liposomes that permit fusion with/crossing of the cell membrane by the negatively-charged DNA. Next, the mixture was added

drop-wise to the cell monolayers and the culture vessel was rocked gently to ensure an even distribution. The cells were returned to the incubator for 6 hours, rocking every hour. At 6 hours they were inspected under EVOS fluorescence microscopy (AMG Group, Invitrogen Life Technologies) to determine initial transfection, and the media was aspirated and replaced with an equal volume of fresh pre-warmed DMEM with 10 % FCS. The transfected cells were returned to the incubator for a further 24 hours before harvesting/fixing. For Brefeldin A (BFA; Sigma-Aldrich) treatment, BFA was added (final concentration 15 µg/mL) 45 minutes prior to fixing.

2.2.6 Creation of Stable Cell Lines

Lentiviral-based vectors are commonly used to create stable cell lines constitutively expressing a novel gene. Stable cell lines were made using green and red NucLight™ and CytoLight™ lentiviral vectors (Essen BioScience Ltd., Welwyn Garden City, UK) encoding green or red fluorescent proteins specifically targeted to the nucleus or cytoplasm respectively and a bleomycin resistence gene to enable stable clone selection. The lentiviral vectors were transfected into low passage number BHK and HeLa cells following the manufacturer's protocol. Cells were seeded in a 24-well tray and allowed to grow to 40 % confluency. The lentivirus was added at an MOI of 3, along with a final concentration of 5 µg/mL Polybrene® (Sigma-Aldrich) to aid in infection. The tray was gently rocked to ensure even dispersal, and then returned to the incubator for 24 hours before the media was replaced with fresh pre-warmed media. The cells were cultured for a further 48 hours before the addition of Zeocin (a bleomycin-class antibiotic) (Invitrogen Life Technologies) to a final concentration of 100 µg/mL for HeLa and 200 µg/mL for BHK cells respectively (concentrations were earlier determined by a kill-curve experiment). Cells were returned to the incubator and incubated for a further 14 days, with a media (supplemented with zeocin) change every 2 days. The tray was examined under EVOS fluorescence microscopy at regular intervals to determine the relative expression/ transfection efficiency/cell-death through zeocin selection. After 14 days of zeocin selection, greater than 90 % of surviving cells displayed the specific fluorescence as expected. The media was aspirated and the wells were washed with trypsin/EDTA (as in routine tissue culture) and detached from the well using 100 µL trypsin transferred into T75 flasks containing 10 mL pre-warmed DMEM with 10 % FCS. Cells were grown until confluent then split as normal. New flasks were established to provide freeze-down stocks.

2.2.7 Cell Extract Preparation

At 30 hours post-transfection the transfections were placed onto ice, the media was removed, and each well/dish was washed with ice-cold PBS (sufficient quantity to cover the cells). Each well/dish was individually scraped using a bent pipette tip and its suspension transferred into a separate microfuge tube. The samples were centrifuged at 2000 rpm, 4 °C for 5 minutes, supernatants were removed, and each pellet re-suspended in 100 µL RIPA cell-lysis buffer (50 mM Tris-HCl, 150 mM NaCl, 1 % [v/v] Nonidet P-40, 0.5 % [v/v] deoxycholic acid, 0.1 % [v/v] SDS, adjusted to pH 7.4) supplemented with 7x protease inhibitor solution (cOmplete, Mini, EDTA-free, Roche Diagnostics Ltd., West Sussex, UK) and left on ice for 25-30 minutes, vortexing occasionally. Then, the samples were centrifuged at 13,000 rpm, 4 °C for 15 minutes. The supernatants were removed into new sterile Eppendorf tubes and stored at -20 °C until required. Each transfection was repeated in triplicate on three successive occasions.

2.2.8 Western Blotting

Protein concentration of pooled cell lysates from triplicate transfections was measured using a NanoDrop 1000 Spectrophotometer (Thermo Fisher), blanking with RIPA buffer. Samples were diluted to 4 µg/µL in 5x SDS-sample buffer (60 mM Tris-HCl pH 6.8, 5 % [v/v] SDS 25 % glycerol [v/v], 0.01 % [w/v] bromophenol blue with 5 % [v/v] β-mercaptoethanol) and sterile water to a total volume of 20 µl. They were subsequently boiled for 4 minutes at 94 °C then cooled on ice. Protein samples (20 µg in 5 µL per well) were loaded and resolved by SDS-PAGE gel electrophoresis (see Section 2.2.3).

Proteins were transferred onto nitrocellulose membranes (Invitrogen Life Technologies) at 20 V for 10 minutes using an iBlot transfer machine (Invitrogen Life Technologies) as described by the manufacturer.

Membranes were washed in PBS-Tween (PBS with 0.1 % [v/v] Tween-20 (Sigma-Aldrich)). Then, blocked for 1 hour at room temperature in PBS-Tween-5 % Milk (PBS with 0.1 % [v/v] Tween-20 and 5 % [w/v] dried skim milk powder) on a shaker/roller. The membranes were probed with primary antibody (either anti-GFP mouse monoclonal antibody (Roche), anti-RFP rabbit polyclonal antibody (Invitrogen Life Technologies), or anti-β-tubulin mouse (Invitrogen Life Technologies) monoclonal antibody (1:1000-2000 dilutions in 5 % PBS-Tween-Milk) overnight. Next day the membranes were washed 3x for 5 minutes per wash in PBS-Tween, then probed with horseradish-peroxidase conjugated anti-rabbit (Invitrogen Life Technologies) or anti-mouse (Invitrogen Life Technologies) antibodies diluted 1:2000 for 1 hour at room temperature. Next, the membrane was washed three times, for 5 minutes, in PBS-Tween. Antibody binding was detected using the EZ-ECL HRP chemi-luminescence detection kit (Biological Industries Ltd., Kibbutz

Beit-Haemek, Israel). The membranes were exposed to Kodak autoradiography film (Thermo Fisher) for 3 seconds to 5 minutes.

2.2.9 Immuno-Dot Blots from Culture Media

The media (up to 3 mL from each well/culture dish) removed from triplicate HeLa cell transfections, prior to cell harvesting, was clarified by centrifugation at 3000 rpm at 4°C for 5 minutes, decanted and stored at -20°C. The supernatants were thawed on ice, then 2.5 µL of pooled supernatant was spotted onto pencil marks made on nitrocellulose membranes, pore size 0.45 µm (Bio-Rad Laboratories Ltd, Hemel Hempstead, Hertfordshire, UK), using a narrow-mouthed pipette tip. Membranes were air-dried at room temperature for 20 minutes before a second application of 2.5 µL supernatant onto the same spot. Two positive controls of 250 ng and 500 ng total protein from *pJN1* transfected cell lysates, each in a total volume of 5 µL were loaded likewise. The membranes were air-dried for 30 minutes before being probed with anti-RFP or anti-β-tubulin following an identical procedure as for Western Blotting (see Section 2.2.8).

2.2.10 Fixing Cells & Deltavision Microscopy

Cells were fixed using the protocol described in Minskaia *et al.*, 2013: the media was aspirated, coverslips were washed 2x with PBS, fixed for 10 minutes in ice-cold 100 % methanol and subsequently washed 2x with deionized water, before mounting using Vectashield (Vector Laboratories Ltd., Peterborough, UK.) mount with DAPI (diamino phenylindole 0.5 µg/mL, used for nuclear staining) on glass microscope slides. Coverslip edges were sealed using clear nail-varnish and the slides were stored in the dark at 4 °C to minimise fluorochrome degradation. Images were obtained using a Deltavision microscope (Applied Precision, Marlborough, UK) fitted with 60x oil-emersion objective Olympus lens and a Photometric CH300 CCD camera. The DeltaVision permits the use of multiple filters and the observation of micro-anatomical details. Fluorescence was detected with an excitation wavelength of 488 nm for eGFP (FITC filter), 587 for mCherryFP (TRITC filter) and 350 nm for DAPI (DAPI filter). Images were acquired and analysed the softWoRx® Resolve 3D software package (Applied Precision).

2.2.11 Mass Spectrometry

Gradient 4-20 % 12 well SDS-PAGE gels (Expedeon) were used for protein separation. A lysate sample containing 20 µg total protein was run out on the gel. The gel was Coomassie stained (Instant Blue, Expedeon) and the band corresponding to 25-30 kDa was excised using a sterile scalpel. The gel chunk was analysed by Dr. Catherine Botting, at the University of St Andrews Mass Spectrometry Facility. It was cut into 1 mm cubes which were subjected to in-gel digestion, using a ProGest Investigator in-gel digestion robot (Genomic Solutions, Ann Arbor, Michigan, USA) using standard protocols (Shevchenko *et al.*, 1996). The gel cubes were destained by washing with acetonitrile and subjected to reduction and alkylation before digestion with trypsin at

37 °C. The peptides were extracted with 10% formic acid and concentrated down to 20 µL using a ThermoSavant SpeedVac (Thermo Fisher).

The peptides were then separated using a nanoLC Ultra 2D plus loading pump and nanoLC AS-2 autosampler equipped with a nanoflex cHiPLC chip based chromatography system (Eskigent AB Sciex UK Ltd, Warrington, Cheshire, UK), using a ChromXP C18-CL trap and column (Eskigent AB Sciex). The peptides were eluted with a gradient of increasing acetonitrile, containing 0.1 % formic acid (5-25 % acetonitrile for 75 minutes, 25-80 % for a further 15 minutes, then 80 % acetonitrile to clean the column, before re-equilibration to 5 % acetonitrile). The eluent was sprayed into a TripleTOF 5600 electrospray tandem mass spectrometer (ABSciex, Foster City, California, USA) and analysed in Information Dependent Acquisition (IDA) mode, performing 250 milliseconds of MS followed by 100 milliseconds MSMS analyses on the 20 most intense peaks seen by MS. The MS/MS data file generated was analysed using the ProteinPilot 4.1 Paragon algorithm (Eskigent AB Sciex) against an internal protein database to which the *pJN132* constructs had been added. Trypsin was set as the cleavage enzyme and carbamidomethyl modification of cysteines. The Mascot algorithm (Matrix Science, Matrix Science Ltd, London, UK) was also used for analysis; this was set for trypsin cleavage at only one end of the peptide (semi-trypsin), carbamidomethyl as a fixed modification of cysteines and methionine oxidation and deamidation of glutamines and asparagines as variable modifications.

2.2.12 Echinoderm Embryo Transfection

Echinoderm embryo transfections were undertaken at the Scottish Institution for Marine Science (SAMS) with the kind collaboration of Dr. Lars Brunner who provided gravid *Psammechinus miliarsis* sea-urchins and UV sterilised seawater. Transfections were undertaken as follows (Bulgakov *et al.*, 2002): gravid adult urchins (3-5 cm test diameter) were agitated to initiate spawning by holding them in air and vigorously shaking them for 3 minutes before placing them into individual beakers containing UV sterilised seawater. Spawning began within 1-2 minutes of placement in the beakers.

Eggs (yellowish-white) or sperm (milky-white) were collected from individual urchins by means of sterile Pasteur pipettes and transferred into sterile 50 mL Falcon tubes. Approximately 500 µL of sperm solution (containing a mix of sperm from 3 individual urchins) was added to 10 mL eggs (again from 3 individuals) and mixed by gently inverting the tube several times. The fertilisation mix was incubated at room temperature, with aliquots checked periodically under a light microscope until the majority of eggs (>95 %) were fertilised (the vitteline layer, the gel-like outer protective coat, had swollen and a few (<5 %) of embryos were already exhibiting their first cleavage division). At this point (after approximately 20 minutes), the fertilisation mix was transferred into 2 mL microcentrifuge tubes and spun for 45 seconds at 500 *g* to concentrate the

embryos. For each tube the supernatant was discarded and the embryos were re-suspended in 1mL sterile seawater and their number counted by transfer into a Sedgewick-Rafter Counting Chamber (Pyser-SGI Ltd., Edenbridge, Kent, UK). The concentration was adjusted to 2500-3000 embryos per mL by adding sterile seawater.

One microgram of plasmid DNA (*pJN132* encoding *STR6*^{wt} or *STR6*^{NAGP}) was added to 500 µL embryo suspension in sterile 15 mL centrifuge tubes, incubated for 1 minute, before the addition of an equal volume (500 µL) of 20 % [w/v] PEG made up in sterile seawater (polyethylene glycol MW 4000 (Sigma-Aldrich). The transfections were incubated for 20 minutes at room temperature, then diluted 10x with sterile seawater (to wash away the PEG) and centrifuged at 500 *g* for 1 minute. The embryos were re-suspended in 10 mL sterile seawater and transferred to sterile T25 cell culture flasks (Griener). They were cultured at 15 °C for one month with their development observed by means of EVOS microscopy. Feeding was twice weekly with 1 mL of *Dunaliella spp.* algal culture (a kind gift from Dr. Brunner).

2.2.13 Plant Infections

Plant expression vectors were created as follows: STR6^{wt}-mCherry-TaV-GFP and STR6^{NAGP}-mCherry-TaV-GFP reporter constructs were amplified from *pJN132* vectors using the forward primer 5'-ATAGCGTTAATTAAAGCTCTAGAACCATGG**ATG**GATTCTG-3' and reverse primer 5'-A**TTA**ATGCGGCCGCCTCGAGTTACTTATACAG-3' (PacI/NotI restriction sites underlined; translation start/stop codons in bold typeface). PCR products were digested with PacI and NotI and ligated into the Tobacco mosaic virus-based overexpression vector *pTRBO* (Lindbo, 2007), similarly restricted.

STR6 and *STR6*^{NAGP} reporter constructs in *pTRBO* were each electroporated into *Agrobacterium tumefaciens* (strain AGL1) cells. Single agrobacteria colonies were grown in liquid culture at 28 °C for 2 days, pelleted and re-suspended in infiltration medium (10 mM MES; 10 mM $MgCl_2$; 15 µM acetosyringone) to an optical density of 0.001 at 600 nm. The agrobacterium suspension was infiltrated into small incisions on the abaxial side of *Nicotiana benthamiana* leaves using a syringe without a needle. Plants were kept at 25 °C, 16 hour light/8 hour dark and imaged at 4 days post-inoculation.

Infiltrated leaves were detached and stuck onto glass microscope slides with the abaxial side facing up using double-sided sticky tape. Leaves were imaged on an upright SP2 confocal laser scanning microscope (Leica) equipped with a 40x water dipping lens. GFP was excited at 488 nm and detected at 495-525 nm, mCherry was excited at 594 nm and detected at 600-630 nm.

Chapter 3. Non-LTR Associated 2As

'I have no data yet. It is a capital mistake to theorise before one has data.

Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.'

A Scandal in Bohemia – Arthur Conan Doyle, 1891

Translational recoding through ribosome skipping via 2A was thought to be a purely viral mechanism until the discovery of 2A-like sequences from the genomes of trypanosomes; single-celled protozoan parasites that are the causative agents of diseases such as sleeping sickness and Chagas disease (Heras *et al.*, 2006). Examination of the gene and protein structure surrounding the trypanosome 2A-like sequences found that the 2As occurred within *L1Tc* non-long terminal repeat (non-LTR) retrotransposon genes. *L1Tc* elements are typical of "young" APE-type retrotransposons in that they possess two ORFS. The first (ORF1) encodes regulatory proteins homologous with viral gag proteins, or proteins with nucleic acid binding ability, and the second (ORF2) encodes the proteins required for translocation, namely apurinic/apyrimidic endonuclease and reverse transcriptase (RT) (Figure 3.1 and Figure 3.3). The 2A sequence was found to be located at/near the N-terminus of ORF2. My colleagues and I have therefore suggested that in these retrotransposons, 2A has replaced the function of the regulatory proteins of ORF1 by regulating the expression of ORF2 (Odon *et al.*, 2013). These trypanosome 2A sequences contained the characteristic C-terminal motif -D[V/I]ExNPG$^{\downarrow}$P-, or a variant -D[V/I]ExHPG$^{\downarrow}$P-, which provided the means of their identification from database probes, moreover they were found to be reasonably active *in vitro* using the polyprotein reporter system (*pSTA1* encoding [GFP-2A-GUS]) developed for viral 2A analysis. The NPG$^{\downarrow}$P sequences were highly active, with a similar activity level to most viral 2As, whereas the HPG$^{\downarrow}$P sequences activities were somewhat reduced (~23% activity) in comparison. (Heras *et al.*, 2006).

Following the detection of ribosome skipping 2A-like sequences within trypanosome non-LTRs (Heras *et al.*, 2006), the Ryan laboratory undertook a search of online genomic and proteomic databases in an attempt to identify any other putative eukaryotic 2A-like sequences. They found a small number of such sequences (reviewed in Luke *et al.*, 2010b; Luke and Ryan, 2013). Given the exponential increase in both the quantity and quality of freely available searchable online genetic sequencing data, it was considered highly likely that these sequences were merely the first to be catalogued. Therefore, it was proposed that a fresh search should be undertaken, the results of which form the focus of this thesis, and are reported here (Chapter 3) and in subsequent Chapters and are listed in Appendix B. However, before reporting on the results of my database search; a short introduction to transposons, particularly non-LTR elements, will be presented in order to place the discovery of 2As in non-LTR retrotransposons into context.

3.1 **Transposons**

Transposable elements (alternatively termed TEs or transposons) are short tracts of DNA that can move position within the genome in their cell of origin. Their movement (transposition) often results in the duplication of the TE, and depending on the site of integration can cause changes in host genotype that can in turn alter its phenotype due to modifications in the transcription and translation of the surrounding genes. It was such phenotypic changes in the leaf colour of maize plants that resulted in the discovery of TEs or "jumping genes" in the 1940s by Barbara McClintock (McClintock, 1950), work that, half a century later, in 1983, would see her awarded a Nobel Prize.

TEs occur in virtually all plant and animal genomes examined to date (the only eukaryotic genome that apparently lacks TEs is that of the malaria parasite *Plasmodium falciparum*), and comprise a large proportion of the genomic DNA (Wicker *et al.*, 2007). For example, in mammals, TEs and degenerate TE-derived sequences comprise around half the genome, and in some plants TEs account for up to 90% of the genome (SanMiguel *et al.*, 1996). Historically, TE sequences were considered as merely "junk" non-coding DNA, of interest to horticulturalists due to the variegated patterns they could cause to form on plant leaves/flowers, but of little overall importance. However, researchers are now beginning to discover the myriad ways in which TEs can either positively direct the evolution of their hosts and/or cause deleterious effects. In humans, TE insertion has been shown to have many effects including regulation of gene expression, increased recombination rate, and unequal crossover during meiosis. TE insertion into genes required in brain development is thought to have influenced human brain evolution, perhaps even being the event that first "made us human". The various unique and inherited patterns of TE insertion in peoples from different geographic areas have proven to be invaluable in identifying a person's origins during genotypic and forensic studies (recently reviewed by Ayarpadikannan and Kim, 2014). TE sequences have also been used in genetic engineering to insert novel genes into embryonic stem cells (Davidson *et al.*, 2009). On the other hand, TE transposition has been implicated in a range of hereditable human diseases including genetic disorders such as haemophilia and Duchenne's muscular dystrophy, and some psychiatric conditions and cancers (reviewed in Ayarpadikannan and Kim, 2014). There is a growing body of evidence that shows eukaryotic genomes are not defenceless against 'attack" or colonisation by TEs. A number of epigenetic mechanisms prevent/minimise TE transposition including differential DNA methylation (Ayarpadikannan and Kim, 2014), blocking TE translation through targeted siRNA silencing (Chung *et al.*, 2008) or enzymatic degradation of foreign nucleic acid sequences (Sawyer *et al.*, 2004).

Despite their hosts' specific defences, and their constant loss through the general processes of DNA repair, recombination, genetic drift, silent mutation and other genomic rearrangements, TEs persist, and are superb examples of successful "selfish genes" (Dawkins, 1976; Doolittle and Sapienza, 1980; reviewed by Orgel and Crick, 1980). There must be sufficient copies of an element to ensure its survival in its current host genome (and potentially maximise the probabilities of it being transferred to a future host through horizontal gene transfer) but, transposition events must be infrequent enough that their host does not incur a loss of fitness. The probabilities of any one TE persisting, its "population dynamics" have been calculated and expressed using mathematical models (Charlesworth and Charlesworth, 1983) similar to those employed in the study of species colonisation and persistence in standard ecological fields such as island biogeography.

Once integrated into the germ-line DNA of their host, TEs are inherited following vertical transmission from parent to offspring in accordance with the rules of standard Mendelian genetics. However, there is a growing body of evidence (reviewed in Silva *et al.*, 2004; Walsh *et al.*, 2013) that some TEs can be and have been acquired through horizontal gene transfer. The precise details of this process, such as the transmission vector(s) have still to be elucidated, but there is a growing body of evidence that parasites such as ticks and viruses are the likely vectors (reviewed in Silva *et al.*, 2004). TEs are classified into two main groups (Figure 3.1) based upon their mode of replication: Class I elements include retrotransposons, group II mitochondrial introns, and endogenous retroviruses. These replicate by means of a "copy and paste" mechanism (Figure 3.1a) involving the use of an RNA intermediate and in general can synthesis their own DNA during replication, whereas Class II or DNA transposons move by means of a "cut and paste" mechanism (Figure 3.1b) or in rare cases, a rolling circle mechanism and are reliant on their hosts replication machinery for DNA synthesis (Wicker *et al.*, 2007). Most TEs are flanked by target site duplications (TSDs) resulting from DNA repair of damage (staggered DNA nicks) generated at the target site by TE insertion (see Figure 3.1) (reviewed in Jurka *et al.*, 2007).

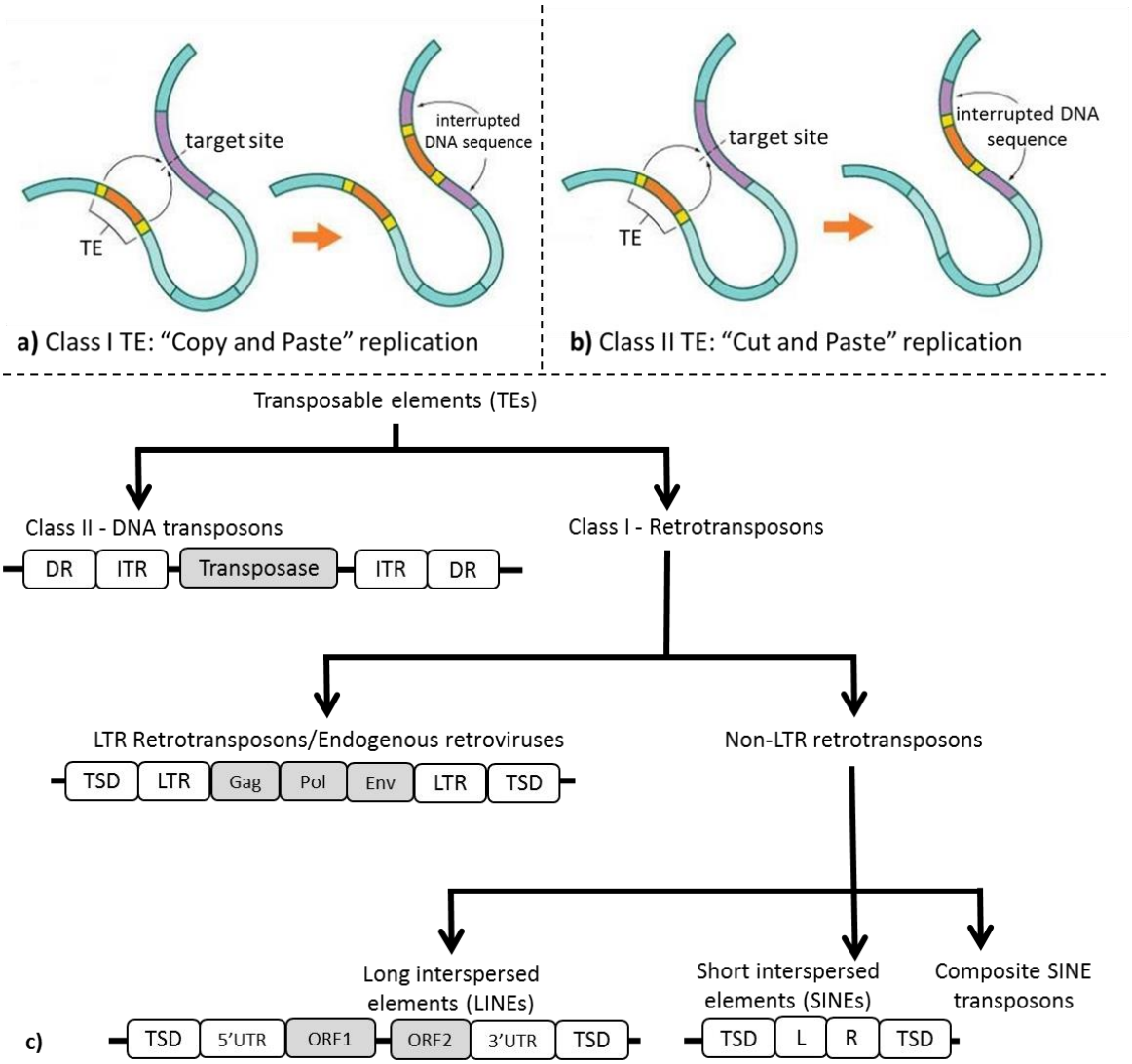


Figure 3.1 Classification, replication and structure of transposable elements
a) Class I "Copy and Paste", **b)** Class II "Cut and Paste" replication, and **c)** Schematic showing the main groups of TEs, representations not to scale. Key to abbreviations: DR = direct repeat; ITR = inverted terminal repeat, TSD = tandem site duplication, LTR = long terminal repeat, UTR = untranslated region, ORF = open reading frame, Gag = Gag/Gag-like proteins, Pol = DNA polymerase, Env = envelope protein, L = left; R = right. Note that some LINE non-LTRs possess a single long ORF. Schematic diagram modified from Ayarpadikannan and Kim, 2014; replication mechanism cartoons modified after image by Lauren Solomon, Broad Institute of MIT & Harvard: https://www.broadinstitute.org/, accessed October 2014.

3.1.1 Non-LTRs

Non-LTR transposons can be subdivided into SINEs (short interspersed elements) and LINEs (long interspersed elements) (Figure 3.1c). SINEs are non-autonomous short mosaic structures derived from transfer RNA (tRNA) or 7S or 5S ribosomal DNA with 3' termini from LINE elements. They possess internal pol III promoters that govern their transcription, but lack reverse transcriptases (RTs) or endonucleases (EN) therefore their retrotransposition is reliant upon a supply of these enzymes encoded by the autonomous LINE elements (reviewed in Jurka *et al.*, 2007).

LINEs are autonomous, encoding all the enzymatic machinery required for their retrotransposition. All possess an internal promoter in their 5' UTR that governs transcription of a single transcript mRNA that encodes either one or two ORFs. They are assigned to one of five groups, termed *R2*, *L1*, *RTE*, *I* and *Jockey* (reviewed in Jurka *et al.*, 2007), which can be further divided to give 15 clades (reviewed in Jurka *et al.*, 2007) or 17 clades (reviewed in Wicker *et al.*, 2007) based on alignment of their reverse transcriptase (RT) domains (Malik *et al.*, 1999). Assuming solely vertical descent, phylogenetic analysis based on synonymous substitution rates suggests that LINEs are as old as eukaryotic life, with origins in the pre-Cambrian paleo-geological era, and that each clade/group is monophyletic descending from a single or group of closely related parental sequences (Malik *et al.*, 1999). The evolutionarily most ancient clades of LINEs (specifically, the *Cre*, *NeSL*, *R2*, *Hero* and *R4* clades) are termed "ancient non-LTRs". These encode a single ORF (here to be termed ORF2) encoding a multi-functional protein comprising both a reverse transcriptase and a restriction enzyme-like endonuclease (REL-endo) domain. One clade *(Dualan*/*Randl*) possesses an additional apurinic/apyrimidinic DNA endonuclease (APE) domain. This clade is thought to represent an intermediate stage that led to the evolution of the more diverse "young APE-type non-LTRs" where the REL-endo domain has been replaced by an APE domain (reviewed in Malik *et al.*, 1999). The 5' region of young APE-type non-LTRs is plastic, in that many of these elements possess two ORFs (ORF1 and ORF2), whereas others lack an ORF (Albalat *et al.*, 2003), for example the trypanosome *L1Tc* elements found to possess 2A-like sequences, lack an ORF1 (Heras *et al.*, 2006)). In young APE-type non-LTRs, ORF1 (when present) encodes a protein domain homologous with viral gag proteins or proteins with nucleic acid binding ability and the second (ORF2) an endonuclease and a reverse transcriptase (RT) domain. A related, albeit little investigated, group of LINEs are the *Penelope-like* elements. These possess an analogous structural arrangement to young non-LTRS with one or two ORFs coded by a single mRNA; however they encode a GIG-YIG type REL-endo domain instead of an APE domain, and their RT domain is encoded upstream of their GIG-YIG endonuclease domain, they may also contain intron sequences that are spliced out of their final mRNA. The exact phylogenetic relationship of *Penelope* elements to young APE-type non-LTRs is presently unknown, partially due to the propensity of these elements to incur 5' and 3' deletions, and to re-integrate into the genome at integration sites previously filled by other TEs (Arkhipova, 2006).

In LINEs possessing both ORF1 and ORF2, both ORFs are encoded by a single polycistronic transcript mRNA, but the exact mechanism through which translation of the second ORF (ORF2) is initiated is presently unknown (Alisch *et al.*, 2006). However, in the case of the *SART1* element, ORFs 1 and 2 are linked by an overlapping stop-start codon (-UAAUG-). Here, the efficiency of ORF2 translation initiation is dependent on a downstream region of RNA secondary structure: increasing the distance between this and the stop/start codon decreases the level of ORF2

translation (Kojima *et al.*, 2005). Overlapping stop-start as a termination-re-initiation strategy is not unique to LINE elements: it is also found in a number of RNA viruses including *Influenza Viruses* (Horvath *et al.*, 1990; Powell *et al.*, 2008), *Respiratory Syncytial Viruses* (Ahmadian *et al.*, 2000; Gould and Easton, 2005), *Pneumoviruses* (Gould and Easton, 2007), and *Caliciviruses* (Meyers, 2003; Meyers, 2007; Luttermann and Meyers, 2007).

The mechanism of LINE retrotransposition has been well characterised. It occurs through a process termed target-primed reverse transcription (TPRT). Briefly, the mRNA strands transcribed from the genomic DNA copy of the LINE element act as targets for their self-encoded RT to transcribe into DNA. cDNA integration is primed by the free 3' hydroxyl group at the target DNA nick introduced by their endonucleases (reviewed in Jurka *et al.*, 2007). In *cis* co-expression of ORFs 1 and 2 are essential for retrotransposition (Moran *et al.*, 1996); however bioinformatic analysis on the ORF1 proteins reveals a range of different proteins encoded across the LINE clades, suggesting an independent origin for the acquisition of each ORF1. In addition, during the retrotransposition process, TEs frequently undergo 5' and 3' end truncation which causes the loss of a functional ORF1; and in some cases also the loss of the initiation codon of ORF2, they also accumulate mutations through copy errors during retrotranspositon, all of which results in their eventual transfer from an active to an inactive coding sequence. For example, although over 20% of the human genome is comprised of LINE1 (*L1*) elements, amounting to over 100,000 of such sequences in various levels of truncation/degradation/mutation (such that a total number is difficult to calculate as many sequences are now so degenerate that classifying them as LINEs proves problematic), a recent study found only 410 human *L1* copies were represented by transcript mRNA (Rangwala *et al.*, 2009). Therefore, finding 2A-like sequences within non-LTR genomic sequences will be no guarantee that these sequences are currently being actively transcribed.

3.2 **Non-LTR 2As – Methodology**

3.2.1 Contributors

The work reported in this Chapter, was undertaken as a collaborative effort between members of the Ryan laboratory: Prof. Martin Ryan and Dr. Andriy Sukhodub conducted the initial search and alignment of the RT domains of APE-type non-LTR sequences; the updated alignments were performed by the author. 2A sequences cloned by the author are listed in Table 3.1 and Table 3.2. Additional sequences were cloned by Dr. Garry Luke and Valerie Odon. TnT reactions, SDS-PAGE gels and the subsequent analyses were undertaken by the author.

3.2.2 *In Silico* - Methodology

Online proteomic and genomic databases were probed for eukaryotic 2A-like sequences using the viral 2A conserved C-terminus motif -D[V/I]ExNPGP-. All sequences discovered, were recorded, along with their flanking sequence and databank accession number to form an in-house 2A-like sequence database (Appendix B). Note that new data is constantly being uploaded to these databases as genome sequencing projects progress; however, conversely, quality control measures since 2011 have resulted in the removal of a large number of sequences, particularly from the sea-urchin sequence database initially compiled by Baylor College of Medicine (now hosted by the Max Planck Institute for Molecular Genetics, available at http://goblet.molgen.mpg.de/cgi-bin/seaurchin-genombase.cgi). Hence, the current dataset (Appendix B) should be considered merely a snapshot of current knowledge at the time of writing and not a definitive list of all eukaryotic 2A sequences. The removed sequences have been omitted from the analyses and Appendix B, excepting those that were selected for *in vitro* analyses prior to the date of their removal, in which case they remain but are noted as potentially redundant in Appendix B.

Sequences in this putative eukaryotic 2A-like database were screened for the presence of TEs in their flanking protein/genes. Sequences were screened by use of the "find conserved domain" function of NCBI BLAST to distinguish reverse transcriptase (RT) or if this were lacking, exo-endonuclease domains (EEP domains) by deign of sequence homology with known TE sequences. Classification of non-LTRs was conducted using the Repbase RTclass1 web server (www.girinst.org/RTphylogeny/RTclass1, (as detailed in Odon *et al.*, 2013) based on core RT domain amino acid homology (Malik *et al.*, 1999)). Sequence alignments were performed using ClustalX2 and visualised using Figtree v1.4.2 or Phylodraw). In the interests of clarity, 2A non-LTR sequences selected for further analyses were given short identification tags following the Repbase nomenclature (non-LTR type/number followed by species identifier suffix derived from their Latin name). Suffixes of the identification tags indicate host species as follows: *_AC Aplysia californica* (sea-slug, mollusc), *_AD Angomonas deanei* (trypanosome, protozoan), *_BF Branchiostoma floridae* (amphioxus/Florida lancelet, cephalochordate) *_BG Biomphalaria*

glabrata, (freshwater aquatic snail, mollusc), *_CE Caerorhabditis elegans* (nematode worm, nematode), *_CGi* (*Crassostrea gigas*, Pacific oyster, mollusc), *_CV* (*Chlorella variabilis* (unicellular green alga, plant) *_LG Lottia gigantean* (owl limpet, mollusc), *_NVe Nematostella vectensis*, (sea anemone, cnidarian), *_OM* Oncorhynchus mykiss (rainbow trout, chordate), *_RP Rhipicephalus pulchellus* (zebra tick, arthropod), *_SK Saccoglossus kowalevskii* (acorn worm, hemichordate), *_SP Strongylocentrotus purpuratus* (purple sea urchin, echinoderm), *_SS Salmo salar* (Atlantic salmon, chordate), and *_XT Xenopus (Silurana) tropicalis* (African claw-toed frog, chordate). These tags together with the accession numbers are used in Appendix B and throughout this Chapter.

3.2.3 *In Vitro* – Methodology

Cloning of putative 2A sequences into the reporter *pSTA1* plasmid (a kind gift from Dr. Garry Luke) was undertaken as described in Chapter 2 and Figure 3.2. Briefly, the 2A sequences were encoded by gene-blocks or cloned by means of long reverse primers. The gene-blocks were initially ordered as part of the 2A activity screen reported in Chapter 8; therefore, a fuller explanation of their design is provided there. The decreasing cost of the synthesis of long primer sequences (100 base pairs and above), permitted their use instead of gene-blocks as these studies progressed. These sequences were cloned by PCR using a forward primer flanking the BamHI site in *pSTA1* and a long reverse primer encoding the novel 2A 5 ’sequence flanked by an XbaI site and on the 3’ end an ApaI site and the C-terminal nucleotides of GFP. The PCR product was BamHI/ApaI digested and ligated into *pSTA1* similarly restricted. An artificial mutant 2A sequence (*STR69*^{mut}) was constructed by means of mutagenesis PCR on the construct *pSTR-69*. The 2A sequences and primers are listed in Tables 3.1 and 3.2. After verification by DNA sequencing using primers *GFPf* and/or *GUS_seq_R* (Table 2.1) the plasmid preparations were used to program TnTs analysis as detailed in Chapter 2.2.2 and in Odon *et al.*, 2013.

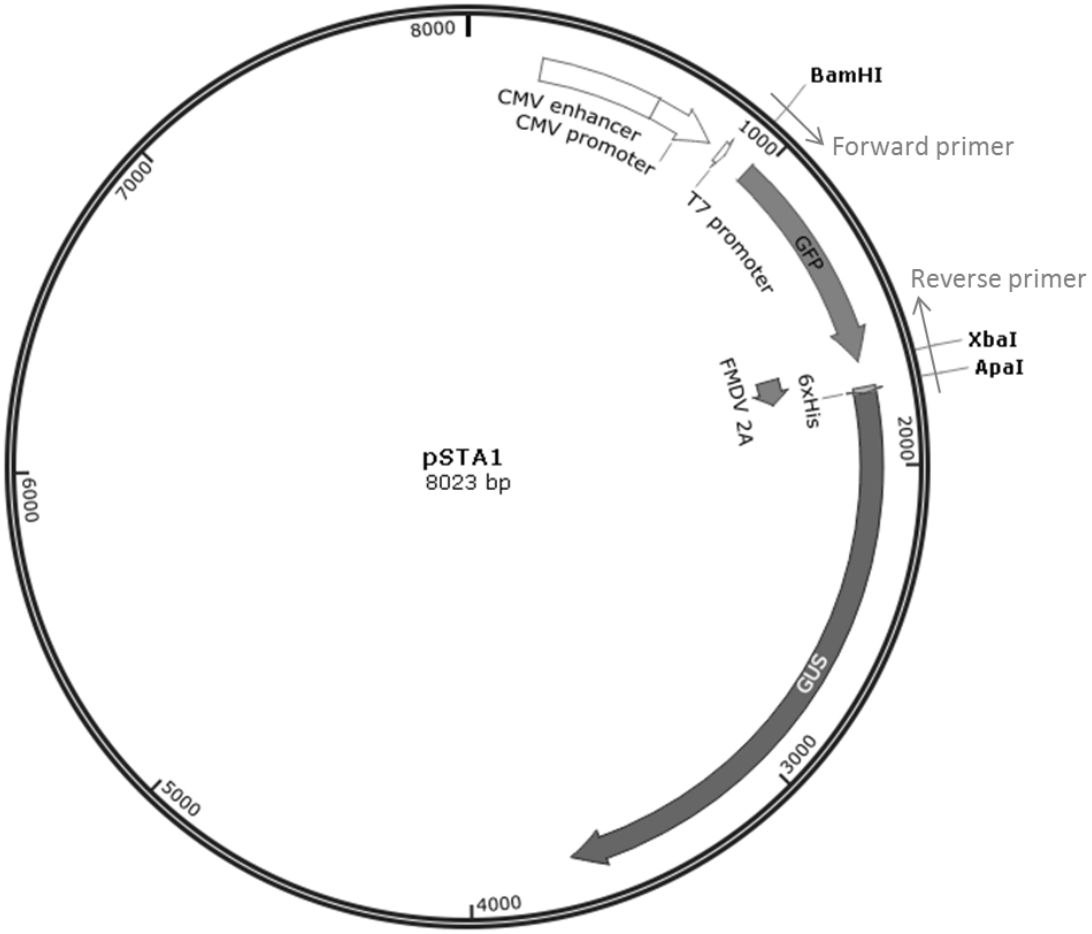


Figure 3.2 *pSTA1* vector and cloning strategy
Schematic detailing the *pSTA1* reporter plasmid showing the relative positions of the 2A sequence (novel 2As of interest were inserted in place of *FMDV 2A*), and the restriction enzyme/PCR primer sites used in cloning.

Table 3.1 Non-LTR 2As cloned by means of PCR.
These sequences were assigned to non-LTR clades on the basis of their downstream in-frame RT domain.

Designation	2A Sequence	Reverse primer sequence
Ingi-1_AC	FLGGQHNPAWLARLLILAGDVEQNPGP	5′GGTGGGGCCCTGGATTCTGTTCTACG TCTCCTGCTAGTATTAGTAGTCGTGCTA GCCATGCTGGATTGTGCTGTCCACCTAG AAATCTAGACCCGGACTT-3′
CR1-CGi	SRHIVVYNFYLQFFMFLLLLCGDIEVNPGP	5′GTGGGGCCCTGGATTTACTTCTATGT CTCCGCAGAGTAGCAGTAGGAACATAAA GAATTGAAGATAGAAGTTATACACTACG ATATGTCTAGACCCGGAC-3′
CR1-1_LG	NDTFSSILYYCFILIIRSGDIELNPGP	5′GGTGGGGCCCTGGGTTTAGTTCTATG TCTCCTGATCGTATTATTAGTATGAAGC AGTAGTACAGTATTGATGAGAATGTGTC GTTTCTAGACCCGGACTT-3′
CR1_BG	FSVRDSRIKYLSLLILIAGDVESNPGP	5′GTGGTGGGGCCCTGGATTTGACTCTA CATCTCCTGCTATTAAGATTAGCAGTGA TAGGTACTTAATTCTGCTGTCTCTGACT GAGAATCTAGACCCGGAC-3′
L2A-1_NVe	KRYPNSTSTFQLTRIAVSGDVSPNPGP	5′GGTGGGGCCCAGGATTTGGACTAACA TCTCCACTAACTGCAATTCGTGTTAGTT GAAATGTACTTGTACTATTAGGATATCG TTTTCTAGACCCGGACTT-3′
Crack-3_NVe	MTKVGICAFSLIILSGDISLNPGP	5′GGTGGGGCCCAGGGTTCAGACTAATA TCTCCACTCAGAATAATAAGACTAAATG CACAAATACCTACTTTAGTCATATAAAT GCTTCTAGACCCGGACTT-3′
Neptune_NVe	MLFVLPILIAKSMDIETNPGP	5′GGTGGGGCCCAGGATTTGTTTCAATA TCCATACTCTTCGCTATAAGAATAGGTA GTACAAATAGCATGAATATTGCTGGTCG ATTTCTAGACCCGGACTT-3′
CR1-L2-1_XT	FKSFSHLLSLSLLLLLAAGDISPNPGP	5′GTGGTGGGGCCCTGGGTTAGGACTAA TATCTCCAGCTGCTAGAAGTAGTAATAG ACTAAGACTAAGTAGATGACTAAATGAC TTGAATCTAGACCCGGAC-3′
RP5_RP	FTSLYADIVHCLCSLLLSGDVELNPGP	5′GTGGTGGGGCCCTGGATTCAGTTCAA CATCTCCACTTAGTAGCAGACTACATAG ACAATGTACAATATCTGCATATAGACTT GTAAATCTAGACCCGGAC-3′
RP8_RP	LSIVVQCCDVIRSLLLLAGDIETNPGP	5′GTGGTGGGGCCCTGGATTGGTTTCAA TATCTCCTGCTAGAAGTAATAGACTTCT AATTACATCACAGCATTGAACTACAATA CTTAGTCTAGACCCGGAC-3′
STR-69^{mut}_SP	CRRIAYYSNSDCTFRLELLKSGDIESNPGP	*FORWARD*: 5′ACTTTTGAAATCAGGCGATATTGAAT CTAACCCTGGTCCT-3′ *REVERSE*: 5′AGGACCAGGGTTAGATTCAATATCGC CTGATTTCAAAAGT-3′

Table 3.2 2A sequences from putative non-LTRs of unknown clade
Sequences identified as putative TEs from the presence of a downstream EEP domain.

2A name	2A Sequence	Nucleotide sequence cloned
SS7_SS	QRSRRPVLIAFSRTLILLLLCSSGDVEVNPGP	5′CAACGATCACGTCGACCAGTGCTAA TAGCATTCTCACGAACACTGATACTAC TGCTGCTCTGCTCCTCAGGAGACGTTG AAGTCAATCCAGGACCT-3′
OM-4_OM	TRRPVILAFSCTLILLLFCSSGDVEVNPGP	5′ACTCGCCGTCCAGTAATCCTAGCAT TCTCATGCACACTAATACTGCTCCTAT TCTGCTCATCCGGAGACGTAGAAGTTA ATCCAGGACCG-3′
CE-1_CE	LCETPSLPHTTFLKRKLLVRSGDVESNPGP	5′CTCTGCGAGACACCATCACTACCAC ACACAACATTCCTAAAACGAAAACTAC TAGTACGATCAGGAGACGTAGAATCAA ACCCAGGACCA-3′
CV-1_CV	LRLPCSCSTTALIKRMKLLLSGDVEENPGP	5′CTACGACTACCATGCTCATGCTCAA CAACAGCACTAATAAAACGAATGAAAC TACTACTATCAGGAGACGTAGAAGAAA ACCCAGGACCA-3′

3.3 **2As in Non-LTRs – Results**

3.3.1 Database Probe

A probe of online databases using the 2A conserved C-terminus motif D[V/I]ExNPGP resulted in over four hundred matches from a wide range of organisms. The complete lists of putative eukaryotic 2A sequences as of September 2014 have been included as Appendix B; the databank accession number and the host organism are provided for each sequence.

Screening the flanking sequence from each newly discovered eukaryotic 2A for the presence of a RT domain indicated that that a substantial number occurred in-frame within non-LTR elements. Not surprisingly, considering the previous report (Heras *et al.*, 2006), these consisted of a number of *Ingi* clade non-LTR sequences from the trypanosomes *T. cruzi*, *T. brucei*, *T. vivax*, and *T. congolense*. There was also an *Ingi* sequence from another parasitic trypanosome, *Angomonas deanei*. However, instances of non-LTRs containing 2A-like sequences were not confined to trypanosomes. The screening revealed 2As in conjunction with non-LTR elements from a cnidarian (*Nematostella vectensis*, sea anemone), an arthropod (*Rhipicephalus pulchellus*, zebra tick), several molluscan species namely, *Aplysia californica* (sea-slug), *Biomphalaria glabrata*, (tropical freshwater aquatic snail), *Crassostrea gigas*, (Pacific oyster) and *Lottia gigantea* (owl limpet), the echinoderm *Strongylocentrotus purpuratus* (purple sea urchin), a hemichordate *Saccoglossus kowalevskii* (acorn worm), the cephalochordate *Branchiostoma floridae* (amphioxus/Florida lancelet), and the chordates *Salmo salar* (Atlantic salmon) and *Xenopus (Silurana) tropicalis* (African claw-toed frog). Through alignment of the downstream RT domain it was possible to assign these sequences to non-LTR clades (Figure 3.3 and Table 3.3).

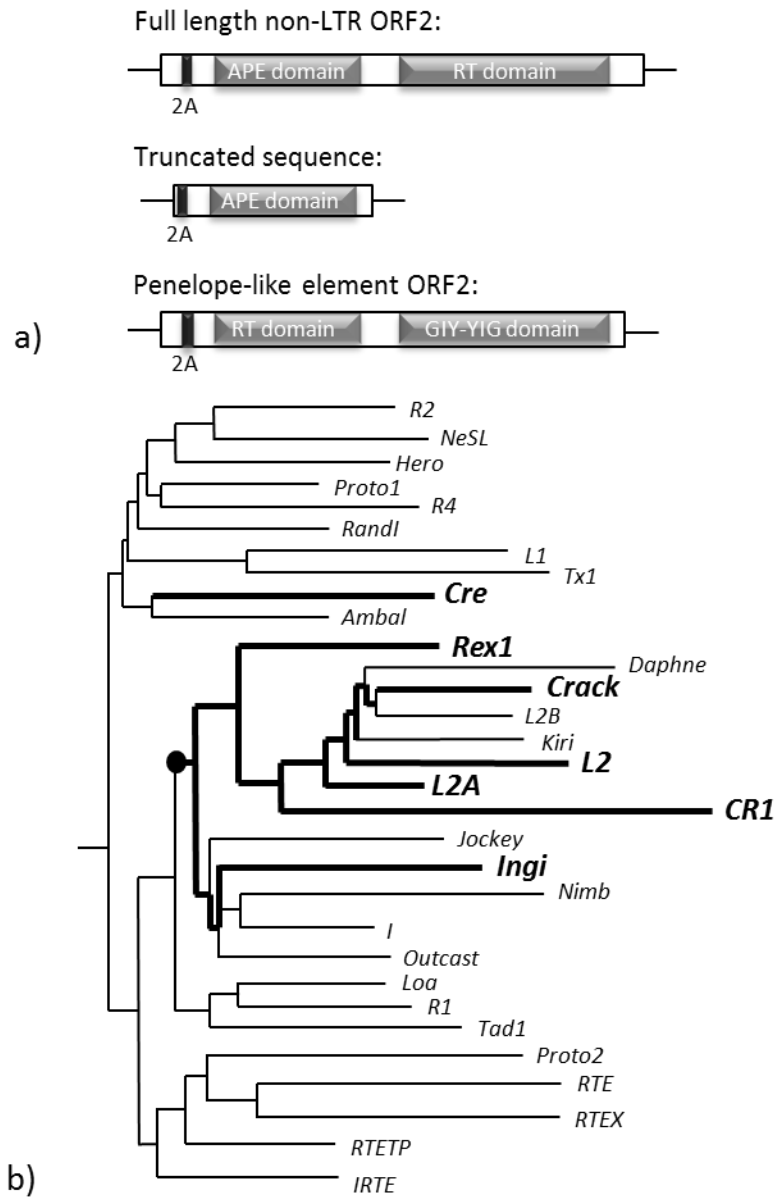


Figure 3.3 2A-like sequences within LINEs
a) Structural organisation diagram detailing the placement of 2A-like sequences within non-LTR ORF2, short TE-like sequences, and *Penelope-like* element ORF2, note that the 2A sequence was encoded in-frame with the downstream domains. **b)** Dendrogram of non-LTR RT domains. The RTclass1 tree file was downloaded from Genetic Information Research Institute (http://www.girinst.org/) and adapted such as each clade is represented by a single line, relative lengths representing the most unrelated element within each clade. Clades with containing non-LTR elements encoding 2A-like sequences (active or inactive at instigating ribosome skipping) are shown in bold and their branches highlighted. Note that the marking of a clade in bold does not indicate that all elements in this group contain 2A-like sequences, but merely that a 2A-like sequence has been identified from one or more elements from this clade. The latest common ancestral node of the main 2A containing cluster is indicated with a closed circle. *Penelope-like* elements are not shown on the dendrogram, as their relationship to non-LTRs is a matter of some controversy.

Table 3.3 Non-LTRs containing 2A sequences

Sequences listed by non-LTR clade. Accession numbers represent their NCBI/GenBank, sea-urchin database GLEAN or Repbase identifiers. 2A translational recoding activity is reported using a semi-quantitative scale calibrated against *FMDV 2A*: +++ highly active, ++ active, + weak, (+) very weak, - no activity, NT not tested.

Non-LTR/ TE Clade	Accession no.	Lab Identifier	2A Sequence	Recoding Activity
Cre	XP_002731766	*SK1_SK*	-RYSKSSCEWMWFLVLLSFIL**SGDIE**V**NPGP**-	NT
	JAA55846.1	*RP1_RP*	-IVLPCPEAVLAPFCSLLLLLC**GDVE**S**NPGP**-	NT
	JAA55762.1	*RP4_RP*	-PNSLVACCSRVLHYFDGLLL**SGDVE**L**NPGP**-	NT
	JAA55454.1	*RP5_RP*	-SFVFTSLYADIVHCLCSLLL**SGDVE**L**NPGP**-	++
	JAA64478.1	*RP6_RP*	-RLLCPVFYDLSVSLGKLLLLA**GDIE**T**NPGP**-	NT
	JAA55744.1	*RP8_RP*	-HCCLSIVVQCCDVIRSLLLLA**GDIE**T**NPGP**-	++
Rex1	GLEAN3_18025	*STR-40_SP*	-KSCISYYSNSTACFNIEIMCC**GDV**K**SNPGP**-	NT
	GLEAN3_24854	*STR-55_SP*	-GARISYHPNTTATFQLRLLVS**GDV**N**PNPGP**-	NT
	GLEAN3_22393	*STR-61_SP*	-GARIRYYNNSSATFQTILMTC**GDV**D**PNPGP**-	++
	GLEAN3 19055	*STR-89 SP*	-GRRIQYYNNSISTFRSELLRC**GDVESNPGP**-	NT
	GLEAN3_26896	*STR-197_SP*	-KHPILYYTNGESSFQIELLSC**GDI**N**PNPGP**-	++
L2	CR1-L2-1_XT	*L2-1_XT*	-HNNFKSFSHLLSLSLLLLLAA**GDI**S**PNPGP**-	+
	L2-2_XT	*L2-2_XT*	-RNHFKSSAHVFSLLFLLLLAA**GDV**S**PNPGP**-	NT
	L2-3_XT	*L2-3_XT*	-KTKTYKSRSHLAFLSFLLLAA**GDI**S**PNPGP**-	NT
	L2-4_XT	*L2-4_XT*	-PRAFKSRSHLLSLTLLLLAA**GDI**S**PNPGP**-	NT
	GLEAN3_22449	*STR-51_SP*	-SRPILYYSNTTASFQLSTLLS**GDIE**P**NPGP**-	++
	GLEAN3_27016	*STR-69_SP*	-CRRIAYYSNSDCTFRLELLKS**GDI**Q**SNPGP**-	++
	Artificial	*STR-69_SP(mut)*	-CRRIAYYSNSDCTFRLELLKS**GDIESNPGP**-	++
	GLEAN3_00868	*STR-133_SP*	-KRRIPYNPNSTASFQLELLHA**GDV**H**PNPGP**-	NT
	GLEAN3_14631	*STR-142_SP*	-KTRIPYSVNSNASFQLELLHA**GDV**H**PNPGP**-	NT
L2A	L2A-1_NVe	*L2A-1_NVe*	-GRIKRYPNSTSTFQLTRIAVS**GDV**S**PNPGP**-	-
Crack	Crack-3_NVe	*Crack-3_NVe*	-LRASIYMTKVGICAFSLIIL**SGDI**S**LNPGP**-	-
	Crack-9_BF	*Crack-9_BF*	-HVKTSVNLAHLCIHTLLLLS**GDV**A**CNPGP**-	NT
	Crack-10_BF	*Crack-10_BF*	-LYHKNLLTEQCNDQVNLICLAF**DI**H**PNPGP**-	NT
	Crack-11_BF	*Crack-11_BF*	-CHVETRVNVVHLCLHTLLLLS**GDV**A**SNPGP**-	NT
	Crack-15_BF	*Crack-15_BF*	-HSVLVCDHCVTVFVVILLLLC**GDI**H**NNPGP**-	(+)
	Crack-16_BF	*Crack-16_BF*	**-DI**Q**TNPGP**-	NT
	Crack-17_BF	*Crack-17_BF*	-AVTSTSVNCVHLCFHTLLILS**GDV**A**VNPGP**-	(+)
	Crack-28_BF	*Crack-28_BF*	-TCTERTERTLNLLVCATLLLA**GDV**S**PNPGP**-	NT
	GU129139.1	*SS8_SS*	-HRRPILIAFSRTLILILLCS**SGDVE**V**NPGP**-	NT
CR1	CR1-1 BF	*CR1-1 BF*	-KKTMIHNDSTKLSLIMILLLS**GDIE**I**NPGP**-	+
	CR1-2_BF	*CR1-2_BF*	-RTSDRLFTCLLYLCSVLMSQAV**DLE**T**NPGP**-	-
	CR1-3_BF	*CR1-3_BF*	-YLRTSDRLCLLYICSVLMAQAV**DLE**T**NPGP**-	NT
	CR1-11_BF	*CR1-11_BF*	-LAPHCRPKFTLFSLTLIILLA**GDVE**L**NPGP**-	-
	CR1-12_BF	*CR1-12_BF*	-PRNPLKSISVSIALLVMLTQS**GDV**H**PNPGP**-	NT
	CR1-18 BF	*CR1-18 BF*	-YLRTSDRLCLLYICSVLMAQAV**DLE**T**NPGP**-	NT
	CR1-31 BF	*CR1-31 BF*	-YLMSRQRLVLLYLTMLLISKSYSP**EP**N**PGP**-	(+)
	XP_797143.2	*STR-1 SP*	-MFVCAFILISVLLLS**GDVE**I**NPGP**-	(+)
	XP_001196407.1	*STR-24_SP*	-MCA**GDV**Q**PNPGP**-	NT
	XP_001179204	*STR-28_SP*	-MGVAESTSLSHLTILLLLS**GQVE**T**NPGP**-	(+)
	XP_001185404.1	*STR-32_SP*	-NSSCVLNIRSTSHLAILLLLS**GQVE**P**NPGP**-	++
	XP_001184905.1	*STR-33_SP*	-LPVNEYRSTSLSHLTILLLLS**GQVE**T**NPGP**-	NT
	XP_001196844.1	*STR-34_SP*	-NSTPAAMFVCVFILISVLLLS**GDVE**IS**PGP**-	NT
	XP_001200466.1	*STR-35_SP*	-NSSCVLNIRSTSHLAILLLLS**GQVE**P**NPGP**-	NT
	CR1-1_CGi	*CR1-1_CGi*	-SRHIVVYNFYLQFFMFLLLLC**GDIE**V**NPGP**-	++
	CR1-1_LG	*CR1-1_LG*	-TLLNDTFSSILYYCFILIIRS**GDIE**L**NPGP**-	++
	CR1-1_BF	*CR1-1_BF*	-KKTMIHNDSTKLSLIMILLL**SGDIE**I**NPGP**-	NT
	CR1-2_BF	*CR1-2_BF*	-ILRTSDRLCLLYLCSVLMSQAV**DLE**T**NPGP**-	NT
	CR1-3 BF	*CR1-3 BF*	-CLKTTDKLCLMYLCSILMAQAA**DLE**T**NPGP**-	NT
	CR1-11 BF	*CR1-11 BF*	-LAPHCRPKFTLFSLTLIILLA**GDVE**L**NPGP**-	NT
	CR1-12_BF	*CR1-12_BF*	-PRNPLKSISVSIALLVMLTQ**SGDV**H**PNPGP**-	NT
	CR1-18_BF	*CR1-18_BF*	-YLRTSDRLCLLYICSVLMAQAV**DLE**T**NPGP**-	NT
	CR1-10_BF	*CR1-10_BF*	-GTDNVSAEFTQWKPAIDLTQHY**DV**H**PNPGP**-	-
	CR1-17_BF	*CR1-17_BF*	-TISFILSIFYSNFLLLLIVLSN**DI**H**PNPGP**-	NT
	CR1-26_BF	*CR1-26_BF*	-NLDIFLSYTTVFISFVVILVA**GDV**H**PNPGP**-	NT
	CR1-36_BF	*CR1-36_BF*	-DKDYGIVIQFMLPFFVLFLIC**GDI**H**PNPGP**-	NT
	CR1-46_BF	*CR1-46_BF*	-TLTICPQCILIFISLIMIILA**GDI**H**PNPGP**-	NT
	CR1-53_BF	*CR1-53_BF*	-HFDIFLLFFPLPVLVVLSLIA**GDI**H**PNPGP**-	-
	CR1-2_NVe	*CR1-2_NVe*	-SAILDSPPTRARLLCGLLLLC**GDI**S**LNPGP**-	NT
	CR1-4_NVe	*CR1-4_NVe*	-FRPRRDFTRPNCYLVGLLLLC**GDVA**S**H**P**GP**-	NT
	CR1-8_NVe	*CR1-8_NVe*	-ITYRFGRTGPSHLVMLLLILG**GDVE**L**NPG**D-	NT
	CR1-19 NVe	*CR1-19 NVe*	-TSAFRKHRTFVSIIPGLLLLC**GDI**I**S**Q**PGP**-	NT
	CR1-20 NVe	*CR1-20 NVe*	-MNVGRSSSEHKHLLLCLLLLG**GDI**Q**LNPGP**-	NT
	CR1-21_NVe	*CR1-21_NVe*	-RKLIAPRSNPSSLAFRLLILS**GDI**P**LNPGP**-	NT
	AC233256.1	*BG1_BG*	-KWKFSIRHSRNKYLSLLILIA**GDVE**S**NPGP**-	(+)
Ingi		*L1Tc (T. cruzi)*	-QRYTYRLRAVCDAQRQKLLLS**GDIE**Q**NPGP**-	++
		Ingi (T. brucei)	-RSLGTCKRAISSIIRTKMLVS**GDVE**E**NPGP**-	++
		Ingi2 (T. brucei)	-LLLCTCERASIGIHRLLLLLS**GDVE**Q**NPGP**-	NT
		Tvingi (T. vivax)	-ILPCTCGRATLDARRLLLLIS**GDVE**R**NPGP**-	NT
		Tcoingi (T. congolense)	-ILPCTCGRATLDARRILLLVS**GDVE**R**NPGP**-	NT
	EPY38571.1	*AD1_AD*	-RPPTFRRMLPRSALMLMLLLG**GDIE**R**NPGP**-	NT
	Ingi-1 AC	*Ingi-1 AC*	**-**PGFFLGGQHNPAWLARLLILA**GDVE**Q**NPGP**-	+++
Penelope-like	*Neptune1_NVe*	*Neptune1_NVe*	-TLHNRPAIFMLFVLPILIAKSM**DIE**T**NPGP**-	-
	Penelope-5_NVe	*Penelope-5_NVe*	-TLHNRPAIFMLFVLPILIAKSM**DIE**T**NPGP**-	-
Positive Control Viral 2A		*FMDV 2A*	-ELYKSGSGACQLLNFDLLKLA**GDVE**S**NPGP**-	++

Interestingly, with the exception of the *Cre* elements (from the tick and acorn worm), the non-LTR clades containing 2A sequences clustered in one monophyletic grouping, indicative of a single common ancestral sequence for these elements. Further to the occurrence of 2As within non-LTRs, there were two instances of a 2A sequence from another distantly related class of retro-element, *Penelope-like* elements, from the sea anemone *Nematostella vectensis*. A number of additional TEs, most likely non-LTRs, also contained in-frame 2A sequences. These were identified through the occurrence of downstream in-frame EEP domains (the sequences were 3' truncated so lacked RT domains (Figure 3.3 and Table 3.4).

Table 3.4 List of novel 2A sequences associated with putative TEs
The table provides a list of entries in GenBank, which contain a 2A-like sequence upstream of an EEP domain, and therefore recognised as potential 2a containing TEs. The entries from the human parasitic trypanosome species known to possess 2As in non-LTRs (*T. brucei, T. brucei gambiense, T. congolense, T. cruzi, T. cruzi marinkellei* and *T. vivax*) have been excluded from this list (but can be found in Appendix B) as it was known these species possessed 2As in *Ingi L1Tc* clade elements (Heras *et al.*, 2006).

Phylogenetic Group	Host Organism	Genbank Accession nos.
Arthropods	*Daphnia pulex*	EFX60676.1
	Ixodes scapularis	GU318570
	Rhipicephalus pulchellus	JAA55160.1, JAA56189.1, JAA64408.1
Cephalochordates	*Branchiostoma floridae*	XP_002587815.1, EEN43826.1
Chordates	*Oncorhynchus mykiss*	CDQ99460.1, CDR01105.1, CDR01026.1, CDQ81228.1,CDQ97352.1, CDQ81639.1, CDQ59560.1, CDQ84132.1
	Salmo salar	GU129139.1, AF256957, EU025709.1, EF427378.1, EF427382.1, EU008541.1, EF467295, GU817337.1, GU129140.1 EU025715.1, HM159469.1
Cnidarians	*Acropora millepora*	EZ041014
	Nematostella vectensis	XP_001627324.1, XP_001621507.1, XP_001630328.1, XM_001627583.1, XM_001639536.1
Green Algae	*Chlorella variabilis*	XP_005851168.1
Molluscs	*Aplysia californica*	XP_005088979.1, XP_005088979.1.1
	Biomphalaria glabrata	AC233255.1
	Lottia gigantea	ESO94951.1, ESP02481.1
	Ostrea edulis	AFA34358.1
Nematodes	*Caerorhabditis elegans*	Z49911.1
Trypanosomes	*Angomonas deanei*	EPY26545.1
	Strigomonas culicis	EPY32652.1

Irrespective of non-LTR clade, the 2A sequences were all located within 40-80 amino acids from the N-terminus of ORF2, the same location as previously reported for the trypanosome 2A sequences from the *Ingi* and *L1Tc* clades (Heras *et al.*, 2006). The *Penelope-like* elements from *N. vectensis* also followed this pattern with the 2A occurring at the beginning of ORF2 (Figure 3.3). The one exception to this rule was the 2A from the *L2A* clade non-LTR element from *N. vectensis*; here the 2A like sequence was encoded by ORF1, but again was located in the N-terminal region (amino acids 72-102 of ORF1).

During successive retro-transposition events TEs may suffer truncation at both 5' and 3' ends, but primarily from the 5' end. If the TE element has undergone 5' truncation resulting in the loss of the authentic ORF2 initiation codon, then bioinformatic search algorithms will report the translation product as beginning with the next in-frame downstream methionine codon, further truncating the protein sequence available in the searchable databases. Therefore, because the 2A sequences are found in this 5' region, these truncation effects necessarily reduced the ability to identify such elements.

3.3.2 Non-LTR 2As – *In Vitro* Translational Recoding Analyses

Selected representatives of 2As from each non-LTR clade were cloned for *in vitro* analyses. The results of these analyses are summarized in Table 3.3 and Figure 3.4.

A number of these 2A sequences were found to possess ribosome skipping abilities, some to the same, or a greater extent, than the viral reference sequence *FMDV 2A*, but the majority possessed lower translational recoding abilities as *FMDV 2A*, possibly due to residue substitutions within their C-terminal motifs (from the canonical -D[V/I]ExNPG$^{\downarrow}$P-) in some instances.

Both the *Cre* element 2A sequences tested (from the tick, *Rhipicephalus pulchellus*) were found to possess similar recoding activity levels to *FMDV 2A* in this assay, and these sequences corresponded to the "classic" viral C-terminus motif.

The *Rex1* 2A sequences chosen for testing were *STR-61_SP* and *STR-197_SP* (from the purple sea urchin, *Strongylocentrotus purpuratus*) possessed a substitution at the same site within their C-termini (E→D, and E→N, respectively); however, both were as active as *FMDV 2A*, although, as with *FMDV 2A* a band of read-through product is also apparent in addition to the ribosome skippage products.

In the case of the *L2* sequences: *STR-51_SP* (from *S. purpuratus*) conformed to the viral C-terminus motif, whereas *STR-69_SP* (also from *S. purpuratus*) possessed the substitution E→Q. Interestingly both were active in mediating ribosome skipping, and in fact mutating Q→E to recover the canonical motif (*STR-69_SP^{mut}*) did not improve skipping activity, if anything, there

was slightly more read-through, and slightly less ribosome skipping products respectively, were observed for *STR-69_SP*^{mut} than for the wild-type (*STR-69_SP*). In contrast, the other *L2* clade 2A tested, *L2-1_XT* (from the frog, *Xenopus tropicalis*) which also had a substitution in the C-terminus motif, E→S, was considerably less active than the viral control sequence. The single *L2A* clade 2A from *N. vectensis*, also possessed this E→S substitution, and displayed no recoding activity.

In the *Crack* clade, the 2As *Crack-15_BF* and *Crack_17_BF* (both from amphioxus *Branchiostoma floridae*) both showed very low activity, and both had a substitution from the E residue within the motif (E→H and E→A, respectively). *Crack-3_NVe* (from sea-anemone *N. vectensis*) also possessed a substitution from E (E→S), and again this sequence was inactive.

The *CR1* clade possessed the largest number of elements with 2A-like sequences, The *CR1* sequences tested *in vitro* were *CR1-1_BF*, *CR1-2_BF*, *CR1-10_BF*, *CR1-31_BF*, and *CR53_BF* from *B. floridae*; *STR-1_SP*, *STR-28_SP*, and *STR-32_SP* from *S. purpuratus*; and three molluscan sequences BG1_BG from *Biomphalaria glabrata* (Caribbean freshwater snail), *CR1-1_CGi* from *Crassostrea gigas* (Pacific oyster) and *CR1-1_LG* from *Lottia gigantea* (owl limpet). All the *CR1 B. floridae* sequences showed little or no recoding ability, despite that they largely conformed to the C-terminal canonical motif. However, it has been shown that a compatible upstream sequence to a length of around thirty amino acids is also a prerequisite for initiating ribosome skipping (Ryan and Drew, 1994; Ryan *et al.*, 1999; Donnelly *et al.*, 2001a; Brown and Ryan, 2010; Sharma *et al.*, 2012) and it may be that these sequences do not conform to this upstream compatibility. In the case of the *S. purpuratus CR1* 2As, both *STR-1_SP* and *STR-28_SP* are N-terminally truncated 2A sequences and it was thought interesting to see if this might affect their activity. *STR-1_SP* conformed to the expected C-terminus motif, but *STR-28_SP* possessed a single substitution at a residue previously shown to be highly important to function (from D→Q). *STR-1_SP* and *STR-28_SP* were found to be largely inactive, however the standard length (30 amino acids) sequence *STR-32_SP*, which also possessed the D→Q substitution was found to be active. All three mollusc sequences were active, but in the case of *CR1-1_LG* an additional band corresponding to an internal initiation product was observed on the gel – a common feature of *in vitro* translation reactions.

As 2A sequences from *Ingi* elements from trypanosomes have previously been tested and shown to be active (Heras *et al.*, 2006) it was decided to test the *Ingi* 2A sequence from the sea-slug *Aplysia californica*. This sequence, *Ingi-1_AC*, was found to be highly active, indeed this sequence could be seen to be more active than the viral sequence used as the positive control (*FMDV 2A*). The 2A sequence from the *Penelope-like* elements in *N. vectensis* was inactive when tested *in vitro*, although this sequence corresponded to the viral 2A C-terminal motif, it contained a methionine

residue immediately upstream in the position generally occupied by glycine, and it may be that this more bulky residue interfered with ribosome skipping.

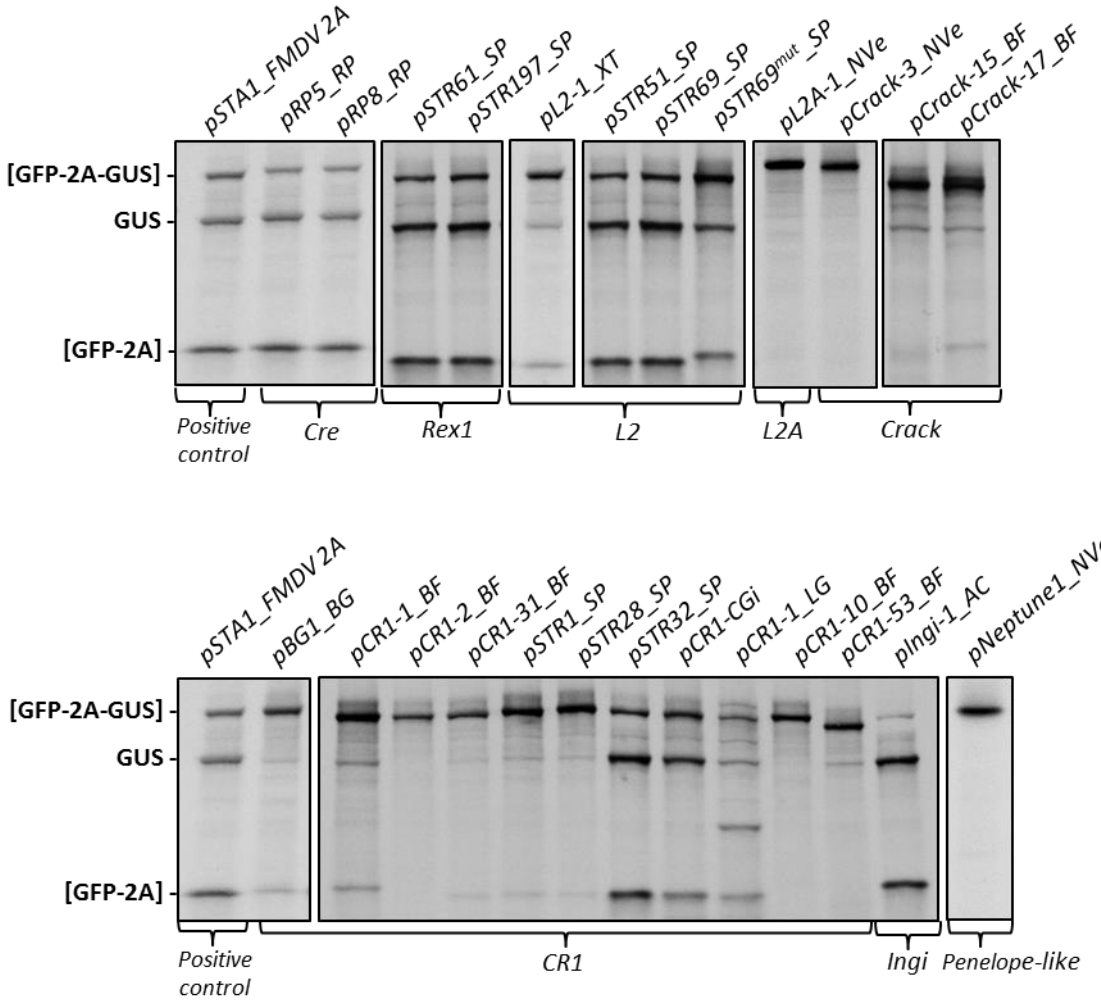
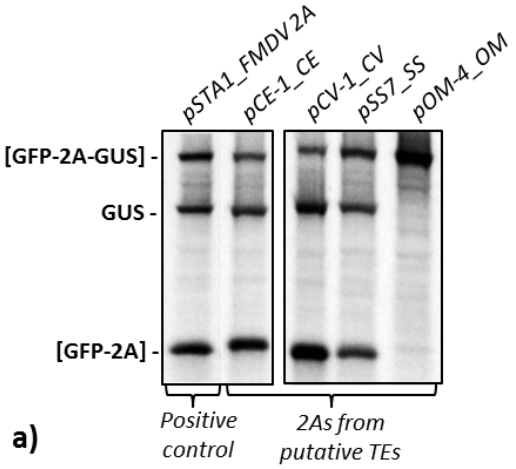


Figure 3.4 Non-LTR 2A sequences – recoding analyses
SDS-PAGE of TnTs programmed with 2A constructs cloned in the reporter *pSTA1* (constructs as labelled). Protein synthesis *de novo* was monitored by the incorporation of S^{35}-methionine. Translational recoding through ribosome skipping was determined by semi-quantitatively evaluating the relative distribution of radiolabel within either the "uncleaved" read-through product [GPF-2A-GUS] or the "cleaved" products [GFP-2A] plus the GUS band on the gel, in comparison to the products generated by a known positive control (*pSTA1*, encoding *FMDV 2A*). Figure is a composite of several gels; white space has been left between lanes derived from each gel. The relative recoding ability of each construct was assessed in comparison to *pSTA1* (*FMDV 2A*) run on each gel, respectively. Sequences are arranged by non-LTR clade, suffixes give the species of origin (see Section 3.2.2). The relative recoding ability and amino acid sequence of each 2A construct is summarized in Table 3.3, where residues corresponding to the viral consensus (SGD[V/I]ExNPGP) are shown in bold, and key N-terminal residues that differ from this consensus and thus may affect ribosome skipping abilities are underlined.

3.3.3 2As from Putative TE elements – *In Vitro* Recoding Analyses

In addition to the 2As from non-LTR elements, four 2A sequences from TEs of unknown type (identified only on the basis of their EEP domains) were subjected to *in vitro* analyses. All but one (*OM4_OM*) were found to be active *in vitro*; with generally similar activity levels to those displayed by the positive control *FMDV 2A* and all (including the inactive sequence *OM4_OM*) possessed the active viral 2A peptide eight residue C-terminal canonical motif.



a)

ID tag	Accession no.	Host Organism	Amino Acid Sequence	Activity
FMDV 2A	AAT01719.1	*Foot-and-Mouth-Disease Virus*	LYKSGSRGACQLLNFLLDKLA**GDVESNPGP**	++
CE-1_CE	AL132860.1	*Caerorhabditis elegans*, nematode	LCETPSLPHTTFLKRKLLVR**SGDVESNPGP**	++
CV-1_CV	XP_005844301.1	*Chlorella variabilis*, unicellular alga	LRLPCSCSTTALIKRMKLLL**SGDVEENPGP**	++
SS-7_SS	GU129139.1	*Salmo salar*, Atlantic salmon	QRSRRPVLIAFSRTLILLLLCS**SGDVEVNPGP**	++
OM-4_OM	EZ854573.1	*Oncorhynchus mykiss*, rainbow trout	TRRPVILAFSCTLILLLFCS**SGDVEVNPGP**	-

b)

Figure 3.5 2As from putative TEs – recoding analyses
a) SDS-PAGE of TnTs programmed with 2A constructs cloned in the reporter *pSTA1* evaluating the relative distribution of radiolabel as in Figure 3.4, again Figure is a composite of two gels. **b)** List of 2A sequences cloned. Residues corresponding to the viral consensus (SGD[V/I]ExNPGP) are shown in bold. The relative recoding ability of each construct is given in comparison to *FMDV 2A* (++=moderately high activity comparable to *FMDV 2A*, +++=higher than *FMDV*, +low activity, (+) very low activity, - =inactive).

3.3.4 Correlation of a Single ORF and the Presence of 2A

Active 2A-like sequences were found to exist within six clades of non-LTR and inactive (or with activity levels too low to be measurable in the TnT assay employed) 2A-like sequences were found in the non-LTR *L2A* clade and *Penelope-like* elements from *N. vectensis*. The non-LTR elements that encoded a 2A sequence were found to be lacking in an ORF1 (the only exceptions being *CR1-26_BF*, *CR1-53_BF* and *CR1-1_LG*). As noted earlier, non-LTRs may undergo 5' truncation during retrotransposition, which can delete ORF1 entirely; this may explain the lack of ORF1. Thus the majority of the active 2A-containing non-LTRs lacked an ORF1, but possessed a 2A sequence at or near the N-terminus of ORF2. These 2A encoding non-LTRs (lacking an ORF1) cluster alongside non-LTR elements (when aligned by RT domain) from other species which encode an ORF1, but lack a 2A. Therefore, this observed correlation (that elements possessing 2A lack ORF1) is unlikely to be merely an artefact, but it could be that 2A has replaced the biological role of ORF1 (regulation of retrotransposition) in these elements.

3.4 **Non-LTR 2As - Discussion**

This study provided the first record, outside of trypanosomes, of active ribosome skipping 2A-like sequences from eukaryotic genomes; these 2As were found from a wide variety of organisms, including single-celled trypanosomes and green algae, from relatively ancient multicellular phyla such as cnidarians, molluscs, arthropods, and echinoderms through to the evolutionary younger hemichordates, cephalochordates and chordates. When the surrounding genome structure was examined, a large number of these 2A-like sequences were found to occur in association with non-LTR retrotransposons, and additionally to occur in a characteristic placement within these elements (at the ORF2 N-terminus in elements lacking an ORF1). Therefore, due to the highly conserved positioning of 2A within non-LTRs, it seems highly likely that an active 2A is performing some useful function in non-LTR biology.

It has been proposed that 2A can function as a translational regulatory element, through directing ineffective ribosome skipping and/or stalling without re-initiation, down-regulating the relative levels of the downstream as compared to the upstream product (Brown and Ryan, 2010). This regulatory role is a proposed function of many viral 2As (discussed in Chapter 1.10). This may be highly relevant in regard to the possible function of 2A in non-LTRs.

Firstly, for optimal retrotransposition activity, an excess of the product encoded by sequences upstream of 2A may be required over that encoded downstream (RT and EEP domains). Upstream of 2A, in ORF1, if present, non-LTRs encode a protein, ORF1p, with low sequence similarity between different non-LTRs. Functional studies have shown that ORF1p is a high affinity RNA-binding protein that forms a ribonucleoprotein (RNP) complex together with non-LTR transcript RNA, and that ORFp1 contains signals required for the nuclear import of this RNP complex

(Matsumoto *et al.*, 2004). Unsurprisingly, ORFp1 is apparently an essential *in cis* requirement for *L1* retrotransposition (reviewed in Martin, 2006). It remains to be established whether the non-LTR elements lacking ORF1 can still initiate successful retrotransposition or whether they become dependent on "hijacking" the use of ORF1p encoded by other non-LTRs active within the cell.

Despite the loss of ORF1p function, there may be advantages in replacement of ORF1 with an N-terminal 2A sequence in ORF2. ORF2 encodes a single multifunctional product (ORF2p) two enzymatic domains required for retrotransposition: a reverse transcriptase and an endonuclease. These can affect the retrotransposition of not only their parental element, but also of additional elements (such as SINEs) within the cell. Therefore as high levels of these factors can prove detrimental as they can facilitate widespread retrotransposition and thus cause irreparable genomic damage. So, disproportionate levels of ORF2p could prove detrimental to the long term persistence of an element within its host genome. Certainly, cells possess a range of mechanisms that have apparently evolved to combat both retrotransposition of TEs and infection with retroviruses (recently reviewed in Ayarpadikannan and Kim, 2014). Therefore, it may be of longer term evolutionary benefit for a non-LTR element to evolve a "self-restraint" mechanism such as using a 2A to down-regulate production of its vital replication proteins. This could provide an explanation for the observation that even though relatively high levels of *L1Tc* mRNA were detectable from trypanosome cells, the levels of ORFp2 proteins were unexpectedly low (Heras *et al.*, 2006). Therefore, our working hypothesis is that these 2A sequences within non-LTR elements are providing a regulatory role in a similar manner to their functioning in viruses.

To date, full genomic data are only available for a few of the organisms in the phyla/subphyla involved in this study. Therefore, interpreting the pattern of 2A distribution within non-LTRs in terms of their likely evolutionary history is at best problematic, both within non-LTR clades and of host species. Nevertheless, the monophyletic grouping (if one excludes the *Cre* elements), of the non-LTR clades with 2As, could argue for an early acquisition of 2As within this group, followed by a pattern of subsequent losses in some copies of these sequences (loss from the high level of 5' deletions in these elements makes this highly probable).

The alternative model is that there were, in a similar manner to the pattern hypothesised for viruses (Luke *et al.*, 2008), multiple independent acquisitions of 2A into non-LTRs during their evolutionary history. The predisposition for TEs to preferentially integrate into certain genomic regions, often in place of another (now lost) TE sequence, coupled with the fact that this process is not "clean" (there can be fragments of 5' and 3' sequence left behind), could mean that a newly acquired TE could integrate in place of an older element, and in doing so inherit its 5' positioned 2A. If possession of such a 2A sequence, in a biologically relevant position, endowed the element with a higher level of evolutionary fitness, then this "selfish gene" would be more likely to spread

copies of itself throughout the genome. In a further fascinating hypothesis, it has been suggested that the TEs that are most able to down-regulate the number of active copies of themselves with a host cell (and thus reduce potential damage to their present host), are more likely to persist within that host, and its descendants, for a longer evolutionary timescale. This gains them a greater opportunity window for "escape" sideways to a fresh host species through horizontal gene transfer (reviewed in Silva *et al.*, 2004). Thus, through their self-down-regulation, non-LTRs with 2A sequences may be especially suited to horizontal gene transfer, which could explain their sporadic distribution throughout the phylogenetic tree of extant organisms.

There are now multiple recorded instances of TEs being acquired by horizontal gene transfer, but at present the mechanics of the process are a mystery. One suggestion is that viruses are the transfer vector (discussed in Silva *et al.*, 2004). Virus particles can encapsidate cellular mRNA transcripts. Indeed, one study investigated the RNA content of highly purified preparations of two closely related insect non-enveloped RNA viruses (flock house virus and *Nudaurelia capensis* omega virus) to find that 5.3% of transcripts packaged by the viruses corresponded to cellular transposon mRNA sequences (Routh *et al.*, 2012). Another possible transfer vector is microvesicles, membrane-bound capsules released by all cell types studied to date. Neoplastic cell microvesicles have been found to contain high levels of retrotransposon mRNA (Balaj *et al.*, 2011). Horizontal gene transfer by these routes might partially explain the predominantly aquatic species distribution of 2A-encoding non-LTR elements. Firstly, the marine environment is known to possess a high viral load, varying from 3×10^6 (in the deep ocean) to 10^8 (in coastal waters) viral particles per mL (Suttle, 2005). The majority of these marine virions are capable of infecting and causing disease in their respective hosts (Wilhelm *et al.*, 1998), therefore there are plenty of potential viral vectors circulating with aquatic habitats. Secondly, as the majority of organisms with 2A-containing non-LTRs are filter feeders with relatively simple digestive systems, ingestion and absorption of virions or microvesicles could deliver the non-LTR mRNA into a new host.

However, delivery of a non-LTR transcript into the somatic cells of a new host species constitutes merely the first step in non-LTR acquisition, for long-term persistence the element must be incorporated into the germline DNA. The mechanisms of this process are a mystery, but from a consideration of the present phylogenetic distribution non-LTR element clades, apparently there have been multiple instances of host transfer, therefore, certainly horizontal gene transfer of non-LTR elements is a possibility, albeit an infrequent one.

No matter how they came to be present, the occurrence of active 2As within non-LTRs, very likely in a regulatory role similar to their functioning in viruses, provides another fascinating parallel between virus genomes and non-LTR retrotransposons.

Chapter 4. Ankyrin-Repeat Associated 2As

'These sponges grew in every shape… with reasonable accuracy they lived up to their nicknames of basket sponges, chalice sponges, distaff sponges, elkhorn sponges, lion's paws, peacock's tails, and Neptune's gloves – designations bestowed on them by fishermen, more poetically inclined than scientists.'

20,000 Leagues Under the Seas - Jules Verne 1871, translator F. P. Walter

4.1 Introduction

The search for additional (non-trypanosome) eukaryotic 2As resulted in the discovery of over 400 such sequences from eukaryotic organisms (Appendix B). It was considered desirable to screen the surrounding sequences in order to ascertain any protein type(s) commonly associated with 2A. As discussed in Chapter 3, a considerable number of the eukaryotic 2A sequences were found in occurrence with non-LTR transposons, but an additional number of 2A-like sequences were found in association with one or more of several types of cellular protein; namely NLR-like innate immune proteins (discussed in Chapter 6), membrane-associated amino acid transporters (discussed in Chapter 5) or ankyrin-repeat proteins. Here the 2As associated with ankyrin-repeat proteins will be examined. Ankyrin-associated 2As were found in the genomes of two marine invertebrates. There were 24 instances from the Australian barrier reef sponge *Amphimedon queenslandica* (formerly named *Reniera sp.*), and two from the North American Pacific coast purple sea-urchin *Strongylocentrotus purpuratus*.

4.1.1 *Amphimedon queenslandica*

Sponges (Phylum *Porifora*) represent one of the earliest types of metazoan life, with a lineage reaching back at least 600 million years, thus they are generally recognised as the oldest surviving metazoan phyletic lineage. Consequently, it has been proposed that investigation of sponge genomics and proteomics will play pivotal roles in the searches for the origins and regulators of metazoan multicellular processes. The evolution of multicellularity from unicellular ancestors required the acquisition of highly regulated pathways coordinating cell growth, division, specialization, adhesion and death. Dysfunction in these pathways leads to diseases such as cancers and autoimmune disorders. Therefore study of their "simplest" extant hosts, namely modern sponges, may aid in the fundamental understanding of these processes (Srivastava *et al.*, 2010).

The extant Australian barrier reef sponge *A. queenslandica* (*Porifera, Demospongiae, Haplosclerida, Niphatidae*) identified and named in 2006, has now been the focus of a genome sequencing initiative making it the first (and only) sponge to have had its genome sequenced, assembled, and annotated (Degnan *et al.*, 2008). *A. queenslandica* possesses a lifestyle and organisation typical of sponges. The adults are immobile filter-feeders (feeding on microbes and particulate matter). They are hermaphroditic spermcast spawners (secreting sperm into the water-

column, but retaining and brooding their eggs after fertilisation) with planktonic larvae. The planktonic larval phase facilitates the wide dispersal of offspring, with the embryos settling and metamorphosing into adults upon receiving appropriate chemical cues from nearby conspecifics. In common with many sponges, adults possess a variety of commensal bacterial partners. Hence, genomic sequencing has been undertaken using sponge embryos in order to reduce bacterial contamination, and it is suspected that the difficulties encountered in culturing *A. queenslandica* in the laboratory might be a consequence of their apparent dependence on a suite of microbial partners (Srivastava *et al.*, 2010).

4.1.2 *Strongylocentrotus purpuratus*

Purple sea-urchins (*S. purpuratus*, *Echinodermata*) are an important keystone species in coastal ecosystems along the Pacific coast of the North American seaboard from British Columbia to New Mexico. They are grazers, feeding on both the micro- and macro-algae, particularly kelp. Overgrazing can reduce areas to "urchin barrens". Adult sea-urchins have few natural predators excepting sea-otters, spiny lobsters and humans. There is a small-scale Californian commercial fishery for *S. purpuratus*, but its main importance is as a model organism for scientific research (Pearse, 2006). Urchin embryos are bilaterally organized, but as adults they possess a radial body plan with an endoskeleton and water vascular system unique to echinoderms. Sea-urchins possess separate sexes (either male or female) that are difficult to distinguish visually (Sodergren *et al.*, 2006). They are broadcast spawners, releasing gametes directly into the water-column. The embryos and larvae form part of the plankton for a period of several months during which they can disperse over many hundreds of miles before their final settlement and metamorphosis into their adult form. The adults of the purple sea-urchin reach test (shell) diameters of 15-20cm and are covered with long purple-black spines (hence the name). Strongylocentrotid urchins can be extremely long-lived, with records of individuals living to over 100 years (Ebert and Southon, 2003).

S. purpuratus has been used as a scientific research subject for over 150 years. Its relevance stems from the fact that the echinoderms (alongside the hemichordates) are believed to be a sister phyla to the chordates, sharing a common ancestor around 540 million years ago. Investigation of sea-urchin embryology and development is thought to be of aid in elucidating these processes in vertebrates, and by extension, humans. Their enormously complex (in terms of receptor repertoire) non-adaptive immune system has been of focus in understanding immune receptor specificity (Sodergren *et al.*, 2006).

S. purpuratus has now undergone whole genome sequencing (Sodergren *et al.*, 2006). The initial sequencing project was led by Baylor College of Medicine, Houston, Texas, USA, but following concerns regarding the annotation of their data-set (Tu *et al.*, 2012) the project is now the

responsibility of an international consortium. To date, the *S. purpuratus* genome is fully sequenced and partially annotated. It represents the first fully sequenced genome from a motile marine invertebrate.

4.1.3 Ankyrin-Repeat Domains

Repeat proteins are the second most abundant protein-protein binding class of protein (after immunoglobins). They are non-globular modular proteins comprising of short, tandem repeating motifs (typically repeating between 20-40 amino acids). Each motif will fold into a distinctive configuration if present as part of a repeat region, but will be unstructured if expressed as a single entity. (Li *et al.*, 2006).

Ankyrin-repeats are one of the most common and widespread types of repeat protein (Li *et al.*, 2006). These were first identified from the yeast Swi6p, Cdc10p and *Drosophila* Notch proteins in 1987 (Breeden and Nasmyth, 1987) and later named for the cytoskeletal ankyrin protein which contains 24 copies of the repeat (Lux *et al.*, 1990).

Each ankyrin-repeat is composed of a ~33 amino acid motif comprising a canonical helix-turn-helix arrangement, in which the two α-helices are arranged in an antiparallel configuration and the loop that they form projects at approximately 90° from the body of the protein (Figure 4.1a). This aids in facilitating the formation of hairpin-like β-sheets with neighbouring loops. There are typically 20-30 such repeats in an ankyrin-repeat domain. These hairpin structures are both flexible and stable within the extended helix bundle (being stabilised through inter and intra-repeat hydrophobic interactions facilitated by conserved non-polar residues situated in the helical regions. The residues at the loop tips (in the linker between the two helices) are the most variable (Mosavi *et al.*, 2002; Li *et al.*, 2006). The structure of ankyrin-repeat domains has been solved. There are some permissible substitutions (in the tip/linker region) that do not affect function, but the overall domain motif is relatively conserved (Mosavi *et al.*, 2002). The interaction of ankyrin-repeats with their binding partners have been described as analogous to the fingers of a hand that can reach out to snatch an interacting protein, holding it in finger-tips in a pincer-like grip. In this model the "finger-tip" (the linker region between helices) residues determine binding specificity, whereas the helical regions stabilise the structure (see Figure 4.1) (Sedgwick and Smerdon, 1999).

Ankyrin-repeat proteins have been found in all multicellular "animal" genomes investigated to date, but are not found in either plants or fungi, therefore they were thought to be a purely "animal" protein class that evolved shortly after the divergence of metazoan phyla (Bennett and Chen, 2001). However, ankyrin proteins have been discovered as part of a class of "eukaryotic-like proteins" secreted by number of marine symbiotic bacteria (with marine invertebrate hosts); it is

thought that these proteins may have been horizontally acquired from their eukaryotic hosts (Fan *et al.*, 2012).

The first ankyrin-repeat proteins to be characterised were found to play a role in stabilizing the cytoskeleton of mammalian erythrocytes. Various ankyrin-repeats are now known to be involved in cytoskeletal organisation, as well as transcription regulation, positive and negative regulation of cell-cycle progression, cell development and organisation, cellular membrane ion channel adaptor proteins and toxic peptides (Sedgwick and Smerdon, 1999; Bennett and Chen, 2001; reviewed by Li *et al.*, 2006) and may possibly aid in endocytosis (Michaely *et al.*, 1999). Mutations within ankyrin-repeats have been implicated in human disease; for example, cells derived from a number of human cancer tumours were found to possess mutations in their genes for tumour suppressors p16 and p18, thus suggesting that ankyrin dysfunction may contribute to tumour formation (Ortega *et al.*, 2002). Ankyrin-R mis-folding is also implicated in hereditary anaemia in mice and humans, and in neural degeneration in mice (reviewed in Bennett and Chen, 2001).

In all instances, ankyrins act as "reaching fingers" enabling protein-protein interaction/binding. There is no evidence that they possess any enzymatic abilities. Ankyrins can be either intra- or extracellular proteins or occur embedded in cellular membranes (Li *et al.*, 2006). In humans (as in all mammals) three ankyrin-repeat genes have been identified: *ankyrin-R* (*ANK1*), *ankyrin-B* (*ANK2*) and *ankyrin-G* (*ANK3*). However, a large number of ankyrin-repeat isoforms are created in a complex pattern of tissue and developmental stage-dependent expression and translation of ankyrin genes. Additionally, in humans, ankyrin-repeats often occur as part of multi-domain proteins. Typical examples are the *ankyrin-B*s (Figure 4.1b); here the ankyrin-repeats tend to occur at or near the protein N-terminus (where they may facilitate binding to membrane components), often followed by a spectrin binding domain (anchors the protein into the cytoskeleton), then near the C-terminus a death domain (targets the protein/complex to the apoptotic pathway). Alternative splicing generates isoforms lacking one or more of the additional domains, or with alternative additional domains. Collectively, these processes permit production of both intra- and extracellular isoforms and of membrane-embedded proteins from the same gene (reviewed in Bennett and Chen, 2001) to create the vast array of ankyrin proteins essential to healthy cell functioning.

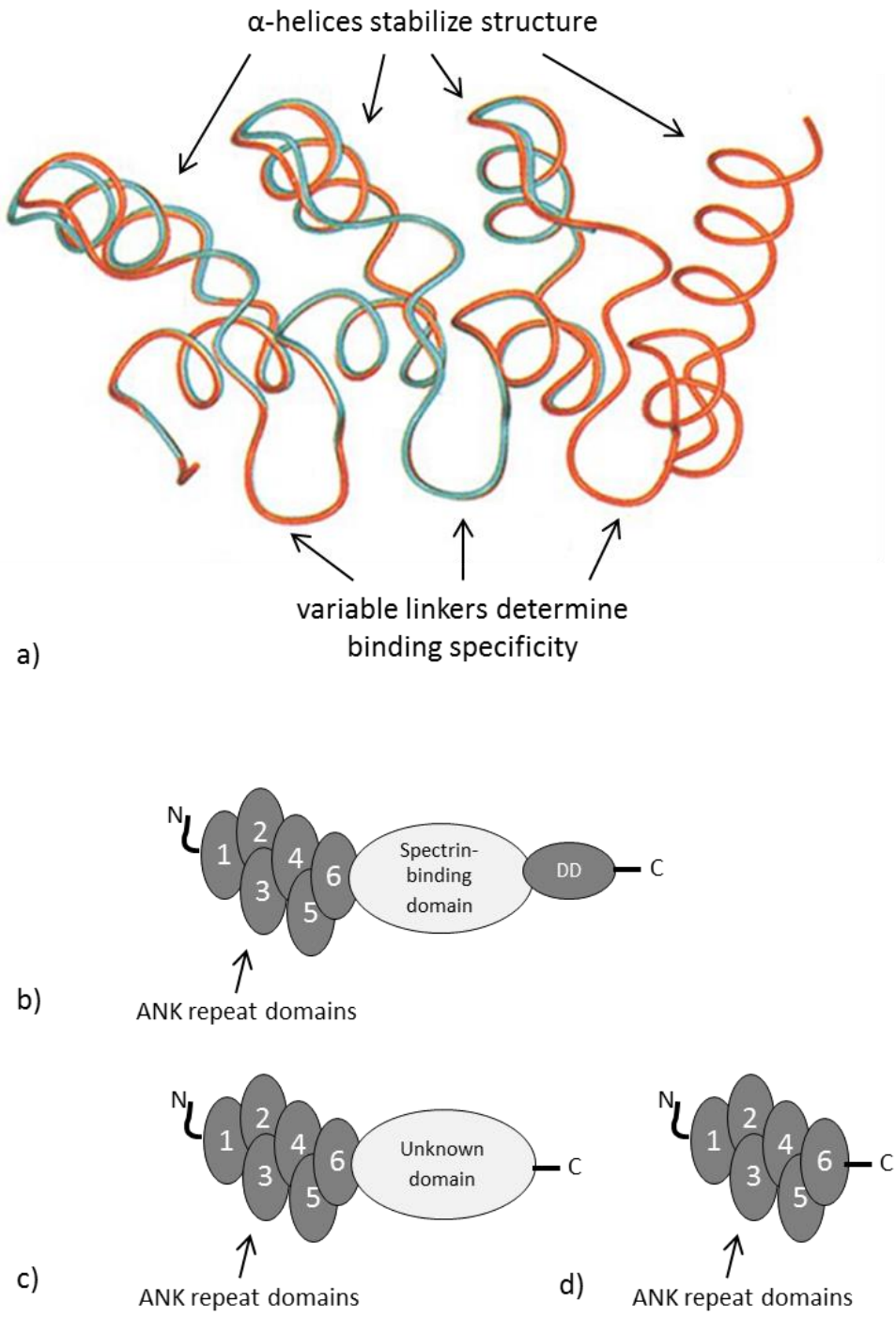


Figure 4.1 Ankyrin-repeat proteins
a) Overlay showing the backbone of two ankyrin-repeat structures (orange and cyan) solved by crystallography (from Mosavi *et al.*, 2002). **b-d)** Various configurations of ankyrin-repeat domains within human ankyrin-containing multi-domain proteins. There can be only a single ankyrin-repeat or up to 20-30 copies of the motif. In humans, **b)** *ankyrin-B*, represents a typical example; here ankyrin-repeats are followed by a spectrin-binding domain, then a death domain. Information from Bennett and Chen, 2001; Mosavi *et al.*, 2002).

4.2 **Methodology**

The work reported in this Chapter was undertaken by the author with the exception of three putative 2A sequences (ID tags: *AQ6, AQ7* and *AQ27*) cloned by Claire Stewart under supervision by the author.

4.2.1 *In Silico* Searches

Online proteomic and genomic databases were probed for eukaryotic 2A-like sequences as described in Chapters 2 & 3 to compile an in-house database (Appendix B). Sequences in this putative eukaryotic 2A database were screened for the presence of conserved protein domains in their flanking proteins/genes by means of the "find conserved domain" function of NCBI BLAST. Sequence alignments were performed using ClustalX2 as detailed in Chapter 2.1.2 and visualised using Figtree v1.4.2 or Phylodraw). In the interests of clarity, 2A sequences selected for further analyses were given short identification tags. These tags together with the NCBI accession numbers are used in Appendix B and throughout this Chapter. In the case of the sponge, *A. queenslandica*, predicted proteins were searched for additional ankyrin sequences (lacking a 2A) by BLAST searching the NCBI database using a typical example of a sponge ankyrin (the conserved ankyrin-domain region from XP_003385788.1). Hits were recorded, aligned using ClustalX2, and visualised using Figtree v1.4.2 or Phylodraw.

4.2.2 *In Vitro* – Methodology

Cloning of putative *A. queenslandica* 2A sequences into the reporter *pSTA1* plasmid were undertaken as detailed in Chapter 2.2.1 and Chapter 3. Briefly, the 2A peptide sequences were translated into nucleic acid sequences and incorporated into gene-blocks or cloned by means of long reverse primers as in Chapters 3 & 8 (for an explanation of cloning procedures see Chapters 2.2.1) The gene-blocks were initially ordered as part of the 2A activity screen reported in Chapter 8, therefore a full explanation of their design can be found there (Chapter 8.2.2). The nucleic acid sequences cloned are listed in Table 4.1. After verification by DNA sequencing using primer *GFPf* and/or *GUS_seq_R,* (Table 2.1), the plasmid preparations were used to program TnTs. TnT analyses were undertaken as detailed in Chapter 2.2.2 and Odon *et al.*, 2013.

Table 4.1 Gene-block/primer sequences for ankyrin 2A cloning
2A sequence *AQ20* was cloned from a gene-block sequence, the other sequences were built into the vector (*pSTA1*) by means of PCR using long reverse primers encoding the 2A of interest.

ID tag	Accession no.	Primer/Gene-block sequence
AQ20	XP_003385788.1	Gene-block fragment: 5'TGCGACACAGTATCATACGCAGTGTATCTTCTACTCTACTTTATGCTGCTCCTACT GCTCTCAGGAGACGTAGAACTGAACCCTGGACCA-3'
STR-37	XP_798371.3	Gene-block fragment: 5'TGACGAACATACTACTACTACGATCAGGAGACGTAGAACGAAACCCAGGACCG-3'
AQ27	XP_003391203.1	Reverse primer: 5'TGGTGGGGCCCGGGATTGATCTCGATGTCGCCGGATAGTAGCAGAAGTAGGGACAC GAGCTTGAATCTAGACACGACTGAGACCATTCTAGACCCGGACTTGTATAGTT-3'
AQ6	XP_003390053.1	Reverse primer: 5'TGGTGGGGCCCAGGGTTAAGTTCCACATCTCCTGATAAGAAGAGGAGGATAAGAAG GCATAAAATAGACAAAACAAGTCTTCTAGACCCGGACTTGTATAGTT-3'
AQ7	XP_003382892.1	Reverse primer: 5'TGGTGGGGCCCGGGATTCAGCTCGACGTCGCCCGAAAGTAGCAGAATTAGGGGTGA CGCCGAGAATACAGCCGTCTCGTCTCTAGACCCGGACTTGTATAGTT-3'

4.3 **Ankyrin 2As – Results**

A screen was undertaken to determine the potential function(s) of proteins containing the newly discovered eukaryotic 2A sequences (Appendix B) and to test the *in vitro* translational recoding abilities of 2As from such proteins. As discussed in Chapter 3., a considerable number of the 2A-like sequences were found in association with non-LTR transposons, but in addition, a number of 2A-like sequences were found in association with ankyrin-repeat, NLR-like (reported in Chapter 6), or membrane-associated amino acid transporter proteins (Chapter 5). This Chapter reports on findings regarding the ankyrin-repeat 2As.

4.3.1 Identification of 2As from Ankyrin Proteins

Ankyrin-repeat proteins containing 2A-like sequences occurred in genome reads from two marine invertebrates. There were 24 instances from the Australian barrier reef sponge *Amphimedon queenslandica* and two from the Pacific purple sea-urchin *Strongylocentrotus purpuratus*. The 30 amino acid 2A-like sequences with their accession numbers from NCBI, and their ID tags used during laboratory analyses are listed in Table 4.2. The *A. queenslandica* 2A-like sequences generally conformed to a viral 2A sequence consensus (to be discussed in Chapter 8) with their C-termini ending in the canonical –SGDVELNPGP-, however there were also instances of sequences ending in –SGDIELNPGP- and -SGDIEINPGP-. The *S. purpuratus* sequences ended in -GDVERNPGP- and -GDVEQNPGP-. All of these motifs are likely to confer activity *in vitro* if accompanied by a suitable upstream sequence (Luke *et al.*, 2008). Virtually all the ankyrin-2A sequences contained an upstream hydrophobic leucine and isoleucine tract characteristic of active 2A sequences (to be discussed in Chapter 8, see Figure 8.21). Therefore, it was thought highly probable that these sequences would prove active *in vitro*. In addition to the 2A sequences found in association with ankyrin-repeats, *A. queenslandica* also possessed four non-ankyrin proteins with 2A-like sequences (listed in grey text in Table 4.2). These were all found in association with P-loop NTPases/NACHT domains (NLR-like proteins, to be discussed further in Chapter 6).

Table 4.2 List of ankyrin-associated 2A-containing proteins

The 30 amino acid 2A sequence is listed, as is the NCBI accession number, the laboratory ID tag (AQ= *A. queenslandica,* STR=*S. purpuratus,* ID tags in bold text correspond to sequences cloned for *in vitro* analyses) the position of the final C-terminal proline residue (indicating placement within the protein), and any conserved domain functions of the protein. Entries in grey text correspond to the four additional *A. queenslandica* 2A-containing proteins that lacked ankyrin-repeats. The lower case lettering corresponds to N-terminal 2As that were extended to the full 30 amino acids by in-frame translation of the upstream gene sequence. It is debateable whether *AQ28* (2A C-terminal atypical motif –NVELNPGP-) constitutes a 2A sequence.

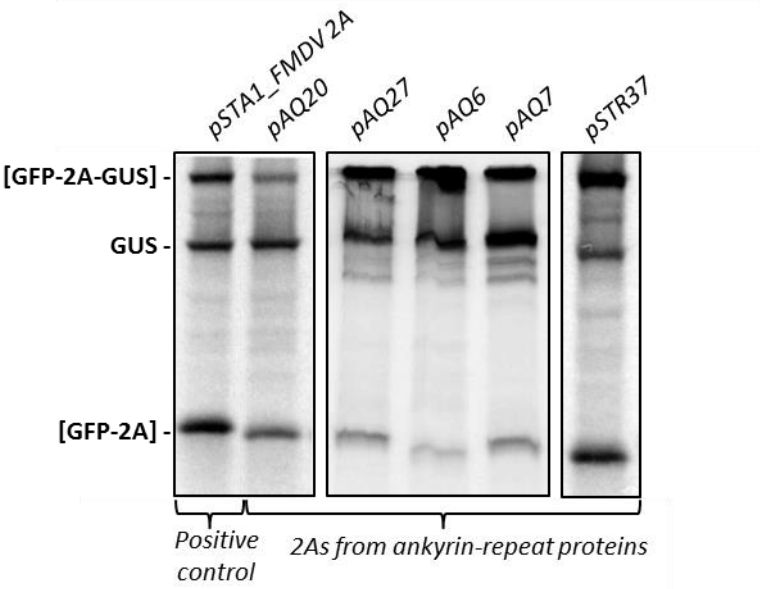
ID tag	Accession no (NCBI)	2A-like sequence: 30 amino acids	2A final P (position from N-terminus of protein)	Conserved domain function(s)
AQ1.1	XP_003390051.1	VKSQPGLILSIFCLLILLFISGDVELNPGP	65	Ankyrin
AQ1.2	XP_003390051.1	QVKTPRLVLSIFCLLILLFISGDVELNPGP	1871	Ankyrin
AQ2	XP_003386731.1	QGKPPKLILSIFCLAILLFISGDVELNPGP	72	Ankyrin, NHL repeat
AQ3	XP_003389958.1	LPGDTINHIFSIISHLLLFLSGDVELNPGP	203	Ankyrin
AQ4	XP_003387964.1	GAGLRKSQHGNSHQTELILLSGDVELNPGP	53	Ankyrin
AQ5	XP_003389878.1	VKSQPGLILSIFCLLILLFISGDVELNPGP	72	Ankyrin
AQ6	XP_003390053.1	QEKGPRLVLSILCLLILLFLSGDVELNPGP	76	Ankyrin
AQ7	XP_003382892.1	HWMNNDETAVFSASPLILLLSGDVELNPGP	177	Ferritin-like; Ankyrin
AQ8	XP_003389930.1	HWRQPSACTSTGNTYCGANGARDVELNPGP	48	Ankyrin
AQ9	XP_003389882.1	QEKGPRLVLSIFCLLILLFISGDVELNPGP	75	Ankyrin
AQ10	XP_003382893.1	AESRKSQHSNSHQPELIILLSGDVELNPGP	55	Ankyrin
AQ11	XP_003389880.1	QVKTPRLALSIFCLLILLFISGDVELNPGP	75	Ankyrin
AQ12	XP_003383551.1	VSDILACFLYSVFVVKLLLLSGDVELNPGP	310	PRE1; Predicted NTPase (NACHT family); P-loop-NTPase
AQ13	XP_003391290.1	PAAIPTVNITAVYSGPNYTSKRDVELNPGP	78	Ankyrin; ZipA
AQ14	XP_003390049.1	QVKTPRLVLSIFCLLILLFISGDVELNPGP	75	Ankyrin
AQ15	XP_003391332.1	qvktprlvlsifcllillfisgDVELNPGP	8	Ankyrin
AQ16	XP_003384088.1	TPETVCFLFYLFLHILLLLLSGDVELNPGP	115	Ankyrin
AQ17	XP_003389053.1	METRPKLILSIFCLFILLFIAGDVELNPGP	80	Ankyrin
AQ18	XP_003390050.1	QEKGPRLVLSIFCLLILLFISGDVELNPGP	75	Ankyrin
AQ19	XP_003390764.1	CDTVLCAVYLMMYFTLLLLLSGDVELNPGP	134	Ankyrin
AQ20	XP_003385788.1	CDTVSYAVYLLLYFMLLLLLSGDVELNPGP	133	Ankyrin
AQ21	XP_003382891.1	CHWMSNKTAAFSTNSLILLLSGDVELNPGP	74	Ankyrin
AQ22	XP_003391287.1	CDTVSYAVNLLLCFMLLLLLSGDVELNPGP	97	Ankyrin
AQ23	XP_003388358.1	SRPANGHHSRSFKAAVESKSDRDVELNPGP	49	TECPR; Ankyrin
AQ24	XP_003390020.1	YTESNQNVCYHHFMFLLLLLAGDIELNPGP	4539	HEPN domain; P-loop-NTPase; predicted NTPase (NACHT family)
AQ25	XP_003390214.1	ASILVCIFLYFVVCRLLLFLSGDIELNPGP	114	P-loop NTPase
AQ26	XP_003390344.1	snliyndlylivclrMLLLLSGDIELNPGP	15	Ankyrin
AQ27	XP_003391203.1	WFFVFMVSVVFKLVSLLLLLSGDIEINPGP	113	DD superfamily; Ankyrin
AQ28	XP_003388278.1	LPQTGVEEAISREEEELRVESANVELNPGP	1146	Chromosome segregation ATPases; P-loop-NTPase
STR-37	XP_798371.3	EADTIKGNDCSDMTNILLLRSGDVERNPGP	310	Ankyrin; Death Domain
STR	XP_003729085.1	TTDDPVMQESTCLPEMLLVKAGDVEQNPGP	828	Ankyrin; Death Domain

4.3.2 Ankyrin 2As - *In Vitro* Recoding Activity Assays

2As from three *A. queenslandica* ankyrin proteins of varying lengths, but all displaying the characteristic N-terminal 2A/C-terminal ankyrin-repeats (see Figure 4.3) were chosen for *in vitro* analysis (namely, *AQ6*, *7* and *20*). In addition, a 2A from an *A. queenslandica* protein with a death domain between the 2A and the ankyrins was picked for examination (*AQ27*), as was the short form of one of the *S. purpuratus* 2As (*STR-37*).

All the 2A sequences displayed some degree of activity when tested for ribosome skipping abilities *in vitro* by use of cell-free coupled transcription-translation (TnT) assays as measured by the incorporation of radio-labelled methionine. Interestingly, sequence *AQ20* was more active than *FMDV 2A* when tested in this system. *STR-37* displayed similar activities to *FMDV 2A* and the remainder of the sequences were less active than *FMDV 2A*. However *AQ6*, a sequence with a canonical C-terminus motif and not *AQ27* (with a non-standard -DIEINPGP- motif) displayed the lowest level of activity. This added credence to the supposition that both the C-terminal motif and an appropriate upstream tract were absolute requirements for efficient 2A-mediated ribosome skipping (Luke *et al.*, 2010b).

a)



b)

ID tag	2A Sequence	Relative Activity
FMDV 2A	LYKSGSRGACQLLNFLLDKLAGDVESNPGP	++
AQ20	CDTVSYAVYLLLYFMLLLLLSGDVELNPGP	+++
AQ27	MVSVVFKLVSLLLLLSGDIEINPGP	++
AQ6	QEKGPRLVLSILCLLILLFLSGDVELNPGP	+
AQ7	HWMNNDETAVFSASPLILLLSGDVELNPGP	++
STR37	MTNILLLRSGDVERNPGP	++

Figure 4.2 Recoding activity analyses
a) SDS-PAGE gel of TnTs run on 2A constructs cloned in the reporter *pSTA1*: image is a composite showing lanes derived from three gels run under identical conditions – see Sections 2.2.2 and 2.2.3. **b)** Table listing the 2As tested and recording their relative recoding ability in comparison to FMDV 2A (++=moderately high activity comparable to *FMDV 2A*, +++=higher than *FMDV*, +low activity). The "extra" bands visible below the [GFP-2A-GUS] band for *AQ27, AQ7* and *AQ6* are due to the presence of additional internal initiation products – a common occurrence in TnT analyses (as found by Odon *et al.*, 2013).

4.3.3 Ankyrin-repeat 2As – Bioinformatic Analyses

4.3.3.1 Ankyrin 2As - Protein Architecture

The positioning of the 2A peptides within the ankyrin-repeat proteins were examined. The two *S. purpuratus* proteins displayed similar domain architectures with ankyrin-repeats at the N-termini (up to 8 repeats) then a tract of approximately 200 amino acids containing no conserved domains, then 2A occurring immediately upstream of a death domain (Figure 4.3b). Whereas, the *A. queenslandica* ankyrin proteins typically possessed a 2A at or near their N-termini, the ankyrin domain was downstream of this, and in some instances the C-terminus of the protein included another (and varied) conserved domain (Figure 4.3a, Figure 4.4). These additional domains tended to possess roles as membrane anchors and/or transcription regulators. Two *A. queenslandica* proteins (*AQ26* and *AQ27*) possessed a death domain immediately downstream of 2A (before a C-terminal ankyrin-repeat region). The placement of 2A in the four non-ankyrin 2A-containing proteins from *A. queenslandica* varied, but in three out of the four instances the 2A occurred between functional domains; in the fourth, the 2A was N-terminal.

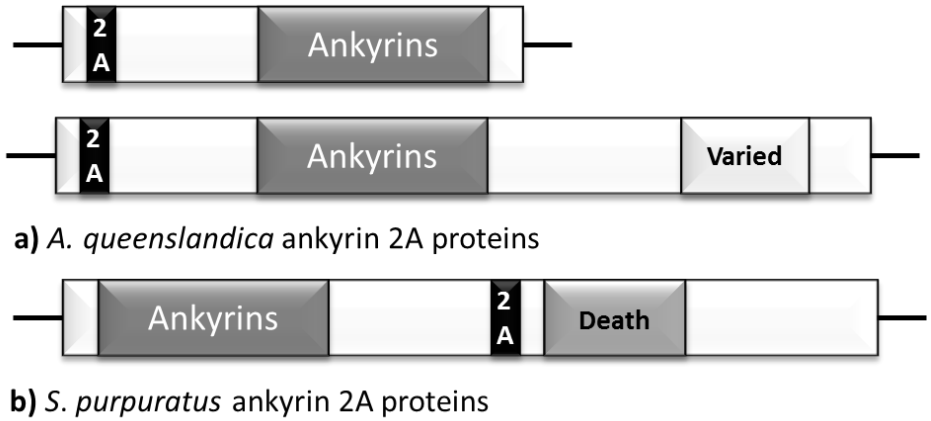


Figure 4.3 Schematic of the protein domain configuration of ankyrin 2A-containing proteins
a) Typical configuration of 2A-ankyrin proteins from sponge, there were two main configurations: -[2A-ankyrin-repeat]- and -[2A-ankyrin-repeat-further downstream domain]- but in both cases 2A occurred at or near the protein N-terminal. **b)** 2A=ankyrin proteins from sea-urchin, here the ankyrin-repeat domain was located upstream of 2A, there was a Death Domain immediately downstream of 2A (not to scale).

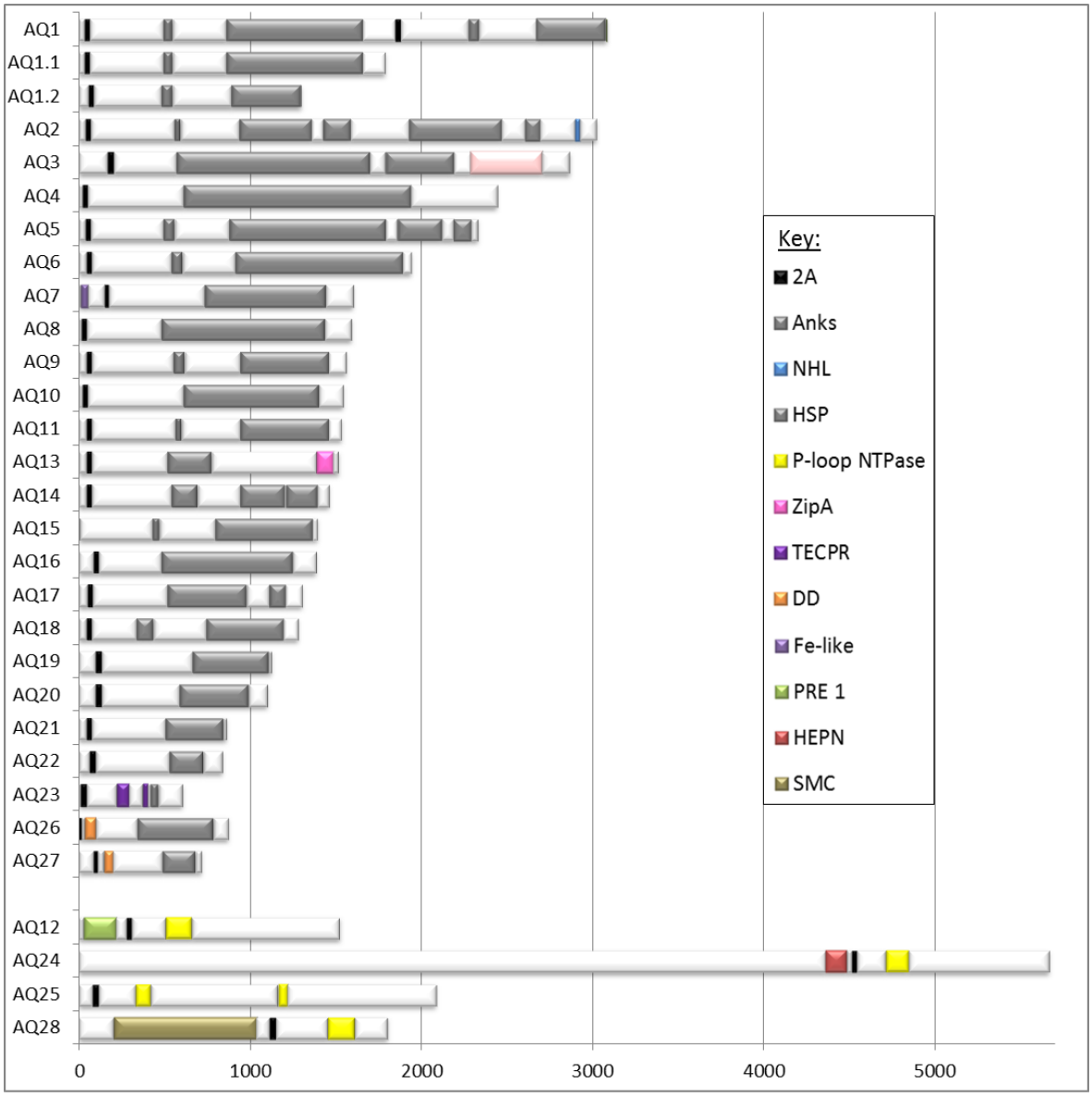


Figure 4.4 Structure of *A. queenslandica* 2A-containing proteins

2A sequences are shown in black, ankyrin-repeat domains in grey, other conserved domains in various colours. Sequences *AQ1* is shown both in its entirety and split into two proteins (as it is suspected that this sequence arises from a gene duplication and should actually consist of two sequences), the second beginning at the methionine residue upstream of the second 2A. Sequences *AQ15* and *AQ26* possess truncated 2A sequences at their N-termini. Key: 2A=2A peptide, Anks=ankyrin-repeat region, NHL=NHL (structural repeat protein), HSP=heat shock protein, P-loop NTPase= P-loop NTPase (ATP-dependent chaperone), ZipA=ZipA membrane-bound cell-division regulator, TECPR=transmembrane domain implicated in phagocytosis, DD= death domain, Fe-Like=ferritin-like domain (Fe-binding), PRE 1=PRE1 transcription factor, HEPN=HEPN (chaperon, possibly nucleotide binding), SMC=chromosomal maintenance protein.

4.3.3.2 Ankyrin 2As - Phylogenetic Relationships

The distribution of 2A sequences within the phylogeny of *A. queenslandica* ankyrin-repeat proteins were examined by alignment of all *A. queenslandica* ankyrin-repeat proteins. A similar investigation was not undertaken for the *S. purpuratus* ankyrins, due to the fact that there were only two occurrences of 2As within ankyrin proteins from this organism. The 2A-containing ankyrins were found to occur within a single monophyletic branch on the sponge ankyrin dendrogram (Figure 4.5). Interestingly, this branch corresponded to the evolutionary youngest section of the tree (as shown by relative branch length from the root on the cladogram, Figure 4.5). However, although the 2As were confined to 4 clades in this single monophyletic grouping, none of the 2A-containing clades consisted of exclusively 2A-possessing proteins but rather all clades with 2As contained a mix of proteins, with and without 2A.

The *A. queenslandica* ankyrin monophyletic grouping containing 2A-like sequences was examined in greater detail (Figure 4.6); however there was no apparent pattern to 2A distribution (Figure 4.6). Next, all 28 *A. queenslandica* 2A sequences, plus the two ankyrin-associated 2As from *S. purpuratus* (length 30 amino acids) were aligned in order to determine their inter-relationship (Figure 4.7). It should be noted that due to the short lengths (30 amino acids), this alignment is not robust.

Remarkably, the *S. purpuratus* 2As clustered with sequences from *A. queenslandica* rather than forming a separate clade. The *A. queenslandica* sequences from ankyrin and non-ankyrin proteins were not in separate clades but interspersed. However, due to the short nature of the sequences aligned it is not apparent if this distribution derives from their phylogenetic relationship or from convergent evolution of short peptide sequences.

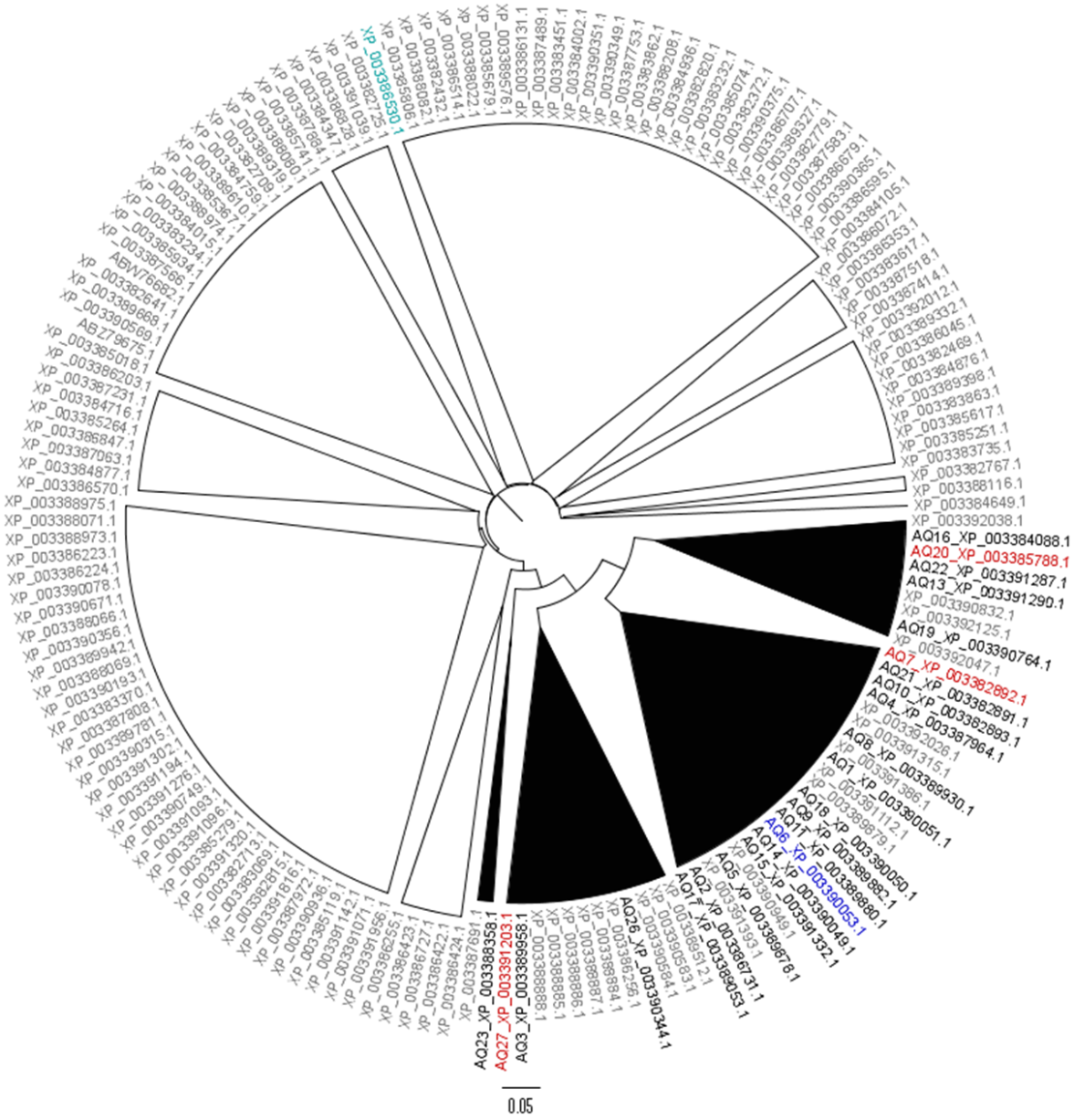


Figure 4.5 Cladogram of *A. queenslandica* ankyrin-repeat proteins

The clades containing 2A sequences have been shaded in black, individual proteins with 2As shown to possess comparable *in vitro* translational recoding properties to *FMDV 2A* are depicted in red, those less active than *FMDV 2A* in blue, while those not tested are in black text. The cyan label denotes the ankyrin protein (NCBI Accession XP_003386530.1) used as an outgroup in Figure 4.6. Rooted cladogram drawn with Figtree v1.4.2 using the default nearest neighbour joining algorithm from a ClustalX2 alignment.

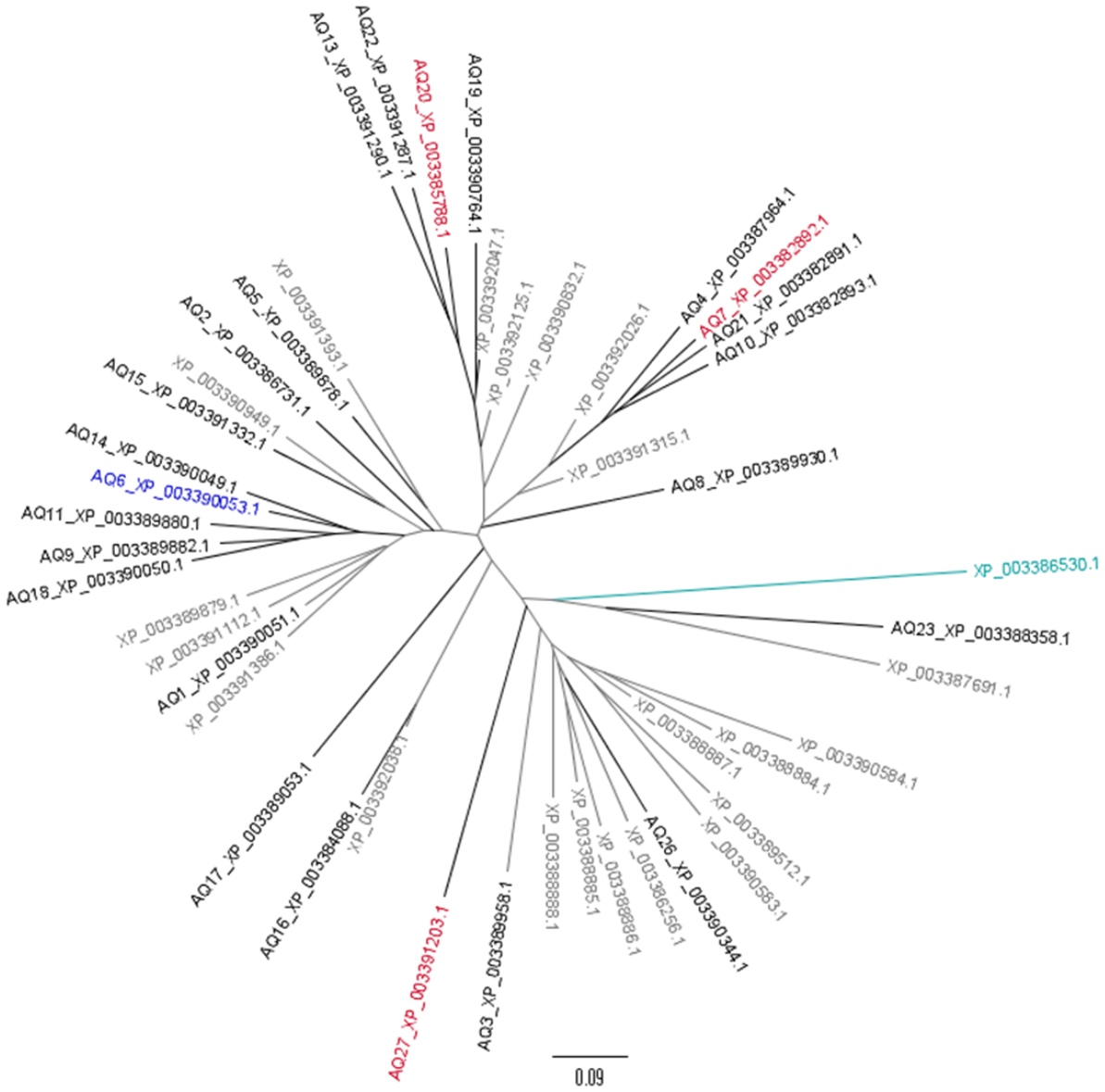


Figure 4.6 Cladogram of *A. queenslandica* 2A-ankyrin proteins

Cladogram of aligned protein sequences from the 4 clades with 2As (shaded black on Figure 4.5). The branches denoting ankyrin proteins with 2A sequences are coloured black and labelled in red if their 2As possess comparable *in vitro* translational recoding properties to *FMDV 2A*, if the 2As were less active than *FMDV 2A* they are labelled blue, while those 2As not tested are labelled in black text. Ankyrin proteins without 2As are shown in grey. The cyan label denotes the ankyrin protein (NCBI Accession XP_003386530.1) used as an outgroup in the alignment. Note that the 2As are distributed throughout the tree. Unrooted tree diagram drawn with Figtree v1.4.2 from a ClustalX2 alignment using the default nearest neighbour joining algorithm.

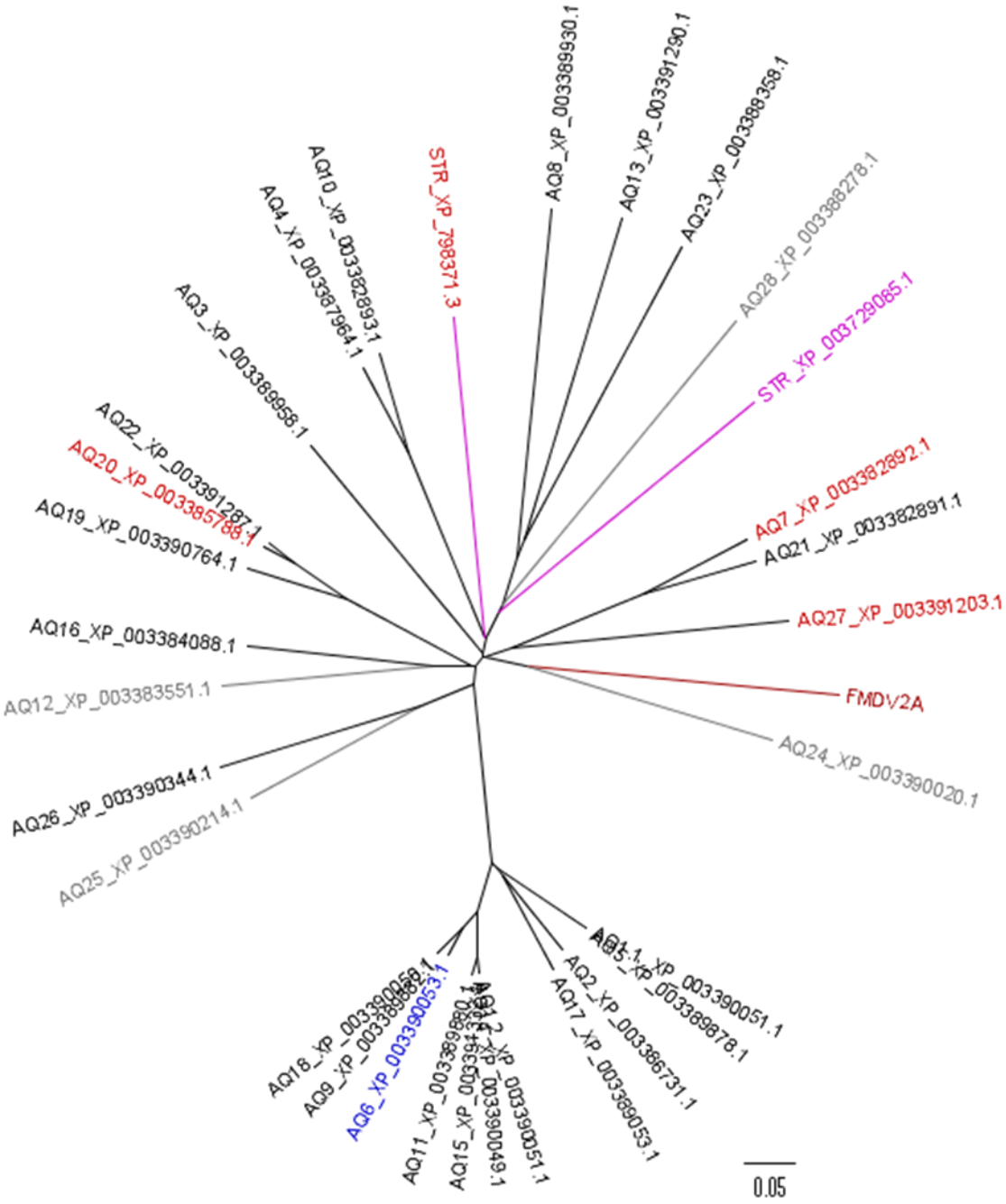


Figure 4.7 Cladogram analysis of ankyrin 2A sequences
Purple branches denote ankyrin protein 2As from *S. purpuratus,* black branches denote ankyrin protein 2As from *A. queenslandica,* while the three 2A sequences from *A. queenslandica* non-ankyrin proteins included for comparison purposes are in shown in grey. *FMDV 2A* (dark red) has been included for comparison purposes. Ankyrin 2As tested *in vitro* and found to possess comparable ribosome skipping abilities to *FMDV 2A* are coloured in red. <30 amino acid 2As aligned using ClustalX2 and visualised with Figtree v1.4.2, the unrooted tree diagram was drawn with using the default nearest neighbour joining algorithm.

4.4 **Ankyrin 2As - Discussion**

4.4.1 Role of 2A in Ankyrin-Repeat Proteins

This is the first report of active 2A sequences from proteins that play a role in eukaryotic cellular metabolism. The conserved positioning of the 2A sequence between active domains (separating the ankyrins from the death domain) in the two *S. purpuratus* proteins suggests the function of 2A could be to generate two different ankyrin protein isoforms from a single mRNA transcript (the full length protein with both ankyrin and death domains, and the truncated version without the downstream death domain). As cells possess multiple different isoforms of each ankyrin-repeat protein, each with a particular specificity/subcellular localisation, but all are transcribed from a single gene (due to alternate splicing), the incorporation of inefficiently "cleaving" 2As into multi-domain ankyrin proteins provides perhaps another way of increasing this isoform number. The *S. purpuratus* 2A tested was efficient at instigating translational recoding, but did not result in complete ribosome skipping, therefore could potentially contribute to increasing ankyrin-repeat protein isoforms *in vivo*.

In the case of the *A. queenslandica* proteins, the 2A occurs at or near the protein N-terminus, similar to the placement of 2A in non-LTRs (previously discussed in Chapter 3). This conserved positioning is also highly suggestive of a common function for 2A in these proteins. However, in this instance 2A is not being used to separate/remove another downstream active domain from the ankyrin domain. Therefore, perhaps here the N-terminal 2A plays a translational regulatory role, similar to its function in viruses and its proposed function in non-LTRs (Odon *et al.*, 2013). Under normal metabolic conditions 2A-directed ribosome skipping results in down-regulation of ankyrin production, but under conditions of cellular stress if 2A activity is impaired, then the full length (ankyrin) protein will be translated. Alternatively, perhaps the 2A may have acquired a secondary role in protein targeting to specific subcellular or extracellular destinations (this supposition will be explored further in Chapter 7).

The functional role and the interacting partners of the *A. queenslandica* and *S. purpuratus* ankyrin proteins have yet to be elucidated. However, it is interesting to note that the non-ankyrin domains from these proteins all function in cell division or membrane attachment. If it were to be found that these particular ankyrin proteins were implicated in cell defence/apoptosis pathways then this would lend credence to a translational regulatory role for 2A in these instances (as 2A may be able to regulate protein levels in response to cellular conditions). Furthermore, the positioning of the 2A within the *S. purpuratus* proteins, in addition to creating greater isoform diversity, hints at a role in translation regulation. In these proteins, under normal conditions, the 2A would cleave the multi-domain protein, separating the ankyrin and death domains, whereas under stress conditions, if the 2A activity was impaired, the multi-domain protein would be left entire, and the binding

partner of the ankyrin domain targeted to the apoptotic pathway through the actions of the attached death domain.

4.4.2 *A. queenslandica* Ankyrin 2As – Phylogeny

Two of the predicted proteins contained N-terminally truncated 2A sequences, but the full 2A sequence was present in the genome, therefore 2As may be being lost through N-terminal deletions caused by loss of the start codon methionine. Losses through N-terminal deletion could partially explain the "patchy" distribution of 2As throughout the latest branch in the *A queenslandica* ankyrin protein phylogenetic tree. However, these questions remain: do the 2As occur only in the latest (in phylogenetic terms) ankyrin clade due to this N-terminal deletion loss process, or was the acquisition of 2A a relatively recent event in the evolutionary history of these proteins? The parsimonious explanation is that 2A was acquired by the ancestral sequence which gave rise to the monophyletic 2A-containing clade, but that subsequent losses have led to the sporadic distribution of 2A throughout this clade.

Therefore, although *A. queenslandica* is an extant representative one of the earliest forms of metazoan life, the 2A sequences within its genome maybe a fairly recent acquisition, and their occurrence therein does not constitute proof that 2As are as old as metazoan life. The ankyrin-associated 2As found in a second marine invertebrate species, the sea-urchin, *S. purpuratus*, do not appear to share a common phylogeny (as evidenced by their different intra-protein domain organisation) with the *A. queenslandica* 2As. Indeed, as approximately 600 million years have elapsed since *S. purpuratus* and *A. queenslandica* shared a common ancestor, it seems problematic to suppose that *S. purpuratus*, but no other eukaryote (from which genomic data is presently available), has retained 2As within its ankyrin genes acquired from a sponge-like predecessor.

There remains the issue of how (and to a lesser degree when) *A. queenslandica* and *S. purpuratus* ankyrin proteins acquired their 2A peptides. It is probable that the unknown vector accountable for the horizontal gene transfer of non-LTR 2As between phyla may also be responsible.

There is accumulating evidence that some sponge species are capable of exchanging functional genes with their bacterial symbionts, including ankyrin-repeat genes, and may have later re-acquired copies of such genes from symbionts (Fan *et al.*, 2012). To add to the confusion regarding source, it is also extremely difficult to differentiate sponge from symbiont nucleic acid sequences when sequencing from tissue samples (Fan *et al.*, 2012). For this reason, the *A. queenslandica* genome sequencing was undertaken using embryos to minimise contamination (Srivastava *et al.*, 2010). The symbiotic bacteria secrete ankyrin-proteins which interfere with sponge amoebic phagocytosis and so protect the bacteria from ingestion (Nguyen *et al.*, 2014), permitting them to colonise the surfaces of the sponge cells. However, there is currently little data pertaining to

whether *A. queenslandica* possesses horizontally transferred genes from any symbionts (Fan *et al.*, 2012). Additionally, as 2A is thought to be a uniquely eukaryotic trait, acquisition of functional 2As from or through bacterial sources seems extremely unlikely. However, the identity of the transfer vector, and the origin of the ankyrin-associated 2A sequences remain a mystery.

Chapter 5. Sodium-Dependent Transporter Associated 2As

'I know of a cure for everything: salt water. *Salt water?* Yes, in one way or the other.

Sweat or tears or the salt sea.'

The Deluge at Norderney - Isak Dinesen, 1934

5.1 Introduction

Screening the newly discovered eukaryotic 2A sequences (Appendix B), revealed that a number of these sequences occurred in association with one or more of several types of cellular protein. This Chapter will describe the 2A and 2A-like sequences discovered in association with membrane-embedded sodium-dependent amino acid transporter proteins closely related to the human SNAT9 protein (encoded by the human *SLC38A9* gene).

5.1.1 Membrane Embedded Transporter Proteins

In humans, membrane protein genes constitute approximately 30% of all identified genes (Lander *et al.*, 2001; Almen *et al.*, 2009). The largest family of phylogenetically-related membrane proteins are the G protein-coupled receptors with around 800 examples in the human genome, and the second largest family are the solute carriers (SLCs) with over 380 human genes. The SLCs mediate transport of a variety of compounds across cell membranes including sugars, amino acids, nucleotides, inorganic ions and small drug compounds. They include genes for passive transporters, ion transporters and exchangers. All SLC families are functionally related in that they all (with a few isolated exceptions) rely on an ion gradient across the cell membrane to drive transportation. However, the SLCs do not include the primary active transporters such as ABC transporters, or ion channels and aquaporins. There are currently 46 recognised families of human SLC genes (Fredriksson *et al.*, 2008) and it has been estimated that roughly a quarter of these are involved in the transport of amino acids. Amino acid transporters are mainly composed of two phylogenetic clusters of SLC genes, the α-family containing Major Facilitator Superfamily genes, and the β-family of Amino acid Polyamine-organoCation (APC) genes, respectively. The β-family APC clan in humans and all other vertebrates contains proteins from three SLC gene families *SLC32*, *SLC36* and *SLC38* that are apparently so closely phylogenetically-related that they should be considered as a single super-family. All members in this superfamily with known function transport amino acids as their primary substrate (reviewed in Schiöth *et al.*, 2013).

The human *SLC32* family is represented by only one gene encoding the vesicular inhibitory amino acid transporter (VIAAT) which mediates H^+ driven uptake of γ-aminobutyric acid (GABA) and glycine into synaptic vesicles in neuronal cells. This was the first member of the β-family APC clan to be discovered, and the *SLC36* and *SLC38* genes were later added to the family on the basis of sequence similarities (reviewed in Schiöth *et al.*, 2013).

The human *SLC36* gene family comprises four genes, *SLC36A1-4*. These all encode for H^+-coupled transporters (termed PAT1-4) for small neutral amino acids, with a preference for L-proline, but also a high affinity for L-glycine and L-alanine, and surprisingly, in the case of PAT4, L-tryptophan. All are found in a wide range of mammalian tissue types.

5.1.2 *SLC38* Gene Family – SNAT Proteins

The *SLC38* family was found to constitute the proteins responsible for the human System A and System N amino acid transport, first described in the 1960s (Christensen *et al.*, 1965). System A was defined as an exclusively Na^+-dependent transport system that could be inhibited by the amino acid analogue 2-methylamino-isobutyric acid, whereas System N had the ability to counter-transport amino acids using either Na^+ or H^+ ions. The positive identification of any of the proteins responsible for these transport systems did not occur until the year 2000, when the sodium-coupled neutral amino acid transport protein 1 (SNAT1, gene *SLC38A1*) was characterised (Varoqui *et al.*, 2000). There are now known to be eleven human *SLC38* genes (*SLC38A1-11*, encoding proteins SNAT1-11, respectively). All the SLC38 genes have been mapped as to chromosomal loci.

Phylogenetic analyses indicated that the *SLC38* family was as old as metazoan life. *SLC38A7-11* were found in the *Trichoplax adherens* (a "simple" metazoan) genome. *SLC38A10* and *SLC38A7/8* were also found in the green alga *Ostreococcus tauri* genome, suggesting they originated before the split of plants and animals; in contrast, *SLC38A9* and *SLC38A11* are found in animals only, apparently post-dating the animal lineage. *SLC38A6* arose after the establishment of the arthropod lineage but before the cephalochordates. *SLC38A1-5* were found purely in vertebrates and were thought to have arisen from *SLC38A6* gene duplication. The putative 2A sequences discovered by the author were found in *SLC38A9* homologues (Schiöth *et al.*, 2013).

Functional analyses of six of the SNAT proteins have been undertaken. All were Na^+ (or H^+) dependent and were able to transport small neutral amino acids. In order to relate the newly identified genes/proteins with the transport activities determined in the early flux studies, they were referred to as possessing either System A or System N properties. The System A transporters were SNAT1, SNAT2, and SNAT 4, and the System N were SNAT3, SNAT5, and SNAT7. The SNAT6 and SNAT8-11 proteins have yet to be functionally characterised (Table 5.1), (see recent review by Bröer, 2014). Therefore, there is currently no data on the substrate specificity of SNAT9, the protein containing putative 2As.

The *SLC38* family SNAT proteins are particularly expressed in proliferating cells, or in cells that undertake significant amino acid metabolism such as liver, kidney and neural cells. They occur in membranes that face intercellular space or blood vessels, but are not present on the apical membrane of absorptive epithelia. They are also found in the placenta where they are significant in

delivering amino acids to the foetus. They are highly regulated in response to amino acid depletion, hypertonicity, and hormonal stimuli, and transport is up-regulated in response to increasing pH in the physiological range from pH6 to pH8 (reviewed in Bröer, 2014). They also play a role in signalling the amino acid state of the cell, and as such have been proposed to act as transceptors independent of their transport function (see review by Hundal and Taylor, 2009).

Table 5.1 Properties of human *SLC38* products
Information from Bröer, 2014. Data pertaining to *SLC38A9* highlighted in bold as this is the gene with the putative 2A sequences.

Gene	Protein	Mechanism*	Substrates†	Functional Class	Expression Profile
SLC38A1	SNAT1	S:1Na^+	(G), A, S, C, N, Q, H, (M)	System A	Ubiquitous
SLC38A2	SNAT2	S:1Na^+	G, P, A, S, C, N, Q, H, M	System A	Ubiquitous
SLC38A3	SNAT3	S:1Na^+/A:1H^+	Q, N, M	System N	Eye, liver, brain, pancreas
SLC38A4	SNAT4	S:1Na^+	G, (P), A, S, C, N, (M), R, K	System A	Liver, bladder
SLC38A5	SNAT5	S:1Na^+/A:1H^+	Q, N, H, A, S	System N	Mouth, cervix, bladder, bone, intestine, kidney, oesophagus, lung, eye
SLC38A6	SNAT6				Oesophagus, cervix, mouth, lung, kidney, muscle
SLC38A7	SNAT7	Na^+ dependent	Q, N, A, H, S	System N	Ubiquitous
SLC38A8	SNAT8				Testis
SLC38A9	**SNAT9**				**Parathyroid, testis, adrenal gland, thyroid**
SLC38A10	SNAT10				Ubiquitous
SLC38A11	SNAT11				Spleen, eye, bone marrow, pharynx

*A=antiport, S=symport
†=Single letter amino acid abbreviations (see Figure 2.1). Letters in brackets denote amino acids transported less frequently.

The topology of SNAT proteins have recently been inferred from sequence comparison with related membrane transport proteins solved by crystallography studies, in particular the high resolution structure of ApcT, a proton-dependent amino acid transporter from *Methanocaldococcus jannaschii* (Shaffer *et al.*, 2009). This sequence shares only 11-12 % sequence homology with SNAT proteins, but the topology plots as assessed by various software programs are remarkably

similar. Therefore, it is strongly supposed that the topology of SNAT proteins will be similar to ApcT (reviewed in Bröer, 2014).

The SNAT protein structure probably consists of an intracellular N terminus, a total of 11 transmembrane helices and an extracellular C-terminus. However an alternative structure has been proposed in which the highly conserved helix 1 is located extracellularly (reviewed in Bröer, 2014). The helices are arranged in a characteristic 5+5 inverted repeat fold configuration with symmetry between helices 1-5 and 6-10. Helices 3-5 and 8-10 are thought to form a rigid scaffold whereas helices 1, 2, 6 and 7 form a bundle that can undergo significant conformational change during transport. Helices 1 and 6 expose hydrogen bond donors and acceptors for substrate and ion binding (Figure 5.1). The intracellular N-terminus is thought potentially to be available to interact with intracellular proteins and the extracellular C-terminus can sense changes in tissue pH (reviewed in Schiöth *et al.*, 2013; Bröer, 2014).

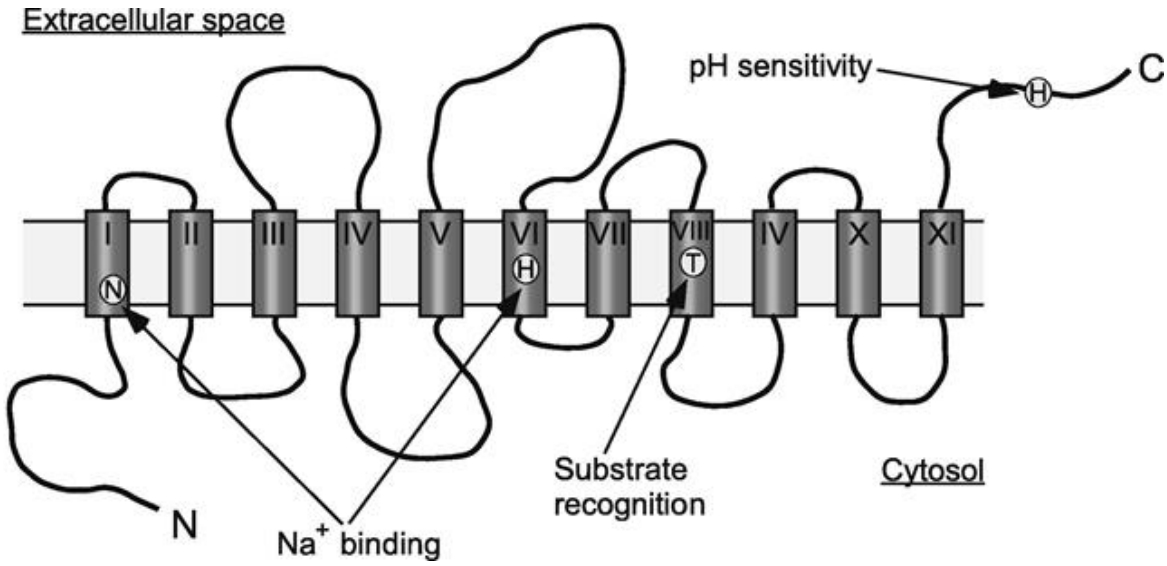


Figure 5.1 SNAT protein topology
Schematic representation of a SNAT protein illustrating the 11 transmembrane helices, a cytosolic N-terminus and an extracellular C-terminus. Amino acid positions with a known function are marked, diagram modified from Schiöth *et al.*, 2013.

5.1.3 SNAT Signalling

Amino acid availability signalling is largely mediated through two pathways, the mammalian target of rapamycin (mTOR) and the general control non-repressed (GCN) pathway. The TOR pathway plays a role in upregulating mRNA translation when amino acid supplies are abundant. Leucine in particular is a potent activator of mTORC1, but leucine is not a known substrate of SNAT transporters. Therefore, an indirect activation method has been proposed whereby glutamine is accumulated in the cytosol (by SNAT2). Glutamine then serves as an exchange substrate for leucine by the general amino acid antiporter 4F2hc/LAT1 (leucine-preferring amino acid

transporter 1). Thus the intracellular leucine is increased and recognised by the TOR pathway. Additionally, mTORC1 and mTORC2 silencing in primary human trophoblast cells altered the distribution of SNAT2 from the plasma membrane to intracellular compartments, suggesting mTOR signalling is essential to maintain surface expression of SNAT2, but SNAT1 and SNAT 4 were unaffected as was the overall quantity of each transporter present (reviewed in Bröer, 2014).

A more direct form of signalling, particularly by SNAT2, has been proposed. Here, SNAT2 may act as a transceptor. This hypothesis follows observations made in a yeast experimental system where a transporter-like protein Ssy1 binds amino acids and forms a complex with two additional proteins, one of which is a protease. When bound in the complex, the protease releases a membrane-bound transcription factor by partial protolysis. The transcription factor is then free to re-locate into the nucleus where it initiates transcription of transporter protein mRNA, which then can provide the cell with additional amino acids. Thus, while Ssy1 is not in itself a transporter, by its amino acid binding, it will upregulate amino acid transport (Hundal and Taylor, 2009; reviewed Bröer, 2014).

An additional mechanism through which SNAT proteins could elicit signals is by depolarisation of the cell membrane due to Na^+ co-transport. This mechanism has been observed in intestinal L cells. L cells are enteroendocrine cells that release glucagon-like peptide (Glp-1) upon exposure to nutrients in the intestine. Glp-1 acts in the pancreas as an insulinotropic hormone. Glutamine causes significant L cell depolarisation and Glp-1 release. Investigation of L cell glutamine transporters revealed high levels of SNAT2 expression (reviewed in Bröer, 2014).

5.1.4 SNAT Regulation

The mechanisms which regulate *SLC38* gene transcription and translation and the intracellular distribution of SNAT proteins are beginning to be investigated. As mentioned above, the mTOR pathway has shown to be important in determining SNAT2 intracellular distribution. It is known that System A transporter (see Table 5.1) activity increases following amino acid deprivation. An amino acid response element previously identified in the promoter region of the asparagine synthetase enzyme gene was apparently lacking in the SNAT2 gene (*SLC38A2*), but was subsequently located in the *SLC38A2* intron 1. It is thought to act as a transcription enhancer. Amino acid depletion activates general control non-derepressible-2 kinase which phosphorylates the translation initiation factor eIF2α which in turn decreases global protein synthesis, but increases translation of a number of selected mRNAs - including an mRNA set which encode proteins that will bind to the SNAT2 amino acid response element described above. The subsequent SNAT2 translation is not mediated by cap-dependent initiation and scanning from the 5’ end of the mRNA, but is instead driven by an IRES (internal ribosome entry site, see Figure 1.2 from Chapter 1) in the 5’ UTR of the SNAT2 mRNA (reviewed in Bröer, 2014). SNAT2 levels are

also regulated through eIF2α during endoplasmic reticulum (ER) stress. Accumulation of unfolded protein in the ER activates kinases which in turn phosphorylate eIF2α, leading to eventual *SLC38A2* upregulation and increased SNAT2 activity. *SLC38A2* upregulation has also been observed as a response to hypertonicity mediated by MAPK pathways. The *SLC38A1/2* promoter regions have also been found to contain a cAMP response element (CRE). SNAT2 mRNA levels increase in the presence of elevated interleukin 6 (IL-6) levels particularly in diabetic and obese individuals (reviewed in Bröer, 2014).

For the System N transporters, SNAT3 regulation has been investigated. Kidney SNAT3 mRNA is upregulated in response to chronic metabolic acidosis followed by an increase in SNAT3 proteins in the basolateral membranes. The upregulation mechanism is unknown, but there is a pH-response element in the 3' UTR of SNAT3 mRNA (reviewed in Bröer, 2014).

Regulation of transport activity of many membrane proteins is mediated by trafficking, and "stores" of many transporter proteins can reside in vesicles under the plasma membrane. A specific stimulus will induce forward trafficking and incorporation of the transporters into the cell membrane to begin functioning. It is not currently known how SNAT proteins are trafficked to the cell membrane. SNAT protein stability and degradation appear to be regulated, at least in part by the stability of their cytosolic N-terminus (reviewed in Schiöth *et al.*, 2013; Bröer, 2014).

5.2 **Methodology**

5.2.1 *In Silico* Searches

Online proteomic and genomic databases were probed for eukaryotic 2A-like sequences as described in Chapter 2.1.1 to compile an in-house lab database (Appendix B). Their flanking regions were then screened for conserved protein domains using the "find conserved domain" function of NCBI BLAST. Sequence alignments were performed using ClustalX2 as detailed in Chapter 2.1.2 and visualised using Figtree v1.4.2 or Phylodraw). In the interests of clarity, 2A sequences selected for further analyses were given short identification tags. These tags together with the NCBI accession numbers are used in Appendix B and throughout this Chapter. The mammalian (bovine) sodium dependent amino acid transporter isoform 9, NCBI Accession NP_001095633.1, was used as a template for a BLAST search. The 500 closest hits were recorded and used for further phylogenetic analyses. Sequences were aligned using ClustalX2, and visualised using Figtree v1.4.2 or Phylodraw.

5.2.2 *In Vitro* – Methodology

Cloning of putative 2A sequences into the reporter *pSTA1* plasmid were undertaken as detailed in Chapter 2.2.1 and Chapter 8.2.2. Briefly, the 2A peptide sequences were translated into nucleic acid and cloned by means of long reverse primers. The primer sequences are reported in Table 5.2. The primers were designed with mixed bases at certain positions in order to obtain multiple versions of the sequence (for instance a potentially active –NPGP- form and a translational recoding inactive -NAGP- mutant used in later experiments, to be reported in Chapter 7) from a single oligonucleotide primer. Plasmid sequences were verified by DNA sequencing with primer *GFPf* and/or *GUS_seq_R* (primer sequences in Table 2.1). Plasmid preparations were used to program in vitro coupled transcription-translation reactions (TnTs) as detailed in Chapter 2.2.2 and Odon *et al.*, 2013.

Table 5.2 SNAT 2A reverse primers
Table details the primers used. Sample NCBI accession numbers are provided for each sequence. Mixed-base (at specific nucleotide bases) oligonucleotide sequences were employed to obtain multiple variants of 2A-like sequences that differed only in one (or two) specific amino acid(s). The residues that varied are given below the 2A-sequence amino acid composition. Residue number is counted back up the sequence from the C-terminal proline (P). XbaI and ApaI restriction enzyme sites used in cloning are underlined.

ID tag	Accession no.		Reverse Primer Sequence
Cow	NP_001095633.1	MANMDSDSRHLLIPEGDHEINPGP (also 5V, 3A)	5′TGGTGGGGCCCAGSGTTAAYT TCATGATCACCTTCTGGAATCAA TAGATGTCGACTATCTGAATCCA TATTAGCCATTCTAGACCCGGAC -3′
Panda	XP_002928125.1	MDSDSRHLLIPEVDHEIINPGP (also 9N, 3A)	5′GTGGTGGGGCCCGGSGTTAAT AATTTCATGATYCACTTCAGGAA TTAATAGATGTCGACTATCTGAA TCCATTCTAGACCCGGAC-3′
Rat	EDM10355.1	MANVDSDSRHLISEVEHEVNPGP (also 3A)	5′GTGGTGGGGCCCGGSGTTAAC TTCATGTTCAACTTCACTAATTA GATGTCGACTATCAGAATCTACA TTAGCCATTCTAGACCCGGAC- 3′
NMR	XP_004848734.1	MTNVDDRHHLISEADHEVNPGP (also 3A)	5′GTGGTGGGGCCCGGSGTTCAC TTCATGATCAGCTTCTGATATTA AGTGATGTCGATCGTCTACGTTA GTCATTCTAGACCCGGAC-3′
2A2 *(2nd 2A)*	NP_775785.2	MNKRIHYYSRLTTPADKALIAPDHVVPAPEE (also 19A/S, 12V)	5′GTGGTGGGGCCCTTCTTCAGG TGCAGGAACTACATGATCAGGTG CAAYTAGTGCTTTATCAGCAGGT GHAGTTAATCGACTATAATAATG AATACGTTTATTCATTCTAGACC CGGAC-3′

5.3 **SNAT9 Associated 2As – Results**

5.3.1 Cataloguing SNAT9 2A-like sequences

A 2A-like sequence was found in SNAT9 homologues (described in Section 5.1.2) from a number of mammalian species (Table 5.3). These sequences were 22-24 amino acids in length, and conformed to the viral 2A consensus (to be further discussed in Chapter 8, see Figure 8.21) in that they possessed an upstream hydrophobic leucine/isoleucine motif, but none displayed the typical viral -D[V/I]ExNPGP- C-terminal motif. Instead their C-termini possessed histidine, asparagine, or leucine in place of the viral valine or isoleucine. In addition to 2A-like sequences ending in -DxxxNPGP-, there were also sequences ending in -ExxxNPGP- or -HxxxNPGP-, as an apparent result of frame-shifting or insertions (Table 5.4). There were also those with a substitution in the C-terminus -NPGP- (Table 5.4). Based on previous studies (Luke *et al.*, 2008; Sharma *et al.*, 2012) it was considered unlikely that these latter sequences would be able to affect ribosome skipping, therefore only sequences retaining the canonical –NPGP- motif were selected for cloning (bold text and grey-shaded background in Tables 5.3 and 5.4). In all cases, the 2A sequence occurred at the N-terminus of the SNAT9 protein.

The first SNAT9 2A sequence identified was from *Bos taurus*, accession NP_001095633.1. This was discovered as part of a database probe collecting hits to the canonical 2A C-terminal motif. A further BLAST search was then undertaken using this sequence (full protein) as the template with the ceiling set at 500 hits. After removal of truncated entries, these were probed for 2A-like sequences (reported in Tables 5.3 and 5.4) and aligned using ClustalX2. The hits contained transporter proteins from other mammals, vertebrates, invertebrates and a number of bacterial sequences; however only placental mammal SNAT9-like proteins were observed to possess 2A-like sequences (Tables 5.3 & 5.4; Figures 5.4 & 5.5).

Table 5.3 SNAT9 2A-like sequences with canonical DxxxNPGP C-termini

The table lists 2A-like sequences identified from SNAT9 proteins. All occurred at the N-termini. None of the sequences possessed the typical viral –D[V/I]ExNPGP- C-terminal motif, but instead had histidine, asparagine, or leucine in place of the viral valine or isoleucine. Sequences selected for *in vitro* activity assays are shown in bold text and are shaded in grey.

Host Organism	Common Name	NCBI Accession no.	SNAT9 2A-like sequence (canonical 2A DxxxNGP C-termini only)
Microtus ochrogaster	prarie vole	XP_005356860.1	MADADSDSRHLISEVDDEVNPGP
Pteropus alecto	black flying-fox	ELK12194.1	MANMDNDSRHLLIPEVDHEINPGP
Condylura cristata	star-nosed mole	XP_004678371.1	MANMDNSSKHLLIPDVDHEINPGP
Bos taurus	**cattle**	**NP_001095633.1**	**MANMDSDSRHLLIPEGDHEINPGP**
Bos taurus	**cattle**	**DAA17954.1**	**MANMDSDSRHLLIPEGDHEINPGP**
Ovis aries	**sheep**	**XP_004017029.1**	**MANMDSDSRHLLIPEGDHEVNPGP**
Felis catus	domestic cat	XP_003981016.1	MANMDSDSRHLLIPEVDHEINPGP
Sus scrofa	pig	XP_003134026.2	MANMDSDSRHLLIPEVDHEVNPGP
Orcinus orca	killer whale	XP_004275195.1	MANMDSDSSHLLIPEVDNEINPGP
Ceratotherium simum simum	white rhino	XP_004422877.1	MANMGSDSRHLLIPEVDHEINPGP
Octodon degus	degu	XP_004623147.1	MANVDDRHDLISEVDHEVNPGP
Trichechus manatus latirostris	Florida manatee	XP_004380528.1	MANVDSDSRHLLISEVDHEINPGP
Cavia porcellus	guinea-pig	XP_003470307.1	MANVHDRHHLISEIDDEVNPGP
Loxodonta africana	African elephant	XP_003408087.1	MANVNKESRHLLISEVDHEINPGP
Mesocricetus auratus	golden hamster	XP_005065531.1	MASMDSDSRPLIPEVDLEVNPGP
Otolemur garnettii	greater galago	XP_003782807.1	MASVDRHHLISEVDHEINPGP
Cricetulus griseus	Chinese hamster	XP_003503482.1	MASVDSDSRHLIPEVDLEVNPGP
Cricetulus griseus	Chinese hamster	EGV96941.1	MASVDSDSRHLIPEVDLEVNPGP
Heterocephalus glaber	**naked molerat**	**XP_004848734.1**	**MTNVDDRHHLISEADHEVNPGP**
Heterocephalus glaber	**naked molerat**	**EHB16430.1**	**MTNVDDRHHLISEADHEVNPGP**
Heterocephalus glaber	**naked molerat**	**XP_004906080.1**	**MTNVDDRHHLISEADHEVNPGP**

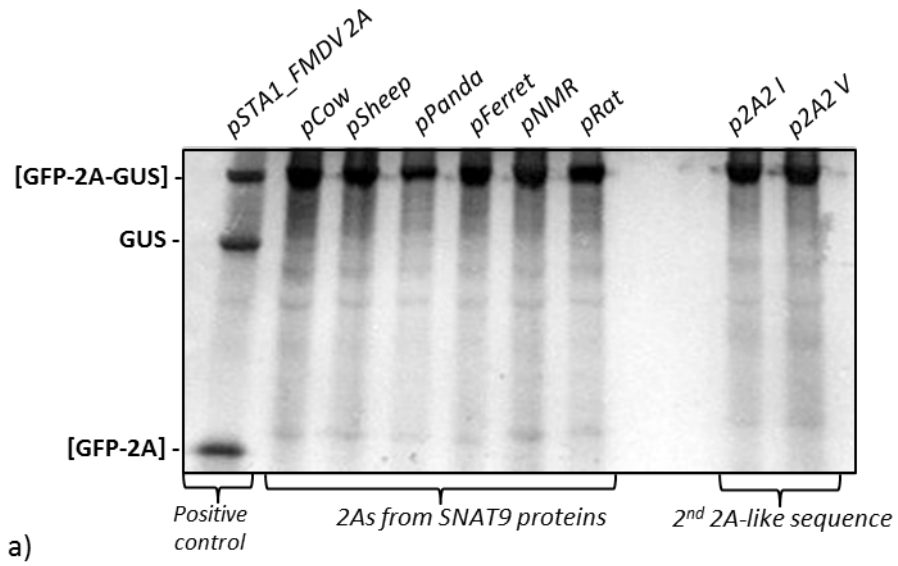
Table 5.4 SNAT9 2A-like sequences with non-canonical C-termini

2A-like sequences identified from SNAT9 homologues. All 2A-like sequences occurred at the protein N-termini. Sequences selected for *in vitro* activity assays are shown in bold text and are shaded in grey.

Host Organism	Common Name	NCBI Accession no.	SNAT9 2A-like sequence
			(-XxxxNPGP- N-termini)
Rattus norvegicus	**brown rat**	**EDM10355.1**	**MANVDSDSRHLISEVEHEVNPGP**
Rattus norvegicus	**brown rat**	**NP_001030328.1**	**MANVDSDSRHLISEVEHEVNPGP**
Ailuropoda melanoleuca	**giant panda**	**XP_002928125.1**	**MDSDSRHLLIPEVDHEIINPGP**
Ailuropoda melanoleuca	**giant panda**	**EFB13106.1**	**MDSDSRHLLIPEVDHEIINPGP**
Mustela putorius furo	**ferret**	**XP_004789844.1**	**MDSDSRHLLIPEVNHEIINPGP**
Mustela putorius furo	**ferret**	**XP_004744702.1**	**MDSDSRHLLIPEVNHEIINPGP**
Odobenus rosmarus divergens	Pacific walrus	XP_004416369.1	MDSDSRHRLVPEVDHEIINPGP
			-Dxxx- then 1 mutation to –NPGP-
Myotis brandtii	Brandt's bat	EPQ12632.1	MADLDSDSRQLLAPEADREVHPGP
Myotis davidii	David's bat	ELK31761.1	MADLDSDSRQLLAPEADREVLPGP
Oryctolagus cuniculus	rabbit	XP_002714085.1	MANMAGDSSHLISEVDPELHPGP
Nomascus leucogenys	white-cheeked gibbon	XP_003265979.1	MANMNSDSRHLGTAKVDHERDPGP
Pan troglodytes	chimpanzee	XP_001145251.1	MANMNSDSRHLGTSEVDHERDPGP
Pan paniscus	bonobo chimp	XP_003827419.1	MANMNSDSRHLGTSEVDHERDPGP
Homo sapiens	human	AAH66891.1	MANMNSDSRHLGTSEVDHERDPGP
Homo sapiens	human	NP_775785.2	MANMNSDSRHLGTSEVDHERDPGP
Homo sapiens	human	AAI01363.1	MANMNSDSRHLGTSEVDHERDPGP
Macaca fascicularis	longtailed macaque	NP_001253166.1	MANMNSDSRHLGTSKVDHERDPGP
Gorilla gorilla gorilla	gorilla	XP_004058897.1	MANMNSDSRHLGTSKVDHERDPGP
Papio anubis	olive baboon	XP_003899730.1	MANMNSDSRHLGTSKVDHKRDPGP
Pongo abelii	Sumatran orang-utan	XP_002815613.2	MANMNSDSRHLGTSQVDHERDPGP
Sorex araneus	common shrew	XP_004608492.1	MANTDSDSRHLLISEVDQEVNTGP
			-Xxxx- then 1 mutation in –NPGP-
Ochotona princeps	pika	XP_004583760.1	MAKMDSDSRHLLTSEGEPEVQPGP
Jaculus jaculus	jerboa	XP_004664993.1	MANMDHDSRCLLTPELEQEVGSGP
Mus musculus	house mouse	NP_848861.1	MASVDGDSRHLLSEVEHEVSPGP
Mus musculus	house mouse	AH52361.1	MASVDGDSRHLLSEVEHEVSPGP
Mus musculus	house mouse	EDL18407.1	MANVDGDSRHLLSEVEHEVSPGP
Mustela putorius furo	ferret	XP_004789848.1	MKLVKMRRDQLLKIMRSLPGP
Mustela putorius furo	ferret	XP_004744706.1	MKLVKMRRDQLLKIMRSLPGP
Echinops telfairi	lesser hedgehog tenrec	XP_004703769.1	MANVDKESVQPLLSEGLPEVNSGP
			mutated
Mus musculus	house mouse	BAC37465.1	MSNFLFNTGKFIFNFIHHINDTDTVLSTNNSNP

5.3.2 SNAT9 2As - *In Vitro* Recoding Activity Assays

All the SNAT-9 2A sequences cloned possessed extremely low ribosome skipping abilities (Figure 5.2), but not as low as those demonstrated by the *Nematostella vectensis* non-LTR 2A sequences, see Chapter 3.4.2) Surprisingly, there were also very faint bands observed from both variants of a second additional (downstream of 2A) 30 amino acid sequence (subsequently termed 2A2, refer to Figure 5.3) when cloned in this reporter system (on the very edge of visibility in this assay), giving some support to the tantalising hypothesis that the active 2As could have evolved from this low activity ancestral sequence.



a)

b)

ID tag	2A Sequence	Relative Activity
FMDV 2A	LYKSGSRGACQLLNFLLDKLAGDVESNPGP	++
Cow	MANMDSDSRHLLIPEGDHEINPGP	(+)
Sheep	MANMDSDSRHLLIPEGDHEVNPGP	(+)
Panda	MDSDSRHLLIPEVDHEIINPGP	(+)
Ferret	MDSDSRHLLIPEVNHEIINPGP	(+)
NMR	MTNVDDRHHLISEADHEVNPGP	(+)
Rat	MANVDSDSRHLISEVEHEVNPGP	(+)
2A2 I	MNKRIHYYSRLTTPADKALIAPDHVVPAPE	(+)
2A2 V	MNKRIHYYSRLTTPADKALVAPDHVVPAPE	(+)

Figure 5.2 SNAT9 2A Recoding activity analyses
a) SDS-PAGE gel of TnTs run on SNAT9 2As cloned in the reporter *pSTA1*. **b)** List of 2As tested and their relative recoding ability *in vitro* compared to *FMDV 2A* (++=moderately high activity comparable to *FMDV 2A*, +=low activity, (+)=very low activity. The "extra" bands visible below the [GFP-2A-GUS] band are due to the presence of additional internal initiation products – a common occurrence in TnT analyses.

5.3.3 SNAT9 2As – Bioinformatic Analysis

5.3.3.1 SNAT9 2A Protein Domain Configuration

When the SNAT9 2A-containing clade was explored in greater detail, it was discovered that all the 2A-possessing proteins possessed a second conserved sequence of 30 amino acids with similarities to 2A downstream of the 2A, but upstream of the first transmembrane helix (Figure 5.3). This second sequence will subsequently be referred to as the 2nd 2A-like sequence or 2A2 to distinguish it from the first 2A or 2A1. The only exceptions to this rule were two naked molerat, *H. glaber*, proteins where an apparent deletion event had erased the later portion of 2A2.

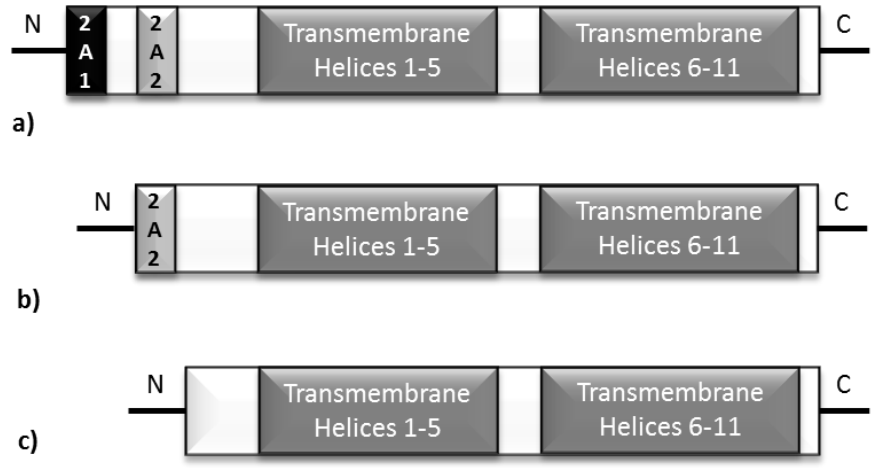


Figure 5.3 SNAT9 protein isoforms
a) Configuration of SNAT9 proteins containing an N-terminus 2A-like sequence (2A1 on the diagram) all possessed a second 2A-like sequence (2A2) downstream of 2A1. This configuration was found in placental mammals only **b)** SNAT9 proteins with the "2nd sequence" 2A2 only occurred from terrestrial vertebrates only **c)** sequences without either of the N-terminus 2A-like conserved sequences. Several mammalian species possessed all three isoforms of SNAT9 (see Figure 5.3).

2A2 was highly conserved with a sequence consisting of:

-MNKRIHYYSRLTTPADKAL[I/V]APDHVVPAPE-

The length of this tract, at 30 amino acids, was highly similar to the majority of 2A-like sequences, and the C-terminus (-DxxxPxPx-) resemblance to the 2A C-terminal motif presented the hypothesis that this was a second, non-active, degenerate 2A possibly present as a result of gene sequence duplication. However, examination of the proteins in the SNAT9 2A-clade without N-terminal 2A1 sequences found that a number without 2A1 possessed this "2nd degenerate 2A" as an N-terminal feature (Figures 5.3 & 5.4). A 2A2 sequence was found in a wider phylogenetic range of organisms than the 2A1 sequence that sometimes preceded it, being present as an N-terminal addition in monotreme, marsupial, avian, reptile, amphibian, but not teleost fish proteins; whereas the 2A1 sequences were found solely from placental mammals. Therefore, it would appear that

2A2 was the first feature in evolutionary terms, predating the acquisition of 2A1 2A-like sequences in these proteins. Therefore, could active 2As have evolved from this sequence? When tested *in vitro*, two common variants of 2A2 proved to be active at instigating ribosome skipping, albeit extremely weakly in comparison to *FMDV 2A* (Figure 5.2).

Interestingly, a number of species possessed multiple isoforms of the SNAT9 protein, with N-terminal 2A1 followed by 2A2, with N–terminal 2A2, and minus both 2A1 and 2A2 (Figures 5.3 & 5.4). The conserved N-terminal position of 2A1 and/or 2A2 is highly suggestive of a conserved role(s) for these sequences in SNAT9 transporter proteins.

5.3.3.2 SNAT9 2A1 Phylogeny

Next, all the SNAT9 2A1 sequences (22-24 amino acids) were aligned in order to determine their inter-relationship (Figure 5.5) and to discover whether this would be in agreement with their host species evolutionary phylogeny (as generally seen for the full SNAT9 alignment, Figure 5.4). It should be noted that due to the short lengths (22-24 amino acids), this alignment is not robust. However, the 2A cladogram shadowed the evolutionary relationships for these species (for example the primate sequences and the ungulate sequences each form a distinct clade, and the rodent sequences formed two clusters). Hence, these data apparently support the hypothesis that the SNAT9 protein isoform with an N-terminal 2A1 was derived from a common ancestral sequence that was acquired in the branch that led to the evolution of placental mammals. An additional point of interest is that the *FMDV 2A* sequence, when included as an outgroup in this analysis, fits not in a basal but in an apical position indicative of an evolutionary "younger" sequence.

Legend to **Figure 5.4** Cladogram of SNAT9 proteins with 2As (Figure on facing page). Branches in grey represent SNAT9 proteins lacking in 2As. Branches in black possess 2A2 only. Branches in red possess 2A1 followed by 2A2. Note the teleost fish SNAT9 proteins (grey, nearest the root of the tree, top and bottom of the image, are lacking 2A1 and 2A2. The amphibian, reptile and bird proteins possess 2A2 only (black clade second from top), whereas the 2A1 sequences (red branches) are scattered through the mammalian portion of the tree. Sequences labelled in blue denote those tested and found to possess extremely low ribosome skipping abilities *in vitro*. Sequences were aligned by ClustalX2 and visualised as a rooted cladogram using the default nearest neighbour joining algorithm in Figtree v1.4.2.

Key to labelling: all sequences with 2A1 are labelled 2A, with 2A2 are labelled 2nd. The species is given by the initials of its Latin nomenclature, then the NCBI accession number of the protein. If more than one species share the same initials, then please refer to the NCBI entry using the accession number to confirm species identity. Species abbreviations (in alphabetical order) are as follows:

Mammals:

AM=*Ailuropoda melanoleuca*, giant panda; BT=*Bos taurus*, cattle; CC=*Condylura cristata*, star-nosed mole; CF=*Camelus ferus,* Bactrian camel; CG=*Cricetulus griseus*, Chinese hamster; CJ=*Callithrix jacchus*, common marmoset; CLF=*Canis lupus familiaris,* dog; CP=*Cavia porcellus*, guinea-pig; CSS=*Ceratotherium simum simum*, white rhino; EC=*Equus caballus*, horse; ET=*Echinops telfairi*, lesser tenrec hedgehog; FC=*Felis catus*, domestic cat; GGG=*Gorilla gorilla gorilla*, Western lowland gorilla; HG=*Heterocephalus glaber*, naked molerat; HS=*Homo sapiens,* human; JJ=*Jaculus jaculus*, jerboa; LA=Loxodonta africana, African bush elephant; MA=*Mesocricetus auratus*, golden hamster; MB=*Myotis brandtii*, bat; MD=*Myotis davidii*, bat, or *Monodelphis domestica,* grey-tailed opossum; MF=*Macaca fascicularis*, long-tailed macaque; MM=*Mus musculus*, house mouse; MO=*Microtus ochrogaster*, prairie vole; MPF=*Mustela putorius furo*, ferret; NL=*Nomascus leucogenys*, northern white-cheeked gibbon; OA=*Ovis aries*, sheep, or *Ornithorhynchus anatinus*, platypus; OC=*Oryctolagus cuniculus*, rabbit; OD=*Octodon degus*, degu; OG= Otolemur garnettii; galago bush-baby; OO=*Orcinus orca*, orca whale; OP=*Ochotona princeps*, pika; ORD=*Odobenus rosmarus divergens*, Pacific walrus; PA=*Papio anubis*, olive baboon or Pongo abelii, Sumatran orang-utan or *Pteropus alecto*, black flying-fox bat; PP=*Pan paniscus,* bonobo chimpanzee; PT=Pan troglodytes, common chimpanzee; RN=*Rattus norvegicus*, brown rat; SA=*Sorex araneus*, shrew; SBB=*Saimiri boliviensis boliviensis*, squirrel monkey; SH=*Sarcophilus harrisii*, Tasmanian devil; SS=*Sus scrofa*, pig; TML=*Trichechus manatus latirostris*, Florida manatee TT=*Tursiops truncates*, bottle-nosed dolphin

Birds: AP=*Anas platyrhynchos*, mallard duck; FA=*Ficedula albicollis*, collared flycatcher; FP=*Falco peregrinus*, peregrine falcon GG=*Gallus gallus*, chicken; MG=*Meleagris gallopavo*, turkey; MU=*Melopsittacus undulates*, budgerigar; TG=Taeniopygia guttata, zebra finch

Reptiles: CM=*Chelonia mydas*, green turtle; CPB=*Chrysemys picta bellii*, painted terrapin; AC=*Anolis carolinensis*, anolis lizard

Amphibians: XL=*Xenopus laevis*, clawed frog; XST=*Xenopus (Silurana) tropicalis*, clawed frog

Fish: DR=*Danio rerio*, zebra fish; MZ=*Maylandia zebra*, cichlid; OL=*Oryzias latipes*, Japanese rice fish; ON=*Oreochromis niloticus*, Nile tilapia; TR=*Takifugu rubripes*, puffer-fish

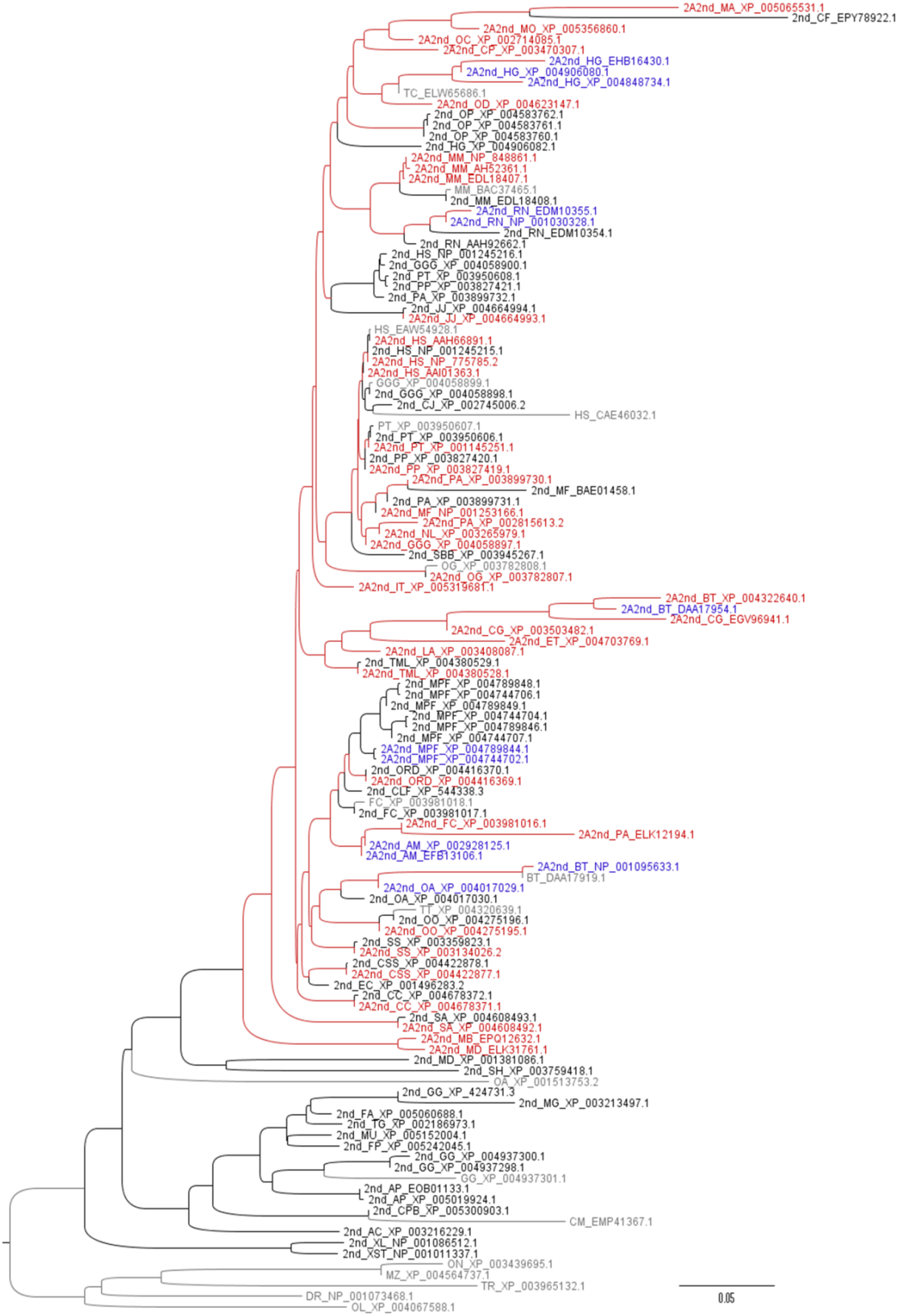


Figure 5.4 Cladogram of SNAT9 proteins with 2As
(Legend on preceding page).

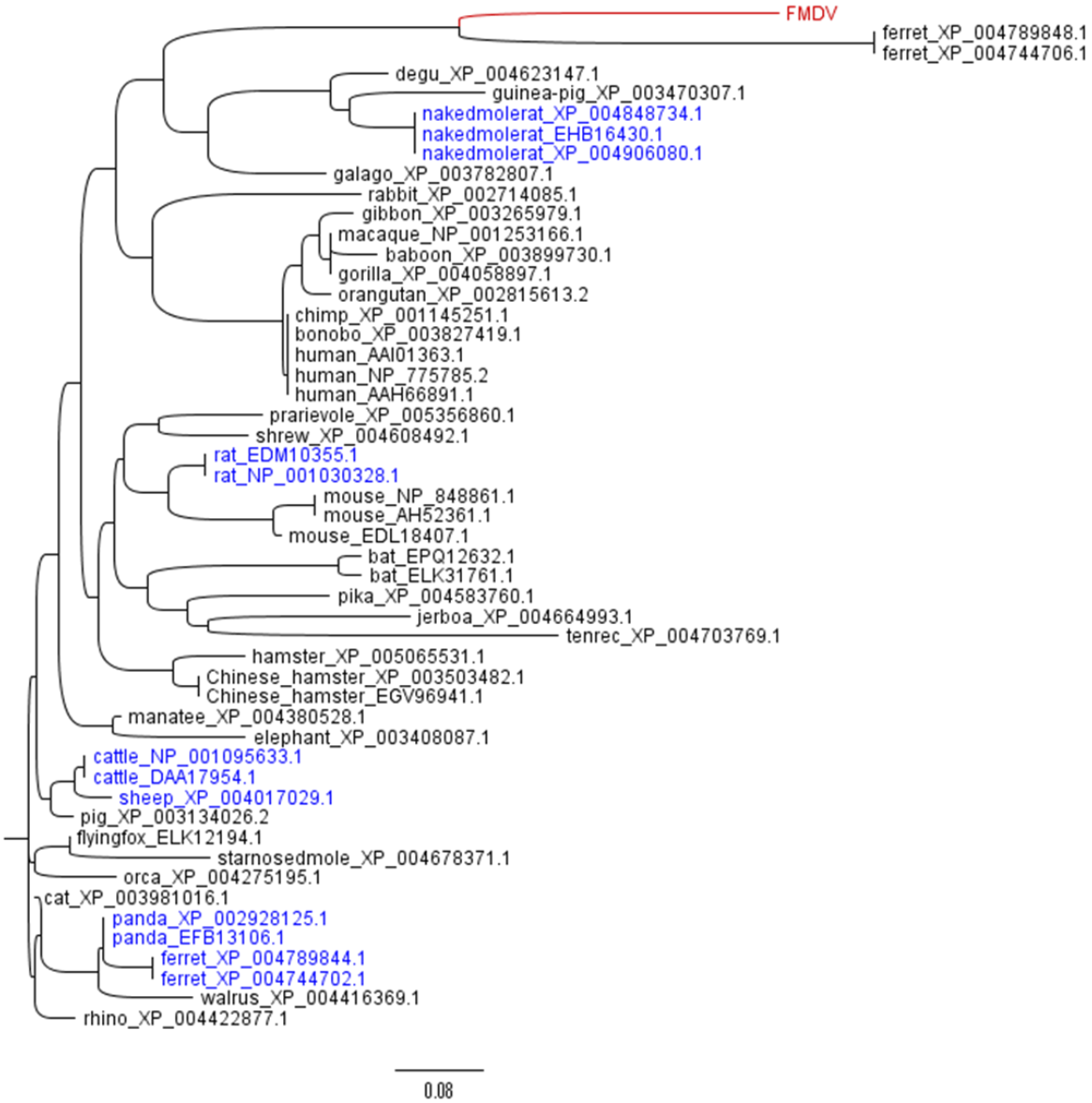


Figure 5.5 SNAT9 2A1 cladogram

All SNAT9 2A1 sequences aligned, note the phylogeny generally follows the evolutionary relationships of the host species; for example the primate and ungulate species each form a distinct clade. Sequences labelled in black denote SNAT9 2A1s not yet tested *in vitro* for translational recoding abilities, sequences labelled in blue found to possess extremely low (in comparison to *FMDV 2A*) ribosome skipping capabilities *in vitro*. 24 amino acid *FMDV 2A* (dark red) has been included for comparison purposes. Sequences aligned with ClustalX2 and rooted cladogram drawn with Figtree v1.4.2 using the default nearest neighbour joining algorithm.

5.4 **Discussion**

This study represents the first report of active (albeit extremely weak) ribosome skipping 2A sequences from mammals. It adds to our understanding of 2A sequence topology, by introducing new possibilities in 2A C-terminal motifs, namely, -DHEINPGP- and -NHEINPGP-, and those with an isoleucine duplication -DHEIINPGP- and -NHEIINPGP-. These 2A sequences were all found associated with SNAT9 proteins, encoded by the *SLC38A9* gene. Little is known about this particular protein, save that it is a plasma membrane-embedded sodium-dependent amino acid transporter protein belonging to the SNAT protein family. This family is monophyletic and evolutionary ancient, with examples from all forms of extant metazoan life and analogues in bacteria. The 2A sequences occurred only in database entries of the SNAT9 protein from placental mammals; no 2A sequences were detected in the monotreme (egg-laying mammal, *Ornithorhynchus anatinus*, duck-billed platypus) nor marsupial (*Otolemur garnettii*, galago bush-baby and *Sarcophilus harrisii*, Tasmanian devil) SNAT9 proteins, nor from those entries from birds, reptiles, amphibians or teleost fish. However, all of these, with the exception of the fish sequences, possessed a 2nd possibly related and/or ancestral "2A2" at the N-termini of one of their SNAT9 isoforms. The relationship of this second 2A2 sequence to the canonical 2A sequence (here termed 2A1) is at present unclear. One hypothesis is that 2A1 may have evolved from duplication of 2A2. This pattern of 2A distribution, with 2A1 occurring only in the placental mammal SNAT9s permits the first potential dating of 2A acquisition. The insertion of 2A1 into the *SLC38A9* gene must presumably have occurred after the evolution of the early mammals but before the branch that led to modern placental mammals. This places this event within the Cretaceous paleo-geological period (145 to 65 million years ago), at a maximum of 129 million and a minimum of 78 million years ago (Wible *et al.*, 2007). This is still a wide timeframe, but it is nevertheless exciting as it the first time it has been possible to place any numbers on the age of 2A.

The conserved positioning of 2A-like sequences at the protein N-terminus, and the presence of multiple isoforms of the SNAT9 protein, both with and without 2As, indicates 2A may function in determining SNAT9 metabolism. In SNAT9 proteins the N-termini are thought to be instrumental in determining protein longevity, albeit that the mechanism is presently unknown (reviewed in Schiöth *et al.*, 2013; Bröer, 2014). Perhaps, in the instances where 2A is "cleaved" off, this increases the speed of degradation. Further experimentation is required to elucidate the function of the 2A-like sequences, 2A1 and 2A2, observed in these SNAT9 proteins. One possible function will be investigated more fully in Chapter 7, as part of a study into dual-purpose 2As with a secondary function as signal peptides. As outlined there (Chapter 7), one working premise is that 2A2 may be a signal peptide targeting the protein to the mitochondrial membrane, whilst 2A1 may be may be a nuclear or mitochondrial import signal. These suppositions will be investigated in Chapter 7.

Chapter 6. NLR-Like Protein Associated 2As

‘You do not mean there is danger of peace?’

Desolation Island - Patrick O'Brian, 1978

6.1 Introduction – NLR-Like Proteins

This Chapter constitutes a report of 2A-like sequences found at the N-termini of NLR proteins from five invertebrate organisms, but predominately from the purple sea-urchin *S. purpuratus*.

6.1.1 Introducing NLRs

The innate immune system recognises non-self threats through pattern recognition receptors (PRRs) that detect and bind to common pathogen-associated molecular patterns (PAMPs) such as the peptidoglycans of bacterial cell walls or viral nucleic acids. The binding of PRR to PAMP initiates signalling cascades. These either up-regulate cellular defence mechanisms through production of immune messengers (such as interferon) and/or apoptotic programmed cell death of infected cells to combat the perceived danger, or can down-regulate the innate immune system into a peace-time standby state by signalling for a “stand-down” of the aforementioned responses. The first PRR system to be discovered was the *Drosophila* cell membrane Toll proteins. Subsequently it was found that mammalian cells possess similar membrane-bound Toll-like receptors (TLRs), plus cytoplasmic RIG-1-like receptors (RLRs), and nucleotide-binding domain and leucine-rich repeat Nod-like receptor (NLR) proteins (reviewed in Lich and Ting, 2007). All these proteins display similarities in architecture, in that they possess a PRR domain, a domain that interacts with the signalling cascade, and commonly a C-terminal leucine-rich repeat domain. It is thought that in the absence of PAMP binding, the protein is configured such that the leucine-repeat region is folded back to conceal the nucleotide-binding and signalling cascade domain and thus inactivate the protein until PAMP binding results in conformation changes that reveal the effector domain and in some instances protein oligomerization (reviewed in Kanneganti *et al.*, 2007).

The NLR proteins are a phylogenetically widespread (spanning both the animal and plant kingdoms) and diverse protein family (reviewed in Lich and Ting, 2007). Various research groups have previously referred to these proteins as CATERPILLAR (caspase activation and recruitment domains [CARD], transcription enhancer, R, [purine]-binding, pyrin, lots of leucine repeats), or NOD-LRR (nucleotide oligomerization domain-lots of leucine repeats) or NLR-LRR (NATCH [NAIP, CIIATA, HET-E, and TP1]-lots of leucine repeats), but the nomenclature is standardizing towards use of NLR. All NLR proteins possess a common topology: their N-termini are highly variable but typically include a signalling cascade effector domain such as a pyrin, CARD or

DEATH domain, this is followed by an in-frame PRR domain (generally an NTPase nucleotide binding domain such as NATCH) and finally a C-terminal leucine rich repeat region (Figure 6.1)

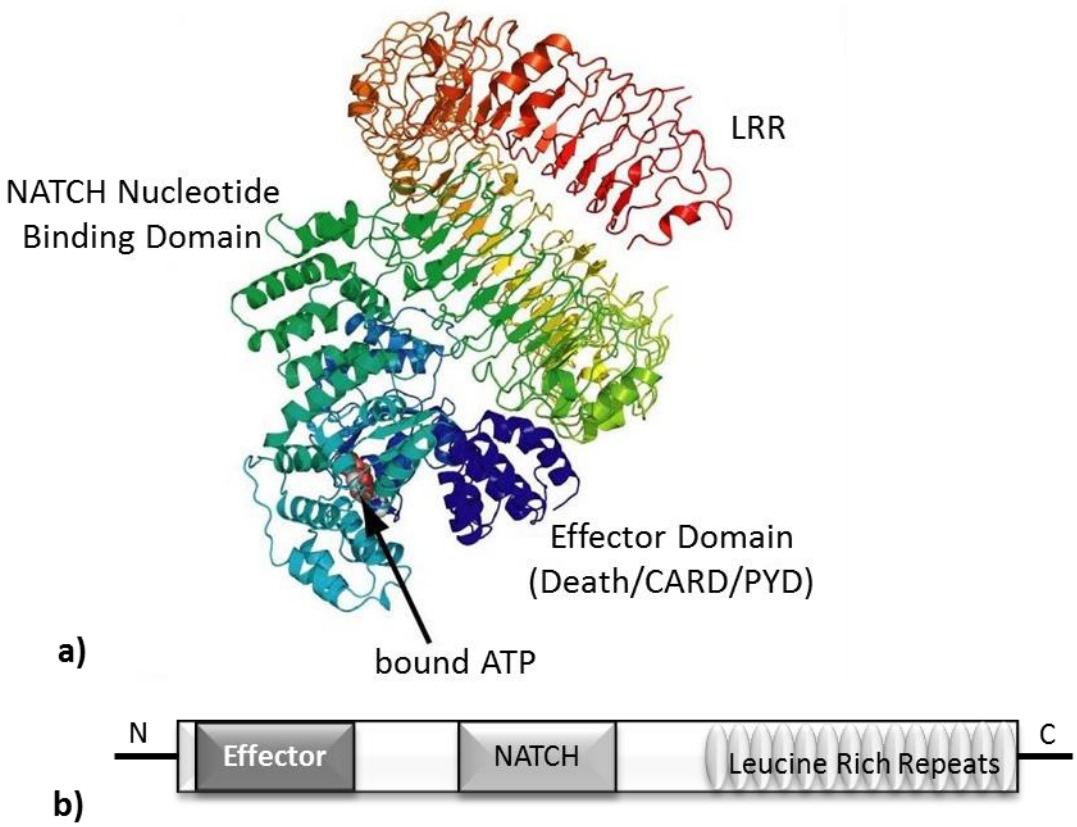


Figure 6.1 NLR protein topology
a) Model illustrating the proposed structure of mammalian NLRC5 (adapted from Neerincx *et al.*, 2010). The N-terminus effector and NACHT domain (darker grey) are shown as is the position of the bound ATP molecule. The C-terminus leucine repeat region is shown in paler grey. Note that the C-terminus LRR "arm" can fold back across the protein effectively blocking the NATCH and/or the effector domain. **b)** Schematic showing the typical domain structure of NLR proteins.

Unlike the transmembrane TLRs, NLRs are thought to localise to the cytoplasm and perhaps the nucleus (Neerincx *et al.*, 2010). NLR proteins are classified into subfamilies on the basis of their N-terminal effector domains. There are two major subfamilies, the CARD subfamily (alternatively termed NLRC, NATCH-LRR or CARD-domain containing proteins), and the pyrin subfamily (also referred to as NLRP, NATCH-LRR, or PYD-containing protein). Both CARD and pyrin domains belong to the death effector domain protein class. Typically they possess a six α-helical structure and signal via homodimerization interactions, for example CARD-CARD or pyrin-pyrin (reviewed in Wilmanski *et al.*, 2008).

There are approximately 20-30 human NLR proteins, 30-40 from mice, but several hundred similar proteins found in both plants and sea-urchins (due to the modular nature of these proteins, and alternative gene splicing, the precise number is unknown) (reviewed in Hibino *et al.*, 2006;Wilmanski *et al.*, 2008). NLRs have been shown to respond to a wide variety of microbial

components. For example, human NOD1 and NOD2 detect bacterial wall components, whereas Ipaf and Naip detect bacterial flagellins, and NLRP1 responds to anthrax toxins. Activation of human NLRs either triggers inflammatory pathway responses mediated by NF-κB, MAPK or Caspase-1 to result in the secretion of pro-inflammatory cytokines (reviewed in Lich and Ting, 2007; Wilmanski *et al.*, 2008) or can act to counter this immune defence by blocking pathway interactions (Schneider *et al.*, 2012).

Not surprisingly, given their critical role in cell self-defence, mutations in human NLR genes have been implicated in contributing to inflammatory disease risk. There is a growing body of evidence to suggest that mutations in the *Card15* gene region (encoding NOD2) correlate with an increased risk of developing Crohn's disease (an inflammatory bowel disorder believed to have a genetic component). Two further inflammatory genetic disorders thought to be a result of mutations to *Card15* are Blau syndrome and early-onset sarcoidosis. In addition, mutations to the *Cias1* gene (also known as NALP3 and PYPAF1), or to *MEFV* which encodes PYRIN result in patients presenting with a range of auto-immune symptoms. A point mutation (L155H) within NALP1 may be the cause of the skin disorder vitiligo. Mutations to the transcription factor for CIITA result in the autosomal recessive hereditary immunodeficiency, bare lymphocyte syndrome (reviewed in Wilmanski *et al.*, 2008).

2A-like sequences in association with NLRs were predominately identified from the sea-urchin *S. purpuratus*. A previous study (Hibino *et al.*, 2006) reported that sea-urchin NLRs closely resembled mammalian NLRC3 and NLRC5 proteins, albeit with one major difference, namely that the sea-urchin proteins possessed an N-terminal Death Domain in place of the NLRC CARD domain. In mammals, the apparent function of NLRC3 is to attenuate TLR pathway responses. NLRC3 inhibits the membrane-bound Toll-like receptor (TLR) NF-κB activation by interacting with TRAF6 – one of the proteins involved in the TLR signalling cascade (Schneider *et al.*, 2012). In contrast, NLRC5 up-regulates the NF-κB and interferon response through binding to intracellular viral dsRNA (Neerincx *et al.*, 2010). Mammalian NLRC3 and NLRC5 were shown to be highly expressed in haematopoeitic cells (Neerincx *et al.*, 2010; Schneider *et al.*, 2012); however, in sea-urchins the highest level of NLR expression was observed from the gut tissues (Hibino *et al.*, 2006).

6.2 **Methodology**

6.2.1 Contributions

All bioinformatics analyses and primer design were the work of the author as were the *in vitro* coupled transcription-translation reaction (TnTs) analyses of *STR6* and *pSTA1_FMDV*. Thanks are due to Dr. Garry Luke for cloning the *S. purpuratus* sequence *STR6*. Cloning and TnT analyses of the *A. queenslandica* NLR sequences were undertaken by Claire Stewart under the author's supervision.

6.2.2 *In Silico* Searches

Online proteomic and genomic databases were probed for eukaryotic 2A-like sequences as described in Chapter 2.1.1. The complied list of 2A-containing proteins (Appendix B) was then screened for conserved protein domains using the "find conserved domain" function of NCBI BLAST. Sequence alignments were performed using ClustalX2 as detailed in Chapter 2.1.2 and visualised using Figtree v1.4.2 or Phylodraw. Multispecies NLR protein alignments were made on the basis of trimmed NATCH/NTPase domain alignment using the NATCH domain from human NLRC3 (NCBI Accession ACP40993.1) as an outgroup. The NTPase (NATCH) domain of sea-urchin 2A-NLR protein NCBI Accession XP_003727629.1 was used as a template for a BLAST search to ascertain the number of NLR-like proteins from *S. purpuratus* currently in the online NCBI databases. Similarly, the NTPase domains from amphioxus *B. Floridae* XP_002614028.1, clam *L. gigantea* ESO82727.1, sponge *A. queenslandica* XP_003390214.1, tick *I. scapularis* AC205634.1 were each used as templates for BLAST searches to ascertain the relative number of NLR-like protein in each species, respectively. In the interests of clarity, 2A sequences selected for further analyses were given short identification tags based on their Latin species names species (AQ=*A. queenslandica*, SP/STR=*S. purpuratus*, BF=*B. floridae*, IS=*I. scapularis*, LG=*L. gigantea*). These tags are presented alongside the NCBI accession numbers throughout this Chapter.

6.2.3 *In Vitro* – Methodology

Cloning of putative 2A sequences into the reporter *pSTA1* plasmid were undertaken as detailed in Chapter 2.2.1. Briefly, the 2A peptide sequences were translated into nucleic acid and cloned by means of long reverse primers as in Chapter 5, except for sequence *STR-37* which was obtained as part of the gene-blocks ordered for the 2A consensus motif investigation reported in Chapter 8. Nucleotide/primer sequences are listed in Table 6.1. Plasmid sequences were verified by DNA sequencing with primer *GFPf* and/or *GUS_seq_R* (Table 2.1). Plasmid preparations were used to program TnTs as detailed in Chapter 2.2.2 and Odon *et al.*, 2013.

Table 6.1 Gene-block/primer sequences for NLR 2As
STR-37 was cloned from a gene-block sequence and *STR6* had previously been cloned by Dr Luke. The *A. queenslandica* sequences were built into the vector (*pSTA1*) by means of PCR using long reverse primers. Underlining denotes the XbaI/ApaI restriction sites used in cloning.

ID Tag	NCBI Accession	2A Amino Acid Sequence Cloned	Nucleic Acid (primer/gene-block/in lab)
AQ24	XP_003390020	QNVCYHHFMFLLLLLAGDIELNPGP	Reverse primer: 5'TGGTGGGGCCCTGGGTTCAGTTCAA TATCGCCCGCCAGCAGCAGCAGCAGAA ACATAAAATGATGATAGCACACGTTCT G TCTAGACCCGGACTTGTATAGTT-3'
AQ12	XP_003383551	ACFLYSVFVVKLLLLSGDVELNPGP	Reverse primer: 5'TGGTGGGGCCCTGGGTTCAGTTCCA CATCGCCGCTCAGCAGCAGCAGTTTCA CCACAAACACGCTATACAGAAAGCACG CTCTAGACCCGGACTTGTATAGTT-3'
STR-37	XP_003730724.1	MTNILLLRSGDVERNPGP	Gene-block: 5'ATGACGAACATACTACTACTACGAT CAGGAGACGTAGAACGAAACCCAGGAC CG-3'
STR6	XP_798533.3	MDGFCLLYLLLILLMRSGDVETNPGP	In lab: 5'ATGGATGGATTCTGTCTTCTCTATC TGCTCCTGATCCTCTTGATGAGGTCTG GTGACGTTGAAACCAATCCAGGTCCT-3'

6.3 **NLR 2As – Results**

6.3.1 Identification of 2A-NLRs

The newly discovered eukaryotic 2A-like sequences (Appendix B) were screened to determine the potential function(s) of their associated proteins and to test their *in vitro* translational recoding abilities. As discussed in Chapter 3, a considerable number of the 2A-like sequences were found in association with non-LTR transposons. There were an additional number found in association with ankyrin-repeat proteins (discussed in Chapter 4), or membrane-associated amino acid transporters (Chapter 5) or NLR-like proteins. The NLR-associated 2As will be discussed here.

2As in NLR-like proteins were identified from organisms in five eukaryotic phyla, namely the echinoderm *S. purpuratus*, the cephalochordate *B. floridae,* the arthropod *I. scapularis*, the mollusc *L. gigantea* and the sponge *A. queenslandica* (Table 6.2). However, the majority of 2A-like sequences discovered (18) were from *S. purpuratus*, in contrast there was only one record each from both *L. gigantea* and *I. scapularis,* and three hits each from *A. queenslandica* and *B. floridae* (Table 6.2). Four of these NLR-associated 2A sequences were selected for *in vitro* translational recoding analyses (grey background in Table 6.2). 2A-containing NLR-like proteins were identified on the basis of their possessing both a 2A-like sequence ending in the canonical 2A C-termini motif -D[V/I]ExNPGP- and a NTPase/NATCH domain. These proteins were similar to the mammalian NLRC3 and NLRC5 proteins, albeit that they possessed an N-terminal Death Domain in place of the CARD domain of NLRC proteins.

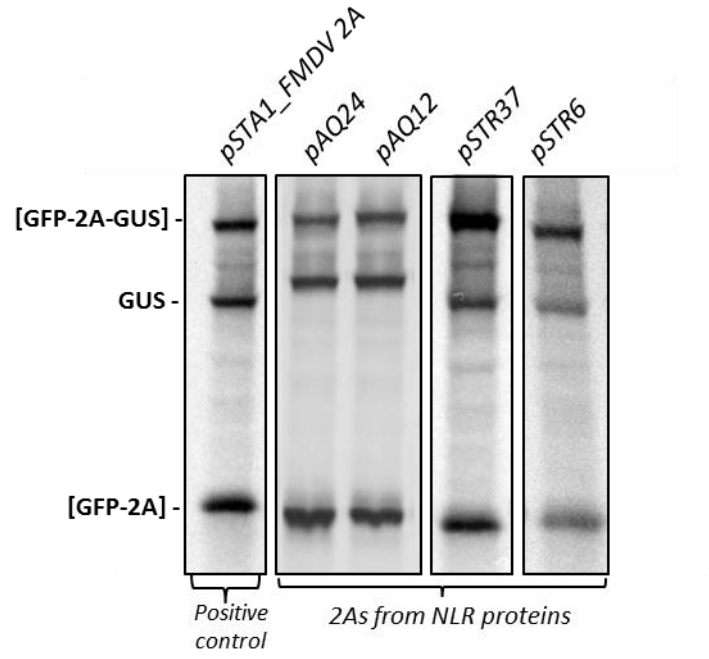
Table 6.2 List of 2A-containing NLR proteins

The 30 amino acid 2A sequences are listed, as are the NCBI accession numbers, and the laboratory ID tag. 2A amino sequences in **bold text** correspond to the length of sequences cloned for *in vitro* analyses), sequences selected for in vitro analyses are presented on a grey background. 2a sequences were categorised as N-terminal if they occurred less than 100 amino acids from the protein N-termini.

ID Tag/Host Organism	NCBI Accession	30aa 2A	2A is N-terminal?
AQ24	XP_003390020.1	YTESN**QNVCYHHFMFLLLLLAGDIELNPGP**	no
AQ12	XP_003383551.1	VSDIL**ACFLYSVFVVKLLLLSGDVELNPGP**	no
AQ25	XP_003390214.1	ASILVCIFLYFVVCRLLLFLSGDIELNPGP	yes
BF	XP_002614028.1	WAEASCLLVWMVISQLMLKLAGDVEENPGP	yes
BF	XP_002609384.1	YVMMSCLLVWFMMVHKLLLQAGDIEPNPGP	yes
BF	XP_002591693.1	WFEAKMCYCSYVILVLLLLMAGDVEVNPGP	no
IS	AC205634.1	NRRKFYYRKKVIFVLLLLLLSGDVETNPGP	no
LG	ESO82727.1	ISTGSLTISTILVIAILLVLSGDVEVNPGP	yes
SP	XP_797335.3	ISCTMDGLCLLYLLLILLMRSGDVETNPGP	yes
STR6	XP_794591.3	TSYT**MDGFCLLYLLLILLMRSGDVETNPGP**	yes
STR6	XP_003729541.1	TSYT**MDGFCLLYLLLILLMRSGDVETNPGP**	no
STR6	XP_001184399.2	RSYT**MDGFCLLYLLLILLMRSGDVETNPGP**	no
SP	XP_003724839.1	MVGFCLLYLLMILLMRSGDVETNPGP	yes
SP	XP_003730118.1	MAGFCLLYLLMILLMRSGDVETNPGP	yes
SP	XP_800245.3	MDGFCLLYLLLILLVRSGDVETNPGP	yes
SP	XP_003727629.1	MDAFCLLYLLMILLMRSGDVETNPGP	yes
SP	XP_003728571.1	MDGFCLLYLLMILLMRSGDVETNPGP	yes
SP	XP_003723697.1	MDGFCLLYLLMILLMRSGDVETNPGP	yes
SP	XP_003729593.1	MDGFCLLYLLMILLMRSGDVETNPGP	yes
SP	XP_003730320.1	MGGFCHLYLLMILLMRSGDVETNPGP	yes
SP	XP_003724840.1	MDGFCLLYLIMILLMRSGDVETNPGP	yes
SP	XP_003730452.1	TTDDPVKEDSACLPEMLLVKAGDVELNPGP	no
SP	XP_788942.3	EADTIKGNDCPDMANILLLRSGDVERNPGP	no
SP	XP_001198729.1	LHPAILCSASLCFRPYLLLMAGDVEPNPGP	yes
SP	XP_003729085.1	TTDDPVMQESTCLPEMLLVKAGDVEQNPGP	no
STR-37	XP_003730724.1	EAYAIKGNDCPD**MTNILLLRSGDVERNPGP**	no
Lacking NACHT Domain:			
SP	XP_785721.3	MDGFCLLYLLMILLVRSGDVETNPGP	yes
STR6	XP_798533.3	ICYT**MDGFCLLYLLLILLMRSGDVETNPGP**	no
SP	XP_003730835.1	EADAIKGNDCPDVTNMLLLRSGDVERNPGP	no
SP	XP_800440.1	KADAIKGNDFPNMTYILLLRCGDVELNPGP	yes
Mutation from canonical -DxExNPGP- 2A C-terminus			
AQ28	XP_003388278.1	LPQTGVEEAISREEEELRVESANVELNPGP	no

6.3.2 NLR 2As – Translational Recoding Assays

Four of the NLR associated 2As were investigated to ascertain their translational recoding abilities. All were found to possess ribosome skipping capabilities comparable with those of the positive control, namely *FMDV 2A* (Figure 6.2). Even the short (18 amino acid) form of *STR-37* displayed relatively high recoding abilities, note that this sequence was previously examined in Chapter 4 as part of the investigation of ankyrin-repeat 2As as its host protein possesses both an NTPase and an ankyrin-repeat domain.



a)

ID tag	2A Sequence	Relative Activity
FMDV 2A	SGSRGACQLLNFLLDKLAGDVESNPGP	++
AQ24	QNVCYHHFMFLLLLLAGDIELNPGP	++
AQ12	ACFLYSVFVVKLLLLSGDVELNPGP	++
STR37	MTNILLLRSGDVERNPGP	++
STR6	MDGFCLLYLLLILLMRSGDVETNPGP	++

b)

Figure 6.2 Recoding activity analyses of NLR-associated 2As
a) SDS-PAGE gel of TnTs run on 2A constructs cloned in the reporter *pSTA1*. (Image is a composite showing representative lanes from different gels run under identical conditions – see Sections 2.2.2 and 2.2.3) **b)** Table listing the 2A tested and recording their relative recoding ability in comparison to *FMDV 2A* (++=moderately high activity comparable to *FMDV 2A*, where +++=higher than *FMDV*, +low activity, (+) very low activity). The "extra" bands visible below the [GFP-2A-GUS] band are due to the presence of additional internal initiation products – a common occurrence in TnT analyses (see Odon *et al.* 2013). STR37 recoding was previously shown in Figure 4.2.

6.3.3 NLR 2As – Bioinformatic Analyses

6.3.3.1 2A-NLR Protein Architecture

The positioning of the 2A peptides within the *S. purpuratus* NLR proteins was examined. In general, all the proteins displayed similar domain architectures with the 2A occurring upstream of the Death Domain (where present) and the NTPase/NATCH domain, and in the majority of instances the 2A was positioned at or close to the protein N-terminus (Figure 6.3). Four additional *S. purpuratus* proteins were discovered that possessed 2A and Death Domains, but that apparently lacked an NTPase domain (Figure 6.3). The architecture of the handful of NLR-like proteins identified from the other invertebrate species was similar to that of the *S. purpuratus* proteins with 2A occurring upstream of the effector and NTPase domains.

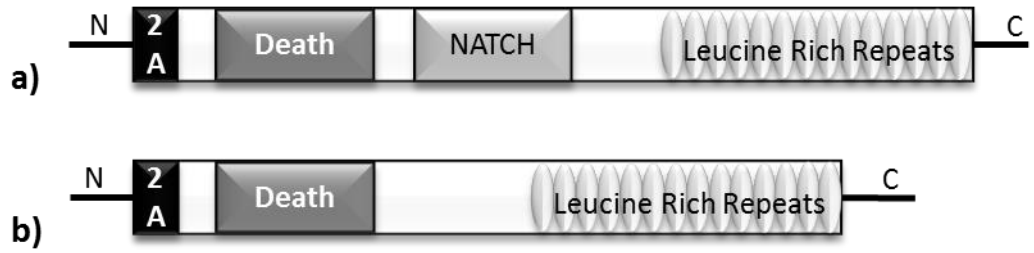


Figure 6.3 Schematic of the protein domain configuration of NLR 2A-containing proteins
a) Typical configuration of 2A-NLR-like proteins from sea-urchin (closest mammalian matches NLRC3 or NLRC5). **b)** Additional sea-urchin proteins with 2A and Death but apparently lacking a NATCH domain (not to scale).

6.3.3.2 NLR 2As - Phylogenetic Relationships

The phylogenetic relatedness of the NLR-like proteins possessing 2A was examined through alignment of their NTPase (NATCH) domains; hence the four *S. purpuratus* sea-urchin proteins apparently lacking NTPase domains were excluded from this analysis. Unsurprisingly, the majority of the sea-urchin sequences formed a discrete clade, as did the sponge sequences (Figure 6.4).

The distribution of 2A sequences within the *S. purpuratus* NLR proteins was examined. A similar investigation was not undertaken for the other species due to the low numbers of 2As in NLRs in these organisms (Table 6.3). An earlier study (Hibino *et al.*, 2006) found 203 NLR proteins from the sea-urchin *S. purpuratus*; but subsequent database consolidation (reported in Tu *et al.*, 2012) resulted in my identification of only 88 NLR-like proteins from *S. purpuratus* (Table 6.3).

Table 6.3 Multi-species comparison of occurrences of NLR-associated 2As
For each species with NLR-associated 2As, the total number of NLR-like proteins is listed, as ascertained by the number of significant hits reported from an NCBI BLAST search (see Section 6.2.2). The number of these NLR proteins with 2A is then given, as is the percentage of NLRs from each species with 2As.

Host Species	Total Number of NLR-like proteins	Number of NLRs possessing 2A-like Sequences	Percentage of NLRs with 2As
S. purpuratus	88*	18	20.5%
A. queenslandica	105	3	2.9%
B. floridae	110	3	2.7%
L. gigantea	26	1	3.8%
I. scapularis	15	1	6.7%

*There were a further four *S. purpuratus* proteins with 2A-Death but apparently lacking in an NTPase domain. These have been omitted from this analysis.

The *S. purpuratus* NLR-like proteins, plus the four *S. purpuratus* proteins containing 2A-Death but lacking NATCH were aligned (Figure 6.5). 2As were observed both in the evolutionary "young" clades (as shown by relative branch length from the root on the cladogram) and in deep-rooted evolutionary older NLR isoforms. In both instances the neighbouring NLR proteins did not possess 2As. This 2A distribution pattern is indicative of either six separate 2A acquisition events during the evolution of *S. purpuratus* proteins, or an early acquisition of 2A followed by subsequent losses.

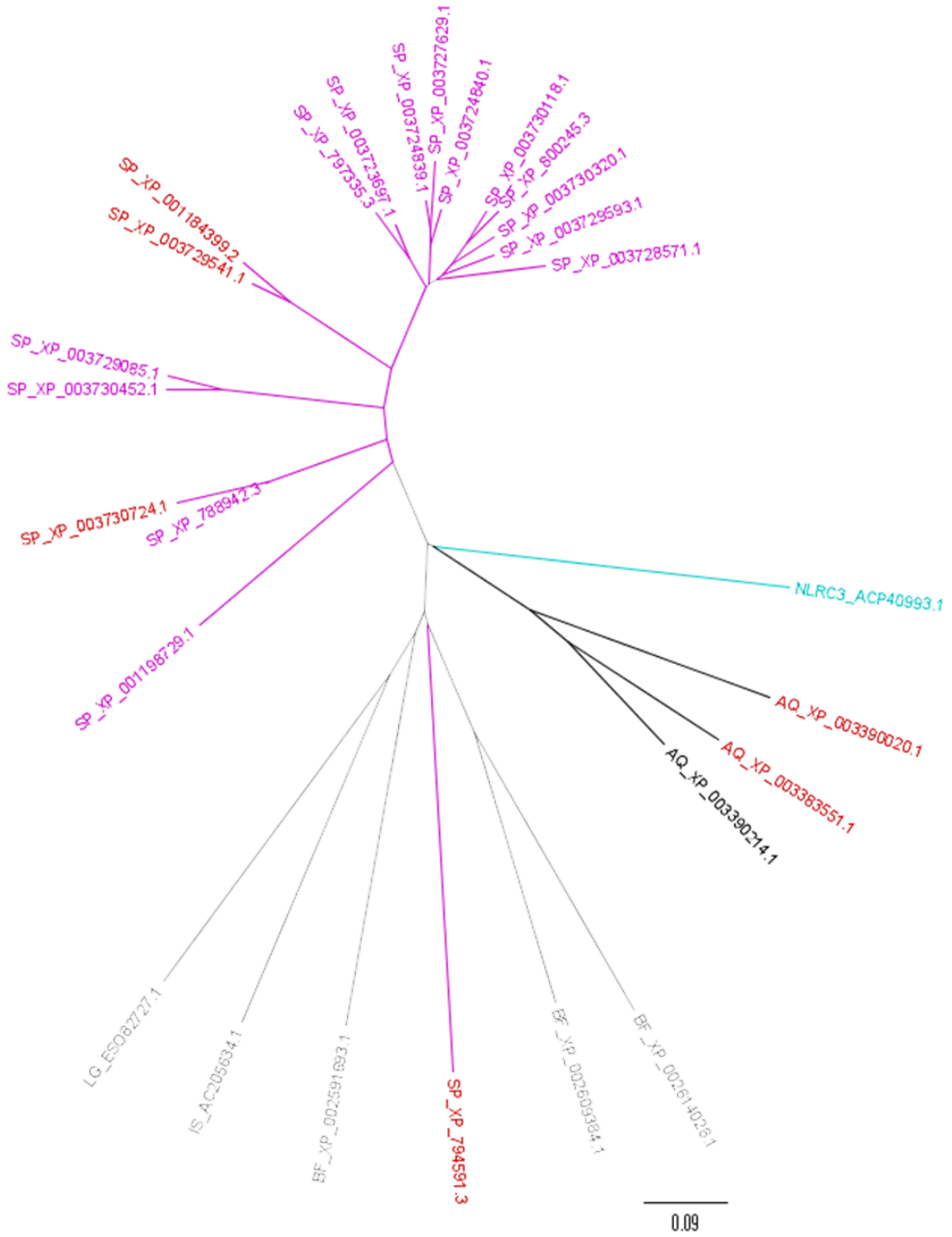


Figure 6.4 Cladogram of aligned NTPase domains from 2A-NLR proteins
Proteins with 2A sequences found to possess ribosome skipping abilities comparable to *FMDV 2A* in *in vitro* assays (Figure 6.2) are coloured red. Sea-urchin *S. purpuratus* proteins with non-tested 2As in purple, non-tested sponge *A. queenslandica* in black, other species in grey, human NLRC NATCH domain is shown in cyan. Proteins are identified by species codes and NCBI accession numbers, unrooted tree drawn with Figtree v1.4.2 using the default nearest neighbour joining algorithm from a ClustalX2 alignment.

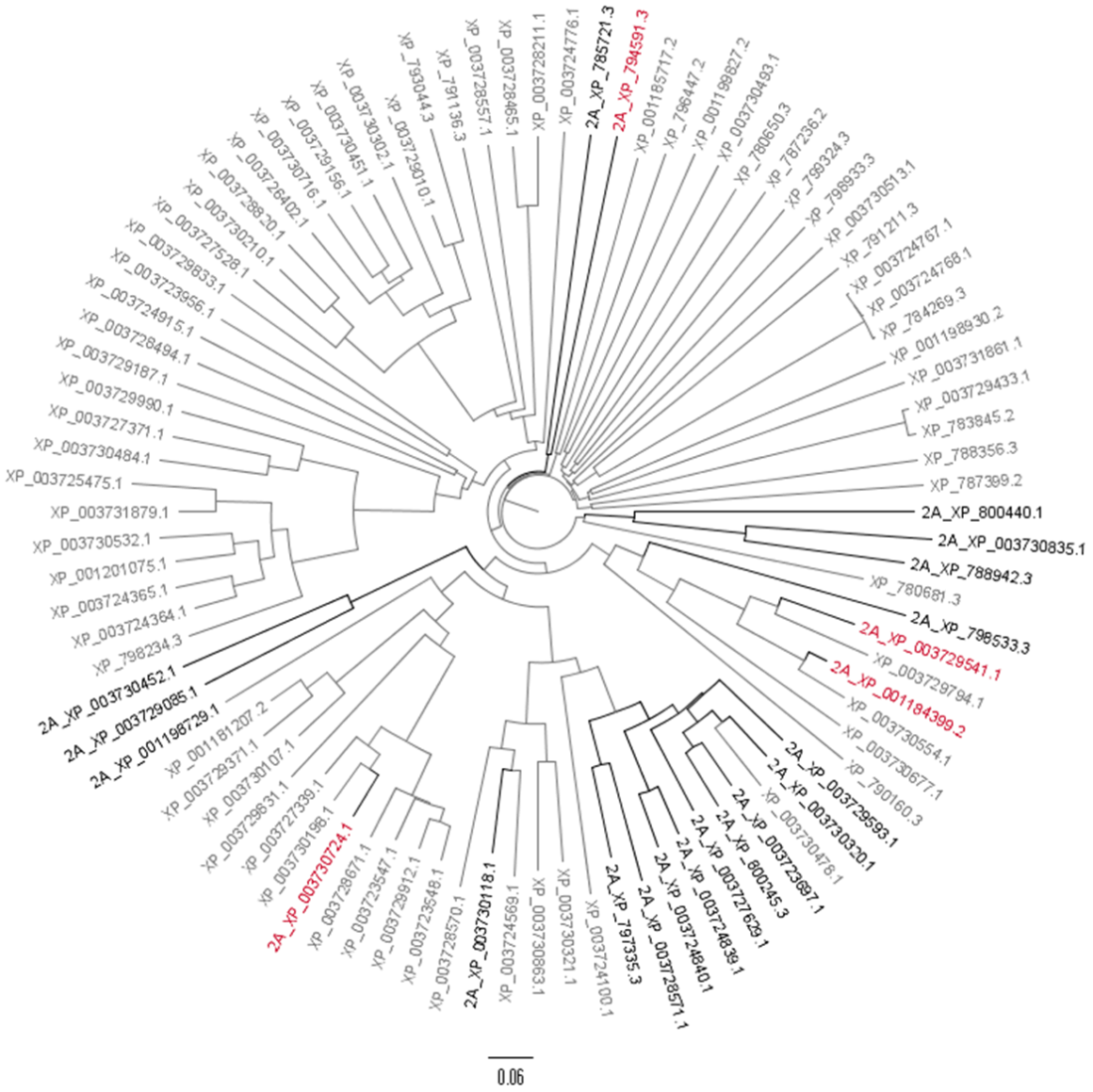


Figure 6.5 Cladogram of *S. purpuratus* NLR proteins
NLR proteins with 2A sequences shown by *in vitro* analyses to possess comparable translational recoding activity to *FMDV 2A* are coloured red, non-tested 2A sequences are shown in black, and NLR proteins without 2As are coloured grey. Additionally, proteins with 2A are tagged by 2A_ preceding their NCBI accession number. Rooted tree drawn with Figtree v1.4.2 using the default nearest neighbour joining algorithm from a ClustalX2 alignment.

Finally, all the NLR 2A sequences were aligned (length <30 amino acids) in order to determine their inter-relationships (Figure 6.6). It should be noted that due to the short lengths, this alignment is not robust. However, a group of the *S. purpuratus* 2As can be seen to form a monophyletic clade that corresponds with the largest *S. purpuratus* clade from the NTPase domain alignment (Figure 6.4) strongly suggestive of a common ancestral 2A-containing NLR protein giving rise to this clade. Interestingly, the 2A sequences similar in amino acid composition to *FMDV 2A* (the *A. queenlandica* 2As tested) and the more divergent sequences (the sea-urchin sequences) all displayed broadly comparable translational recoding abilities when tested (Figure 6.2).

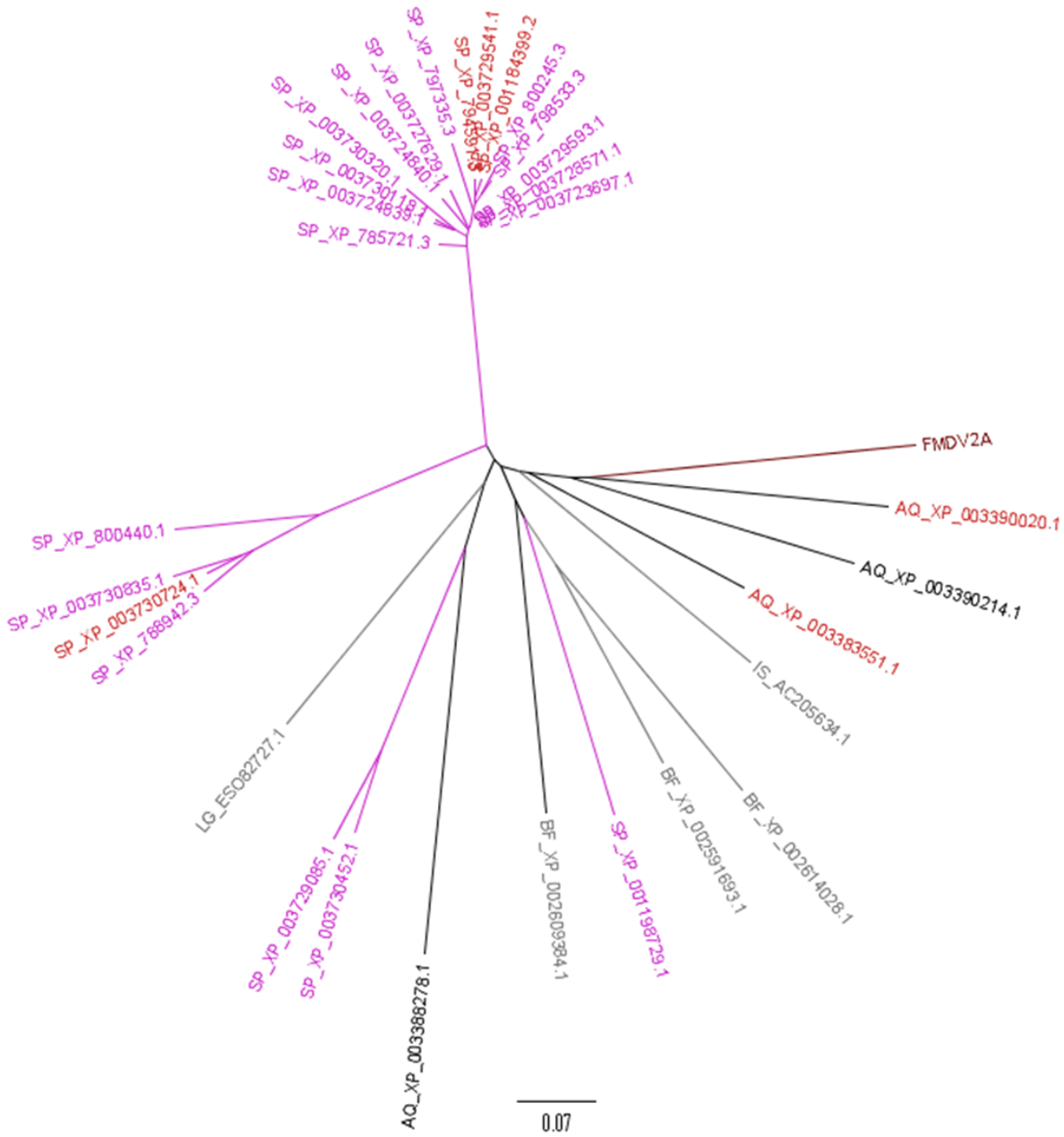


Figure 6.6 Cladogram analysis of 2A sequences from NLR proteins
Purple branches denote NLR protein 2As from *S. purpuratus,* whilst black denotes NLR protein 2As from *A. queenslandica,* while 2A sequences from *B. floridae, I. scapularis*, and *L. gigantea* are shown in grey. *FMDV 2A* (dark red) has been included for comparison purposes. NLR 2As tested *in vitro* and found to possess comparable ribosome skipping abilities to *FMDV 2A* are coloured in red. <30 amino acid 2As aligned with ClustalX2 and unrooted tree diagram drawn with Figtree v1.4.2 using the default nearest neighbour joining algorithm.

6.4 **NLR 2As - Discussion**

6.4.1 Role of 2A in NLR Proteins

This is the first report of active 2A sequences from proteins that play a role in eukaryotic innate immunity. The conserved positioning of the 2A sequence upstream and in-frame with the active NLR domains, and generally at the protein N-termini is indicative of a common function for 2A in these proteins. The 2A-containing NLR proteins were found to be similar to human NLRC3 and NLRC5. The exact number of mature NLR proteins possessed by any one species is at present unknown due to the modular nature of these proteins, and alternative gene splicing forming numerous isoforms that either lack one of the active domains or possess active domain duplications. It is thought that possession of a multiple NLR isoforms can aid in responding to a range of microbial threats. Incorporation of 2As into multi-domain NLRs therefore provides another way of increasing isoform number. The NLR 2As tested were efficient at instigating translational recoding, but did not mediate ribosome skipping in 100% of instances, therefore could generate greater isoform diversity *in vivo*.

Generally, the 2A occurred at, or near, the protein N-termini in a similar manner to that seen for non-LTR, ankyrin-repeat, and SNAT9 amino acid transporter 2A-containing proteins. My colleagues and I have previously suggested that the N-terminal 2As in non-LTRs might play a translational regulatory role, similar to their proposed function in some viruses (Odon *et al.*, 2013). Alternatively, perhaps the 2A may have acquired a secondary role in protein targeting to specific subcellular or extracellular destinations (this idea will be explored further in Chapter 7).

The functional role and the interacting partners of the 2A-NLR proteins have yet to be elucidated. The reporting of high levels of NLR expression in the sea-urchin gut tissues (Hibino *et al.*, 2006) is indicative of sea-urchin NLRs playing a crucial role in immune defence in this organism, as sea-urchins are grazers with a simple “leaky” gut that provides the main entry point for pathogens. However, it is not currently known if, similar to NLRC3, their role is in down-regulation, or if like NLRC5 their role is to up-regulate the innate immune response, or whether (as seems highly probable) the diversity of sea-urchin NLR proteins contains both positive and negative modulators of the innate immune response.

6.4.2 NLR 2A Phylogeny

Only one of the NLR proteins from each of *L. gigantea* and *I. scapularis*, and three instances from both *B. floridae* and *A. queenslandica*, possessed 2A peptides - thus suggestive of a recent acquisition of 2A in NLR proteins in these organisms. In contrast, there were 18 instances of 2As in NLRs from *S. purpuratus*, plus a further four proteins containing 2A-Death but lacking an

NTPase (NATCH) domain. One clade of *S. purpuratus* 2A-NLR proteins appeared monophyletic, but the relationship of the remainder of the 2A-NLR proteins remains uncertain.

Notably *S. purpuratus*, *B. floridae*, and *L. gigantea* also possessed 2As in association with non-LTR sequences (reported in Chapter 3). There are no current reports of non-LTR 2As from *I. scapularis*, however its genome sequencing is still ongoing, and the author has identified 2As in association with unclassified non-LTRs in the closely related tick species *Rhipicephalus pulchellus* (listed in Chapter 3). Active non-LTRs spread throughout the genome through self-duplication and re-integration, particularly at sites of DNA damage/repair. In sea-urchin, NLR gene duplication and exon shuffling is thought to have contributed to the increased repertoire of NLR proteins (Hibino *et al.*, 2006). Consequently, if NLR gene structure is such that it is a hotspot of gene reorganisation, it is conceivable that the 2A sequence could have been acquired through "capture" from a truncated non-LTR transcript. No non-LTRs with 2A have been reported for *A. queenslandica*, but this sponge possesses modular ankyrin-repeats with N-terminal 2As, therefore it is possible that the sponge NLRs could have acquired 2As from ankyrin gene re-arrangement (or *vice versa*).

How and where and when the NLR-associated 2A peptides originated remains unknown. Again, as with the other protein classes with 2As, the most likely transmission agency remains the same unidentified horizontal gene transfer vector responsible for transmitting non-LTR 2As between phyla. However, the identity of this hypothetical vector and of the ancestral host where these 2A sequences first evolved are unknown at present.

Chapter 7. A Dual Role for 2As as Signal Peptides?

'I have a right to be blind sometimes… ...I really do not see the signal… …Damn the signal! Keep mine for closer battle flying! That's the way I answer signals! Nail mine to the mast!'

-Attributed to Nelson at Copenhagen, in *The Life of Horatio Lord Nelson* – Robert Southey, 1813

7.1 Introduction

Two general observations were made regarding the newly discovered eukaryotic 2As. Firstly, the 2A tended to be an N-terminal feature of proteins; secondly, many of the 2A sequences possessed a leucine-rich hydrophobic tract. An online protein database probe and concurrent literature search raised the possibility that as these two tendencies, namely N-terminal and hydrophobic are also typical properties of signal peptides, therefore, some of the newly discovered 2As might, under certain conditions, play a dual role as signal peptides. The properties of typical signal peptides will now be described before the presentation of a report of the author's investigations concerning the signal properties of selected eukaryotic N-terminal 2As in an *in vitro* system.

7.1.1 Signal Peptides – ER/Transmembrane Trafficking

The majority of secreted, and many membrane-bound, proteins possess cleavable N-terminal signal sequences "signal peptides" that facilitate their targeting to and trafficking across the ER (in eukaryotes) or the cytoplasmic cell membrane (in prokaryotes). Signal peptides typically start within 10 amino acids of the protein N-terminus and are between 20-30 amino acids in length. They consist of a core 6-15 residue hydrophobic "H-domain" with a marked preference for leucines or alanines. This core is flanked on the N-terminal side by positively charged polar residues (the "N-domain") and on the C-terminal side by neutral, but polar residues ("C-domain"). Signal peptides are remarkably tolerant of individual residue substitutions as long as a central hydrophobic tract is maintained (although hydrophobicity alone is not sufficient for signal function) (reviewed in Stroud and Walter, 1999; and Hegde and Bernstein, 2006). Signal peptides are diverse, lacking any sequence homology apart from their conserved three domain structure (N, H and C-domains), but the signal sequence recognition mechanisms are evolutionary conserved, as individual sequences have been shown to maintain their function in evolutionary distant organisms. For example, a bacterial sequence was found to be active in mammalian cells (Talmadge *et al.*, 1980) and *vice versa* (Muller *et al.*, 1982).

The "signal hypothesis" that a specific N-terminal tract was responsible for the translocation of proteins across subcellular membranes was first proposed in the 1970s (Blobel and Dobberstein, 1975). It has been subsequently shown that this N-terminal signal mediates co-translational protein trafficking by binding to a signal recognition particle (SRP) a ribonucleoprotein complex that binds to the signal peptide sequence as it exits the ribosome tunnel during translation. The signal-SRP binding slows translation and targets the complex to the ER/cell membrane through docking with

the SRP receptor. The nascent protein is then translocated across the ER membrane by means of the Sec61p (in eukaryotes) or SecY (in prokaryotes) translocon pore complex before a membrane-bound protease removes the signal from the mature translocated protein. Chaperone binding to the N-terminal signal can also facilitate translocation (Figure 7.1). In addition to proteins targeted through the ER membrane into the ER lumen and subsequently to the exocytic pathway, transmembrane proteins are trafficked in a similar manner – often the first transmembrane domain acts as the SRP/chaperone signal, but in these instances the signal is not cleaved off after docking with the membrane as it forms part of the mature protein (reviewed in Stroud and Walter, 1999; and Hegde and Bernstein, 2006).

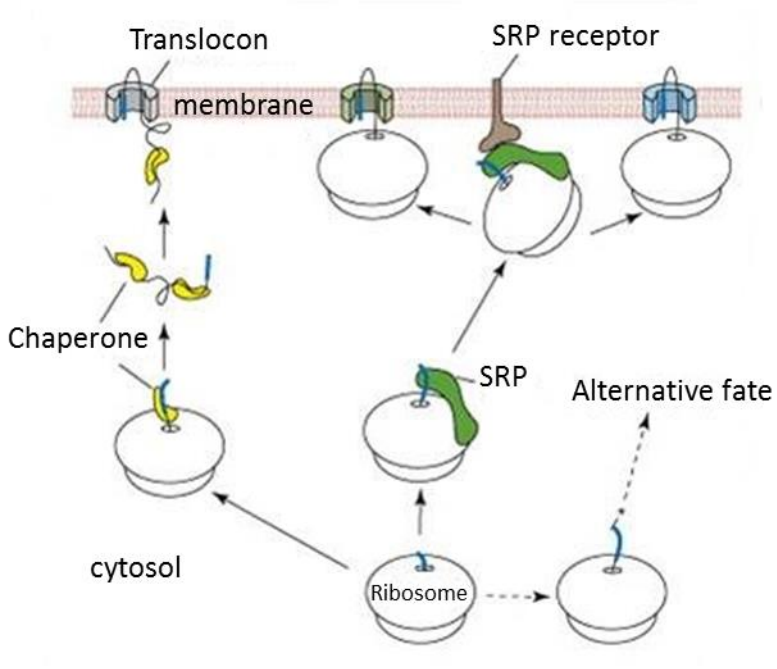


Figure 7.1 Signal sequence mediated protein trafficking
A nascent peptide (black) containing a signal sequence (blue) emerges from the ribosome. The signal can interact with SRP (green) or chaperones (yellow); or modifying enzymes (not shown). After proteins arrive at the ER (or cytoplasmic) membrane, the specific signal sequence can also influence their delivery to various translocon complexes (pore shown in different colours on diagram). The diagram represents a composite from several experimental systems with the intention of conveying the concepts of signal peptide directed targeting rather than to depict the specific pathways from any one organism. Diagram modified from Hegde and Bernstein, 2006.

Enzymatic modification of the signal sequence can interfere with the trafficking pathway and alter the destination of the mature protein. For example, in mammalian cells myristoylation of the N-domain of the NADH cytochrome b_5 reductase inhibits its recognition by SRP (Colombo *et al.*, 2005) and the protein is instead trafficked to the mitochondrial outer membrane. There is preliminary data to suggest that signal sequence phosphorylation may similarly regulate SRP binding (Anandatheerthavarada *et al.*, 1999).

SRP-binding alone is not sufficient for successful translocation. Recognition by the translocon requires the interaction of accessory factors (in mammals the TRAM and/or the TRAP complex) however the details of the interaction have yet to be elucidated (reviewed by Hegde and Bernstein, 2006 and more recently by Denks *et al.*, 2014). One translocated, the signal sequence is generally cleaved from the mature peptide by membrane-bound signal peptidase, but this event does not necessarily mark the end of the signal peptide playing a biological role. For example, human HLA class I histocompatibility antigen, alpha chain E (HLA-E) also known as MHC class I antigen E presents a nine-residue peptide on the cell surface that is derived from a conserved sequence found in the signal sequences of several major histocompatibility complex (MHC) class I proteins. This acts as an inhibitory signal to natural killer cells. Hence, down-regulation of MHC class I expression during viral infection results in a decrease in HLA-E signal peptide-derived antigen expression and a consequent increase in destruction by natural killer cells. In a remarkable twist, *Human Cytomegalovirus* has evolved a glycoprotein with a signal sequence that contains the same nine-residue sequence as the MHC class I signal. This allows the virus to avoid detection by the natural killer cell mediated immune pathway, while concurrently down-regulating the MHC mediated response (reviewed in Hegde and Bernstein, 2006).

Breakdown or inefficiency in the signal sequence-SRP trafficking pathway may have pathological consequences, one example being prion protein (PrP), a cell-membrane glycoprotein, where alterations in trafficking are implicated in a range of neurodegenerative conditions. PrP is targeted to the ER, but a small proportion (between 5-15%) is not translocated entirely and remains either as a transmembrane (^{Ctm}PrP) or a cytosolic form (cyPrP). Increased quantities of both these forms have shown to contribute to neurodegenerative pathologies. Biochemical investigations indicate that the increase of the pathogenic forms of PrP is due to a slight, but measurable, increase in the inefficiency of SRP binding to the signal sequence. Indeed, replacement of the native PrP signal sequence with a more efficient version was found to enhance the survival of cell cultures that overexpressed PrP *in vitro* (Rane *et al.*, 2004).

7.1.2 Signal Peptides – Intracellular Targeting

In eukaryotic cells, signal peptides are also used to target proteins to intracellular destinations such as the nucleus, mitochondria or chloroplasts. In the case of nuclear import, the protein must possess a nuclear localisation signal (NLS); this must consist of one or two tracts of exposed positively charged lysine or arginine residues. These motifs are recognised by the importin pathway protein, importin α, which acts to chaperone the protein to the nuclear membrane and mediate its interaction with importin β at a nuclear import complex. Nuclear import is thought to be highly conserved between phyla (reviewed in Dingwall and Laskey, 1991). Mitochondrial trafficking is also highly conserved. Mitochondrial and (in plants) chloroplast trafficking is post-translationally

mediated by an N-terminal signal sequence that is recognised by any one of a variety of cytosolic chaperone proteins. Chaperone binding halts protein folding and maintains the nascent protein in an unstructured state in which it can be passed through first the outer membrane complex (in plants the TOM complex, an analogous complex has been identified from yeast) then traverse the inner membrane complex (TIM). The process is highly ATP-dependent. Once inside the mitochondria the signal sequence is cleaved by a specific protease. In plants, the mitochondrial signal sequence is typically around 30 amino acids in length; however, sequences of over 100 amino acids have been reported from yeast. There is little conservation of sequence homology between mitochondrial signals, however all are rich in basic and hydrophobic residues, especially arginine, alanine, leucine and serine, and their N-terminal region is predicted to fold into an amphiphilic a-helix This is thought to both facilitate chaperone binding, and to aid in threading the protein through the membrane into the organelle. In plants, the mitochondrial signal may also function as a chloroplast targeting (reviewed in Duby and Boutry, 2002).

7.1.3 Signal Sequence Prediction

The generalities in signal sequence composition, positioning within proteins, and potential to form secondary structure have led to the creation of a number of computer algorithms which if presented with a protein query claim to be able to recognise any signal peptides and predict signal peptidase cleavage sites. These programs depend heavily on the early pioneering investigations of von Heijne and colleagues (for examples see von Heijne, 1986; Nielsen *et al.*, 1997; Nielsen *et al.*, 1999). This group has produced and maintained SignalP, a successful predictor of extracellular and transmembrane proteins which predominately analyses hydrophobicity and the presence/absence of signal peptidase cleavage sites in order to report the probability of a given sequence possessing an N-terminal signal peptide. SignalP is now on its fourth release version (Petersen *et al.*, 2011) and is readily available online (http://www.cbs.dtu.dk/services/SignalP/). An alternate program PSORT, identifies potential signals based on their similarity (both residue sequence and likely topology) to known signal sequences. It is not as accurate as SignalP, but possesses the ability to identify not only ER and transmembrane proteins but also those targeted to a number of different intracellular locations including the mitochondria, chloroplasts and nucleus (Nakai and Horton, 1999). PSORT II can be accessed at http://psort.hgc.jp/cgi-bin/runpsort.pl. Refinement of the PSORT algorithm led to the creation of WoLF PSORT. This program was shown to be more accurate than PSORT II (Horton *et al.*, 2007). It was hosted at http://wolfpsort.org/, however, that website is currently offline as the algorithm was acquired by Genscript in 2013. It is hosted at http://www.genscript.com/psort/wolf_psort.html, but the webpage is frequently unavailable and the results page often reports only the outcome of a PSORT II analysis not of WoLF PSORT. As each of these programs (SignalP, PSORT, WoLF PSORT) possessed unique inherent benefits, and limitations on their use, all three were employed in the analyses reported in this Chapter.

7.2 **Methodology - Overview**

The eukaryotic 2A sequences that potentially functioned as signal peptides were determined by use of SignalP, PSORT II and WoLF PSORT as detailed in Chapter 2.1.4. A selection of promising candidates were cloned into reporter plasmids based on *pJN123* or *pEMX*, (details in Chapter 2.1.1, Appendix A, and to follow). Several reporter plasmids were trialled in order to ascertain the most efficient reporter assay system for future investigations. Plasmid sequences are given in Appendix A. Cloning PCR primers are listed in Table 7.1. Intracellular localisation of transfected proteins was visualised by microscopic analyses as described in Chapter 2.2.10. Where appropriate, co-transfection with plasmid *pSTR6-GFP* or *pMITO* was used to mark the cytoplasm or mitochondria, respectively, as were transfections using the *CytoLight*™ stable cell lines (Chapter 2.2.6) to visualise either the cytoplasm or nucleus. Counter-staining with DAPI was also used for nuclear visualisation (Chapter 2.2.10). Transfected protein levels were verified by Western Blotting as detailed in Chapter 2.2.8. Brefeldin A was employed to ascertain Golgi localisation (Chapter 2.2.10). 2A translational recoding was determined by *in vitro* cell-free coupled transcription-translation (TnT) analyses as described in Chapter 2.2.2 and Odon *et al.*, 2013. Sequences were confirmed by DNA sequencing using specific primers as listed in Table 2.1., prior to *in vitro* and/or *in vivo* analyses.

Table 7.1 PCR primers used in the work reported in Chapter 7.

Primer	Primer Sequence (5' to 3')	Notes
11Rmut F	GCTCCTGATCCTCTTGATGGCATCTGGTGACGTTGAAACC	Mutagenesis primers targeting *STR6* in *pJN132*
11Rmut R	GGTTTCAACGTCACCAGATGCCATCAAGAGGATCAGGAGC	
12Mmut F	GCTCCTGATCCTCTTGGTGAGATCTGGTGACGT	
12Mmut R	ACGTCACCAGATCTCACCAAGAGGATCAGGAGC	
13Lmut F	CTCTATCTGCTCCTGATCCTCGNNATGAGATCTGGTGACGTTGAA	
13Lmut R	TTCAACGTCACCAGATCTCATNNCGAGGATCAGGAGCAGATAGAG	
14Lmut F	CTTCTCTATCTGCTCCTGATCGNNTTGATGAGATCTGGTGACGTT	
14Lmut R	AACGTCACCAGATCTCATCAANNCGATCAGGAGCAGATAGAGAAG	
16Lmut F	ATTCTGTCTTCTCTATCTGCTCGNNATCCTCTTGATGAGATCTGGTG	
16Lmut R	ATTCTGTCTTCTCTATCTGCTCGNNATCCTCTTGATGAGATCTGGTG	
21Lmut F	AGAACCATGGATGGATTCTGTGNNCTCTATCTGCTCCTGATCCTC	
21Lmut R	GAGGATCAGGAGCAGATAGAGNNCACAGAATCCATCCATGGTTCT	
25Dmut F	CTAGCTCTAGAACCATGGCTGGATTCTGTCTTCTCTA	
25Dmut R	TAGAGAAGACAGAATCCAGCCATGGTTCTAGAGCTAG	
CtoG F	TGGTGACGTTGAAACCAATGCCGGGCCCAT	Mutagenesis primer converting $STR6^{wt}$ to $STR6^{NAPG}$
CtoG R	ATGGGCCCGGCATTGGTTTCAACGTCACCA	
GtoC F	TGGTGACGTTGAAACCAATCCCGGGCCCAT	Mutagenesis primer converting $STR6^{NAPG}$ to $STR6^{wt}$
GtoC R	ATGGGCCCGGGATTGGTTTCAACGTCACCA	
pEMmut F	GTTTAAACTTAAGCTTGGTACCGAGCTCTCTAGAATGACTAATGCCCTTCTATTGAGATC	to insert XbaI site upstream of 2A in *pEM* thuds creating *pEMX*
pEMmut R	GATCTCAATAGAAGGGCATTAGTCATTCTAGAGAGCTCGGTACCAAGCTTAAGTTTAAAC	
AQ27extract F	GAGCTCTCTAGAATGGTCTCAGTCGTGTTC	to amplify *AQ27* before cloning into *pEMX*
AQ27extract R	ATCGATGGGCCCGGGATTGATC	
STR6 F	GAGCTCTCTAGAATGGATGGATTCTGTC	to amplify *STR6* before cloning into *pEMX*
STR6 R	ATCGATGGGCCCAGGACCTGGA	
SS7 F	GAGCTCTCTAGAATGCAACGATCACGTC	to amplify *SS7* before cloning into *pEMX*
SS7 R	ATCGATGGGCCCAGGTCCTGGA	
AQ27 P3A F	GCGACATCGAGATCAATGCCGGGCCCA	To make NAGP mutants in *pEMX*
AQ27 P3A R	TGGGCCCGGCATTGATCTCGATGTCGC	
SS7 P3A F	GGAGACGTTGAAGTCAATGCAGGACCTGGGCC	
SS7 P3A R	GGCCCAGGTCCTGCATTGACTTCAACGTCTCC	
STR6 P3A F	CTGGTGACGTTGAAACCAATGCAGGTCCTGGG	
STR6 P3A R	CCCAGGACCTGCATTGGTTTCAACGTCACCAG	
Mitosig R	TAAGATGGGCCCAGTTGCTACAGGTGGATCACCTAATGAATGTATCTTAGCTCTTGGTACTGGTAATCGTCTTGCTGAACCTGTAAATCCTCTTAATAGAAGTGGAGTTAATACTGACATTCTAGACCCGGAC	Reverse primer to insert mitochondrial signal sequence in *pSTR6-GFP* to create *pMito*

7.2.1 Contributors

All experiments were undertaken by the author excepting the initial *STR6* mutagenesis which was undertaken by three Honours Students under the close supervision of the author. The subsequent plant transfections were undertaken by Jonathon Cope supervised by Dr. Jens Tilsner. John Nicholson and Ashley Pearson assisted with the Deltavision microscopy. Plasmids *pEM* and *p132* were a gift from Dr. Garry Luke and *pJN1* was a gift from John Nicholson. Dr. Catherine Botting undertook the mass spectrometric analysis from a gel-slice prepared by the author.

7.2.2 *STR6* Mutagenesis Investigations

These experiments relied on the co-expression of [putative_signal_2A-mCherryFP] and eGFP from a signal plasmid vector, namely *pJN132*. This vector was created through ligation of *p132* and *pJN1* (Figure 7.2). A series of *STR6* mutants in *pJN132* were created through site-directed mutagenesis PCR with mixed base primers (to potentially obtain multiple mutants from each round of mutagenesis) as described in Chapter 2.2.1.4.1. Selected clones were subjected to *in vitro* and *in vivo* investigation. The *pJN132* clones were ApaI digested (to remove mCherryFP-TaV) and subsequently re-ligated to create *STR6-GFP* constructs (*pSTR6*-GFP, see Figure 7.3) that were used for *in vitro* cell-free coupled transcription-translation (TnT) assays to evaluate the relative recoding abilities of the *STR6* 2A sequences.

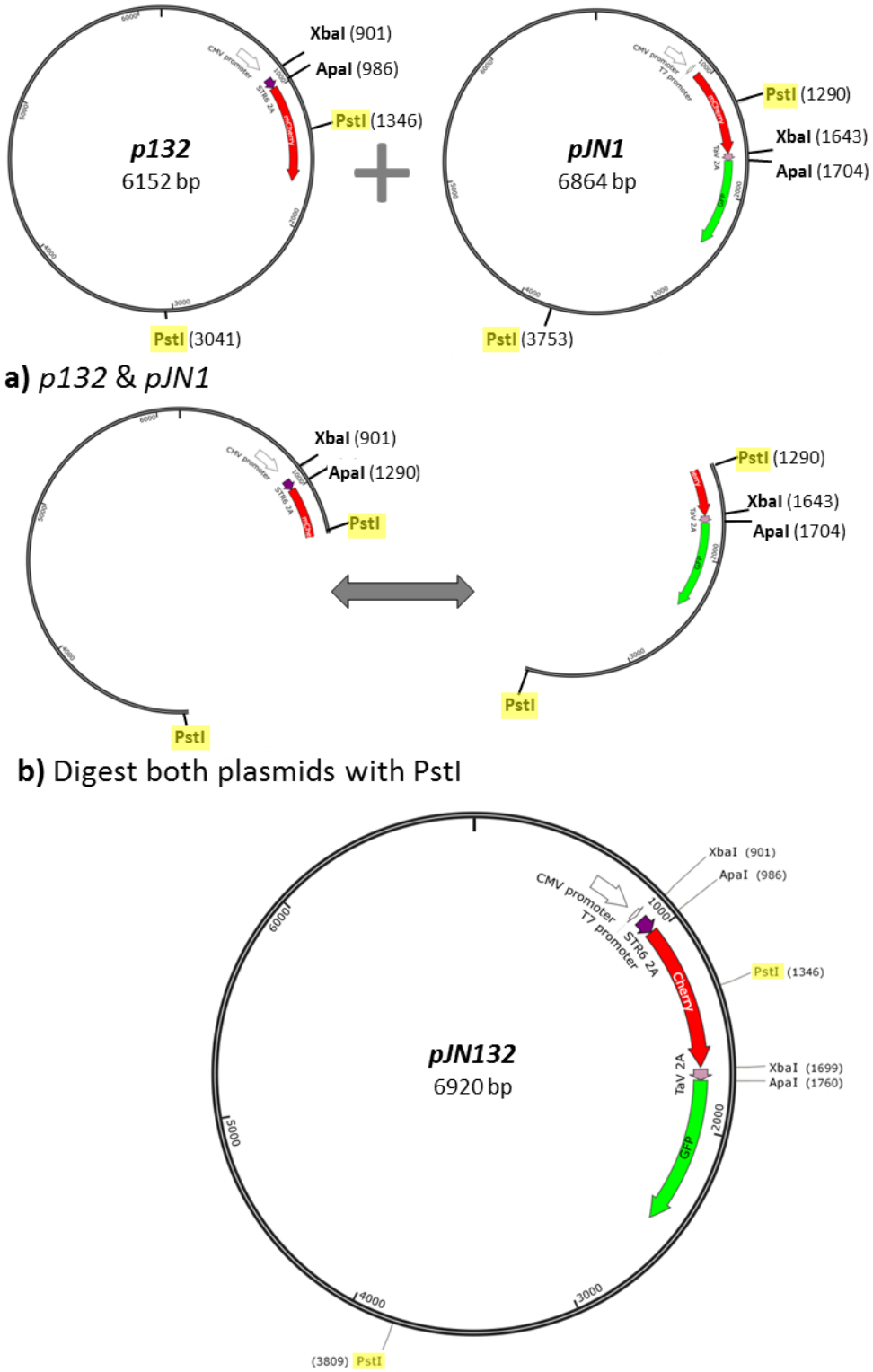


Figure 7.2 Creation of *pJN132*
Plasmid *pJN132* encodes STR6-mCherryFP-TaV2A-eGFP in a *pcDNA* backbone. It was created through PstI (sites marked in yellow) restriction digest of *p132* and *pJN1* and subsequent ligation. Dark purple arrows denote putative signal 2As, light purple *TaV 2A*, white promotor regions, red mCherryFP, and green eGFP.

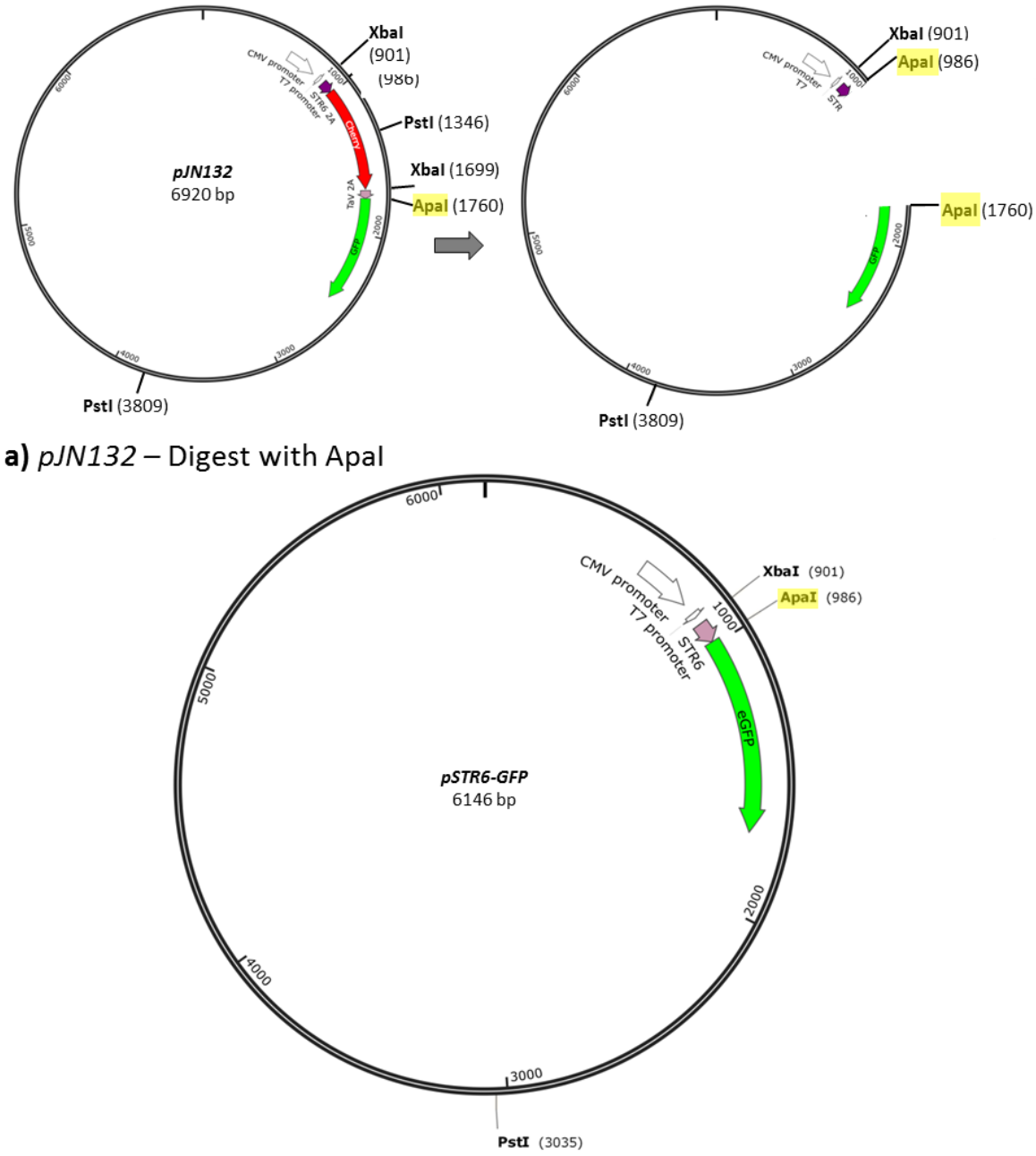


Figure 7.3 Creation of *pSTR6-GFP*
Plasmid *pSTR6-GFP* used for *in vitro* translational recoding analyses (TnTs) on the *STR6-mutant* clones. *pSTR6-GFP* was created through ApaI (sites marked in yellow) restriction deletion to *pJN132*. Dark purple arrows denote putative signal 2As, light purple *TaV 2A*, white promotor regions, red mCherryFP, and green eGFP.

7.2.3 *STR6* Transfections *In Vivo*

STR6^{wt} and *STR6*^{NAGP} constructs in *pJN132* or *pTRBO* were transfected into echinoderm (green sea-urchin, *Psammechinus miliarsis*) or inoculated into plant leaves (tobacco, *Nicotiana benthamiana*) as detailed in Chapter 2.2.12 and 2.2.13, respectively.

7.2.4 *AQ27*^{NAGP} & *SS7*^{NAGP} – Signals

The putative signal 2As *AQ27*^{NAGP} (sponge, *Amphimedon queenslandica*, ankyrin-repeat protein, discussed in Chapter 4) and *SS7*^{NAGP} (from a salmon, *Salmo salar*, putative transposon, *SS7_SS* reported in Chapter 3) were cloned into *pEMX* (signal2A-mCherryFP in *pcDNA3.1* backbone, full sequence provided in Appendix A).

pEMX was created through mutagenesis PCR (Chapter 2.2.1.4) on *pEM* (a kind gift from Dr. Garry Luke) to add an XbaI site immediately upstream of the 2A. Putative signal 2As were amplified from *pSTA1* through PCR with specific primers. PCR product ends were prepared through double XbaI/ApaI digestion and ligated into *pEMX* similarly restricted. Ribosome skipping inactive mutants (NAGP mutants) were created through mutagenesis PCR Constructs in *pEMX* were transfected into *CytoLight*™ stable HeLa cells or co-transfected with *pMITO* to provide eGFP labelled cytoplasm or mitochondria respectively. *pMITO* was created by inserting (by means of PCR) the human cytochrome c oxidase subunit mitochondrial signal sequence (as used in the commercial Clontech DSRed plasmid) upstream of eGFP to replace *STR6* in *pSTR6-GFP*. Plasmids were transfected into HeLa cell monolayers and visualised as described in Chapter 2.2.4 and 2.2.10.

7.2.5 Amino Acid Transporter SNAT9 2As – Signals

Amino acid transpoter SNAT9 2As (described in Chapter 5.) were cloned into *pSTR6-GFP* (replacing *STR6*) concurrent with their *pSTA1* cloning (by XbaI/ApaI digest of the PCR product to release the 2A) for the analyses reported in Chapter 5. These SNAT9 2As were then inserted into *pJN132* (SNAT9_2A-mCherryFP-TaV2A-eGFP) by repair of their *pSTR6-GFP* vector to form *pJN132* through ApaI restriction and subsequent re-ligation. Briefly, the ApaI-bound fragment (mCherryFP-TaV2A) digested from *pJN132* and ligated into the *pSTR6-GFP*-based vectors which had been similarly restricted by ApaI digest. This created *pJN132*-based constructs (SNAT9_2A-mCherryFP-TaV2A-eGFP). Plasmids were transfected into HeLa cell monolayers and visualised as described in Chapters 2.2.4 and 2.2.10.

7.3 **Analyses of Dual Function Signal Peptide 2As**

All sequences in the in-house newly identified eukaryotic 2A database were screened with SignalP, as was the in-house list of viral 2As. No viral 2As were detected as potential signals. The SignalP combined score (D-value) for each eukaryotic sequence is reported in Appendix B if greater than 0.350, a Signal-P D-value greater than 0.450 indicates a strong probability that the protein was targeted to the exocytic pathway. A selection of N-terminal 2A sequences with high Signal-P values were subjected to further *in vitro* analyses, the results of which are reported in this Chapter.

7.3.1 Identification of the Signal Properties of *STR6* and Related Sequences

A group of closely related sea-urchin NLR-protein N-terminal 2A sequences as typified by the 2A sequence *STR6* were identified as putative extracellular pathway signals (as shown by SignalP analyses, see

Table **7.2** and Figure 7.4). Wild-type *STR6* (*STR6*^{wt}) was previously shown to be active in instigating ribosome skipping *in vitro* (see Chapter 8, Figure 6.2) Therefore, if, in the majority of instances *STR6* was actively orchestrating ribosome skipping and so "cleaving" itself from the downstream protein; then, consequently, it would be unable to function as an attached N-terminal trafficking signal. To discover whether *STR6*^{wt} (and related 2As) could (if "uncleaved") act as signals, an artificial mutant version, *STR6*^{NAGP}, was created by mutating P3A (numbering counting back towards the 2A N-terminus, where the active 2A C-terminal proline equated to P1). This single alanine substitution had previously been shown to abrogate ribosome skipping (reviewed in Luke *et al.*, 2008; Luke and Ryan, 2013). However, the P3A alanine substitution also had the unexpected side-effect of lowering the Signal-P D-value (from a relatively high value of D=0.542 to D=0.475) albeit above the threshold for potentially possessing signal properties (Figure 7.4).

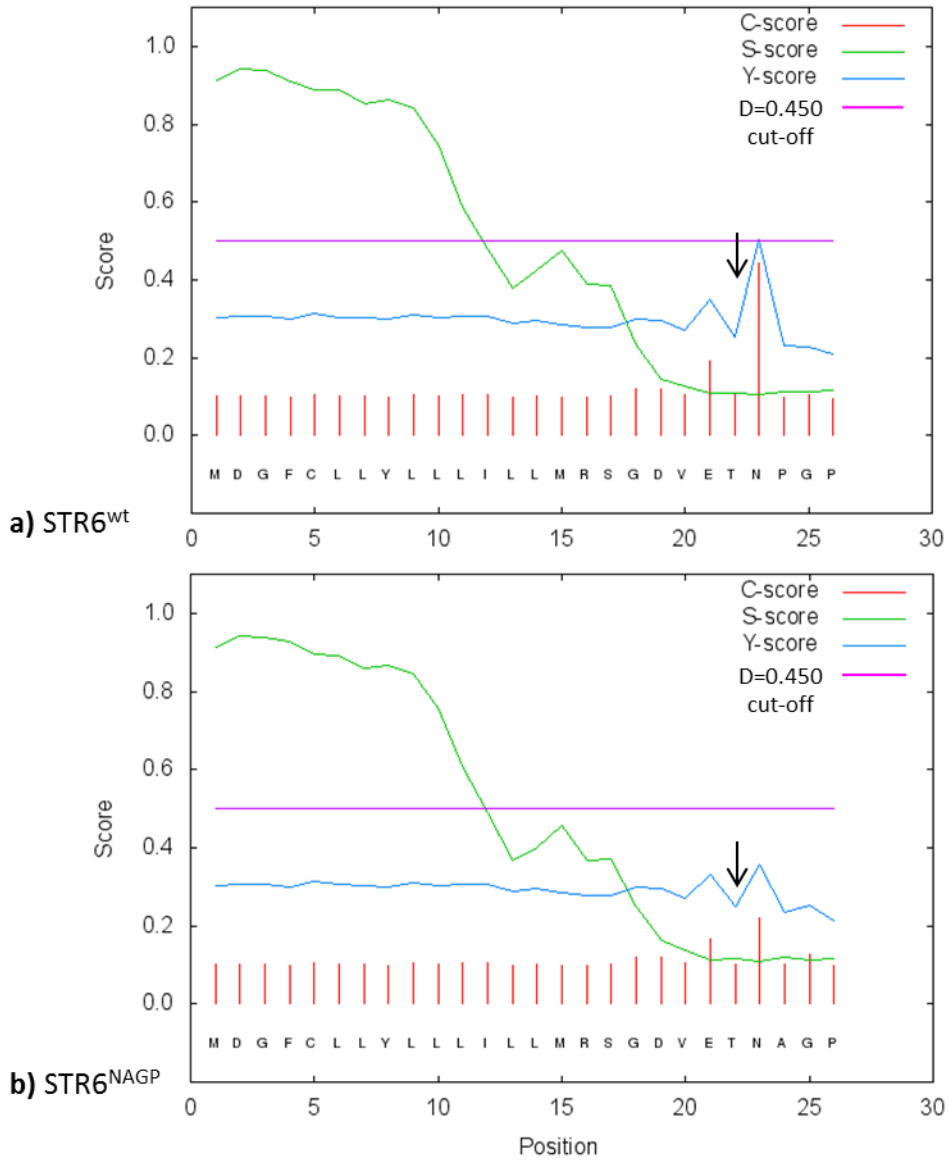


Figure 7.4 STR6^{wt} & STR6^{NAGP} Signal-P analyses

Signal-P plots, purple line corresponds to the D-value (combined CSY score) 0.450 cut-off. Plots above this value indicate potential signal sequences. Both *STR6* and *STR6*^{NAGP} scored highly as potential exocytic pathway signals, especially note the high S-value (signal potential) of the N-terminal portions of the sequences, and the signal peptidase cleavage site (marked by black arrows) within the conserved C-terminal motif (at DVET$^{\downarrow}$NPGP or DVET$^{\downarrow}$NAGP). The arrows indicate the signal peptidase cleavage sites. Full details of the analyses follow:

a) STR6^{wt}: Max C-value 0.444 at position 23, max Y-value =0.505 also at position 23, max S-value 0.942 at position 2, mean S-value positions 1-22 =0.574, D-value positions 1-22 = 0.542, signal peptidase cleavage site between positions 22-23, no transmembrane region detected.

b) STR6^{NAGP}: Max C-value 0.219 at position 23, max Y-value = 0.355 also at position 23, max S-value 0.942 at position 2, mean S-value positions 1-22 =0.576, D-value positions 1-22 = 0.475, signal peptidase cleavage site between positions 22-23, no transmembrane region detected.

Table 7.2 List of high scoring potential exocytic pathway signal NLR-2As
First the naturally occurring NLR-2A sequences are listed, ordered by Signal-P D-score, then the list of artificial mutants used for *in vitro* investigations. Grey highlighting indicates residues differing from STR6^{wt} chosen as the reference NLR-2A signal 2A peptide. Sequences in **bold text** represent naturally occurring similar sequences to STR6, and their respective NAGP mutants selected for *in vitro* analyses. Note the change in D-values for the NAGP mutants in comparison with the wild-type sequences ending in –NPGP-. Numbering of mutants (for example M12V) corresponds to the amino acid position reading towards the N-terminus from the 2A C-terminal proline (P).

Sequence	Example Accession no.	Amino Sequence	SignalP score
		Naturally Occurring Sequence Variants	
SP	XP_785721.3	MDGFCLLYLLMILLVRSGDVETNPGP	D = 0.557
SP	XP_797335.3	MDGLCLLYLLLILLMRSGDVETNPGP	D = 0.557
STR6^{wt}	**XP_001184399.2**	**MDGFCLLYLLLILLMRSGDVETNPGP**	**D = 0.542**
SP	XP_003724839.1	MVGFCLLYLLMILLMRSGDVETNPGP	D = 0.534
SP	XP_003730118.1	MAGFCLLYLLMILLMRSGDVETNPGP	D = 0.531
STR6-M12V	**XP_800245.3**	**MDGFCLLYLLLILLVRSGDVETNPGP**	**D = 0.528**
SP	XP_003727629.1	MDAFCLLYLLMILLMRSGDVETNPGP	D = 0.483
STR6-L16M	**XP_003728571.1**	**MDGFCLLYLLMILLMRSGDVETNPGP**	**D = 0.473**
SP	XP_003730320.1	MGGFCHLYLLMILLMRSGDVETNPGP	D = 0.444
SP	XP_003724840.1	MDGFCLLYLIMILLMRSGDVETNPGP	D = 0.438
		Artificial mutants	
STR6-R11A^{NAPG}		MDGFCLLYLLLILLMASGDVETNAGP	D = 0.627
STR6-L13V^{NAPG}		MDGFCLLYLLLILVMRSGDVETNAGP	D = 0.608
STR6-M12V^{NAPG}		**MDGFCLLYLLLILLVRSGDVETNAGP**	**D = 0.596**
STR6-L13A^{NAPG}		MDGFCLLYLLLILAMRSGDVETNAGP	D = 0.596
STR6-L16A^{NAPG}		MDGFCLLYLLAILLMRSGDVETNAGP	D = 0.584
STR6-D25A^{NAPG}		MAGFCLLYLLLILLMRSGDVETNAGP	D = 0.522
STR6^{NAPG}		**MDGFCLLYLLLILLMRSGDVETNAGP**	**D = 0.475**
STR6-L21V^{NAPG}		MDGFCVLYLLLILLMRSGDVETNAGP	D = 0.450
STR6-L21A^{NAPG}		MDGFCALYLLLILLMRSGDVETNAGP	D = 0.435
STR6-L16M^{NAPG}		**MDGFCLLYLLMILLMRSGDVETNAGP**	**D = 0.418**
STR6-L14G		MDGFCLLYLLLIGLMRSGDVETNAGP	D = 0.406
STR6-L14D^{NAPG}		MDGFCLLYLLLIDLMRSGDVETNAGP	D = 0.387

7.3.2 *STR6* - Protein Localisation

7.3.2.1 Microscopy

STR6^{wt} and *STR6*^{NAGP} in *pJN132* vectors were transfected into HeLa cells. Transfection of this construct was expected to result in protein expression as illustrated in Figure 7.5: briefly, if functional, the upstream putative signal would direct mCherryFP to the exocytic pathway resulting in ER/Golgi mCherry localisation, whereas eGFP expression would be cytoplasmic. The transfected cultures were examined at 5, 8, 20, 24, 30, 36 and 48 hours post-transfection. The highest expression with a low background of cell-death was observed 24-30 hours post-transfection. This is in agreement with earlier work from our laboratory (Minskaia *et al.*, 2013). Therefore, 30 hours incubation was used in all subsequent experiments.

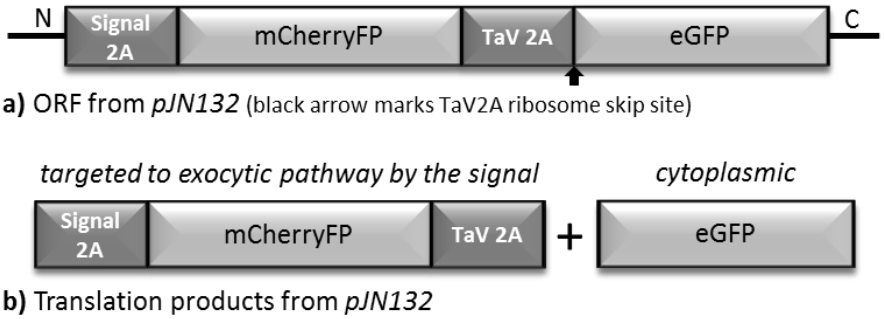


Figure 7.5 Translation products from *pJN132*
Diagram illustrates the expected translation products from *pJN132*. Note that mCherryFP is expected to be targeted to the exocytic pathway, whereas eGFP will be cytoplasmic.

There was a visible difference in the expression pattern of mCherryFP between *STR6*^{wt} and *STR6*^{NAPG} (Figure 7.6) cloned in *pJN132* constructs. In the case of *STR6*^{wt}, both mCherryFP and eGFP were observed throughout the cytoplasm, whereas for *STR6*^{NAPG} there was an accumulation of mCherryFP near/surrounding the nucleus in a pattern characteristic of Golgi stack localisation. Brefeldin A treatment confirmed the Golgi localisation of mCherryFP (Figure 7.6)

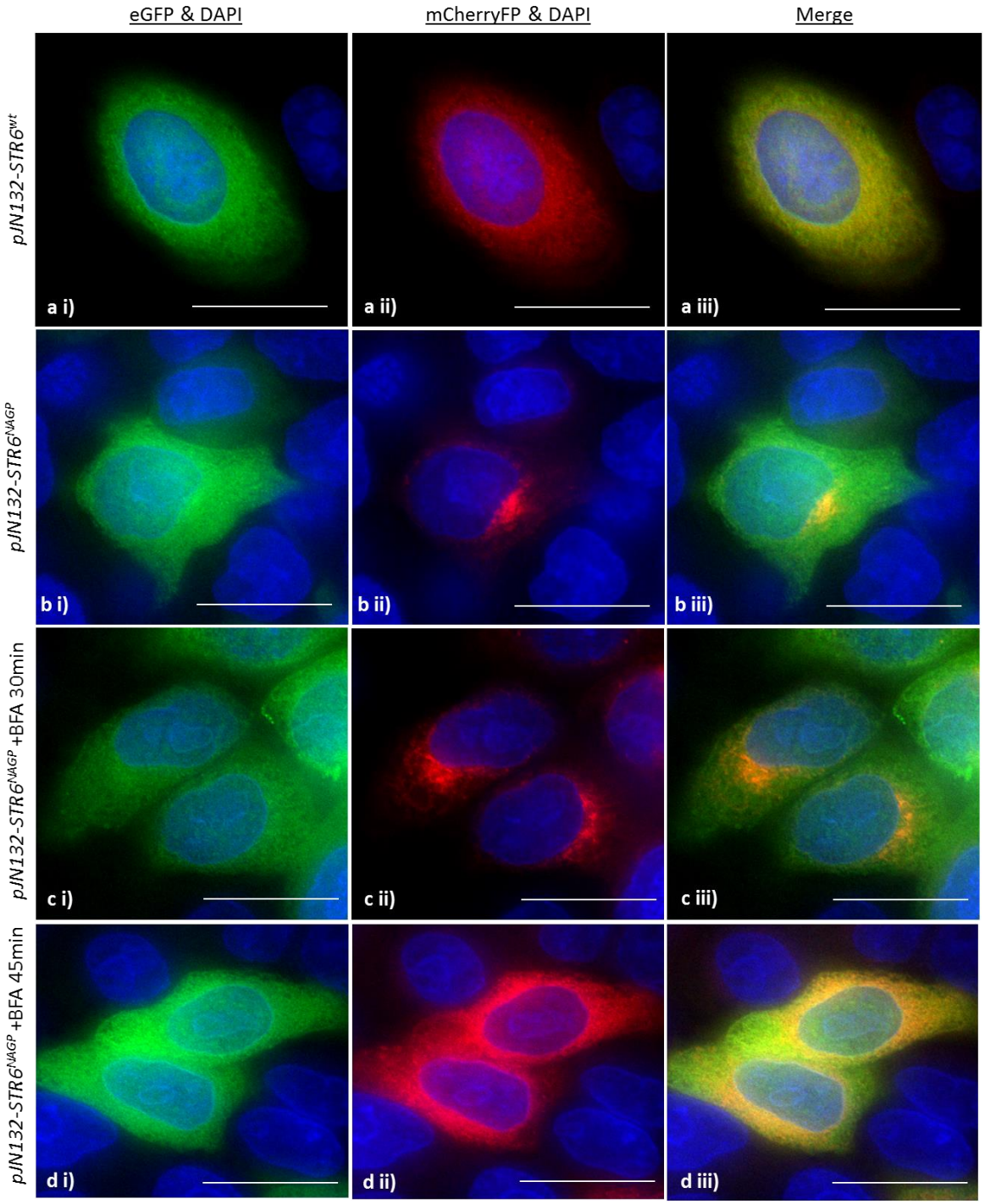


Figure 7.6 Deltavision microscopy of transfected *pJN132* constructs
a) *pJN132-STR6^{wt}* transfections, note the co-localisation of eGFP (green) and mCherryFP (red) throughout the cytoplasm, blue corresponds to nuclear DAPI staining **b)** *pJN132-STR6^{NAGP}*, note the mCherry localisation to the perinuclear/Golgi region **c)** *pJN132-STR6^{NAGP}* transfections with Brefeldin A (15 µg/mL) added 30 minutes prior to fixation, the Golgi is showing signs of disintegration **d)** *pJN132-STR6^{NAGP}* with Brefeldin A (15µg/mL) added 45 minutes prior to fixing, here the Golgi has disintegrated. All transfections fixed at 30 hours post-transfection. Scale bar=15 µm, BFA=Brefeldin A, images are single de-convolved Z-stack dual/triple channel images obtained from Deltavision microscopy (see Section 2.2.10) .

7.3.2.2 *STR6* Immuno-blotting

Western Blot analysis of transfected cell lysates revealed comparable protein expression levels for each fluorescent protein between constructs, thus confirming the different mCherryFP localisation patterns were not merely a result of transfection efficiency and/or differing transfected protein expression levels between constructs (Figure 7.7). The comparable β-tubulin levels observed throughout confirmed that the fluorescent protein localisation observed by microscopy was not a result of abnormal cell growth or cell death in the transfected cells.

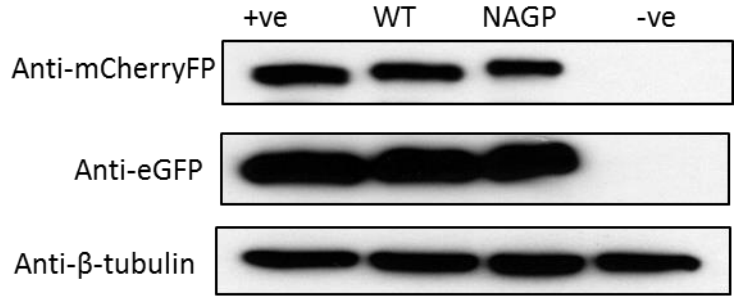


Figure 7.7 Western blot analysis of *STR6* transfections
Western blots showing comparable protein expression levels between *pJN1* (eGFP-TaV2A-mCherryFP), labelled as positive (+ve) control on blot, and *pJN132* constructs (STR6-eGFP-TaV2A-mCherryFP) encoding either $STR6^{WT}$ (marked as WT on blot) or $STR6^{NAGP}$ (NAGP on blot). The cytoskeletal protein β-tubulin was used as positive control for cell growth, and the negative control for fluorescent protein expression (labelled -ve on blot) was non-transfected HeLa cells. Pooled cell extracts from confluent monolayer HeLa cells harvested at 30 hours post-transfection (3x 30 mm dishes) using the protocols outlined in Chapters 2.2.5, 2.2.7 and 2.2.8. were loaded into each lane. The experiment was repeated 3x, the Figure shows the results of a typical replicate. The respective protein types were identified by both immuno-staining, and by molecular weight comparison of the immuno-stained protein band (using Pageruler Plus (ThermoFisher) molecular weight ladder) with the expected weights of each protein.

In order to further ascertain whether mCherryFP was indeed targeted to the exocytic pathway, an attempt was made to isolate and quantify extracellular mCherryFP from the culture media. Acetone precipitation and centrifugal separation were both trialled, but both were unsuccessful. However, the semi-quantitative method of immuno-dot blotting yielded results. Dot blots revealed that there was a more than two-fold increase in the mCherryFP quantities released from cells transfected with *pJN132-*$STR6^{NAGP}$ than *pJN132-*$STR6^{wt}$ (Figure 7.8), but *pJN132-*$STR6^{wt}$ displayed higher supernatant mCherryFP than *pJN1*. The β-tubulin control verified that the higher level of mCherryFP observed in the supernatant from the $STR6^{NAGP}$ transfected HeLa cells was not due merely to increased cell death in these transfections. Thus, it would appear that when not removed by ribosome skipping the STR6-peptide functions as a protein trafficking signal directing the newly synthesised protein to the ER and Golgi and so delivering it to the exocytic pathway

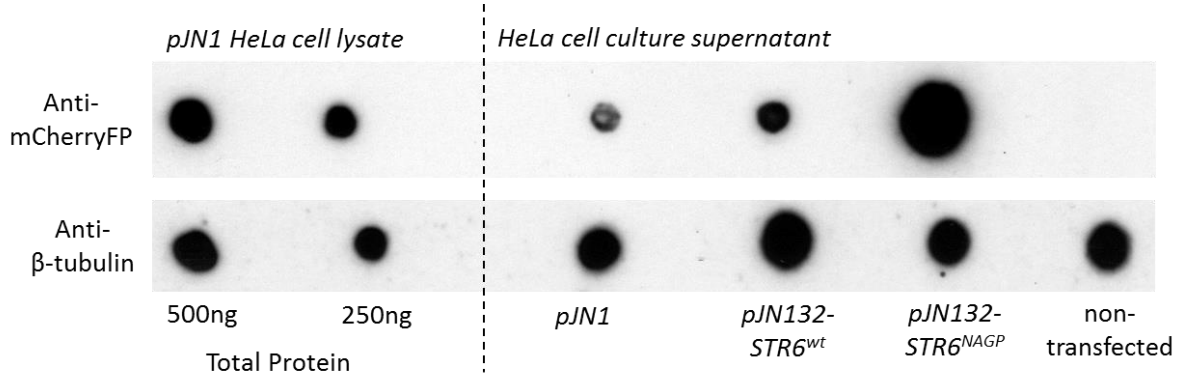


Figure 7.8 Supernatant mCherryFP detection
The supernatant from three replicate transfections of each construct were pooled, spotted onto nitrocellulose membranes then probed with anti-mCherryFP or anti-β-tubulin. Diluted cell lysates (to spot 250 ng and 500 ng total protein, respectively) from *pJN1* (mCherryFP-TaV2A-eGFP) transfected HeLa cells were employed as positive control for the assay, and to semi-quantify the levels of protein present in the supernatants. The experiment was repeated 3x, the blots shown in this Figure represent one such replicate and are representative of the results. Supernatant from cells transfected with *pJN1* contained less mCherryFP per µl than the mCherryFP present as part of 50 ng total protein from cell lysate of the same. In contrast, the supernatant from *STR6*^{wt} transfected cells possessed similar levels of mCherryFP per µL as 50 ng total protein from *pJN1* cell lysate, whereas supernatant from *STR6*^{NAGP} transfections contained more mCherryFP per µL than 100 ng total protein from *pJN1* cell lysate.

7.3.2.3 STR6-Peptides: Mass Spectrometry

Mass spectrometry was used to investigate whether STR6^{NAPG} was recognised and cleaved by signal-peptidases (signalases) post-trafficking as predicted by SignalP (Figure 7.4). Tryptic peptides derived from STR6^{NAGP} (MDGFCLLYLLLILLMRSGDVET▼NAGP; ▼=signalase cleavage site between threonine and aspargine) were detected from mass spectrometry on gel-bands corresponding to the molecular weight of STR6-mCherryFP. Both the Mascot and ProteinPilot search algorithms matched, with high confidence, MSMS spectra to a series of peptides which were non-tryptic at their N-termini and corresponded to cleavage after amino acid 22 (counting from the N-terminus) of STR6^{NAGP}, the site predicted for signalase (Figure 7.9b-f). Tryptic peptides with serine 17 (counting from the STR6^{NAGP} N-terminus) as their N-termini were also present (Figure 7.9g-h) corresponding to material not recognised by signalase. However, the signalase "cleaved" forms of STR6^{NAGP} (Figure 7.9b-f) comprised the majority of the STR6^{NAGP}-dervived peptides detected. This analysis confirmed that the signal peptidase cleavage site was probably recognised by endogenous signalases; but that cleavage was not 100% efficient, or, that the mass spectrometry analysis was sensitive enough to be able to detect the subpopulation of material corresponding to the nascent protein prior to signalase cleavage.

Legend to **Figure 7.9** (Figure on following page)

a) Pie-chart summarising the relative abundance of each STR6^{NAGP}-derived tryptic peptide (data presented in the traces **b-h**, identified through mass spectrometric analysis of material from a Coomassie-stained protein band corresponding to the predicted molecular weight of [STR6^{NAGP}-mCherryFP]. Red shading represents tryptic peptides with N-termini corresponding to the predicted signalase cleavage site (signalase "cleaved"), blue shading the uncleaved (by signalase) peptides.

b-h) Extracted ion chromatograms (XIC). Screen grabs from mass spectrometry software. Traces show only the extracted peak(s) of the mass to charge ratio (m/z) of the peptide of interest. Plots are time (x-axis) against intensity (y-axis). The arrows identify the peak(s) corresponding to the peptide of interest. Amino acid sequences are listed below each trace image. N-terminus amino acids in **bold** correspond to C-terminal residues from the STR6^{NAGP} peptide; the downstream residues correspond to the in-frame N-terminal residues from mCherryFP. <u>Underlined</u> residues have been post-translationally modified (oxidated or deaminated, as stated). Trace images were obtained from Dr. Catherine Botting, and the piechart was generated in Excel using the counts from the run data held on the Mascot server.

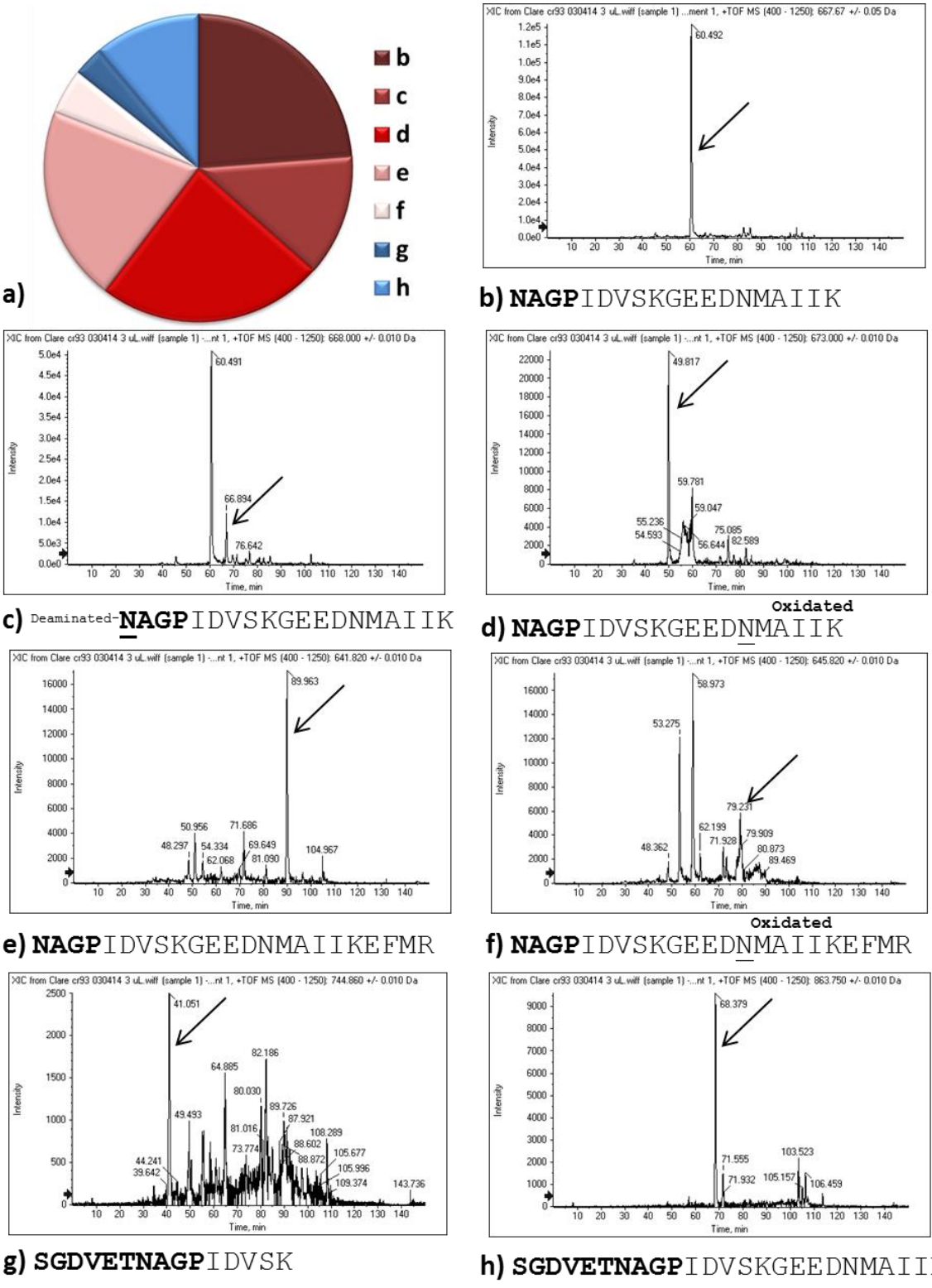


Figure 7.9 STR6^{NAGP} - mass spectrometry analyses
Full legend on preceding page.

7.3.3 *STR6* - *In Vivo* Investigations

The signal targeting pathway is believed to be highly evolutionarily conserved between phyla. Certainly, in the *in vitro* experiments described above, an echinoderm 2A-sequence (*STR6*) was found to be capable of protein targeting in mammalian cells. However, as plant biotechnology is a major utilizer of ribosome skipping 2As it was considered important to discover whether *STR6* constructs could direct protein targeting in plant cells. Additionally, transfection into echinoderm embryos was attempted with the goal of determining whether the STR6-peptide did indeed act as a protein targeting signal in its original host (sea-urchin).

7.3.3.1 *STR6* - Tobacco Leaf Infections

In tobacco, *Nicotiana benthamiana,* leaves inoculated with either $STR6^{wt}$ or $STR6^{NAPG}$ in a *pTRBO*-based construct encoding STR6-mCherryFP-TaV2A-eGFP, there was a marked difference in mCherryFP expression patterns between leaves inoculated with $STR6^{wt}$ or $STR6^{NAPG}$. For $STR6^{wt}$ both mCherryFP and eGFP were located throughout the cytoplasm, whereas in leaves inoculated with $STR6^{NAGP}$ an accumulation of mCherryFP could be observed around the outer perimeter of the cells, indicative of extracellular trafficking (Figure 7.10).

7.3.3.2 *STR6* - Echinoderm Transfections

STR6 constructs were transfected into echinoderm embryos (North American *S. purpuratus* embryos were unavailable; therefore a native Scottish sea-urchin species, *Psammechinus miliarsis* was substituted). It was hoped that the expression patterns *in vivo* would address whether STR6 acted as a signal peptide in its native host.

Successful transfection was found to be rare (5-10 % of larvae) and in observed cases, concurrent with larval death. Therefore, while embryo transfection was achieved, as transfection proved rapidly fatal to the developing larvae, it was impossible to differentiate between intra- and extracellular mCherryFP (Figure 7.11). Larval death was attributed to transfection, not sub-optimal incubation conditions, as non-transfected larvae developed normally and metamorphosed into juvenile urchins by 28 days (Figure 7.11). This study set the groundwork for future investigation by showing that it was possible to transfect developing echinoderms embryos with *STR6* constructs and that the mammalian virus-derived *CMV* promotor could be used to drive protein expression in echinoderm cells.

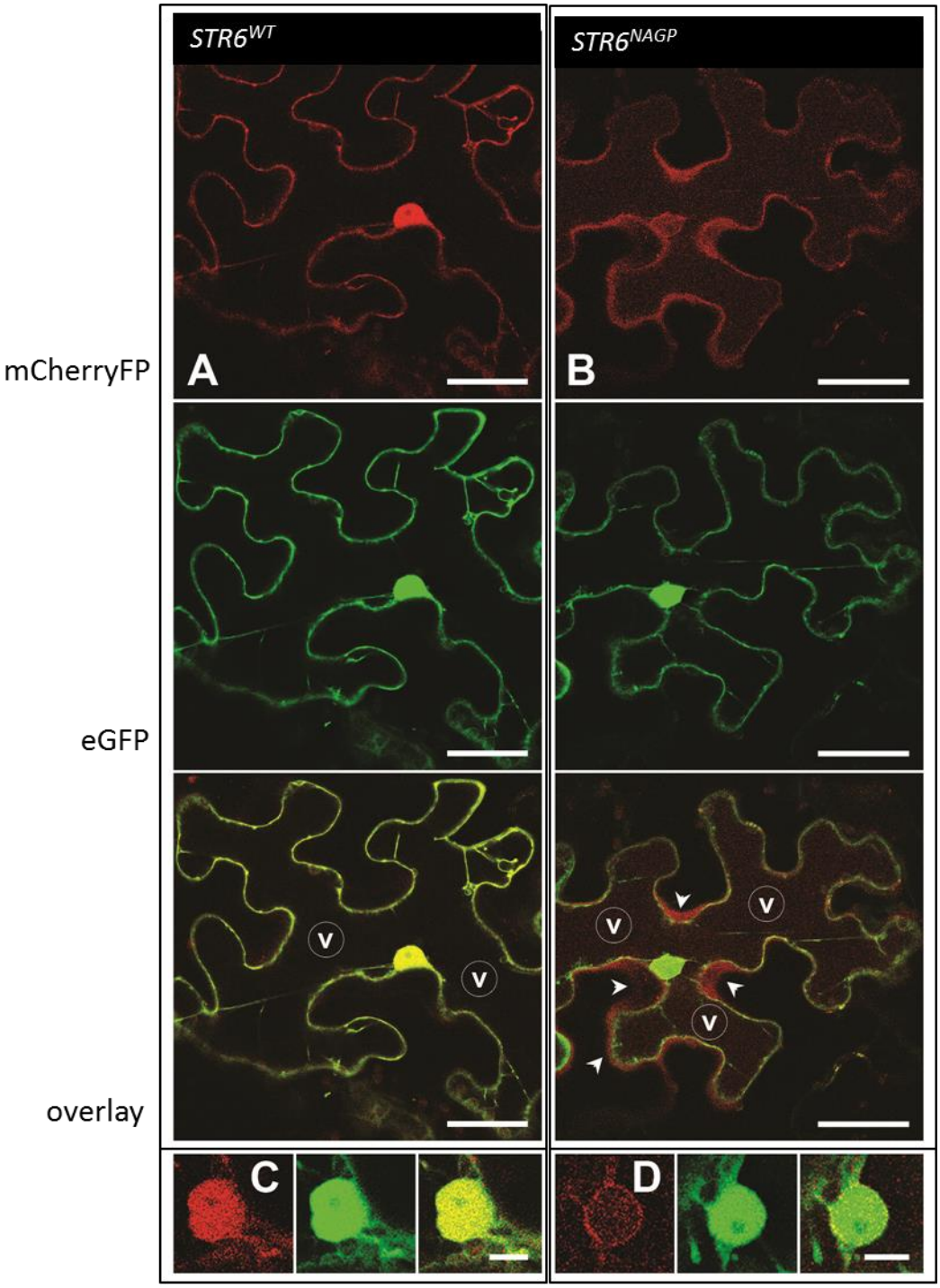


Figure 7.10 Tobacco leaves inoculated with *STR6* constructs
A & C) Single section confocal micrographs of a *N. benthamiana* leaf epidermal cell inoculated with $STR6^{wt}$ and **B & D)** of a cell similarly inoculated with $STR6^{NAGP}$. Top to bottom panels of **A** and **B** show mCherryFP, eGFP and merged channels, respectively. Note the halo (indicated by arrowheads) of red mCherryFP surrounding the cell wall (indicative of extracellular trafficking) that is absent in $STR6^{NAGP}$ inoculations (**A**). **C & D)** Intracellular vacuoles visualised in red and green and overlay image, again note the co-localisation of eGFP and mCherryFP in the vacuole for the $STR6^{WT}$ inoculations (**C**) but the halo effect again surrounding the vacuole for the $STR6^{NAGP}$ inoculations (**D**), whereas eGFP is evenly distributed throughout the vacuole in both cases. The cells show no fluorescence in the central vacuole due to the acidity denaturing the fluorescent proteins. Scale bars in **A & B** correspond to 50 µm, **C & D** scale bars are 10 µm, v= central vacuole. Images supplied by Jonathon Cope and Dr. Jens Tilsner.

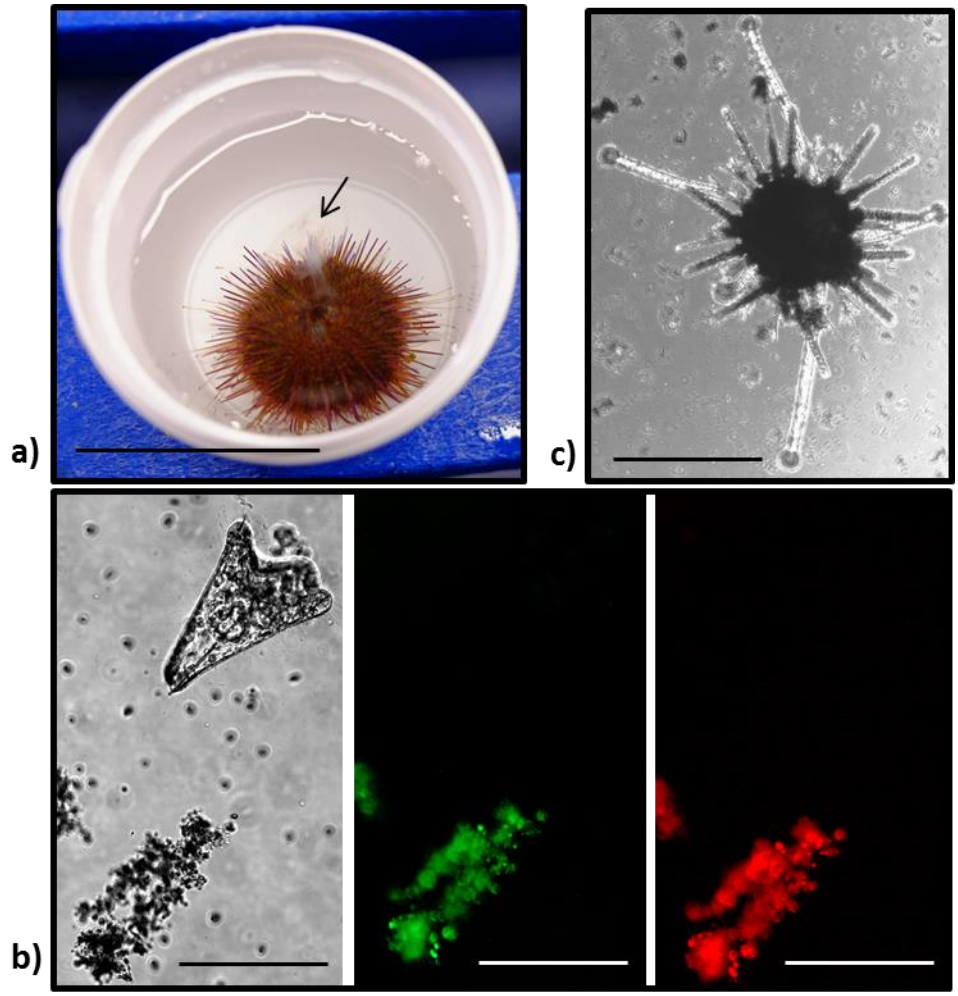
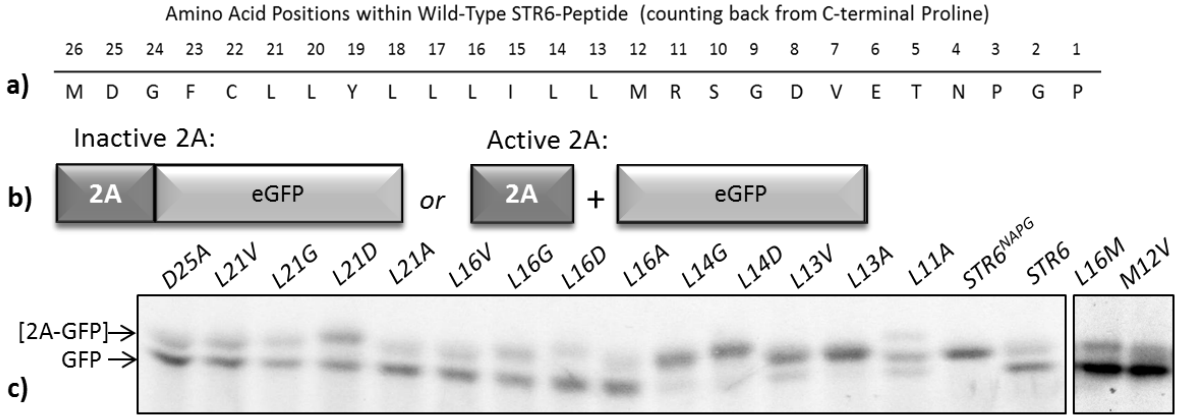


Figure 7.11 *STR6* Sea-urchin transfection
a) Spawning adult female *Psammechinus miliarsis* urchin. Note the milky cloud of eggs (highlighted by the arrow) being released from the dorsal pore of the animal. Scale-bar=5 cm. **b)** Bright-field, green and red channel EVOS microscopy images of *pJN132-STR6*^{NAGP} transfected *P. miliarsis* larval cultures at 48 hours post transfection. The non-transfected larva (top right) appeared healthy, whereas the transfected larva (bottom left) expressing eGFP and mCherryFP was dead and disintegrating. The fluorescence is believed to be from transfected protein expression not auto-fluorescence as auto-fluorescence in both the red and green channels is very rare. The small (1-20 µm diameter) black dots observed on the bright-field image are single-celled *Dunaliella spp.* algae added as food for the developing urchin larvae. Scale-bars=200 µm. **c)** by 28 days post-transfection the non-transfected *P. miliarsis* larvae had metamorphosed into juvenile urchins. Scale-bar=500 µm.

7.3.4 *STR6* Mutants

Following the identification of *STR6* as a dual function sequence, able to both instigate ribosome skipping, or, if left attached the downstream protein function as a signal peptide, and recognising the potential of such a dual use construct for biotechnology, a series of *STR6* point mutants were created with altered Signal-P D-values (both increasing and decreasing their extracellular pathway signal potential). These were tested for "cleavage" abilities *in vitro* (Figure 7.12) using coupled transcription-translation reactions (TnTs). Here, it was found that substitutions to the 2A C-terminal 14 amino acids resulted in a reduction or cessation of ribosome skipping abilities, but that *STR6* was remarkably tolerant to substitutions in its N-terminal portion (amino acids 19-25).



Construct	2A Peptide Sequence	Activity
STR6	MDGFCLLYLLLILLMRSGDVETNPGP	++
STR6^{NAGP}	MDGFCLLYLLLILLMRSGDVETNAGP	-
R11A	MDGFCLLYLLLILLMASGDVETNPGP	(+)
L13A	MDGFCLLYLLLILAMRSGDVETNPGP	-
L13V	MDGFCLLYLLLILVMRSGDVETNPGP	+
L14D	MDGFCLLYLLLIDLMRSGDVETNPGP	(+)
L14G	MDGFCLLYLLLIGLMRSGDVETNPGP	(+)
L16A	MDGFCLLYLLAILLMRSGDVETNPGP	++
L16D	MDGFCLLYLLDILLMRSGDVETNPGP	++
L16G	MDGFCLLYLLGILLMRSGDVETNPGP	++
L16V	MDGFCLLYLLVILLMRSGDVETNPGP	++
L21A	MDGFCALYLLLILLMRSGDVETNPGP	++
L21D	MDGFCDLYLLLILLMRSGDVETNPGP	++
L21G	MDGFCGLYLLLILLMRSGDVETNPGP	++
L21V	MDGFCVLYLLLILLMRSGDVETNPGP	++
D25A	MAGFCLLYLLLILLMRSGDVETNPGP	++
M12V	MDGFCLLYLLLILLVRSGDVETNPGP	++
L16M	MDGFCLLYLLMILLMRSGDVETNPGP	++

d)

Figure 7.12 *STR6-mutants* recoding activity analyses
a) *STR6-mutant* amino acid numbering system. **b)** Schematic showing the expected translation products for an inactive 2A (2A-GFP) or active (2A and GFP) **c)** SDS-PAGE gels of TnTs run on 2A constructs cloned in *pSTR6-GFP* **d)** List of constructs tested with relative recoding ability compared to wild-type *STR6* (++=moderately high activity comparable with *STR6*, +=low activity, (+)=very low activity, -=inactive. The "extra" bands observed for *R11A* are due to the presence of additional internal initiation products (see Odon *et al.* 2013).

Ribosome skipping inactive forms (NAGP-mutants, if the sequence was recoding active) of the *STR6-mutants* in plasmid *pJN132* (Figure 7.5) were transfected into HeLa cells (Figure 7.13) and observed using Deltavision microscopy. All the *STR6* mutants could direct protein targeting, as all displayed the characteristic pattern of mCherryFP localisation to the Golgi stack/perinuclear region/ER as seen for $STR6^{NAGP}$. Using this microscopic technique, it proved impossible to assess whether the mutant *STR6* constructs were of greater or lesser effectiveness as signal peptides than $STR6^{NAGP}$ as all displayed similar mCherryFP localisation patterns (for a representative selection of images see Figure 7.13) However, the fact that all could mediate protein trafficking serves to confirm the robustness of N-terminal signals to single amino acid substitutions.

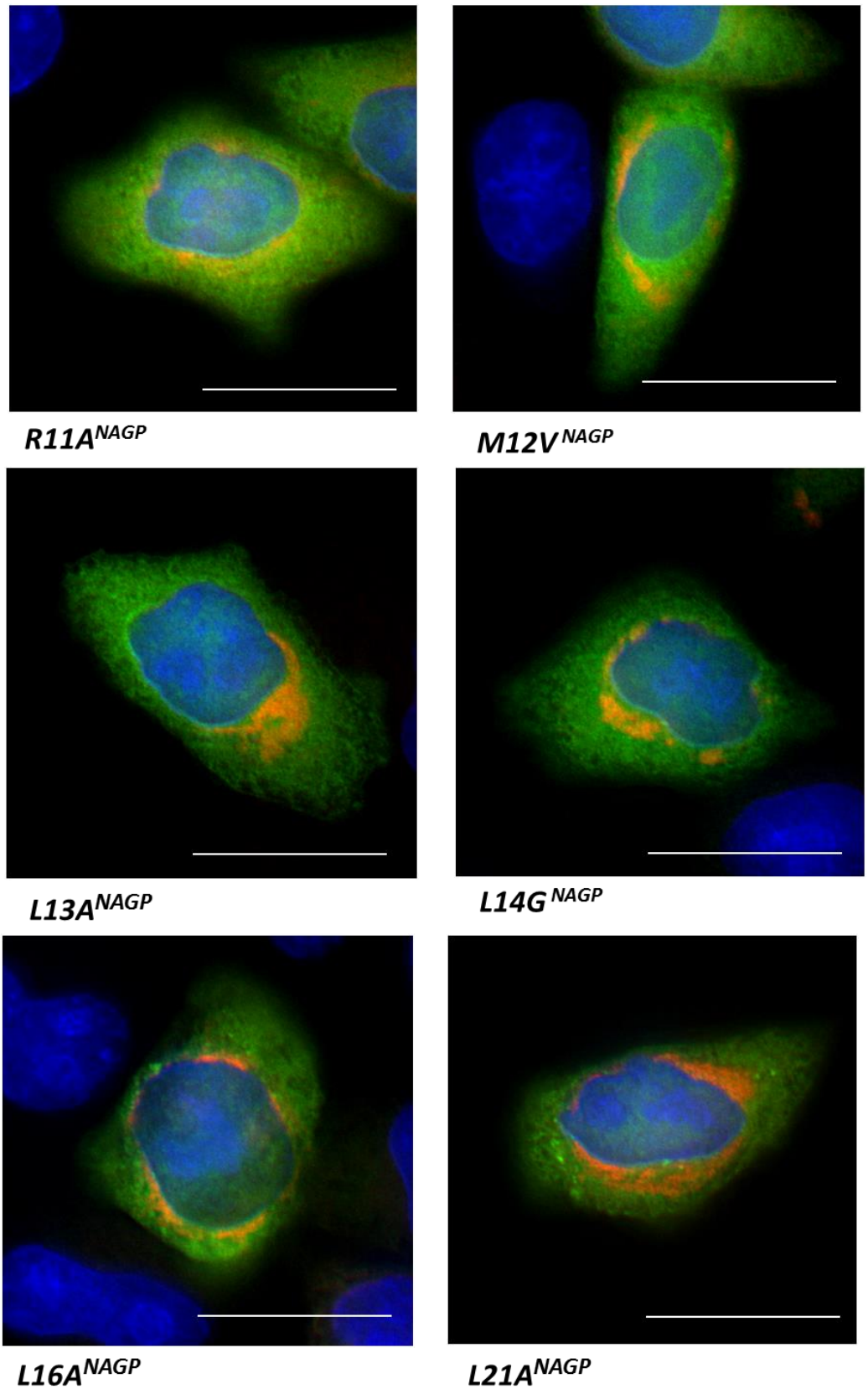


Figure 7.13 Deltavision microscopy of transfected *STR6-mutants* in *pJN132*
Multi-channel de-convolved single Z-stacks (see Section 2.2.10) displaying a representative sample of the *STR6^{NAGP}-mutants* in *pJN132*. Amino acid sequences are given in Table 7.2. Note the targeting of mCherryFP to the Golgi/ER, in contrast to the cytoplasmic distribution of eGFP. Blue=nuclear DAPI staining. All transfections fixed at 30 hours post-transfection. Scale bar=15 µm.

7.3.5 *AQ27*^{NAGP} & *SS7*^{NAGP}- Mitochondrial Signal Sequences?

A number of eukaryotic 2As (in addition to *STR6* and related sea-urchin NLR-derived signal 2As), displayed high values when investigated for signal potential using SignalP (SignalP values reported in Appendix B). Translational recoding inactive (-NAGP-) mutant versions of two of these, namely *SS7* and *AQ27* (from salmon and sponge, respectively) were further analysed with PSORT II, which also reported signal potential, although the PSORT II *k*NN values comparing sequence composition to known signal sequences in the PSORT dataset was indicative of nuclear or mitochondrial destination, not of targeting to the exocytic pathway through trafficking via the ER/Golgi as was the case with *STR6*^{NAGP} (Table 7.3).

To ascertain whether *AQ27*^{NAGP} and/or *SS7*^{NAGP} might be exocytic (as indicated by SignalP), or nuclear or mitochondrial signal sequences (as indicated by PSORT) each was cloned in-frame, upstream of mCherryFP in a reporter construct (*pEMX*, see Figure 7.14) and transfected into a green (eGFP) cytoplasmic HeLa stable cell line (created with *CytoLight*™). The eGFP intensity of *CytoLight*™ infected cells was highly variable between individual cells, and unfortunately, it was found that cells displaying lower green intensity transfected more readily than their brighter green counterparts. However, both *SS7*^{NAGP} and *AQ27*^{NAGP} transfections displayed patterns characteristic of perinuclear/Golgi and/or mitochondrial localisation (Figure 7.15). However, for both constructs the localisation patterns between cells varied considerably (more so than with the *STR6* constructs), with some cells displaying predominately cytoplasmic mCherryFP expression rather than mCherryFP targeting. Further co-transfections of *SS7*^{NAGP} or *AQ27*^{NAGP} with *pMITO* (mitochondrial signal sequence-eGFP) confirmed partial mitochondrial localisation (Figure 7.15). Therefore, it would appear that both *SS7* and *AQ27* can act, albeit inefficiently, as protein targeting signals to traffick proteins to multiple subcellular compartments, and that the PSORT II *k*NN values were able to correctly predict the likely trafficking destination in each instance.

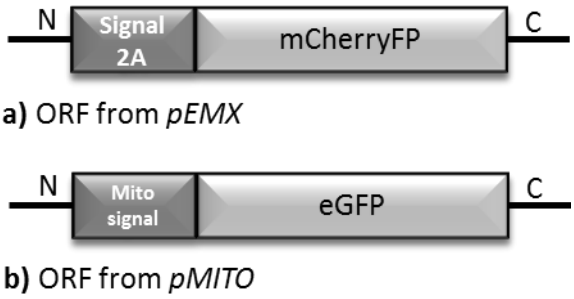


Figure 7.14 Plasmids used in *SS7*^{NAGP}/*AQ27*^{NAGP} transfections
Schematic showing the expected translation products from **a)** *SS7*^{NAGP}/*AQ27*^{NAGP} transfections, if the 2A were acting as a signal then the mCherryFP would be targeted to a specific subcellular location, and **b)** the expected translation product from *pMITO*, here the mitochondrial signal sequence remains attached to eGFP and so targets it to the mitochondria.

Table 7.3 Signal peptide analyses of AQ27^{NAGP} and SS7^{NAGP}
Results from SignalP and PSORT II analyses on AQ27^{NAGP} and SS7^{NAGP} peptides, STR6^{NAGP} data is included for comparison. SignalP D-values greater than 0.450 indicate a high probability of the sequences functioning as signal peptides, D-values are reported in **bold text**. The PSORT II *k*NN values report the likely destination of each peptide by comparing its amino acid composition to those of known signal sequences targeted to each sub-cellular location. *k*NN values are reported as percentage probability of the sequence targeting to each site. Percentages greater than 20 % are reported in **bold.**

AQ27^{NAGP}	MVSVVFKLVSLLLLLSGDIEINAGP
	SignalP output: max. C-value 0.132 at position 24, max. Y-value 0.300 at position 11, max. S-value 0.903 at position 3, mean S-value (position 1-10) = 0.846 D-value (positions 1-10) **= 0.595** Signalase cleavage site between positions 10-11 **PSORTI II *k*NN analysis: 52.2 %: cytoplasmic** **21.7 %: mitochondrial** 8.7 %: cytoskeletal 8.7 %: plasma membrane 8.7 %: nuclear
SS7^{NAGP}	**MQRSRRPVLIAFSRTLILLLLCSSGDVEVNAGP**
	SignalP output: max. C-value 0.214 at position 32, max. Y-value 0.318 at position 15, max. S-value 0.950 at position 2, mean S-value (positions 1-14) = 0.908, D-value (positions 1-14) = **0.637** Signalase cleavage site between positions 15-16 **PSORTI II *k*NN analysis: 39.1 %: mitochondrial** 17.4 %: cytoplasmic 17.4 %: Golgi 13.0 %: endoplasmic reticulum 8.7 %: nuclear 4.3 %: vacuolar
STR6^{NAGP}	**MDGFCLLYLLLILLMRSGDVETNAGP**
	SignalP output: Max C-value 0.219 at position 23, max Y-value = 0.355 also at position 23, max S-value 0.942 at position 2, mean S-value (positions 1-22) = 0.576, D-value (positions 1-22) **= 0.475**, Signalase cleavage site between positions 22-23 **PSORTI II *k*NN analysis: 33.3 %: endoplasmic reticulum** **22.2 %: Golgi** **22.2 %: cytoplasmic** 11.1 %: nuclear 11.1 %: mitochondrial

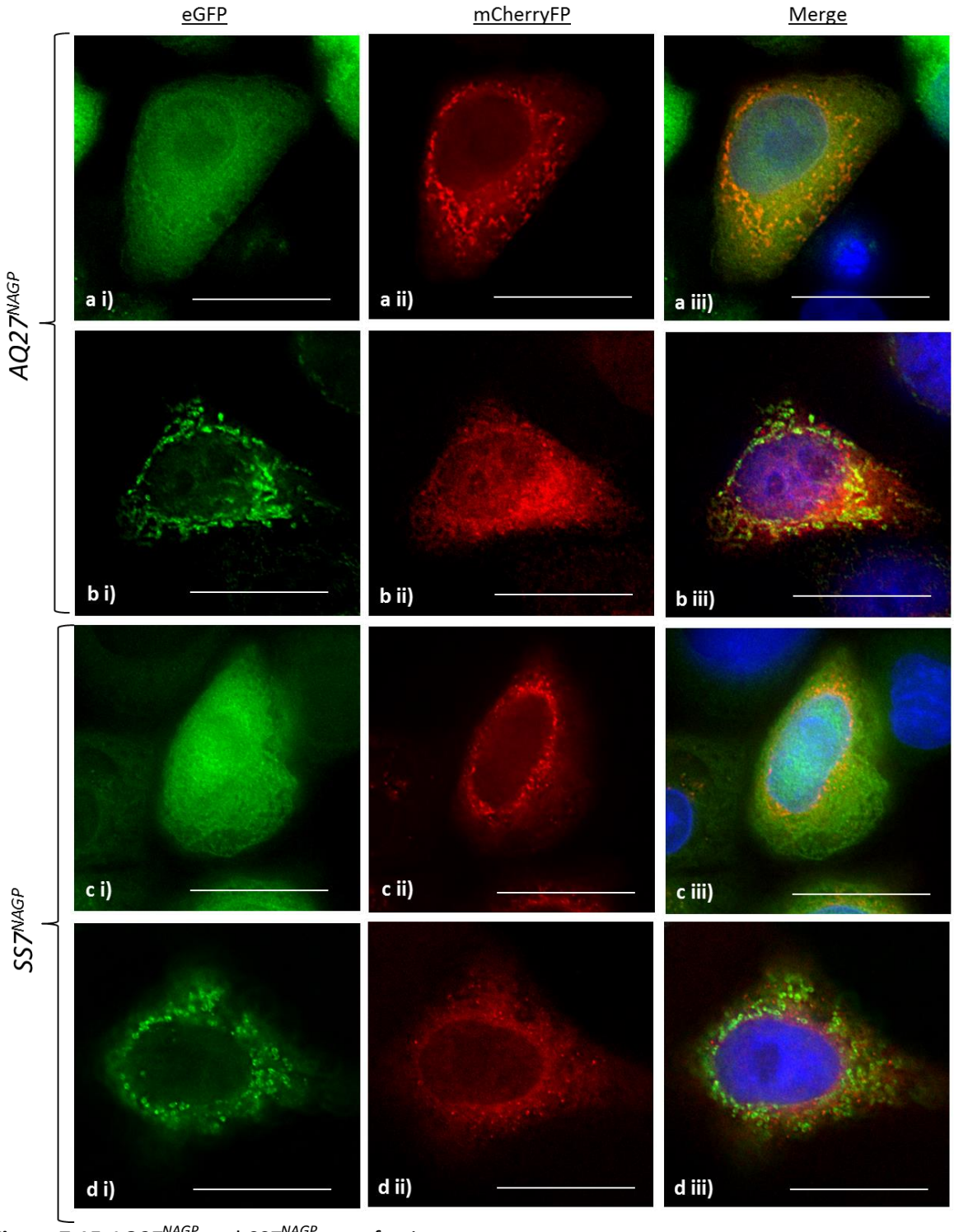


Figure 7.15 *AQ27*^{NAGP} and *SS7*^{NAGP} transfections

Images are single de-convolved Z-stack single/multi-channel images obtained from Deltavision microscopy (see Section 2.2.10). **a)** *AQ27*^{NAGP} in *pEMX* (2A-mCherryFP) transfected into *CytoLight™* HeLa cells. Note the mCherryFP expression predominately localised to the perinuclear region. **b)** *AQ27*^{NAGP} in *pEMX* co-transfected with *pMITO* (mitochondrial signal-eGFP) into wild-type HeLa cells, so the mitochondria of transfected cells are eGFP labelled. Note the partial co-localisation of eGFP and mCHerryFP. **c)** *SS7*^{NAGP} in *pEMX* transfected into *CytoLight™* HeLa cells. Again, note the mCherryFP expression predominately localised to the perinuclear region. **d)** *SS7*^{NAGP} in *pEMX* co-transfected with *pMITO* Note the partial co-localisation of eGFP and mCherryFP.

7.3.6 Amino Acid Transporter SNAT9 N-Terminal 2As – Signals?

Typically, signal peptides are an N-terminal addition to proteins, and the SNAT9 sodium dependent amino acid transporter 2A1 sequences (discussed in Chapter 5.) were found to occur as N-terminal features (Figure 7.16). Therefore, could these SNAT9 2A1s be undertaking an additional role as signal peptides? SignalP analyses reported low (non-signal) D-values suggesting that this was unlikely; but, PSORT II *k*NN analyses reported an approximately 50% probability that the SNAT9 2A1s were nuclear in destination as judged by similarity in amino acid composition to known signal sequences in the PSORT II database (Table 7.4). The SNAT9 2A2 sequences also reported a low SignalP D-value, but PSORT II *k*NN analyses reported a greater than 50 % probability of the 2A2 sequence being targeted to the nucleus, and approximately 40 % probability of the SNAT9 2A2 sequence being targeted to the mitochondria (Table 7.4).

Figure 7.16 ORF of mammalian SNAT9 sodium-dependent amino acid transport proteins
Schematic diagram illustrating the positioning of the putative 2A/signal peptide sequences with the ORF of mammalian SNAT9 proteins. Not to scale.

I suspected that the SNAT9 2As might be regulating intracellular levels of SNAT9 protein by acting as signal peptides. I proposed that the SNAT9 2A1 might be acting as a nuclear signal effectively removing available SNAT9 from the cytoplasm by sequestering it in the nuclear/perinuclear region or acting and/to transfer a proportion of the SNAT9 into a functional role as an amino acid transporter imbedded in the nuclear membrane. Whereas 2A2 (exposed when N-terminal, or upon 2A1 ribosome skipping causing a break in the nascent chain) might be either a nuclear or mitochondrial signal, additionally acting to target some of the newly synthesised SNAT9 to a functional role as a mitochondrial transmembrane transporter. To test these suppositions, four SNAT9 2A1s, namely *Cow*, *Rat*, *Panda*, and *NMR* (full details in Chapter 5.) and 2A2 were cloned into *pJN132* (signal2A-mCherryFP-Tav2A-eGFP) constructs (see Figure 7.5) in place of the *STR6* signal 2A. These constructs were then transfected into HeLa cells and visualised using Deltavision microscopy (Figure 7.17).

The microscopic analyses revealed an apparent absence of signalling properties for any of the SNAT9 2As, this despite their positive identification as putative signals by PSORT *k*NN analyses. PSORT II is known to have a success-rate of 57 % in predicting potential eukaryotic signal peptides, with a propensity for false positives (Horton and Nakai, 1997); hence it would appear that in this instance the program was reporting false positives, and these N-terminal 2As were not signal peptides.

Table 7.4 Signal peptide analyses of SNAT9 amino acid transporter 2As
Results from SignalP and PSORT II analyses on four SNAT9 2A1 peptides and the SNAT9 2A2 sequence, a SignalP D-value greater than 0.450 would indicate a high probability of the sequence functioning as signal peptides whereas a value between 0.0 and 0.1 indicates no signal potential. The PSORT II *k*NN values report the likely destination of each peptide by comparing its amino acid composition to known signal sequences targeted to each sub-cellular location. *k*NN values are reported as percentage probability of the sequence targeting to each site. Percentages greater than 20 % are reported in **bold.**

Cow	**MANMDSDSRHLLIPEGDHEINPGP**
	SignalP D-value = 0.099 No signalase cleavage site **PSORTI II *k*NN analysis: 56.5 %: nuclear** **21.7 %: cytoplasmic** 17.4 %: cytoskeletal 4.3 %: mitochondrial
Panda	**MDSDSRHLLIPEVDHEIINPGP**
	SignalP D-value = 0.097 No signalase cleavage site **PSORTI II *k*NN analysis: 65.2 %: nuclear** **17.4 %: cytoplasmic** 13.0 %: cytoskeletal 4.3 %: mitochondrial
NMR	**MTNVDDRHHLISEADHEVNPGP**
	SignalP D-value = 0.098 No signalase cleavage site **PSORTI II *k*NN analysis: 47.8 %: nuclear** **21.7 %: cytoplasmic** 17.4 %: cytoskeletal 8.7 %: mitochondrial 4.3 %: peroxisomal
Rat	**MANVDSDSRHLISEVEHEVNPGP**
	SignalP D-value = 0.097 No signalase cleavage site **PSORTI II *k*NN analysis: 56.5 %: nuclear** **21.7 %: cytoplasmic** 13.0 %: cytoskeletal 4.3 %: peroxisomal 4.3 %: mitochondrial
SNAT9 2A2	**MNKRIHYYSRLTTPADKALIAPDHVVPAPE**
	SignalP D-value = 0.105 No signalase cleavage site **PSORTI II *k*NN analysis: 52.2 %: mitochondrial** **39.1 %: nuclear** 4.3 %: cytoplasmic 4.3 %: endoplasmic reticulum

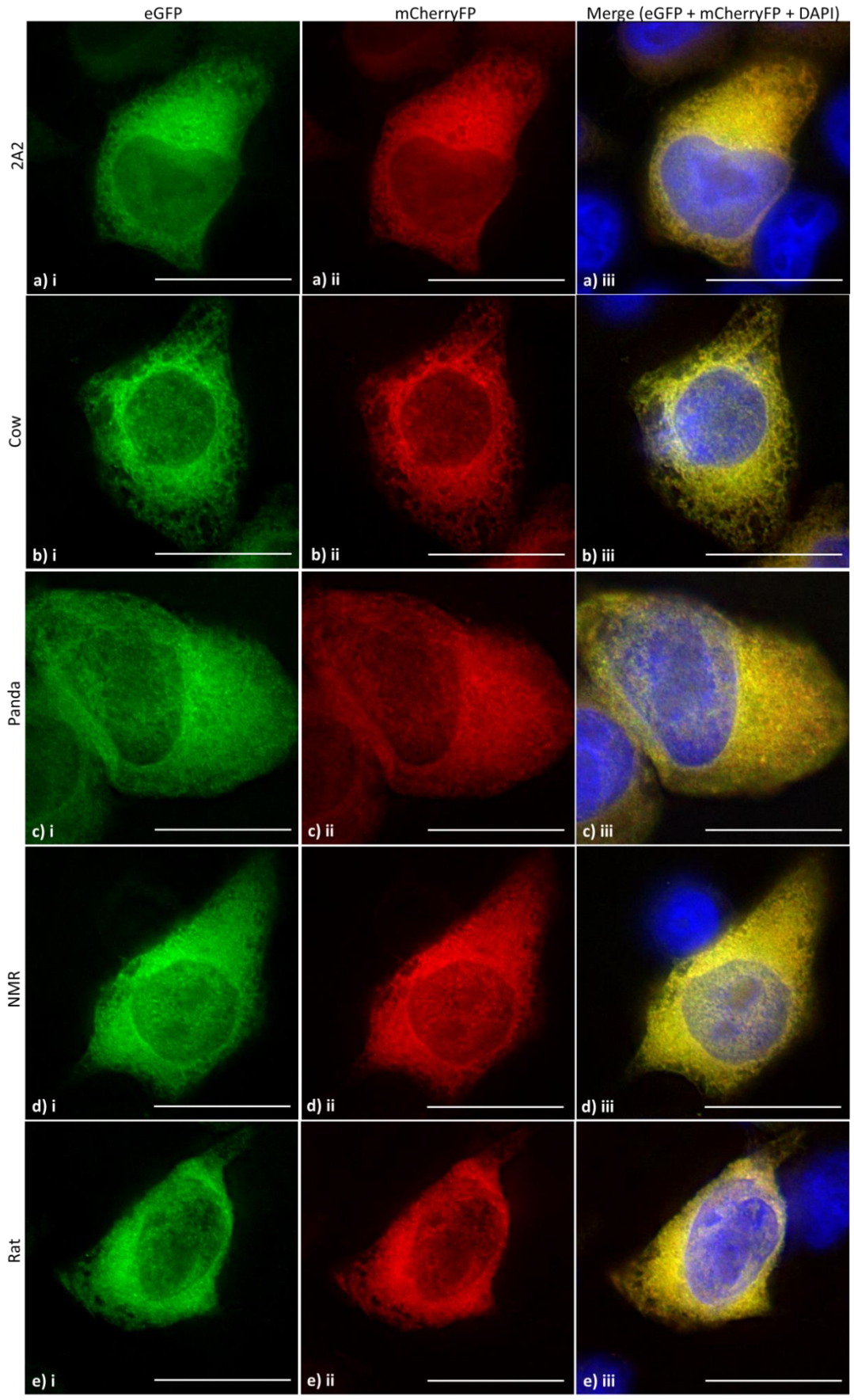


Figure 7.17 SNAT9 2A transfections
(Legend on following page)

Legend for **Figure 7.17:**
HeLa cells transfected with *pJN132*-based (signal2A-mCHerryFP-TaV2A-eGFP) constructs encoding SNAT9 2As in the signal 2A position **a)** SNAT9 2A2, **b)** Cow, **c)** Panda, **d)** NMR, **e)** Rat. Amino acid sequences can be found in Table 7.4. Note the co-localisation of mCherryFP and eGFP on the merge images, indicating no signalling abilities for the 2A preceding mCherryFP. Due to their extremely low translational recoding abilities *in vitro* (Chapter 5, Figure 5.2) the wild-type -NPGP- C-terminal forms of the 2A sequences were used. Images were obtained using Deltavision microscopy (see Section 2.2.10) and are single /triple channel de-convolved single Z-stacks.

7.4 **Signal 2As – Discussion & Future Experiments**

This Chapter reports that a subset of the newly identified eukaryotic 2A translational recoding sequences could also function as signal peptides in the instances when they remained attached to the downstream protein. Here, recoding inactive forms of such eukaryotic 2As were observed to act as signal peptides targeting nascent proteins to specific subcellular locations. There are numerous co- and post-translational mechanisms through which cells can direct newly synthesised proteins to specific intracellular destinations including alternative transcription initiation sites, alternate splicing of mRNA to give transcripts with/without the signal, low affinity signals that are not always recognised by their binding partners, and protein folding to mask the signal sequence (for a detailed review see Karniely and Pines, 2005). The dual use of eukaryotic semi-efficient 2A translational recoding sequences as signal peptides represents a novel mechanism of dual protein targeting that was both identified and described here for the first time.

Interestingly, no viral 2As were identified as possessing signal potential (using SignalP) suggesting that dual purpose signal/ribosome skipping 2As were a purely eukaryotic trait. Given the current 2A dataset, it is not possible to determine the sequence of trait acquisition, namely whether translational recoding 2As evolved signalling abilities or *vice versa*. 2As have been found that can function as protein trafficking signals in one of two different trafficking pathways, namely exocytic (*STR6* and related sequences) or mitochondrial (*AQ27* & *SS7*).

7.4.1 Biological Function of Dual Purpose 2As - Signal Peptides?

Translational recoding inactive versions of *STR6* and related echinoderm 2A sequences all showed highly efficient protein targeting to the exocytic pathway by means of export through the Golgi/ER. Given that $STR6^{wt}$ does not affect complete ribosome skipping, *in vivo* it could facilitate dual protein targeting by re-directing a proportion of the newly synthesised *STR6*-linked protein from the cytoplasm to the ER. This could function to regulate intracellular levels of the 2A-linked downstream protein, and/or target a proportion of the downstream protein to an extracellular destination.

If, as my colleagues and I have previously suggested for viral 2As, cellular stress conditions result in inefficient ribosome skipping by 2A, then this would result in a higher proportion of 2A-

peptides remaining attached to their downstream protein and available to act as protein targeting signals. The result, for *STR6*, would be a larger proportion of downstream protein being targeted to the exocytic pathway, or in the case of *AQ27/SS7*, to the mitochondria. The biological significance of this change in downstream protein destination is not immediately apparent. In the case of the *STR6*-related sequences, a greater extracellular concentration of NLR proteins in times of stress may aid in innate immune responses. Likewise, for *AQ27*, a higher quantity of death domain/ankyrin-repeat protein targeted to the mitochondria maybe beneficial. Death domain and ankyrin proteins have been implicated in governing scheduled apoptosis, as has mitochondrial integrity. Targeting to the mitochondria might therefore predispose the cell to undergo/refrain from apoptosis. In the case of *SS7* the 2A/mitochondrial signal sequence is upstream of a putative retrotransposon. Perhaps by targeting a proportion of the retrotransposon proteins to the mitochondria, this regulates the quantity available for nuclear import, and thus down-regulates retro-transposition. There were also a number of retrotransposons (for example *CR1-18_BF* from *Branchiostoma floridae*, amphioxus) which contained, in-frame, an N-terminal signal sequence (as identified by SignalP) followed by a 2A that was also a exocytic signal (again as identified by SignalP) then the exo-endonuclease and reverse transcriptase domains. In these instances, successful 2A-directed ribosome skipping would remove the signal from the mature protein, and so limit/negate protein trafficking. Again, inefficient 2A ribosome skipping would up-regulate trafficking. Hence, it would appear that N-terminal 2A/signals function to regulate cellular processes by controlling the synthesis/transport of their downstream proteins.

7.4.2 Evolutionary Conservation of Extracellular Signals

The *STR6 in vivo* studies reaffirmed the view that extracellular pathway signal peptides are highly evolutionarily conserved (Talmadge *et al.*, 1980; Muller *et al.*, 1982) with a echinoderm-derived signal sequence displaying functionality in plant cells. Therefore, dual purpose 2A/extracellular signal sequences could be active (for both functions) in any eukaryotic system and so could be potentially valuable in genetic engineering utilities.

7.4.3 SNAT9 Amino Acid Transporter 2As – Signals?

The SNAT9 sodium-dependent neutral amino acid transporter 2As were identified as putative signals on the basis of their PSORT II *k*NN analyses. However, when transfected into HeLa cells, no signal activity was observed, this could be due to a false positive report by the PSORT program; which happens in 43% of instances (Horton and Nakai, 1997), or due to HeLa cells lacking the specific chaperone(s)/co-factors necessary for recognition of the SNAT9 2As as nuclear/mitochondrial import signals. SNAT9 proteins are encoded by the *SLC38A9* gene. This is differentially expressed amongst body tissues, with the highest expression recorded for parathyroid, testis, adrenal gland, and thyroid tissue samples. Perhaps, transfection of a cell-line

derived from one of these tissues (as opposed to cervical cancer-derived HeLa) would show protein targeting; therefore, it is possible that the lack of signal activity in this instance could have been due to the cell type employed.

7.4.3.1 Future Directions

Time constraints limited the number of eukaryotic 2A sequences that could be investigated for signal function. There were a number that scored highly on SignalP D-values (listed in Appendix B) and it would be interesting to investigate a selection of these further, by cloning for *in vitro* analyses. It would also be desirable to test our long-standing supposition that cellular stress conditions influence the translational recoding abilities of ribosome skipping 2As. If cellular stress does indeed influence 2A function, then this would validate the proposal that the dual purpose signal 2As are being utilised to differentially target proteins to different distinct sub-cellular locations based on their current metabolic requirements.

Investigation of dual purpose ribosome skipping 2As-signals in their original hosts, for example, using echinoderm embryo transfections, would aid in increasing our understanding of 2A-signal function. The PEG transfection study reported here demonstrates proof-of-principle, including that it was possible to transfect echinoderm embryos with plasmid DNA preparations where translation was controlled by a mammalian CMV promotor. Regrettably, transfection resulted in embryo death. Single embryo micro-injection is an alternative transfection technique which could be utilised. Used on sea-urchin embryos since the 1980s for dye insertion (Pochapin *et al.*, 1983) micro-injection is now used routinely for RNA and DNA transfection (Stepicheva and Song, 2014) as it provides both higher transfection efficiency and lower transfected embryo death rates than PEG-mediated transfection due to requiring substantially less nucleic acid per embryo. The author used PEG due to low cost and high portability, but if this line of inquiry were to be continued, to discover whether *STR6*-related 2A sequences could act as signals in their original host organism, transfection through micro-injection would need to be employed.

Studies investigating the use of semi-efficient translational recoding 2As expressed *in cis* from a signal construct followed by signal 2As to direct dual protein targeting through partial signal masking by the first translational recoding 2A in artificial constructs encoding [moderately efficient 2A-signal 2A-reporter protein], where the signal would be revealed in only a proportion of cases (when translational recoding occurred), are currently ongoing with encouraging preliminary results. It is hoped that such constructs, making use of these dual function ribosome skipping/signal 2As will shortly constitute another valuable 2A “tool” when publicised and made available to the biotechnology sector.

Chapter 8. 2A Phylogeny & Consensus Sequence Modelling

'There is grandeur in this view of life… …from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved.'

On the Origin of Species by Means of Natural Selection… - Charles Darwin, 1859

8.1 Introduction

Following the compilation of the in-house eukaryotic 2A database (Appendix B), these sequences were sorted by their C-terminal consensus motifs in order to establish the frequency of each motif occurring in eukaryotic 2A sequences and whether this was in accordance with viral 2As. Sharma *et al.* (2012) had reported that each DxxxNPGP motif required a different upstream sequence for optimal activity, therefore the upstream region of the 2As were analysed to discover if there were any common patterns in N-terminal 2A sequence composition associated with each common C-terminal DxxxNPGP motif. Only the 2A sequence itself (up to 30 amino acids) was considered.

Representative 2A-like sequences were cloned into the *pSTA1* reporter vector and used to program cell-free coupled transcription-translation reactions (TnTs) to determine translational recoding ability. The effect of single amino residue substitutions on translational recoding ability was investigated by comparing the relative recoding abilities of pairs of naturally occurring 2A-like sequences, and through the creation of artificial intermediates by means of mutagenesis PCR.

8.2 Methodology

8.2.1 *In Silico* Searches and Analyses

An exhaustive search of major online databases was undertaken in order to compile a comprehensive in-house database of all putative eukaryotic 2A and viral 2A sequences (Appendix B). Online databases were probed for novel ribosome skipping 2A-like sequences as detailed in Section 2.1.1, using the 2A conserved motif D[V/I]ExNPGP. The search was performed in December 2011, with an update of the eukaryotic sequence search in September 2014.

For the purposes for this study, a potentially active 2A was considered to span the sequence from the final proline residue of the canonical C-terminal motif to the next methionine residue "up" towards the N-terminal, the protein N-terminus, or taking the sequence to 30 amino acids in length: whichever of these resulted in a peptide around 30 amino acids length. For brevity, the sequences were labelled with short "tags" in place of their database accession numbers. Tag labels were composed of letters corresponding to the host species name then a number, but the numbers do not necessarily correspond to the current number of sequences from each host as online database consolidation since 2010 has resulting in the removal of a large number of sequences, particularly from the purple sea-urchin. Both tags and accession numbers are listed in Table 8.1 and Table 8.3.

The composition of the 2A-like sequences was analysed. Firstly, the database was trimmed by including only one representative in the case of multiple identical protein/nucleic acid sequences from the same virus/organism entered under different accession numbers in the same database. For each eukaryotic 2A DxxxNPGP motif with more than 20 total entries in the dataset, namely DVESNPGP, DVEENPGP, DIETNPGP, DVETNPGP, DVELNPGP, DVEVNPGP, and DVERNPGP, plus DVTINPGP the most abundant viral motif, the most frequently occurring amino acid (the "consensus sequence") at each position "up" the peptide chain for a total of 30 amino acids towards the N-terminus was identified. For each position, a particular residue was considered the consensus residue if it were: firstly, the most frequently occurring, and secondly, occurred more than 3 times in sequences from the databases. N-terminally truncated 2A-like sequences (typically only the canonical C-terminus motif) were excluded from these composition analyses with 2A sequences being included only if they consisted of 13 or more amino acids "upstream" from the C-terminal proline of the 2A active sequence.

The dataset was probed to find a direct match or the closest match possible, for each of the most frequently occurring viral and eukaryotic cellular 2A consensus sequences, irrespective of host organism from which the sequence derived. These consensus sequences were modelled using PEP-FOLD 2011 and visualised and aligned to *FMDV 2A* using PyMOL v1.3 as detailed in Chapter 2.1.3.

8.2.2 Cloning & *In Vitro* Translational Recoding Analyses

Representatives of the consensus 2As, together with wildcard sequences, were translated into nucleic acid sequences and incorporated into gene-blocks (Table 8.1; Table 8.2) or cloned by means of long reverse primers (Figure 8.1, Table 8.3). The wildcard sequences *ME-1*, *OM-4*, *CE-1/3/5*, *CV-1* and *AM-1* (from *Mytilus edulis*, common shore mussel; *Oncorhynchus mykiss*, rainbow trout; *Caenorhabditis elegans*, nematode worm; *Chlorella variabilis*, unicellular green alga and *Acropora millepora*, reef-building hard coral, respectively) were included in order to broaden the phylogenetic host range of sequences tested. The *FMDV* DVES sequence (present in the wild-type vector *pSTA1*) and the sea-urchin sequences *STR-1*, *STR6*, *STR-81*, and *STR-140* were included in the *in vitro* analyses as they had already been cloned in the laboratory (Table 8.1).

Gene-blocks were designed such that no block exceeded 376 bp (the maximum permitted by the synthesising company, see Chapter 2) Alternate codon usage was employed to generate nucleic acid sequences encoding the peptides of interest, but free of substantial sequence repeats (which impair synthesis). The 2A sequences were extracted from the gene-blocks by specific sequential enzymatic digestion (Figure 8.1) finishing with XbaI/ApaI digestion and ligation into *pSTA1* similarly digested.

Table 8.1 2A sequences incorporated in gene-blocks

Dxxx	Name	Source/host: Accession no. Latin name, common name	Amino Acid Sequence	Nucleic Acid Sequence	Gene block (G)/ Insert (I)
DVEL	*IS-1*	DS876754 *Ixodes scapularis*, tick	MFLVLLLLLSGDVELNPGP	ATGTTCCTCGTACTTCTACTCCTGCTTTCAGGAGACGTTGAACTAAACCCAGGCCCA	G1I1
DVEL	*AQ20*	XP_003385788.1 *Amphimedon queenslandica* sponge	CDTVSYAVYLLLYFMLLLLLSGDVELNPGP	TGCGACACAGTATCATACGCAGTGTATCTTCTACTCTACTTTATGCTGCTCCTACTGCTCTCAGGAGACGTAGAACTGAACCCTGGACCA	G1I2
DVEL	*IS-68*	DS667985 *Ixodes scapularis*, tick	MFSLCCQCFDVLSQVLLMSGDVELNPGP	ATGTTCTCACTGTGCTGTCAATGCTTCGACGTACTATCACAGGTACTTCTCATGTCCGGAGATGTAGAACTTAACCCTGGACCA	G1I3
DVET	*DCV*	ADF56663.2 *Drosophila C Virus*	MTQGIGKKNPKQEAARQMLLLLSGDVETNPGP	ATGACACAAGGAATCGGTAAGAAGAATCCTAAGCAGGAAGCAGCACGACAGATGCTACTTCTGCTCTCAGGAGACGTAGAGACTAATCCAGGTCCA	G2I1
DVEV	*SS7*	GU129139.1 *Salmo salar*, Atlantic salmon	QRSRRPVLIAFSRTLILLLLCSSGDVEVNPGP	CAACGATCACGTCGACCAGTGCTAATAGCATTCTCACGAACACTGATACTACTGCTGCTCTGCTCCTCAGGAGACGTTGAAGTCAATCCAGGACCT	G2I2
DVEV	*OM-4*	EZ854573.1 *Oncorhynchus mykiss*, rainbow trout	TRRPVILAFSCTLILLLFCSSGDVEVNPGP	ACTCGCCGTCCAGTAATCCTAGCATTCTCATGCACACTAATACTGCTCCTATTCTGCTCATCCGGAGACGTAGAAGTTAATCCAGGACCG	G2I3
DIET	*Cardio virus*	ACG61135.2 *Human Cardiovirus D*	FTGFFKAVRDYHASYYKQRLQHDIETNPGP	TTCACAGGATTCTTCAAAGCAGTACGAGACTACCACGCATCATACTACAAACAACGACTACAACACGACATAGAAACAAACCCAGGACCA	G3I1
DVES	*AF-180*	EO35FTK01B7DS5 *Allocentrotus fragilis*, sea urchin	QNIDVKEADKRHITQTLLTRAGDVESNPGP	CAGAACATCGACGTAAAAGAAGCAGACAAACGACACATAACACAAACACTACTAACACGAGCAGGAGACGTAGAATCAAACCCAGGACCA	G3I2
DVES	*CE-1/3/5*	AL132860.1 *Caerorhabditis elegans*, nematode worm	LCETPSLPHTTFLKRKLLVRSGDVESNPGP	CTCTGCGAGACACCATCACTACCACACACAACATTCCTAAAACGAAAACTACTAGTACGATCAGGAGACGTAGAATCAAACCCAGGACCA	G3I3
DIEL	*IS-34*	DS730003 *Ixodes scapularis,* tick	MVFPCRQVLFLCSCERISGLKLLLSGDIELNPGP	ATGGTCTTCCCATGCCGACAAGTACTATTCCTATGCTCATGCGAACGAATATCAGGACTAAAACTACTACTATCAGGAGACATAGAACTAAACCCAGGACCA	G4I1
DIET	*SK-45/ 53*	Contig114660 *Saccoglossus kowalevskii*, acorn worm	MRSNLVLIRDIETNPGP	ATGCGATCAAACCTAGTACTAATACGAGACATAGAAACAAACCCAGGACCA	G4I2
DVEE	*CV-1*	XP_005844301.1 *Chlorella variabilis*, unicellular alga	LRLPCSCSTTALIKRMKLLLSGDVEENPGP	CTACGACTACCATGCTCATGCTCAACAACAGCACTAATAAAACGAATGAAACTACTACTATCAGGAGACGTAGAAGAAAACCCAGGACCA	G4I3
DVER	*STR-37*	EMJFY7Z01C829N *A. fragilis & S. franciscanus,,*urchin	MTNILLLRSGDVERNPGP	ATGACGAACATACTACTACTACGATCAGGAGACGTAGAACGAAACCCAGGACCG	G4I4
DIEL	*HRVC*	AEJ21074.1 Human Rotavirus C	GTGYPLIVANSKFQIDKILISGDIELNPGP	GGAACAGGATACCCACTAATAGTAGCAAACTCAAAATTCCAAATAGACAAAATACTAATATCAGGAGACATAGAACTAAACCCAGGACCA	G5I1
DIEL	*ME-1*	AM878476.1 *Mytilus edulis*, shore mussel	YKISLLLLTNSSDIELNPGP	TACAAGATCTCACTACTACTACTAACAAACTCATCAGACATAGAACTAAACCCAGGACCA	G5I2
DVTI	*RV*	BAJ14093.1 *Bovine, Porcine & Panda Rotaviruses*	LTNSYTVNLSDEIQEIGSAKSQDVTINPGP	CTCACGAACTCATACACAGTAAACCTATCAGACGAAATACAAGAAATAGGATCAGCAAAATCACAAGACGTAACAATAAACCCAGGACCC	G5I3
DVEE	*TB-2*	AFN16295.1 *Trypanosoma brucei*, trypanosome	RSLGTCQRAISSIIRTKMLLSGDVEENPGP	CGCTCACTCGGAACGTGCCAACGAGCAATATCTTCTATCATACGTACCAAGATGCTGCTATCGGGAGACGTAGAAGAGAACCCAGGACCA	G6I1
DVEE	*PTV*	AAK12385.1 *Porcine Teschovirus 6*	MTTMSFQGPGATNFSLLKQAGDVEENPGP	ATGACAACGATGTCATTCCAAGGTCCTGGAGCAACGAACTTCTCGCTACTCAAACAAGCCGGTGACGTAGAAGAAAACCCAGGACCA	G6I2
DVEE	*AM-1*	EZ007780.1 *Acropora millepora*,coral	MFMMVFYNAYIPLLIRQANDVEENPGP	ATGTTCATGATGGTATTCTACAACGCATACATACCACTGCTAATACGACAAGCAAACGACGTTGAAGAAAACCCAGGTCCG	G6I3

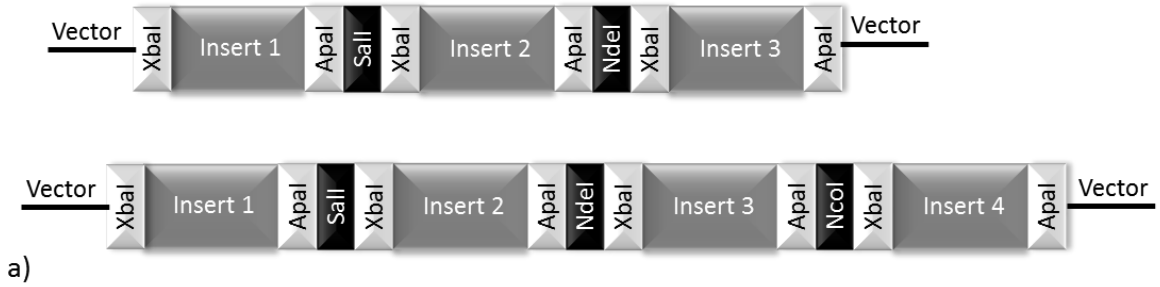
Table 8.2 2A Gene-blocks

2A nucleic acid sequences are presented in black, restriction enzyme linker sequences are underlined in grey text. For the amino acid sequence and host of each 2A, please refer back to Table 3.1.

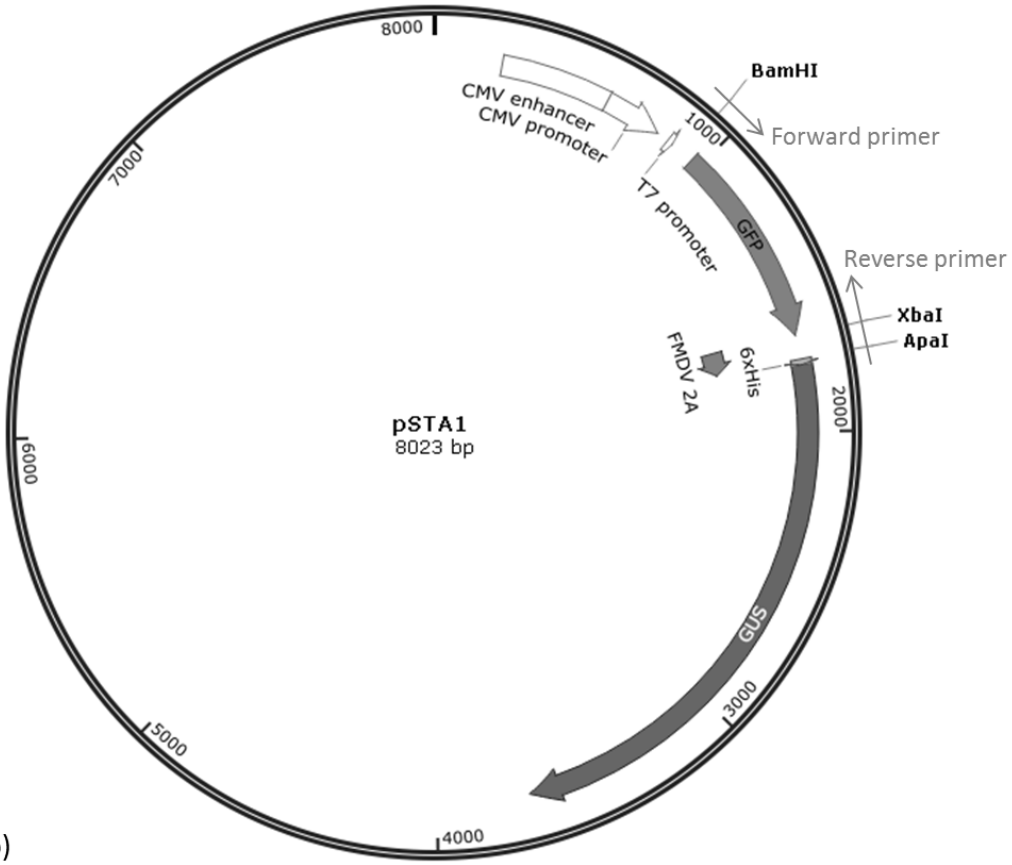
Gene block	2As	Full Nucleic Acid Sequence (cloned into *pBluescript* vector)
1	*IS-1, AQ20, IS-68*	5′TCTAGAATGTTCCTCGTACTTCTACTCCTGCTTTCAGGAGACGTTGAACTAAACC CAGGCCCAGGGCCCGTCGACTCTAGATGCGACACAGTATCATACGCAGTGTATCTTC TACTCTACTTTATGCTGCTCCTACTGCTCTCAGGAGACGTAGAACTGAACCCTGGAC CAGGGCCCCATATGTCTAGAATGTTCTCACTGTGCTGTCAATGCTTCGACGTACTAT CACAGGTACTTCTCATGTCCGGAGATGTAGAACTTAACCCTGGACCAGGGCCC-3′
2	*DCV, SS7, OM-4*	5′TCTAGAATGACACAAGGAATCGGTAAGAAGAATCCTAAGCAGGAAGCAGCACGAC AGATGCTACTTCTGCTCTCAGGAGACGTAGAGACTAATCCAGGTCCAGGGCCCGTCG ACTCTAGACAACGATCACGTCGACCAGTGCTAATAGCATTCTCACGAACACTGATAC TACTGCTGCTCTGCTCCTCAGGAGACGTTGAAGTCAATCCAGGACCTGGGCCCCATA TGTCTAGAACTCGCCGTCCAGTAATCCTAGCATTCTCATGCACACTAATACTGCTCC TATTCTGCTCATCCGGAGACGTAGAAGTTAATCCAGGACCGGGGCCC-3′
3	*CardioV, AF-180, CE-1/3/5*	5′TCTAGATTCACAGGATTCTTCAAAGCAGTACGAGACTACCACGCATCATACTACA AACAACGACTACAACACGACATAGAAACAAACCCAGGACCAGGGCCCGTCGACTCTA GACAGAACATCGACGTAAAAGAAGCAGACAAACGACACATAACACAAACACTACTAA CACGAGCAGGAGACGTAGAATCAAACCCAGGACCAGGGCCCCATATGTCTAGACTCT GCGAGACACCATCACTACCACACACAACATTCCTAAAACGAAAACTACTAGTACGAT CAGGAGACGTAGAATCAAACCCAGGACCAGGGCCC-3′
4	*IS-34, SK-45, CV-1, STR-37*	5′TCTAGAATGGTCTTCCCATGCCGACAAGTACTATTCCTATGCTCATGCGAACGAA TATCAGGACTAAAACTACTACTATCAGGAGACATAGAACTAAACCCAGGACCAGGGC CCGTCGACTCTAGAATGCGATCAAACCTAGTACTAATACGAGACATAGAAACAAACC CAGGACCAGGGCCCCATATGTCTAGACTACGACTACCATGCTCATGCTCAACAACAG CACTAATAAAACGAATGAAACTACTACTATCAGGAGACGTAGAAGAAAACCCAGGAC CAGGGCCCCCATGGTCTAGAATGACGAACATACTACTACTACGATCAGGAGACGTAG AACGAAACCCAGGACCGGGGCCC-3′
5	*HRVC, ME-1, RV*	5′TCTAGAGGAACAGGATACCCACTAATAGTAGCAAACTCAAAATTCCAAATAGACA AAATACTAATATCAGGAGACATAGAACTAAACCCAGGACCAGGGCCCGTCGACTCTA GATACAAGATCTCACTACTACTACTAACAAACTCATCAGACATAGAACTAAACCCAG GACCAGGGCCCCATATGTCTAGACTCACGAACTCATACACAGTAAACCTATCAGACG AAATACAAGAAATAGGATCAGCAAAATCACAAGACGTAACAATAAACCCAGGACCCG GGCCC-3′
6	*TB-2, PTV, AM-1*	5′TCTAGACGCTCACTCGGAACGTGCCAACGAGCAATATCTTCTATCATACGTACCA AGATGCTGCTATCGGGAGACGTAGAAGAGAACCCAGGACCAGGGCCCGTCGACTCTA GAATGACAACGATGTCATTCCAAGGTCCTGGAGCAACGAACTTCTCGCTACTCAAAC AAGCCGGTGACGTAGAAGAAAACCCAGGACCAGGGCCCCATATGTCTAGAATGTTCA TGATGGTATTCTACAACGCATACATACCACTGCTAATACGACAAGCAAACGACGTTG AAGAAAACCCAGGTCCGGGGCCC-3′

Table 8.3. List of 2A sequences cloned by means of PCR

Dxxx	Name	Source/host: Accession no., Latin name, & common name	Amino Acid Sequence	Nucleic Acid Sequence (5' to 3')	In lab/ primer/ mutated
DVES	*FMDV*	AAT01719.1 FMDV 2A	MDELYKSGSRGACQLL NFLLDKLAGDVESNPG P	ATGGACGAACTATACAAGTCCGGGTCT AGAGGAGCATGCCAGCTGTTGAATTTT GACCTTCTTAAGCTTGCGGGAGACGTC GAGTCCAACCCCGGGCCC	In lab
DVEL	*STR-81*	GLEAN3_21478 *Strongylocentrotus purpuratus,* urchin	SKTDLISGQIPPLSEL LLLKSGDVELNPGP	TCAAAGACAGATTTGATATCTGGACAA ATTCCTCCTCTCTCCGAACTACTTCTC TTGAAATCTGGTGATGTAGAGCTCAAC CCAGGGCCC	In lab
DVET	*STR6*	XP_798533.3 *Strongylocentrotus purpuratus,* urchin	MDGFCLLYLLLILLMR SGDVETNPGP	ATGGATGGATTCTGTCTTCTCTATCTG CTCCTGATCCTCTTGATGAGGTCTGGT GACGTTGAAACCAATCCAGGTCCT	In lab
DVEI	*STR-1*	XP_797143.2 *Strongylocentrotus purpuratus,* urchin	MFVCAFILISVLLLSG DVEINPGP	ATGTTTGTGTGCGCGTTTATTCTGATT AGCGTGCTGCTGCTGAGCGGCGATGTG GAAATTAACCCGGGGCCC	In lab
DVET	*MO-1*	GAA99414.1 *Mixia osmundae IAM 14324,* fern fungus	AAHGQVVLKTNKQGDK YYPDVETNPGP	GCAGCTCATGGCCAAGTAGTTCTTAAA ACTAATAAACAAGGCGATAAGTACTAT CCGGATGTAGAAACTAATCCAGGGCCC	On reverse primer
DVEL	*FMDV DVEL mut*	Artificial sequence N-terminus *FMDV A* serotype, DVEL from *Asia* serotype	MDELYKSGSRGACQLL NFLLDKLAGDVELNPG P	ATGGACGAACTATACAAGTCCGGGTCT AGAGGAGCATGCCAGCTGTTGAATTTT GACCTTCTTAAGCTTGCGGGAGACGTC GAGCTCAACCCCGGGCCC	Mutated
DVER	*STR-37m I15A*	Artificial Sequence	MTNALLLRSGDVERNP GP	ATGACGAACGCACTACTACTACGATCA GGAGACGTAGAACGAAACCCAGGACCG	Mutated
DVEL	*STR-140*	GLEAN3_26442 *Strongylocentrotus purpuratus,* urchin	MTNALLLRSGDVELNP GP	TCTAGAATGCCCTTCTATTGAGATCTG GTGATGTTGAACTGAACCCTGGGCCC	In lab
DIET	*SK-45 L*	Contig114660 *Saccoglossus kowalevskii*, acorn worm	MYDNKNWTFALYLYHC RMRSNLVLIRDIETNP GP	ATGTATGATAATAAAAACTGGACTTTT GCACTTTATTTGTATCATTGTCGTATG CGATCAAACCTAGTACTAATACGAGAC ATAGAAACAAACCCAGGGCCC	Primer
DVEV	*SS7 mutA*	Artificial Sequence	QRSRRPVLIAFSRTLI LLLFCSSGDVEVNPGP	CAACGATCACGTCGACCAGTGCTAATA GCATTCTCACGAACACTGATACTACTG CTGTTCTGCTCCTCAGGAGACGTTGAA GTCAATCCAGGACCT	Mutated
DVEV	*SS7 mutB*	Artificial Sequence	QRSRRPVLIAFSCTLI LLLLCSSGDVEVNPGP	CAACGATCACGTCGACCAGTGCTAATA GCATTCTCATGTACACTGATACTACTG CTGCTCTGCTCCTCAGGAGACGTTGAA GTCAATCCAGGACCT	Mutated
DVEV	*SS7 mutC*	Artificial Sequence	QRSRRPVLIAFSCTLI LLLFCSSGDVEVNPGP	CAACGATCACGTCGACCAGTGCTAATA GCATTCTCATGTACACTGATACTACTG CTGTTCTGCTCCTCAGGAGACGTTGAA GTCAATCCAGGACCT	Mutated
DVEV	*SS7 mutD*	Artificial Sequence	QRSRRPVILAFSRTLI LLLLCSSGDVEVNPGP	CAACGATCACGTCGACCAGTGATACTA GCATTCTCACGAACACTGATACTACTG CTGCTCTGCTCCTCAGGAGACGTTGAA GTCAATCCAGGACCT	Mutated
DVEV	*SS7 mutE*	Artificial Sequence	QRSRRPVILAFSRTLI LLLFCSSGDVEVNPGP	CAACGATCACGTCGACCAGTGATACTA GCATTCTCACGAACACTGATACTACTG CTGTTCTGCTCCTCAGGAGACGTTGAA GTCAATCCAGGACCT	Mutated
DVEV	*SS7 mutF*	Artificial Sequence	QRSRRPVILAFSCTLI LLLLCSSGDVEVNPGP	CAACGATCACGTCGACCAGTGATACTA GCATTCTCATGTACACTGATACTACTG CTGCTCTGCTCCTCAGGAGACGTTGAA GTCAATCCAGGACCT	Mutated
DVEV	*SS7 mutG*	Artificial Sequence	QRSRRPVILAFSCTLI LLLFCSSGDVEVNPGP	CAACGATCACGTCGACCAGTGATACTA GCATTCTCATGTACACTGATACTACTG CTGTTCTGCTCCTCAGGAGACGTTGAA GTCAATCCAGGACCT	Mutated



a)



b)

Figure 8.1 2A cloning strategy

a) Design of three and four insert gene-blocks, inserts were extracted from their *pBluescript* vector by insert specific, then XbaI/ApaI digest. **b)** Plasmid *pSTA1* showing the relative positions of the 2A sequence (novel 2A sequences were inserted in place of *FMDV 2A*), and the restriction enzyme and/or PCR primer sites used in cloning.

The decreasing cost of the synthesis of long primer sequences (100 base pairs and above), permitted their use instead of gene-blocks as the study progressed, the longer version of the acorn worm *SK-45* sequence and the fungal sequence *MO-1* were cloned by this method (Table 8.2). These sequences were cloned by PCR using a forward primer flanking the BamHI site in *pSTA1* and a long reverse primer containing an ApaI site, the novel 2A, the XbaI site, and the C-terminal nucleotides of GFP. The PCR product was BamHI/ApaI digested and ligated into *pSTA1* similarly restricted. Artificial mutant 2A sequences were constructed by means of mutagenesis PCR (Table 8.4).

Mutagenesis PCR (see Chapter 2.2.1.4.2) was used to generate artificial sequences intermediate between two naturally occurring 2A sequences. This technique was used to create a composite *FMDV* sequence with the N-terminus from *Serotype A* but the DVEL motif from an *Asia* serotype; it was also used to create series of intermediates between two sea-urchin sequences (*STR-37* and *STR-140*), and between two salmonid sequences (*SS7* and *OM-4*) (Figure 8.2).

After verification by DNA sequencing using primer *GFPf* (Table 2.1) the plasmid preparations were used to program cell-free coupled transcription-translation reactions (TnTs) as described in Chapter 2.2.2 and Odon *et al.*, 2013.

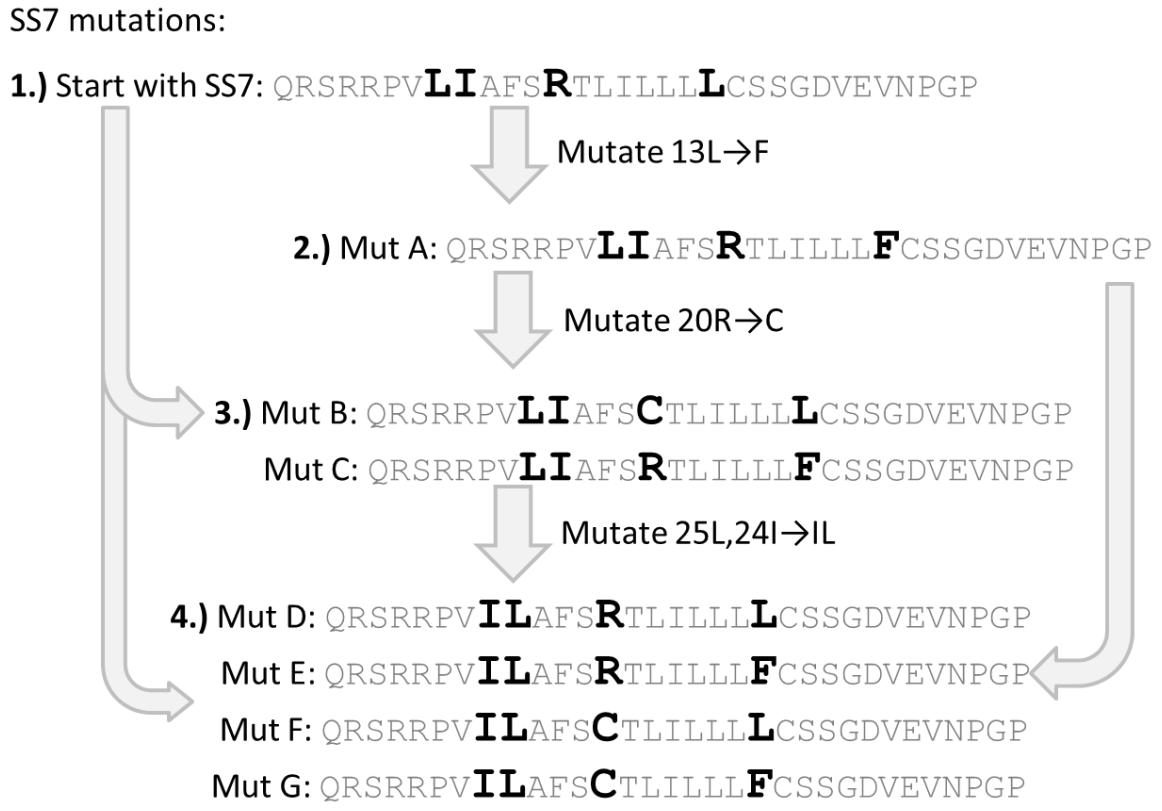


Figure 8.2 *SS7* mutagenesis cloning strategy
Flow diagram detailing the stepwise mutagenesis required for generation of a series of intermediate mutant sequences between *SS7* and *OM-4*, residues to be mutated in **bold black** typeface.

Table 8.4 Primers used in Chapter 8 cloning

Table of primers used in cloning the 2A sequences analysed in this Chapter. Primers were designed and synthesised as stated in Section 2.1.5. XbaI and ApaI restriction sites used in cloning are underlined.

Primer	Sequence (5' to 3')	Notes
pSTA1_F	GGGCTCGGATCCACCATGGGGCACCACC	Forward PCR primer
MO-1 R	GTGGTGGGGCCCTGGATTAGTTTCTACATCCGGATAGTACT TATCGCCTTGTTTATTAGTTTTAAGAACTACTTGGCCATGA GCTGCTCTAGACCCGGAC	Reverse MO1 PCR primer
SK-45L R	GTGGTGGGGCCCTGGGTTTGTTTCTATGTCTCGTATTAGTA CTAGGTTTGATCGCATACGACAATGATACAAATAAAGTGCA AAAGTCCAGTTTTTATTATCATACATTCTAGACCCGGAC	Reverse SK-45L PCR primer
DVES-DVEL F	GGGAGACGTCGAGCTCAACCCCGGGCC	Mutagenesis PCR F primer
DVES-DVEL R	GGGCCCGGGGTTGAGCTCGACGTCTCCC	Mutagenesis PCR R primer
STR-140 mutI→A F	CGGGTCTAGAATGACGAACGCACTACTACTACGATCAGGA	Mutagenesis PCR F primer
STR-140 mutI→A R	TCCTGATCGTAGTAGTAGTGCGTTCGTCATTCTAGACCCG	Mutagenesis PCR R primer
STR-140 mutR→L F	CAGGAGACGTAGAACTAAACCCAGGACCGGG	Mutagenesis PCR F primer
STR-140 mutR→L R	CCCGGTCCTGGGTTTAGTTCTACGTCTCCTG	Mutagenesis PCR R primer
SS7mut L→F F	CACTGATACTACTGCTGTTCTGCTCCTCAGGAGAC	Mutagenesis PCR F primer mutates L→F at 3' end
SS7mut L→F R	GTCTCCTGAGGAGCAGAACAGCAGTAGTATCAGTG	Mutagenesis PCR R primer mutates L→F at 3' end
SS7mut R→C F	CGTCGACCAGTGCTAATAGCATTCTCATGTACACTGATACT ACT	Mutagenesis PCR F primer mutates R→C mid sequence
SS7mut R→C R	AGTAGTATCAGTGTACATGAGAATGCTATTAGCACTGGTCG ACG	Mutagenesis PCR R primer mutates R→C mid sequence
SS7mut LI→IL F	CGATCACGTCGACCAGTGATACTAGCATTCTCACGAACAC	Mutagenesis PCR F primer mutates LI→IL at 5' end
SS7mut LI→IL F	GTGTTCGTGAGAATGCTAGTATCACTGGTCGACGTGATCG	Mutagenesis PCR R primer mutates LI→IL at 5' end
SS7 2mut F	CGATCACGTCGACCAGTGATACTAGCATTCTCATGTACAC	Mutagenesis PCR F primer mutates LI→IL at 5' end of R→C mutants
SS7 2mut R	GTGTACATGAGAATGCTAGTATCACTGGTCGACGTGATCG	Mutagenesis PCR F primer mutates LI→IL at 5' end of R→C mutants

8.3 **Results: 2A Phylogeny & Sequence Composition**

8.3.1 *In Silico* Searches

A probe of online databases using the 2A conserved C-terminus motif D[V/I]ExNPGP resulted in over four hundred matches from a wide range of viruses and eukaryotic organisms. The complete lists of putative eukaryotic sequences as of September 2014, and for comparison purposes, viral sequences as of December 2011, have been included as Appendix B. For each putative 2A sequence the databank accession number and the host organism/virus is stated. New data is constantly being uploaded as genome sequencing projects progress; therefore these lists represent the known 2A-like sequences at the time of search. In addition, quality control measures since 2011 have resulted in the removal of a large number of sequences from the databases, particularly from the sea-urchin sequence database maintained by Baylor. In general, these sequences have been omitted from Appendix B, excepting those that had been cloned for *in vitro* analyses prior to their removal, in which case they remain but are noted as removed in Appendix B.

8.3.2 Phylogenetic Distribution of Eukaryotic 2A Sequences

The eukaryotic phylogenetic groups containing 2A-like sequences were mapped onto a super-tree of all extant organisms (Figure 8.3). The host organisms with 2A-like sequences have an apparent polyphyletic distribution, pointing to multiple losses or acquisitions of 2A during evolutionary time. 2A-like sequences were found from organisms as varied as unicellular protists, evolutionarily ancient multicellular organisms such as sponges and cnidarian, through to molluscs, arthropods, hemichordates and chordates. Interestingly, virtually all 2A-like sequences were identified from *Animalia*, the only exceptions being the sequences from the unicellular green algae, *Chlorella variabilis*, and a few problematic fungal sequences that although possessing of canonical C-terminal motifs do not otherwise resemble “classic” 2A sequences in that they lack the upstream leucine/isoleucine tract characteristic of active viral 2As.

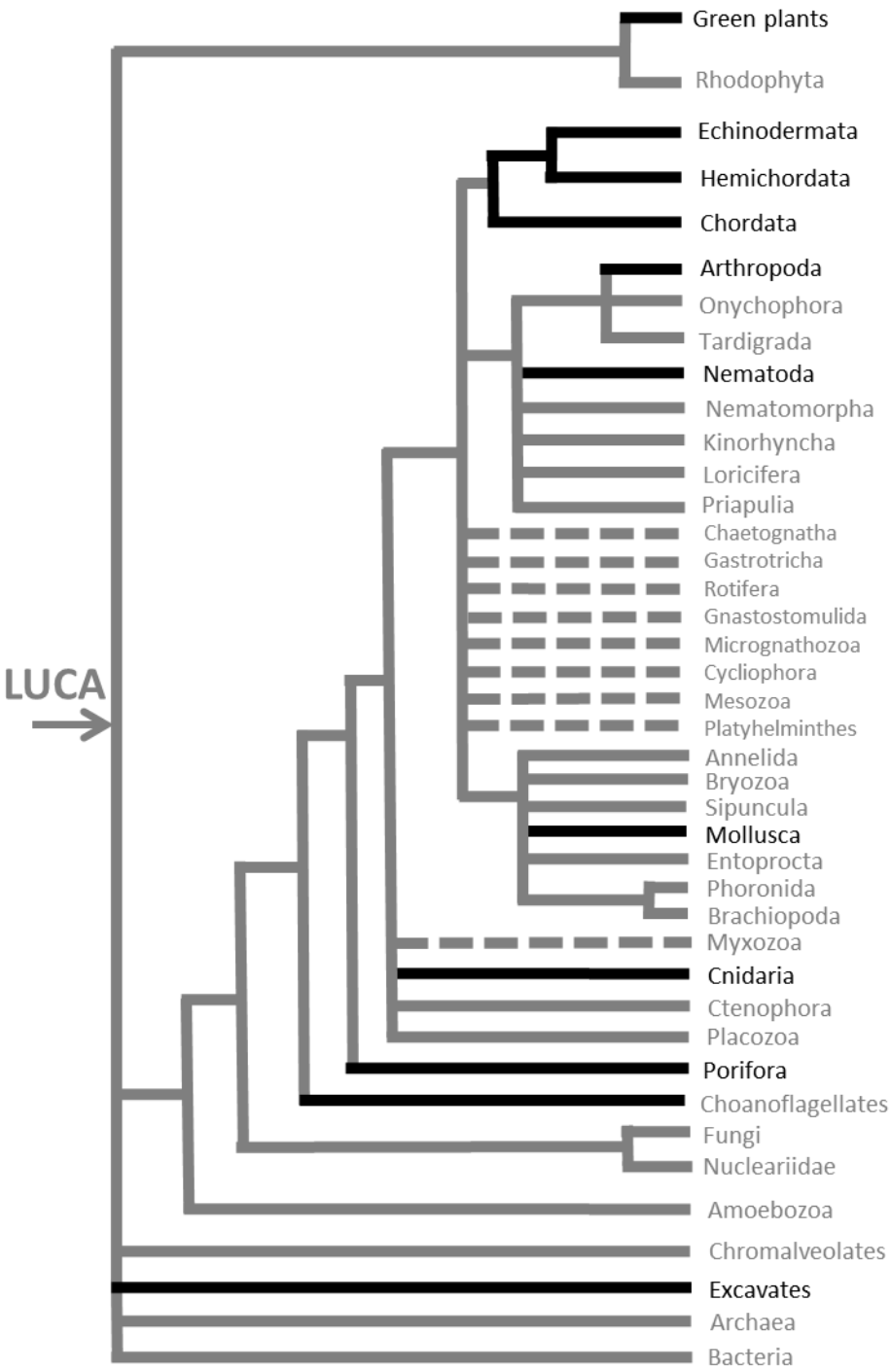


Figure 8.3 The extant phyla with 2A-like sequences.
Phylogenetic super-tree showing all major extant phyla, information from Tree of Life Web Project http://tolweb.org/tree/, not to scale. Phyla with members possessing 2A-like sequences (identified on basis of C-terminal sequence homology with the canonical DxxxNPGP motif) are marked with black lines. Dashed lines mark the most probably positioning of extant phyla whose phylogenetic relations currently undetermined/disputed. Note that phyla marked black do not indicate that all organisms in the phyla contain 2A-like sequences, but merely that a 2A-like sequence has been identified from one or more species in this phylum. Fungi have been marked as grey on this tree as the handful of DxxxNPGP motif containing sequences from this group do not otherwise resemble 2A-like sequences, and where tested have proved inactive *in vitro*. LUCA= last universal common ancestor

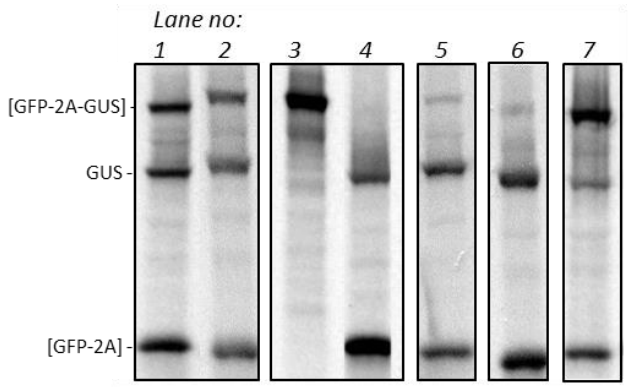
8.3.3 Translational Recoding Assay Results

Putative 2A sequences were tested for translational recoding abilities using the artificial polyprotein reporter *pSTA1* (see Figure 1.11). The constructs were evaluated for relative translational recoding activity relative to the positive control *FMDV 2A* (moderate to high activity). Constructs were scored as either high activity (+++), moderate to high (++), active (+) or low/inactive (-). If the sequences were highly active then two major bands would appear on the SDS-PAGE gel (corresponding to [GFP-2A] and GUS). If intermediate in activity, then there would be three bands (a band of [GFP-2A-GUS] "uncleaved" product would also be apparent in addition to the [GFP-2A] and GUS bands), and if inactive, then one band of "uncleaved" product [GFP-2A-GUS]. Putative 2A sequences selected for testing consisted of the most abundant eukaryotic and viral consensus sequences (details of the selection criteria to follow in Section 8.3.7) and the wildcard sequences listed in 8.2.2.

8.3.4 Viral 2As - Translational Recoding Assays

All but one (namely the DVTINPGP *Rotavirus* sequence) of the viral 2A sequences tested were active *in vitro* (Figure 8.4), albeit with varying levels of recoding activity. Interestingly, despite its widespread occurrence (Figure 8.9) it was the *Rotavirus* DVTINPGP sequence that displayed little/no activity. Further database searching found that this sequence occurred <40 amino acids from the N-terminal of the outer capsid protein VP4 haemagglutinin domain, in a chain that formed part of the crucial haemagluttinin cell membrane binding structure (an EMBL-EBI hosted structure can be viewed at http://pfam.xfam.org/family/PF00426), therefore if this sequence had instigated ribosome skipping, then VP4 would have lost the ability to direct haemagglutination. All other viral sequences tested were at least as active as the positive control *FMDV*. Switching the C-terminal motif of *FMDV 2A* from DVESNPGP to DVELNPGP had no effect on activity levels, and the *Human Rotavirus* DIELNPGP sequence also possessed activity levels on a par with *FMDV*. The *Porcine Teschovirus* DVEENPGP, *Cardiovirus* DIETNPGP and *Drosophila C Virus* DVETNPGP sequences all possessed higher translational recoding capabilities than *FMDV 2A*. These findings were in agreement with those of a previous study where twenty amino acid versions of the *Drosophila C Virus* and *Porcine Teschovirus* 2As were found to instigated higher levels of ribosome skipping than *FMDV 2A* (Donnelly *et al.*, 2001a).

a)



b)

Lane	Name	DXXX	Sequence	Activity
1	*FMDV*	DVES	MDELYKSGSRGACQLLNFLLDKLAGDVESNPGP	++
2	*FMDV mut DVEL*	DVEL	MDELYKSGSRGACQLLNFLLDKLAGDVELNPGP	++
3	*RV*	DVTI	LTNSYTVNLSDEIQEIGSAKSQDVTINPG	-
4	*PTV*	DVEE	MTTMSFQGPGATNFSLLKQAGDVEENPGP	+++
5	*CardioV*	DIET	FTGFFKAVRDYHASYYKQRLQHDIETNPGP	+++
6	*DCV*	DVET	MTQGIGKKNPKQEAARQMLLLLSGDVETNPGP	+++
7	*HRVC*	DIEL	GTGYPLIVANSKFQIDKILISGDIELNPGP	++

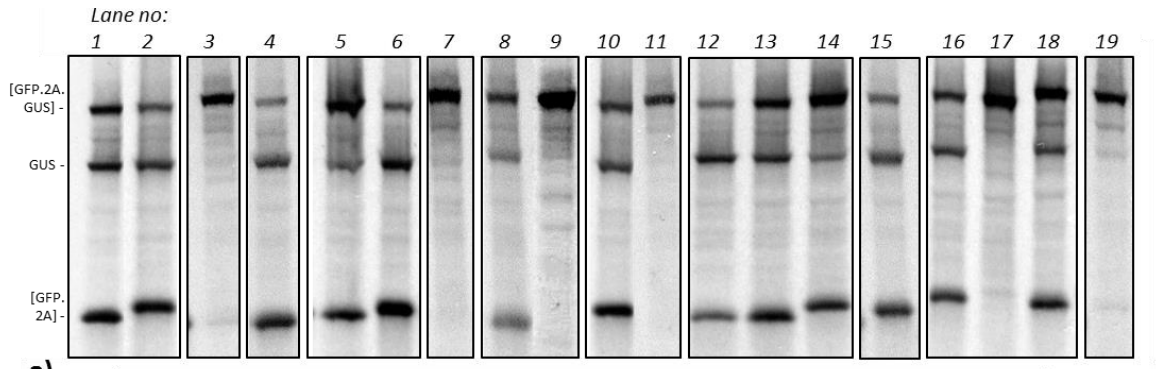
Figure 8.4 Translational recoding analyses of a selection of viral 2As
a) SDS-Page gels of TnTs run on 2A constructs cloned in the reporter *pSTA1*. (Image is a composite showing lanes derived from two gels run together under identical conditions – see Chapter 2., sections 2.2.2 and 2.2.3) **b)** Table listing the constructs by lanes on the gel and recording their relative recoding ability in comparison to *FMDV 2A*. NCBI accession numbers are listed in Table 8.1.

8.3.5 Eukaryotic 2As - Translational Recoding Assays

The majority of eukaryotic 2A sequences tested displayed similar levels of translational recoding capabilities to *FMDV 2A* (Figure 8.5). In contrast to the viral sequences, none facilitated a markedly greater level of ribosome skipping than *FMDV*, although the DVEENPGP sequences from trypanosome (*TB-2*) and algae (*CV-1*), and the DVELNPGP sequences from sponge (*AQ20*) and sea-urchin (*STR-81*) all displayed marginally higher activity than *FMDV* as evidenced by their slightly fainter [GFP-2A-GUS] bands coupled with increased product in their [GFP-2A] and GUS bands, respectively (Figure 8.5).

Six sequences displayed virtually no activity when tested by TnT assay. These were the sea-urchin DVESNPGP sequence *AF-180* and DVEINPGP sequence *STR-1*, the acorn-worm DIETNPGP sequence *SK-45*, the shore mussel DIELNPGP sequence *ME-1*, the fungal DVETNPGP sequence *MO-1*, and the rainbow trout DVEVNPGP sequence *OM-4* (Figure 8.5). Neither sequence length nor C-terminal motif were a reliable predictor of likely translational recoding abilities. The inability to determine 2A activity solely by possession of an appropriate C-terminal motif was

evidenced by the DVESNPGP sequences where *CE-1* was active but *AF-180* inactive, plus the DVEVNPGP sequences where *SS7* displayed moderate to high activity but *OM-4* was inactive. A longer sequence length might have been thought more conducive to high(er) activity (Minskaia *et al.*, 2013). However, one of the shortest sequences tested, *STR-37* at 18 amino acids was active whereas the longer, 30 amino acid length sequences *AF-180* and *OM-4* were inactive. Indeed, the high activity of an 18 amino acid viral 2A, namely *TaV 2A* from *Thosea asigna Virus*, has led to its routine use in genetic engineering (reviewed in Luke and Ryan, 2013).



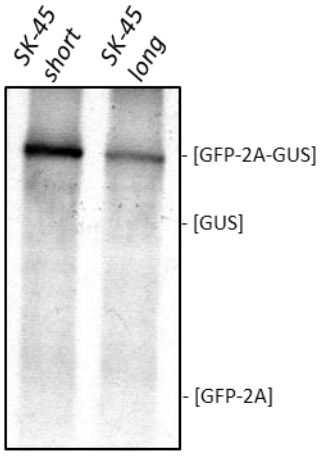
a)

b)

Lane	Name	DXXX	Sequence	Activity
1	*FMDV*	DVES	MDELYKSGSRGACQLLNFLLDKLAGDVESNPGP	++
2	*CE-1*	DVES	LCETPSLPHTTFLKRKLLVRSGDVESNPGP	++
3	*AF-180*	DVES	QNIDVKEADKRHITQTLLTRAGDVESNPGP	-
4	*TB-2*	DVEE	RSLGTCQRAISSIIRTKMLLSGDVEENPGP	++
5	*AM-1*	DVEE	MFMMVFYNAYIPLLIRQANDVEENPGP	++
6	*CV-1*	DVEE	LRLPCSCSTTALIKRMKLLLSGDVEENPGP	++
7	*SK-45*	DIET	MRSNLVLIRDIETNPGP	-
8	*STR6*	DVET	MDGFCLLYLLLILLMRSGDVETNPGP	++
9	*MO-1*	DVET	AAHGQVVLKTNKQGDKYYPDVETNPGP	-
10	*IS-34*	DIEL	MVFPCRQVLFLCSCERISGLKLLLSGDIELNPGP	++
11	*ME-1*	DIEL	YKISLLLLTNSSDIELNPGP	-
12	*AQ20*	DVEL	CDTVSYAVYLLLYFMLLLLLSGDVELNPGP	+++
13	*IS-1*	DVEL	MFLVLLLLLSGDVELNPGP	++
14	*IS-68*	DVEL	MFSLCCQCFDVLSQVLLMSGDVELNPGP	++
15	*STR-81*	DVEL	SKTDLISGQIPPLSELLLLKSGDVELNPGP	++
16	*SS7*	DVEV	QRSRRPVLIAFSRTLILLLLCSSGDVEVNPGP	++
17	*OM-4*	DVEV	TRRPVILAFSCTLILLLFCSSGDVEVNPGP	-
18	*STR-37*	DVER	MTNILLLRSGDVERNPGP	++
19	*STR-1*	DVEI	MFVCAFILISVLLLSGDVEINPGP	-

Figure 8.5 Eukaryotic 2As - translational recoding assays
a) SDS-Page gel of TnTs run on 2As in the reporter *pSTA1.* Image is a composite with lanes derived from three gels run under identical conditions (Chapter 2.2.2 and 2.2.3). Constructs *CE-1*, *CV-1*, *SS7* and *OM4* were previously shown in Chapter 3 (Figure 3.5), *AQ20* and *STR-37* in Chapter 4 (Figure 4.2), and *STR6* and *STR-37* in Chapter 6 (Figure 6.2). In this Chapter, *SK45* will be shown again in Figure 8.6, *STR-37* in Figure 8.7, and *SS7* and *OM4* in Figure 8.8. **b)** Relative recoding ability compared to *FMDV 2A*. For NCBI Accession numbers please refer to Tables 8.1 & 8.3.

To test whether, when presented with a (relatively) short (18 amino acid) inactive sequence, increasing sequence length would confer the ability to cause translational recoding, a short inactive sequence (*SK-45*) was lengthened by in-frame translation of its genomic sequence to the next upstream methionine (34 amino acids). However, it was found increasing length did not confer any translational recoding activity as both the short and long version proved inactive (Figure 8.6).



a)

b)

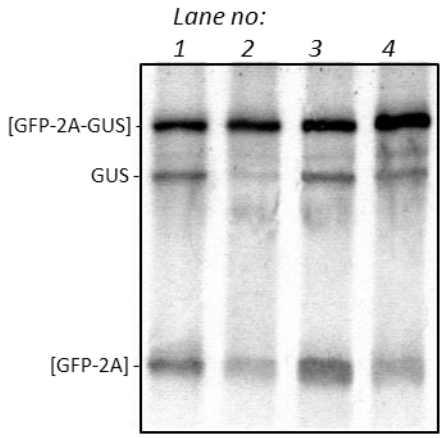
Name	Sequence	Activity
SK-45 short	MRSNLVLIRDIETNPGP	-
SK-45 long	MYDNKNWTFALYLYHCRMRSNLVLIRDIETNPGP	-

Figure 8.6 *SK-45* effect of increasing 2A sequence length
Increasing the length of the acorn worm DIETNPGP sequence did not confer translational recoding activity at 34 amino acids to a sequence that was inactive at 18 amino acids. **a)** TnT products resolved by SDS-PAGE gel electrophoresis from 2A constructs cloned in *pSTA1* vectors, *SK-45 short* was previously investigated by TnT analysis (prior gel analysis shown in Figure 8.5) **b)** comparison of the short and the long *SK-45* sequences. For accession number of *SK-45* refer to Table 8.1.

8.3.6 Mutagenesis on 2A Sequences – Artificial Intermediates

Mutagenesis PCR was employed to create artificial 2A sequences intermediate between two naturally occurring sequences. Firstly, a pair of short (18 amino acid) sea-urchin sequences both displaying moderate/high recoding abilities was selected and the effect of switching their DxxxNPGP motifs examined. It was found that the DVERNPGP motif could be switched for DVELNPGP (R5L) with no impairment to function; however, switching in the DVELNPGP motif with the upstream DVERNPGP sequence (construct I15A) resulted in a slight decrease in ribosome skipping abilities (Figure 8.7).

a)



b)

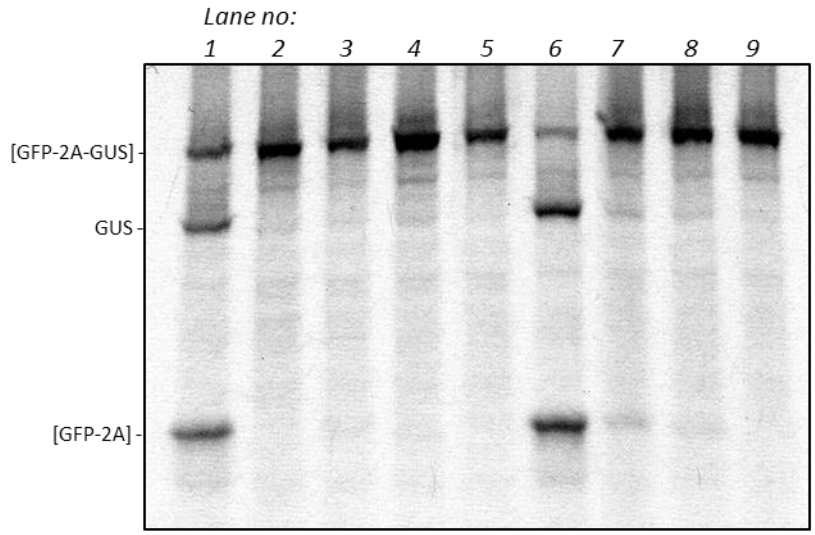
Lane	Name	DXXX	Sequence	Activity
1	*STR-37*	DVER	MTN**I**LLLRSGDVE**R**NPGP	++
2	*I15A*	DVER	MTN**A**LLLRSGDVE**R**NPGP	+
3	*R5L*	DVEL	MTN**I**LLLRSGDVE**L**NPGP	++
4	*STR-140*	DVEL	MTN**A**LLLRSGDVE**L**NPGP	++

Figure 8.7 *STR-37* to *STR-140* artificial intermediate sequences
a) TnT products resolved by SDS-PAGE gel from 2A constructs cloned in *pSTA1* vectors, *STR-37* was first investigated in Chapter 4 (Figure 4.2) and again in Chapter 6 (Figure 6.2). **b)** Table detailing the relative recoding abilities of the sequences tested, using the activity scoring system as described previously. *STR-37* & *STR-140* databank accession numbers are provided in Tables 8.1 & 8.3.

Next, mutagenesis PCR was used to create intermediate between two naturally occurring salmonid 2A sequences, one possessing moderate/high activity (*SS7*) and one virtually inactive (*OM-4*) in an attempt to ascertain which residues were essential to function.

In general, altering any of the residues that varied between the active (*SS7*) and inactive (*OM-4*) sequences, resulted in a loss of activity (Figure 8.8). The only exceptions to this rule were *Mutant D* where LI at positions 25 and 24 (counting back from the N-terminal proline) in the wild-type *SS7* sequence were switched to IL. This substitution increased activity above that of the wild-type (as evidenced by the lesser amount of product in the read-through product [GFP-2A-GUS] band), and *Mutant E* which possessed both the IL switch and L13F. This mutant possessed a very low level of activity with bands of ribosome skippage products ([GFP-2A] and GUS), only just visible as extremely faint bands on the SDS-PAGE gel (Figure 8.8).

a)



b)

Lane	Name	Sequence	Activity
1	*SS7*	QRSRRPV**LI**AFS**R**TLILLLL**L**CSSGDVEVNPGP	++
2	*OM-4*	TRRPV**IL**AFS**C**TLILLLL**F**CSSGDVEVNPGP	-
3	*Mutant A*	QRSRRPV**LI**AFS**R**TLILLLL**F**CSSGDVEVNPGP	-
4	*Mutant B*	QRSRRPV**LI**AFS**C**TLILLLL**L**CSSGDVEVNPGP	-
5	*Mutant C*	QRSRRPV**LI**AFS**C**TLILLLL**F**CSSGDVEVNPGP	-
6	*Mutant D*	QRSRRPV**IL**AFS**R**TLILLLL**L**CSSGDVEVNPGP	+++
7	*Mutant E*	QRSRRPV**IL**AFS**R**TLILLLL**F**CSSGDVEVNPGP	+
8	*Mutant F*	QRSRRPV**IL**AFS**C**TLILLLL**L**CSSGDVEVNPGP	-
9	*Mutant G*	QRSRRPV**IL**AFS**C**TLILLLL**F**CSSGDVEVNPGP	-

Figure 8.8 *SS7* to *OM-4* artificial intermediate mutants
a) TnT products resolved by SDS-PAGE gel from constructs cloned in *pSTA1* vectors, *SS7* and *OM-4* were previously investigated (Chapter 3, Figure 3.5) and earlier in this Chapter (Figure 8.5) from earlier TnT/SDS-PAGE gel analyses. **b)** Table listing the relative recoding abilities of the constructs. Bold, underlined residues correspond to the residues subjected to substitution by mutagenesis. Wild-type *SS7* and *OM-4* were used as the positive and negative controls, respectively. *SS7* & *OM-4* databank accession numbers can be found in Table 8.1.

8.3.7 2A Sequences – Commonalities in Amino Acid Composition?

8.3.7.1 Frequency of Each C-Terminal DxxxNPGP Motif

Analysis of viral and cellular 2A-like sequences revealed disparities in the frequency of occurrence of each DxxxNPGP C-terminal motif between viral and eukaryotic sequences (Figure 8.9). For example, the motif DVESNPGP, the second most common viral motif, was eight-fold more abundant in viruses than eukaryotic cellular sequences (100 as opposed to 12 counts); whereas DVELNPGP, the most frequent cellular motif, occurred twelve-fold more frequently in cellular than viral sequences (47 as opposed to 4 counts) (Figure 8.9). DVTINPGP was the most commonly occurring viral motif, but there was not one instance of this sequence being recorded from cellular sequences, indeed, motifs DVTINPGP, DVTVNPGP, DIEENPGP, DIEANPGP, DVEKNPGP and DVEMNPGP were solely viral, whereas twenty-four different DxxxNPG$^{\downarrow}$P motifs with greater than three occurrences in the dataset were uniquely cellular in origin. Virtually all the C-terminal motifs followed the canonical DxxxNPGP organisation, the exceptions being QVETNPGP, HEIINPGP, NHEINPGP and EVEVNPGP; the sequences with these motifs were categorised as 2A-like by virtue of their N-terminal sequence homology with known 2A sequences.

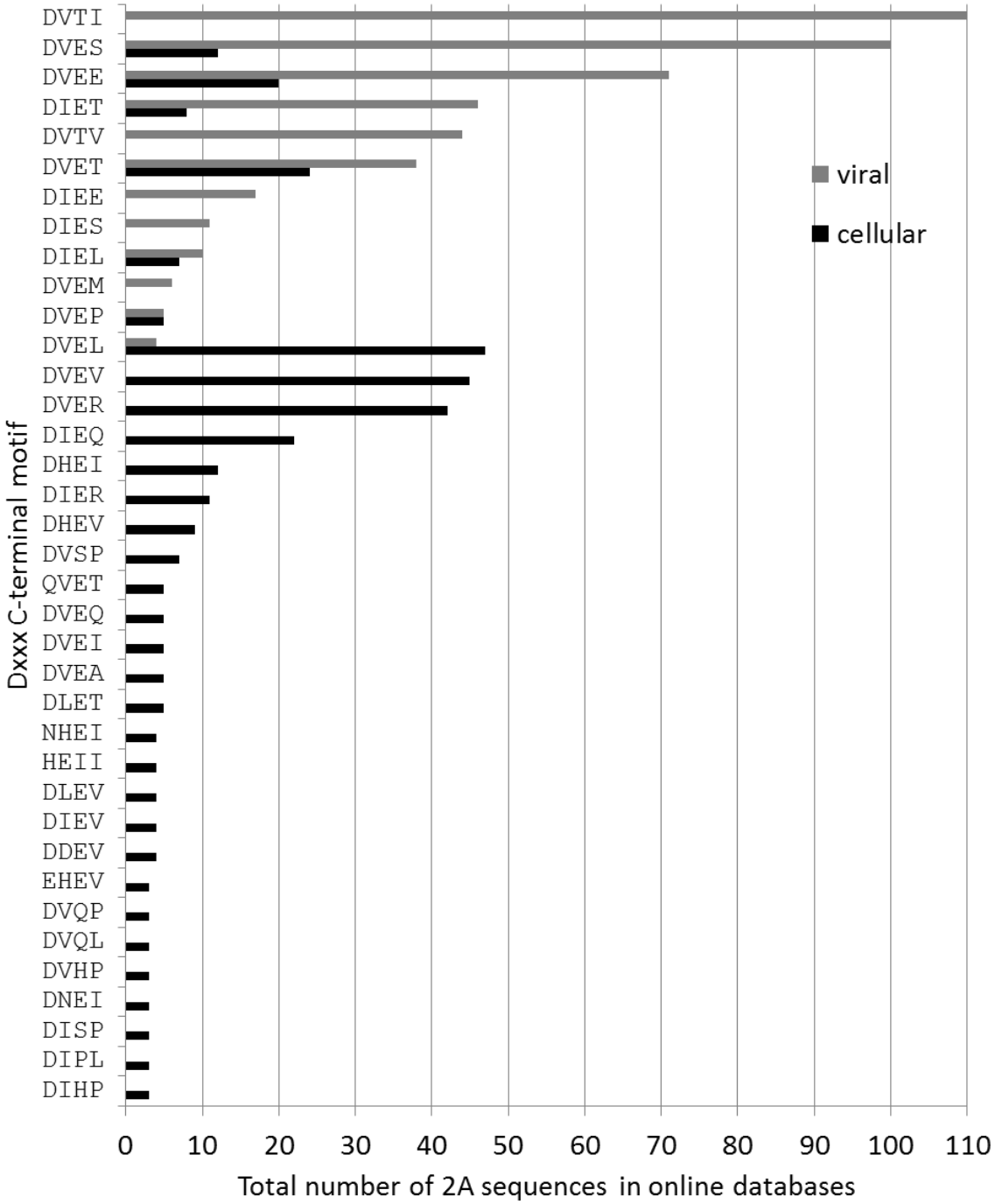


Figure 8.9 The most frequently occurring 2A C-terminal DxxxNPG$^{\downarrow}$P motifs

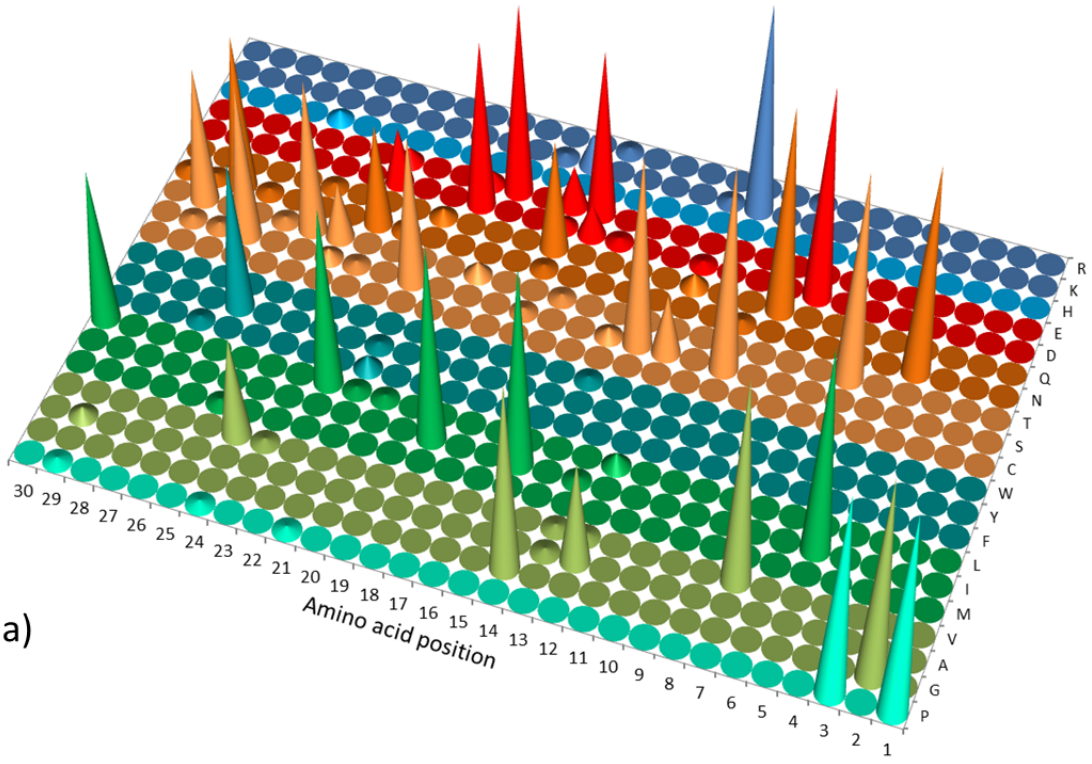
From both in viral and cellular 2A and 2A-like sequences (grouped by the most frequent viral motifs), all motifs occurring in three or more sequences are included. Cellular sequences are shown in black and viral in grey. Only sequences ending in the canonical NPG$^{\downarrow}$P have been included in this analysis. Total counts are shown here in order for the reader to appreciate the number of sequences present in the databases at the time of search (September 2014 for eukaryotic, December 2011 for viral). When expressed as percentage of the total count a similar trend was revealed. Out of interest the author repeated the searches in February 2015, this time slightly altering the search criteria by recording each discrete 30 amino acid sequence from each organism/virus only once, regardless of the number of different proteins it was found to be present within in its host. Remarkably, the overall trend and the most prevalent 2A C-terminal motifs were not found to have varied since these prior analyses, therefore it is felt that this Figure still represents a true indicator of the variety and distribution of 2A sequences recorded.

N-terminal sequence composition was examined in order to determine the upstream 2A consensus sequence associated with each common DxxxNPGP N-terminal motif. Viral and eukaryotic sequences were analysed separately. The results of these analyses are reported as follows:

8.3.7.2 C-Terminal DVTINPGP 2A Sequences

DVTINPGP was the most frequently occurring C-terminal motif (Figure 8.9); however, this motif was unique to viral sequences. It occurred exclusively in the N-terminal region of proteins annotated as VP4 capsid protein from various rotavirus strains. After exclusion of identical sequences (see Chapter 8.2 for exclusion criteria), 46 discrete DVTINPGP sequences were entered into the analysis to find a consensus sequence (Figure 8.10). The reported consensus was: LTNSYTVNLSDEIQEIGSAKSQDVTINPGP. There were 38 direct matches to this sequence, from various *Rotavirus* strains, in the online databases (full list in Appendix B). Therefore this sequence was selected for *in vitro* cloning (TnT gel image in Figure 8.4), where it was found to be inactive. This inactivity was in accordance with the findings of a previous investigation (Luke *et al.*, 2008) which found that the C-terminal glutamate in the C-terminal motif of active 2As (D[V/I]**E**xNPGP) was both highly conserved and essential for 2A function. Hence the substitution of threonine for glutamate may have negated the translational recoding abilities of this sequence. Nevertheless, the inactivity was surprising considering the high number of sequences bearing this motif. However, when the flanking protein was examined, the putative 2A was found to be sited near the N-terminal (within <40 amino acids of the N-terminus) of the outer capsid protein VP4 within one of the two "fingers" of the cell-membrane interaction haemagglutination domain. Translational recoding at this position would cause the virus to lose the ability to cause haemagluttination, interfere with virus entry into host cells, and cause misfolding of VP4 and so interfere with viral packaging. Given the ways which ribosome skipping at this position would result in fitness loss to the virus, negating both its ability to infect host cells, and to successfully package new virions, it is therefore unlikely that this sequence constitutes a presently active 2A sequence *in vivo*.

a)



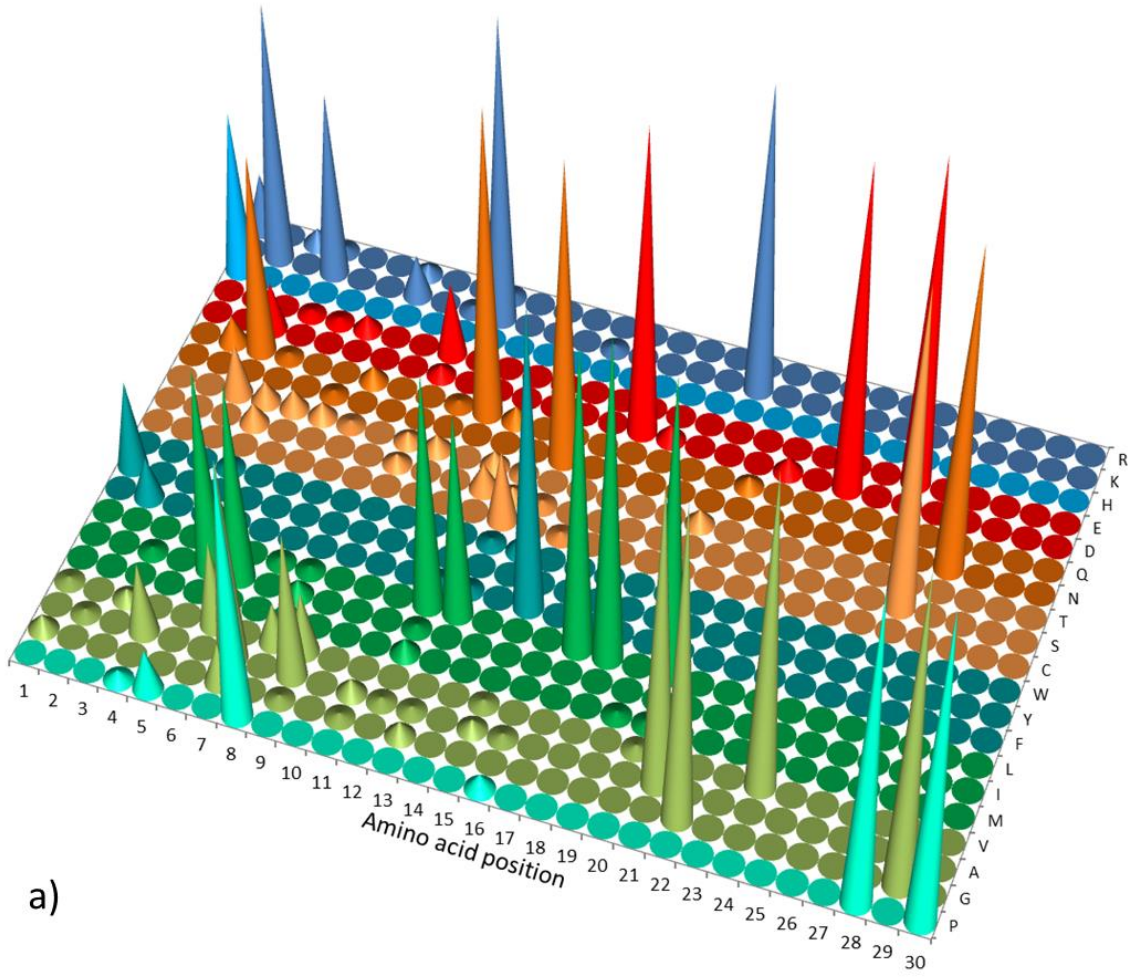
b)

	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
P	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	46	0	46
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	42	0	0	0	0	0	0	0	0	0	0	0	46	0
A	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	23	0	0	0	0	0	0	0	0	0	0	0
V	0	0	0	0	0	0	22	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	46	0	0	0	0	0	0
M	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	43	0	0	44	0	1	0	0	0	0	0	0	0	46	0	0	0	0
L	33	0	0	0	0	0	0	0	39	1	1	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	1	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	32	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	1	1	33	0	0	2	1	0	32	0	0	0	1	0	0	2	42	13	0	45	0	0	0	0	0	0	0	0	0
T	0	29	0	0	1	33	12	0	0	0	0	3	0	0	1	0	0	0	0	0	0	0	0	0	46	0	0	0	0	0
N	1	1	33	1	0	0	0	22	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	46	0	0	0
Q	0	0	0	0	0	1	0	0	0	2	0	0	0	24	0	0	0	0	3	0	0	45	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0	0	12	0	0	36	0	0	2	7	1	0	0	1	0	0	0	46	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	3	0	0	3	41	0	9	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	0	0	1	7	0	0	0	0	1	45	0	0	0	0	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Consensus Sequence:																														
	L	T	N	S	Y	T	V	N	L	S	D	E	I	Q	E	I	G	S	A	K	S	Q	D	V	T	I	N	P	G	P

Figure 8.10 Determining the viral DVTI consensus 2A sequence
a) Cone graph summarising the relative frequency of occurrence of the various amino acids in each position in the 2A sequence for 30 amino acids from the 2A C-terminal proline. **b)** Data-table used to draw the graph, the dataset analysed comprised the DVTI viral sequences with 2A regions longer than 13 residues (see Appendix B).

8.3.7.3 C-Terminal DVESNPGP 2A Sequences

Overall, DVESNPGP was the most abundant C-terminal motif. It was also the second most abundant viral C-terminal motif. However, the number of occurrences was not equally split between viral and eukaryotic cellular sequences, with an 8-fold greater occurrence in viral than eukaryotic sequences. The results of DVESNPGP consensus sequence analysis are reported in Figure 8.11 and Figure 8.12. The viral DVELNPGP consensus was HKQKIIAPAKQLLNFDLLKLAGDVESNPGP. This sequence occurs frequently, with 33 hits from *FMDV* viral sequences, and the C-terminus 14 residues are identical to those from the *FMDV* sequence in *pSTA1*, therefore it was decided to use this *pSTA1 FMDV* sequence for in *vitro analysis* as a representative of viral DVESNPGP sequences. The DVESNPGP eukaryotic sequences were highly variable; therefore the creation of a consensus sequence spanned only the C-terminal 18 amino acids. The consensus was: LLSLLL(V/L)(L/R)(C/S)GDVESNPGP. DVESNPGP cellular sequences LCETPSLPHTTFLKRKLLVRSGDVESNPGP, from the nematode worm *C. elegans*, and QNIDVKEADKRHITQTLLTRAGDVESNPGP, from sea-urchin were shown to be active *in vitro* (Figure 8.5).



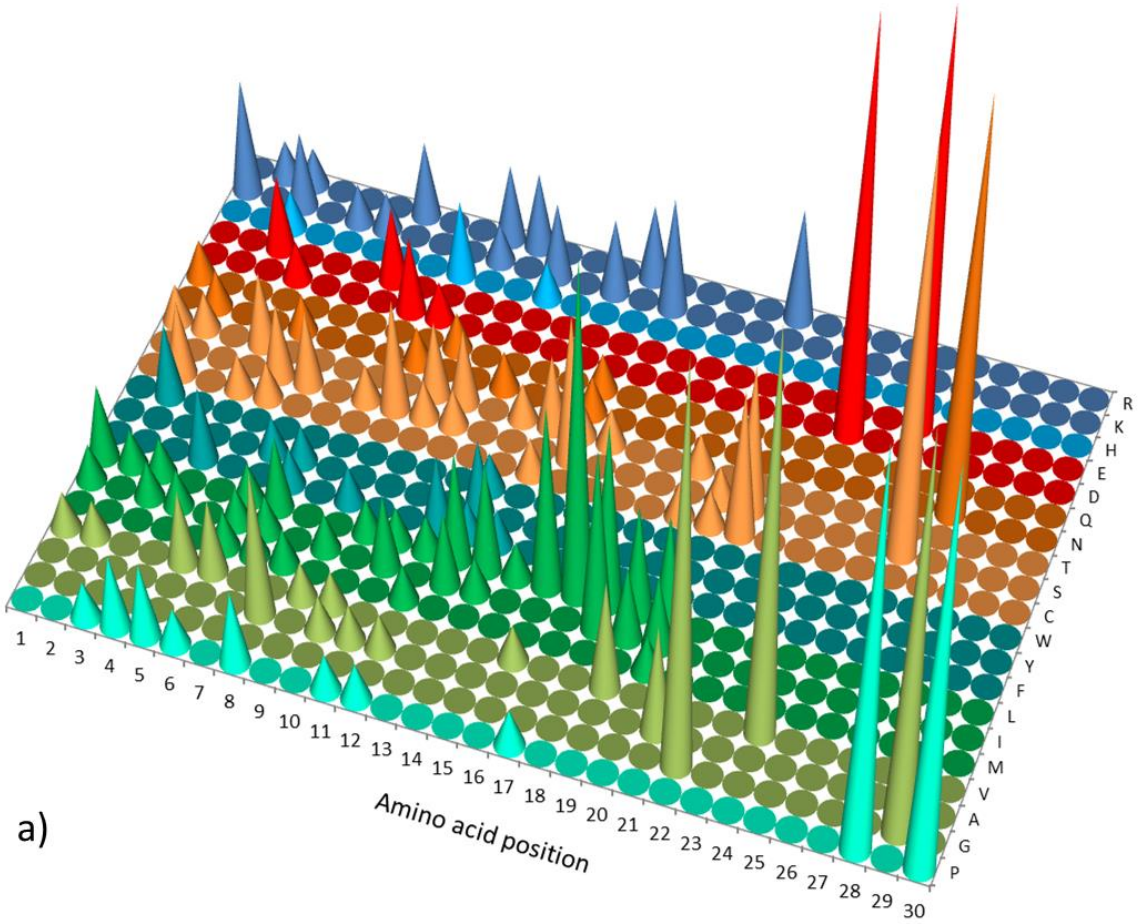
a)

b)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
P	0	0	0	2	8	0	0	51	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	63	0	63
G	3	0	0	0	0	0	8	0	1	0	1	0	3	0	0	0	0	0	0	0	0	63	0	0	0	0	0	0	63	0
A	0	1	0	14	0	0	40	0	29	0	3	1	1	0	2	1	0	0	0	0	60	0	0	0	0	0	0	0	0	0
V	1	0	2	0	0	17	0	8	12	0	0	0	0	0	1	0	0	0	0	1	0	0	0	63	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
I	0	0	1	0	40	39	0	3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	10	0	1	1	0	0	0	46	40	0	0	0	60	63	0	59	0	0	0	0	0	0	0	0	0	0
F	0	8	0	0	0	0	0	0	0	0	0	1	1	0	59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	0	0	0	1	0	0	0	0	0	0	0	0	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	14	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	4	0	0	0	0	2	0	0	5	2	1	0	0	0	0	0	0	0	0	0	0	0	63	0	0	0	0
T	0	0	10	4	4	3	1	0	2	3	0	3	1	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	1	0	0	0	0	0	0	0	58	0	0	0	0	0	0	0	0	0	0	0	0	63	0	0	0
Q	0	5	38	1	0	0	3	0	0	1	59	3	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
D	0	0	9	0	0	0	0	0	2	0	0	0	0	0	59	3	0	0	0	3	0	63	0	0	0	0	0	0	0	0
E	0	0	0	1	1	3	0	0	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	63	0	0	0	0	0
H	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	0	49	0	35	0	0	8	0	1	58	0	0	0	1	0	0	0	0	59	0	0	0	0	0	0	0	0	0	0	0
R	11	0	3	1	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Consensus Sequence:																														
	H	K	Q	K	I	I	A	P	A	K	Q	L	L	N	F	D	L	L	K	L	A	G	D	V	E	S	N	P	G	P

Figure 8.11 Determining the viral DVESNPGP consensus 2A sequence

a) Cone graph summarising the relative frequency of occurrence of the various amino acids in each position in the 2A sequence for 30 amino acids from the 2A C-terminal proline. **b)** Data-table used to draw the graph, the dataset analysed comprised the DVESNPGP viral sequences with 2A regions longer than 13 residues (see Appendix B).



b)

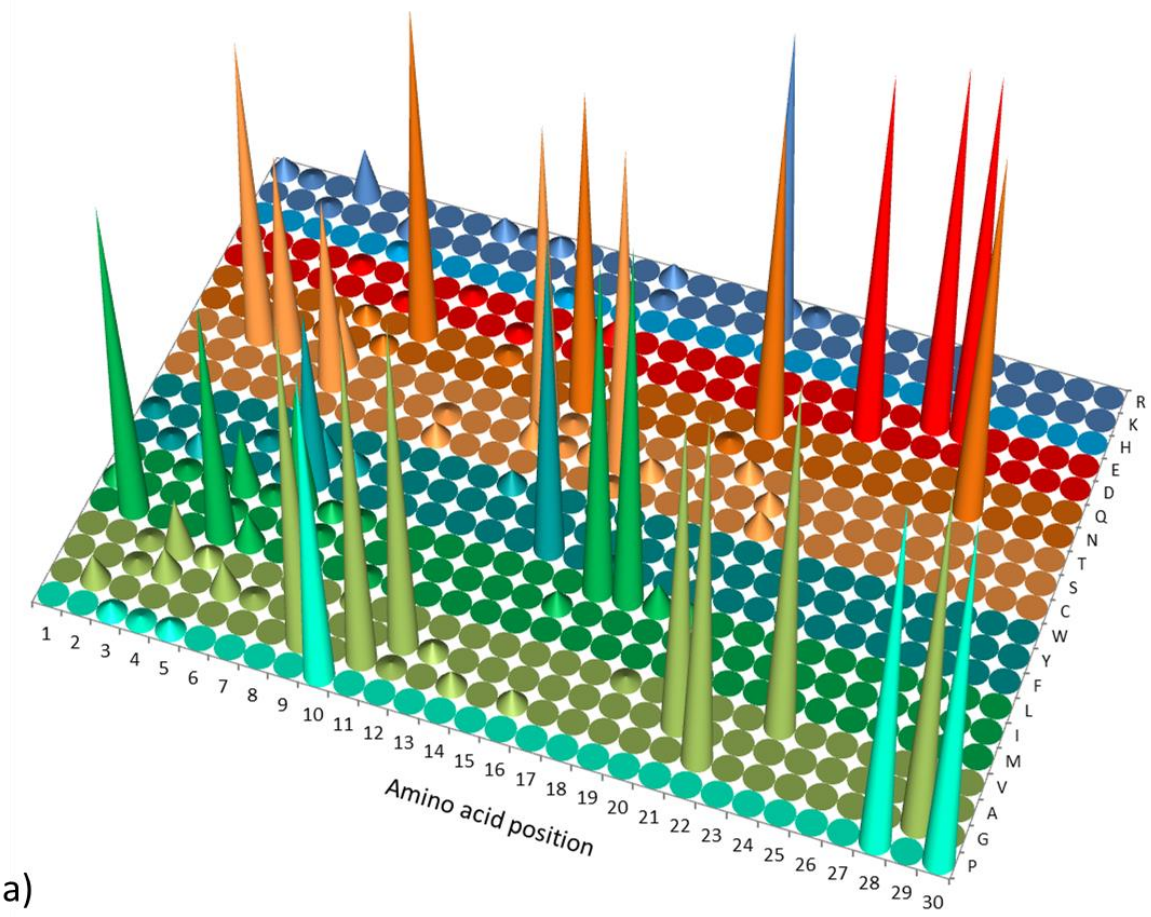
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
P	0	0	1	2	2	1	0	2	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	11	0	11
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	11	0
A	0	0	0	0	0	0	0	4	0	1	1	1	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0
V	1	1	0	0	2	2	0	0	1	1	0	0	0	0	0	1	0	0	3	0	0	0	0	11	0	0	0	0	0	0
M	0	0	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
I	1	0	1	1	0	1	1	0	1	0	2	0	2	1	0	0	0	5	2	2	0	0	0	0	0	0	0	0	0	0
L	2	1	1	0	0	1	2	0	0	1	1	1	3	3	1	5	9	5	3	3	0	0	0	0	0	0	0	0	0	0
F	0	0	0	2	0	0	1	0	1	0	0	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	1	1	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	0	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	2	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	4	0	0	0	0	0	0	0	0	0
S	0	0	0	0	2	2	0	1	3	1	1	0	0	0	4	0	0	0	0	1	4	0	0	0	0	11	0	0	0	0
T	1	1	0	2	0	0	0	0	0	2	2	0	1	2	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
N	0	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0
Q	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	1	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0
E	0	0	2	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0
H	0	0	1	0	0	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	3	0	2	0	1	1	0	0	0	1	0	2	0	2	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R	0	1	1	0	0	0	2	0	0	2	2	0	0	0	2	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0
Consensus Sequence:																														
	K								S				L	L	S	L	L	L	V/L	L/R	C/S	G	D	V	E	S	N	P	G	P

Figure 8.12 Determining the eukaryotic DVESNPGP consensus 2A sequence
a) Cone graph summarising the relative frequency of occurrence of the various amino acids in each position in the 2A sequence for 30 amino acids from the 2A C-terminal proline. **b)** Data-table used to draw the graph, the dataset analysed comprised the DVES eukaryotic sequences with 2A regions longer than 13 residues (see Appendix B).

8.3.7.4 C-Terminal DVEENPGP 2A Sequences

DVEENPGP was present in both eukaryotic and viral sequences, but there were over three-fold more viral than eukaryotic cellular sequences with this motif. The viral consensus sequence was MTTMSFQGPGATNFSLLKQAGDVEENPGP (Figure 8.13). There were direct matches to this sequence from a number of different *Porcine Rotavirus* strains (Appendix B). The cellular consensus was RSLGTCQRAISSIIRTKMLLSGDVEENPGP (Figure 8.14). There were seven matches to this sequence from *Trypanosoma brucei* and *Trypanosoma brucei gambiense* (Appendix B). The viral and eukaryotic DVEENPGP consensus sequences were analysed *in vitro* (Figures 8.4 & 8.5) and were found to be as active (in the case of the trypanosome sequence *TB-2*) or more active (viral sequence *PTV*) than *FMDV 2A.*

a)

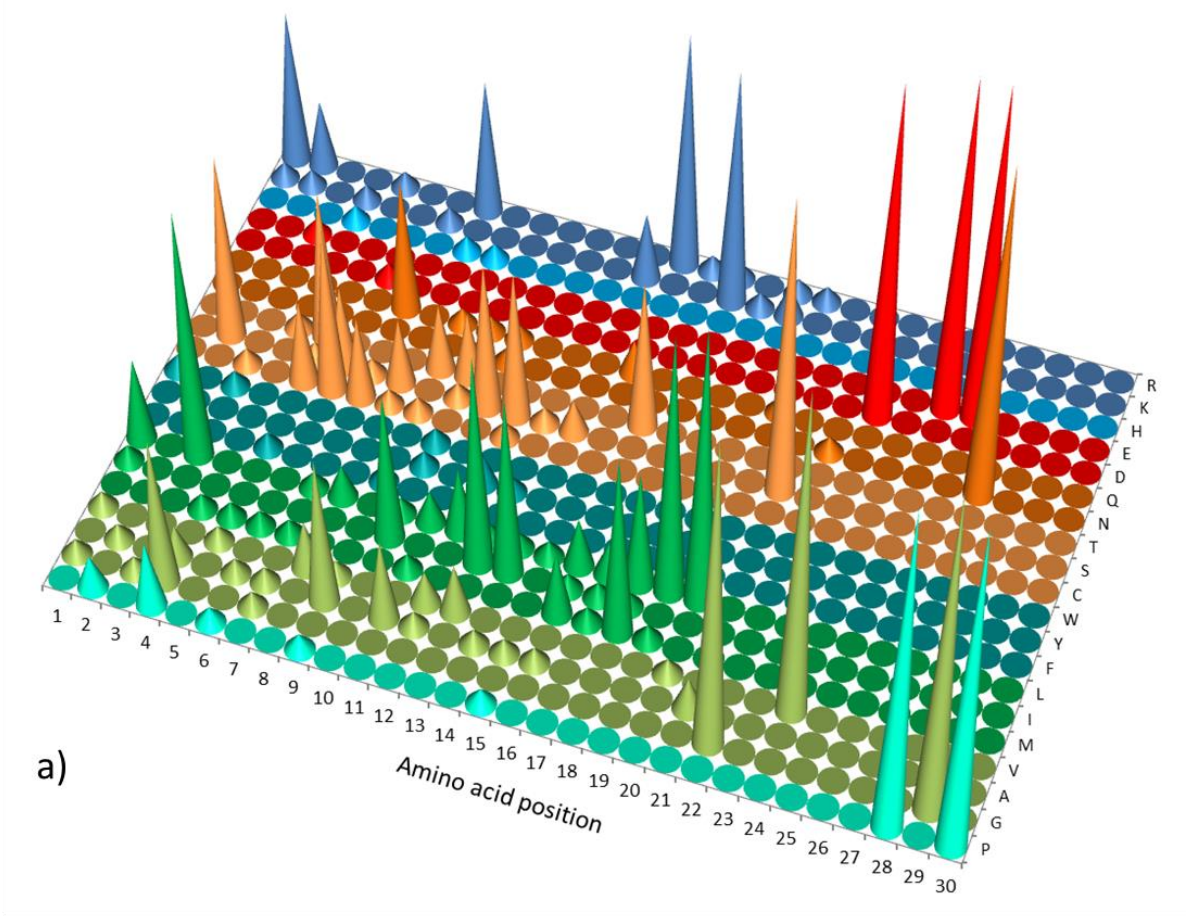


b)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
P	0	0	2	1	2	0	0	0	0	58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	66	0	66
G	0	5	0	0	0	0	0	0	61	0	62	1	0	3	0	3	0	0	0	0	0	66	0	0	0	0	0	0	66	0
A	0	0	1	5	0	6	1	0	0	0	0	61	2	0	0	0	0	0	0	0	59	0	0	0	0	0	0	0	0	0
V	0	0	1	10	2	0	0	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	0	66	0	0	0	0	0	0
M	0	57	0	0	44	6	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	3	0	0	0	1	0	0	1	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	1	3	12	1	23	2	1	0	0	0	0	0	0	0	62	66	5	3	0	0	0	0	0	0	0	0	0	0
F	0	1	3	0	0	1	33	0	0	0	0	0	0	0	59	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	1	0	0	0	1	0	6	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	1	0	0	0	0	4	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	35	0	0	0	1	0	0	4	1	3	59	3	0	0	0	3	0	0	0	0	0	0	0	0	0
T	0	0	55	36	0	11	0	0	0	0	0	0	55	1	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	3	0	2	0	0	0	0	0	0	58	0	0	0	0	1	0	0	0	0	0	0	0	66	0	0	0
Q	0	0	0	0	0	2	0	60	0	0	1	0	0	0	0	0	0	0	0	59	0	0	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0	1	0	0	0	1	3	0	0	0	0	0	0	0	0	0	0	66	0	0	0	0	0	0	0
E	0	0	0	0	1	0	0	0	1	0	0	0	2	3	0	0	0	0	0	0	0	0	0	0	66	66	0	0	0	0
H	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	1	0	0	3	0	0	0	0	0	0	0	0	1	0	0	0	56	0	0	0	0	0	0	0	0	0	0	0
R	3	1	0	9	0	0	0	0	3	1	2	0	0	0	3	0	0	0	3	1	0	0	0	0	0	0	0	0	0	0
Consensus Sequence:																														
	R/I	M	T	T	M	S	F	Q	G	P	G	A	T	N	F	S	L	L	K	Q	A	G	D	V	E	E	N	P	G	P

Figure 8.13 Determining the viral DVEENPGP consensus 2A sequence

a) Cone graph summarising the relative frequency of occurrence of the various amino acids in each position in the 2A sequence for 30 amino acids from the 2A C-terminal proline. **b)** Data-table used to draw the graph, the dataset analysed comprised the DVEENPGP viral sequences with 2A regions longer than 13 residues (see Appendix B).



b)

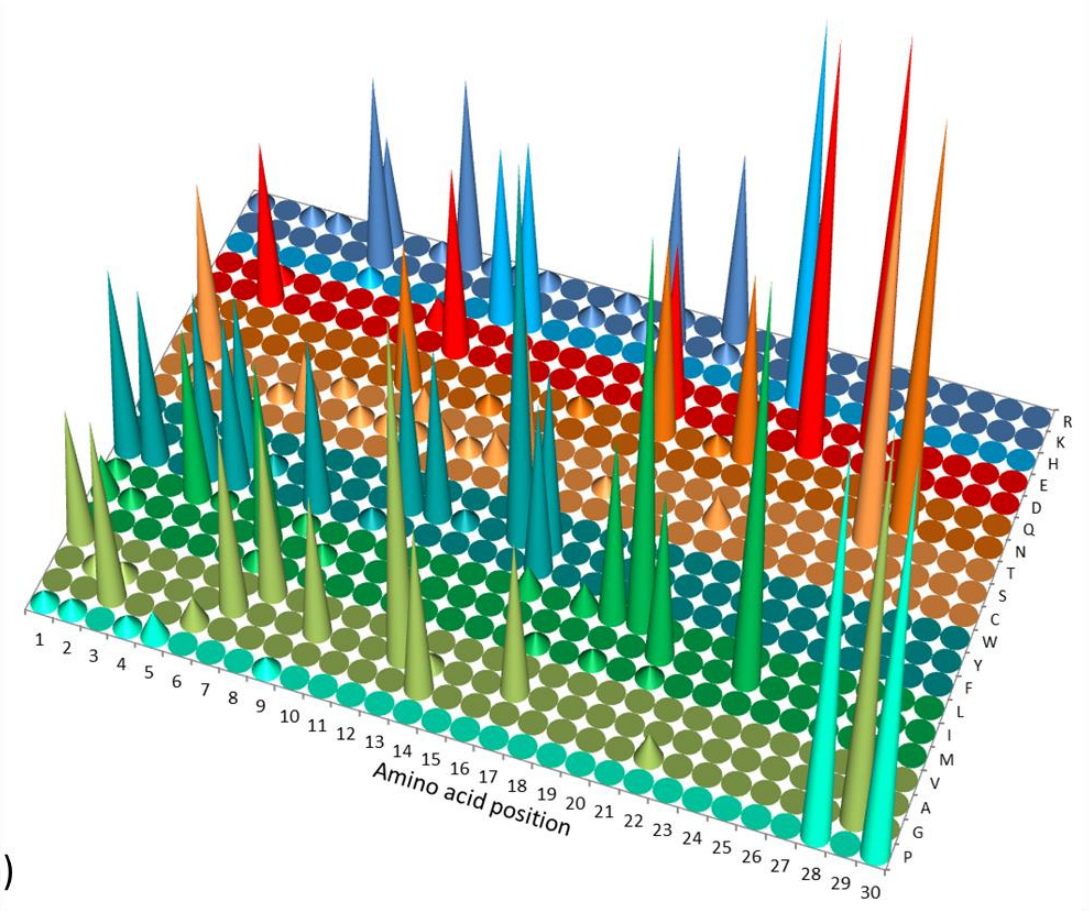
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
P	0	2	0	4	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	20	0	20
G	1	0	1	9	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	20	0
A	0	1	0	2	0	1	1	0	9	0	5	1	0	1	1	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0
V	1	0	0	0	1	0	0	3	1	0	0	2	3	0	0	0	0	0	0	1	0	0	0	20	0	0	0	0	0	0
M	0	0	0	1	1	1	1	0	0	0	1	0	0	0	0	4	1	11	1	0	0	0	0	0	0	0	0	0	0	0
I	1	0	0	0	0	0	0	0	0	9	0	0	13	11	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
L	5	0	15	0	0	0	1	2	0	1	2	4	1	1	1	3	2	7	16	17	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	1	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	1	0	5	10	5	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	11	0	0	1	5	1	4	0	1	9	9	1	2	0	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0
T	0	0	0	1	9	0	0	0	4	4	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	20	0	0	0
Q	0	0	0	0	0	0	8	0	1	1	1	0	0	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
D	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0
E	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	20	0	0	0	0
H	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	1	1	0	1	0	0	1	0	0	0	0	0	0	4	0	0	14	1	2	0	0	0	0	0	0	0	0	0	0	0
R	9	4	0	0	1	0	0	8	0	0	0	0	0	0	14	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0
Consensus Sequence:																														
	R	S	L	G	T	C	Q	R	A	I	S	S	I	I	R	T	K	M	L	L	S	G	D	V	E	E	N	P	G	P

Figure 8.14 Determining the eukaryotic DVEENPGP consensus 2A sequence
a) Cone graph summarising the relative frequency of occurrence of the various amino acids in each position in the 2A sequence for 30 amino acids from the 2A C-terminal proline. **b)** Data-table used to draw the graph, the dataset analysed comprised the DVEENPGP eukaryotic sequences with 2A regions longer than 13 residues (see Appendix B).

8.3.7.5 C-Terminal DIETNPGP 2A Sequences

DIETNPGP was present in eight eukaryotic 2A sequences and forty-six viral sequences. The viral consensus sequence was F(T/F)GF(F/Y)K(A/I)VRDYHASYYKQ(R/L)LQHDIETNPGP (Figure 8.15). There was a direct match to one version of this sequence: FTGFFKAVRDYHASYYKQRLQHDIETNPGP, from *Human CardiovirusD/VI2229/2004* (Genbank Accession ACG61135.2), therefore it was selected for *in vitro* investigation, and in addition sequence *SK-45* from the hemichordate acorn-worm, *Saccoglossus kowalevskii*, was chosen to represent cellular DIET sequences for *in vitro* analysis. The *Cardiovirus* 2A was found to be highly active (higher translational recoding activity levels than *FMDV 2A*, refer to Figure 8.4), but the eukaryotic sequence *SK-45* was inactive (Figure 8.6).

a)



b)

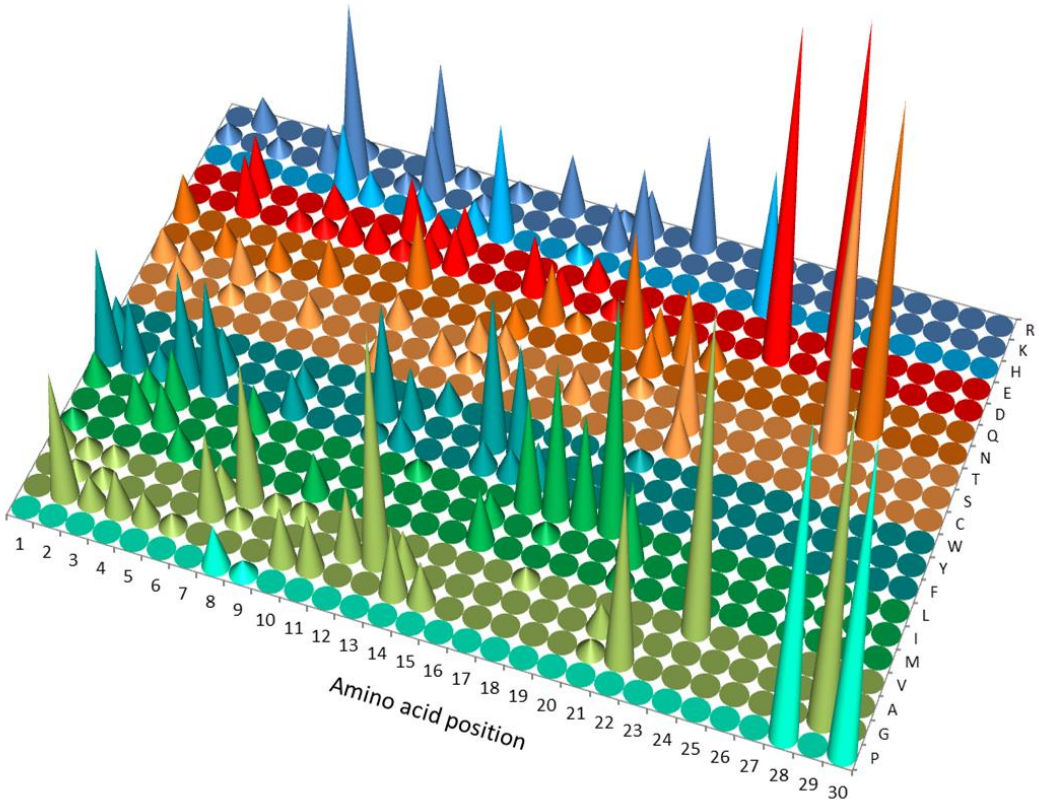
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
P	1	1	0	1	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30	0	30
G	0	0	14	0	0	2	0	0	0	0	0	0	0	13	0	0	0	0	0	0	0	2	0	0	0	0	0	0	30	0
A	0	1	1	0	0	0	14	0	0	11	0	0	26	1	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0
V	10	0	0	0	0	0	0	14	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0
I	3	1	0	0	0	0	14	0	1	0	0	0	0	0	0	0	0	0	0	0	13	0	0	30	0	0	0	0	0	0
L	1	0	0	13	1	0	0	1	0	0	0	0	0	0	0	2	0	2	14	29	0	0	0	0	0	0	0	0	0	0
F	14	13	0	14	13	0	0	13	0	1	0	0	0	0	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	13	1	0	0	0	0	14	13	1	0	28	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	1	0	0	1	4	0	1	1	1	2	1	2	15	0	0	1	0	0	0	2	0	0	0	0	0	0	0	0	0
T	0	13	0	0	0	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30	0	0	0	0
N	0	0	0	0	0	0	0	0	11	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	15	0	1	14	0	0	0	0	0	0	0	0	0
D	0	0	12	0	0	0	0	0	0	14	0	0	0	0	0	0	0	13	0	0	0	0	30	0	0	0	0	0	0	0
E	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30	0	0	0	0	0
H	0	0	0	0	0	1	0	0	0	0	13	14	0	0	0	0	0	0	0	0	0	28	0	0	0	0	0	0	0	0
K	0	0	0	0	0	14	0	0	0	0	0	0	0	1	0	1	15	0	1	0	0	0	0	0	0	0	0	0	0	0
R	1	0	1	1	0	8	0	1	14	1	1	1	0	0	1	0	1	0	14	0	0	0	0	0	0	0	0	0	0	0
Consensus Sequence:																														
	F	T/F	G	F	F/Y	K	A/I	V	R	D	Y	H	A	S	Y	Y	K	Q	R/L	L	Q	H	D	I	E	T	N	P	G	P

Figure 8.15 Determining the viral DIETNPGP consensus 2A sequence
a) Cone graph summarising the relative amount of the various amino acids in each position in the 2A sequence for 30 amino acids from the C-terminal proline. **b)** Data-table used to draw the graph, the dataset analysed comprised the DIETNPGP viral sequences with 2A regions longer than 13 residues (see Appendix B).

8.3.7.6 C-Terminal DVETNPGP 2A Sequences

DVETNPGP was present in both viral and eukaryotic sequences (38 and 24 sequences respectively). For viral sequences the consensus was: FG(D/E)FFKAVRQYHAGYYLL(R/L)LSGDVETNPGP, (Figure 8.16). None of the variants of this consensus sequence occurred in nature, therefore the closest match from a natural sequence, MTQGIGKKNPKQEAARQMLLLLSGDVETNPGP, from *Drosophila C Virus* was cloned and found to be more efficient than *FMDV 2A* at instigating ribosome skipping *in vitro* (Figure 8.4). For eukaryotic sequences the consensus was R(R/S)YTMDGFCLLYLLMILLMRSGDVETNPGP, (Figure 8.17). A highly similar sequence *STR6*, MDGFCLLYLLLRSGDVETNPGP, from the sea-urchin *S. purpuratus* had previously been found to possess ribosome translational recoding abilities *in vitro* (see Chapters 6. & 7.).

a)

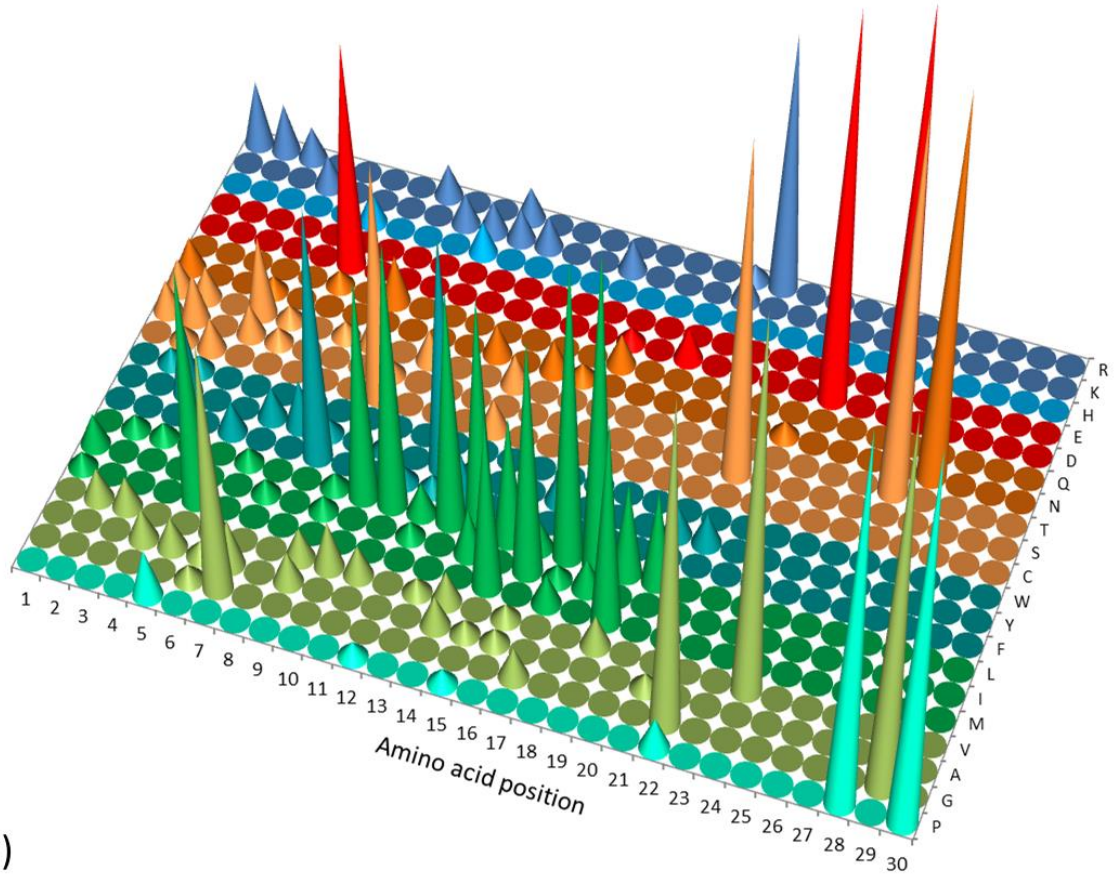


b)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
P	0	0	0	0	0	0	0	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	23
G	0	9	2	3	2	1	0	0	0	4	4	0	0	5	3	0	0	0	0	0	1	13	0	0	0	0	0	0	23	0
A	0	1	1	0	0	0	6	1	0	0	0	5	17	3	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
V	2	1	1	0	0	0	2	10	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	23	0	0	0	0	0	0
M	1	0	0	0	2	0	0	0	0	3	0	0	0	0	0	4	0	0	0	0	2	0	0	0	0	0	0	0	0	0
I	0	0	3	3	0	0	2	0	0	0	0	0	0	0	0	2	0	1	0	0	6	0	0	0	0	0	0	0	0	0
L	2	0	2	4	0	0	3	0	0	0	0	0	1	0	0	0	8	11	8	17	0	0	0	0	0	0	0	0	0	0
F	8	5	2	8	8	0	0	3	0	0	1	2	0	0	2	2	3	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	3	0	0	0	3	0	0	2	0	0	8	4	0	0	11	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0
S	0	2	0	1	0	0	2	0	0	0	0	2	1	3	0	0	2	0	0	0	7	0	0	0	0	0	0	0	0	0
T	2	2	0	3	1	0	0	0	0	2	0	0	2	2	0	0	0	0	1	0	0	0	0	0	0	23	0	0	0	0
N	0	0	2	0	2	0	3	0	0	0	0	0	0	2	0	0	0	0	3	0	0	0	0	0	0	0	23	0	0	0
Q	3	0	0	0	0	0	0	0	0	5	0	0	0	0	4	1	0	8	3	5	2	0	0	0	0	0	0	0	0	0
D	0	0	4	0	1	1	2	2	1	2	3	0	0	4	2	0	1	2	0	0	0	0	23	0	0	0	0	0	0	0
E	0	0	4	0	0	2	0	0	4	2	3	0	0	0	0	2	0	0	0	0	0	0	0	0	23	0	0	0	0	0
H	0	0	0	0	0	5	2	0	2	0	2	8	0	0	1	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0
K	1	0	1	0	3	12	0	1	5	0	1	0	0	0	0	3	6	0	0	0	0	0	0	0	0	0	0	0	0	0
R	0	2	0	0	0	1	0	0	8	1	0	1	0	4	0	1	3	0	8	0	0	0	0	0	0	0	0	0	0	0
Consensus Sequence:																														
	F	G	D/E	F	F	K	A	V	R	Q	Y	H	A	G	Y	Y	L	L	R/L	L	S	G	D	V	E	T	N	P	G	P

Figure 8.16 Determining the viral DVETNPGP 2A consensus sequence
a) Cone graph summarising the relative frequency of occurrence of the various amino acids in each position in the 2A sequence for 30 amino acids from the 2A C-terminal proline. **b)** Data-table used to draw the graph, the dataset analysed comprised the DVETNPGP viral sequences with 2A regions longer than 13 residues (see Appendix B).

a)



b)

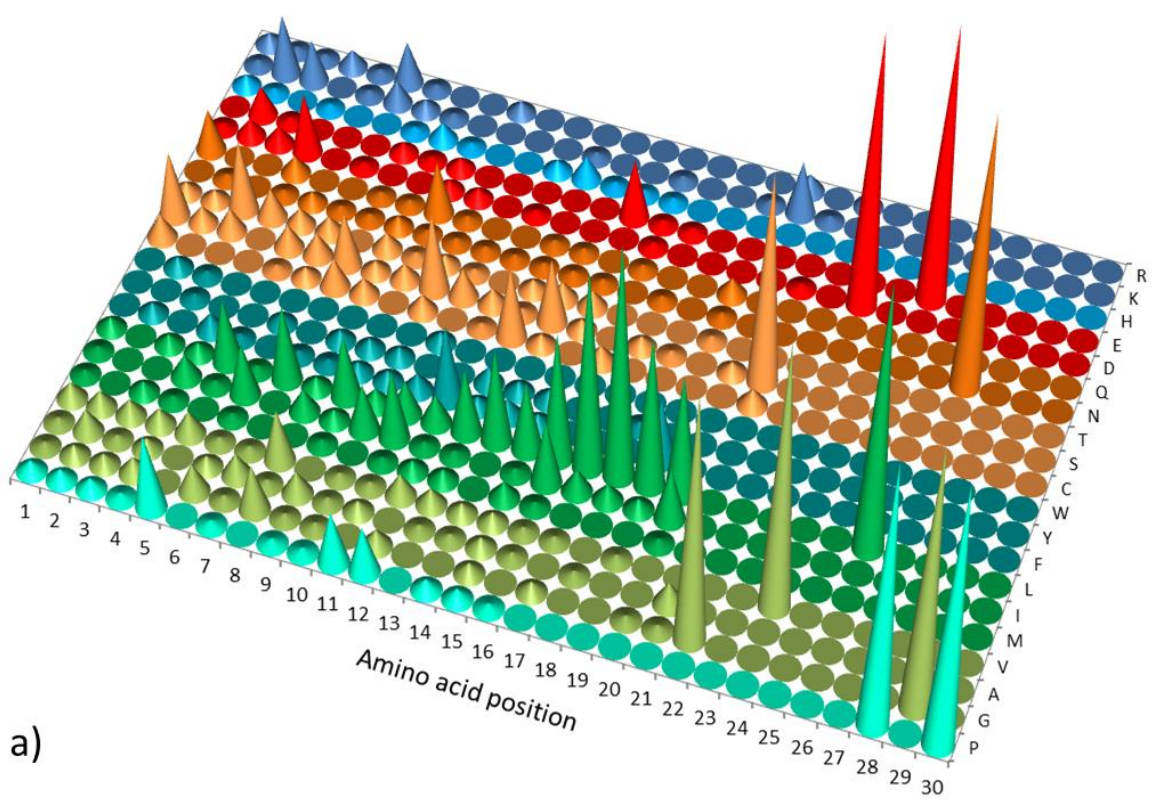
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
P	0	0	0	0	3	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	2	0	0	0	0	0	24	0	24
G	0	0	0	0	0	1	16	0	0	0	0	0	0	0	0	0	2	0	0	0	0	21	0	0	0	0	0	0	24	0
A	0	0	0	2	2	1	3	0	2	0	0	0	0	2	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
V	0	2	2	0	0	2	0	0	2	3	2	0	1	2	0	1	0	0	2	0	0	0	0	24	0	0	0	0	0	0
M	1	0	0	0	15	0	0	0	0	0	0	0	0	0	11	0	2	0	13	0	0	0	0	0	0	0	0	0	0	0
I	2	0	0	0	0	0	1	0	1	0	0	1	0	3	0	15	1	2	0	0	0	0	0	0	0	0	0	0	0	0
L	0	1	1	1	0	1	0	0	1	14	17	2	15	15	8	2	19	20	6	6	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	2	0	0	16	0	0	1	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	5	0	0	2	3	1	0	1	0	15	1	0	0	2	0	0	0	2	2	0	0	0	0	0	0	0	0	0
W	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	1	2	0	0	0	0	0	15	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	2	3	0	2	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	21	0	0	0	0	0	0	0	0	0
T	2	2	0	5	1	0	1	2	0	2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	24	0	0	0	0
N	2	0	0	1	0	0	0	0	0	0	0	2	0	2	1	0	0	0	0	0	0	1	0	0	0	0	24	0	0	0
Q	0	0	0	0	0	1	0	3	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	0	0	14	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	24	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24	0	0	0	0	0
H	0	0	0	0	0	2	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	2	0	0	0	0	2	2	2	2	0	0	2	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
R	4	3	2	0	0	0	0	2	0	0	2	0	0	0	0	0	0	0	1	16	0	0	0	0	0	0	0	0	0	0
Consensus Sequence:																														
	R	R/S	Y	T	M	D	G	F	C	L	L	Y	L	L	M	I	L	L	M	R	S	G	D	V	E	T	N	P	G	P

Figure 8.17 Determining the eukaryotic DVETNPGP consensus 2A sequence
a) Cone graph summarising the relative frequency of occurrence of the various amino acids in each position in the 2A sequence for 30 amino acids from the 2A C-terminal proline. **b)** Data-table used to draw the graph, the dataset analysed comprised the DVET eukaryotic sequences with 2A regions longer than 13 residues (see Appendix B).

8.3.7.7 C-Terminal DVELNPGP 2A Sequences

DVELNPGP was the most frequent eukaryotic DxxxNPGP motif (Figure 8.9); it was also present albeit at a low frequency in viruses (one occurrence from *Equine Rhinitis B Virus 1*, three from *FMDV Asia Strain*). A sequence made by mutating the DVESNPGP to DVELNPGP motif on the *FMDV* sequence from *pSTA1* was chosen to represent DVELNPGP virus sequences for *in vitro* analysis where it was discovered to possess comparable ribosome skipping activity levels to the DVESNPGP *FMDV 2A* sequence. The eukaryotic DVELNPGP consensus sequence was: SKTDPILVLSIFCLLLLLLLSGDVELNPGP, (Figure 8.18). There were no direct matches to this sequence in the dataset, therefore two highly similar sequences: *IS-1*: MFLVLLLLLSGDVELNPGP from the tick *Ixodes scapularis*, and *AQ20*: CDTVSYAVYLLLYFMLLLLLSGDVELNPGP, from the sponge *Amphimedon queenslandica*, respectively, were selected for cloning. Both were moderately active *in vitro*, with comparable levels of ribosome skipping to *FMDV 2A*.

a)



b)

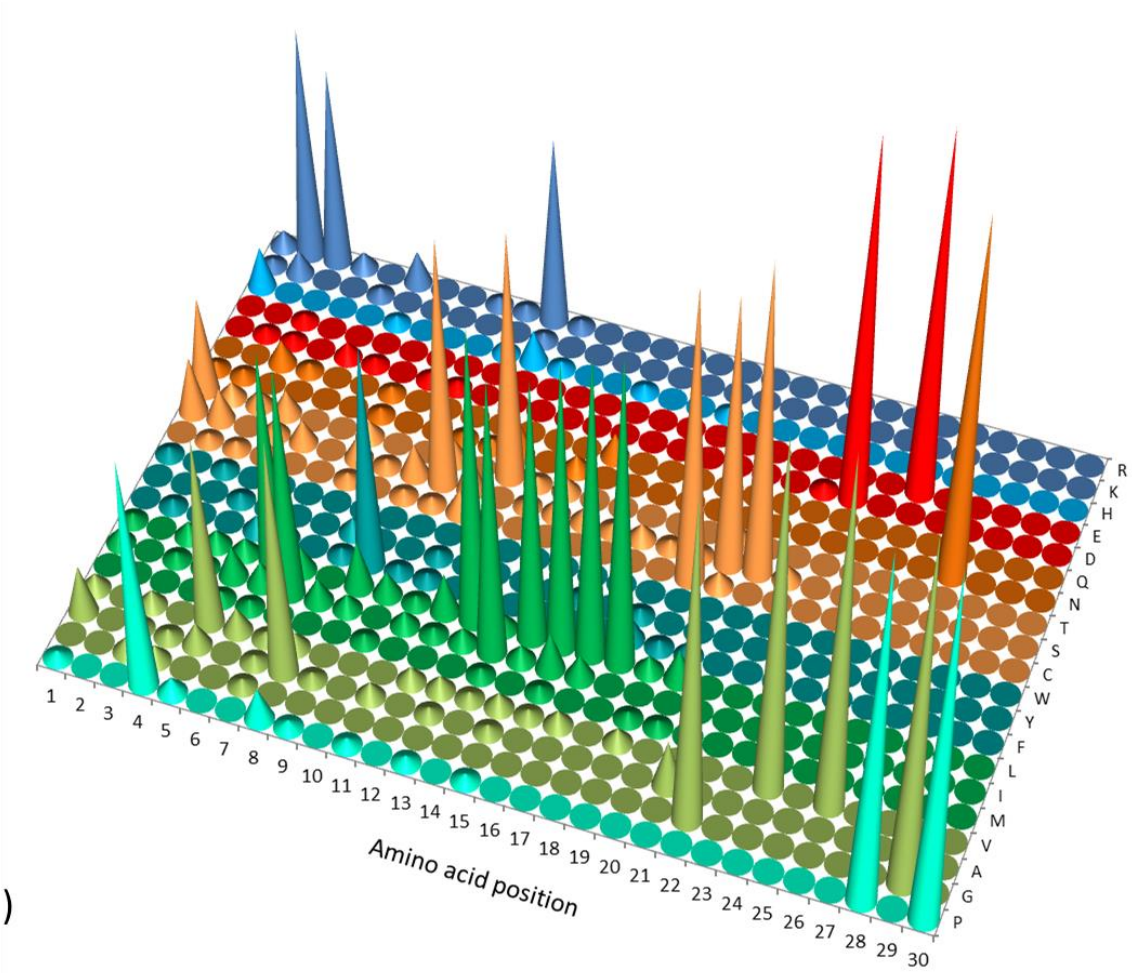
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
P	2	2	2	1	13	0	1	0	1	1	9	8	0	2	2	1	0	0	0	0	0	0	0	0	0	0	0	46	0	46
G	1	2	2	3	0	5	1	8	1	1	0	3	0	0	2	0	2	0	0	1	1	43	0	0	0	0	0	0	46	0
A	1	4	1	2	0	2	4	1	4	1	2	0	1	1	2	1	0	1	0	0	4	0	0	0	0	0	0	0	0	0
V	3	3	3	3	4	1	1	9	0	1	1	4	3	2	2	1	1	0	1	0	0	0	0	46	0	0	0	0	0	0
M	1	0	2	1	0	1	0	0	1	0	1	2	2	0	3	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
I	1	0	0	4	1	9	0	4	2	8	12	1	1	0	1	11	4	3	2	9	0	0	0	0	0	0	0	0	0	0
L	2	0	2	2	11	1	13	2	11	3	6	5	11	16	12	22	34	40	26	21	0	0	0	0	0	46	0	0	0	0
F	0	2	0	3	1	0	1	1	2	2	3	13	2	3	2	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	1	0	0	2	0	2	2	2	0	0	3	2	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
W	0	2	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	4	0	0	0	1	2	5	3	0	3	0	1	12	2	0	3	0	0	0	0	3	0	0	0	0	0	0	0	0	0
S	11	3	4	0	4	4	9	2	3	13	5	3	1	11	2	0	2	1	0	2	36	0	0	0	0	0	0	0	0	0
T	2	2	12	5	3	3	0	3	1	2	1	0	2	3	3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
N	0	1	0	1	2	1	1	2	1	1	1	1	1	1	1	0	1	0	1	0	0	0	0	0	0	0	46	0	0	0
Q	7	0	0	3	0	0	1	1	9	1	0	0	1	1	0	1	0	0	3	0	0	0	0	0	0	0	0	0	0	0
D	1	3	3	10	0	1	0	0	2	2	0	1	0	0	0	1	0	0	0	0	1	0	46	0	0	0	0	0	0	0
E	0	5	1	0	0	0	2	1	2	0	0	0	0	0	10	1	1	0	0	0	0	0	0	0	46	0	0	0	0	0
H	2	1	0	1	0	0	1	3	1	0	0	2	4	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	1	10	7	0	1	4	2	1	0	0	0	0	1	0	0	1	0	0	1	9	1	0	0	0	0	0	0	0	0	0
R	2	1	1	3	1	7	1	0	0	2	0	0	0	0	0	0	0	0	0	3	0	3	0	0	0	0	0	0	0	0
Consensus Sequence:																														
	S	K	T	D	P	I	L	V	L	S	I	F	C	L	L	L	L	L	L	L	S	G	D	V	E	L	N	P	G	P

Figure 8.18 Determining the eukaryotic DVELNPGP consensus 2A sequence
a) Cone graph summarising the relative frequency of occurrence of the various amino acids in each position in the 2A sequence for 30 amino acids from the 2A C-terminal proline. **b)** Data-table used to draw the graph, the dataset analysed comprised the DVELNPGP eukaryotic sequences with 2A regions longer than 13 residues (see Appendix B).

8.3.7.8 C-Terminal DVEVNPGP 2A Sequences

DVEVNPGP was the second most frequent eukaryotic C-terminal motif, but there were no instances of this motif from viral sequences. The eukaryotic consensus sequence was TRRPVLIAFSRTLILLLLCSSGDVEVNPGP (Figure 8.19). There was a direct match to this sequence from Atlantic salmon, *Salmo salar*, (GenBank Accession GU129140.1). It was decided to clone two sequences, both similar, but not identical to this consensus: namely, *SS7*: QRSRRPVLIAFSRTLILLLLCSSGDVEVNPGP, from *S. salar*; and *OM-4*: TRRPVILAFSCTLILLLFCSSGDVEVNPGP from rainbow trout, *Oncorhynchus mykiss* as these two sequences were ideally suited to permit the creation of a series of artificial intermediate mutant sequences (Figure 8.2 & Figure 8.8) in order to investigate the effect of single residue substitution on 2A translational recoding activities. Interestingly, *SS7* proved active whilst *OM-4* was inactive as a translational recoding sequence (Figure 8.8).

a)



b)

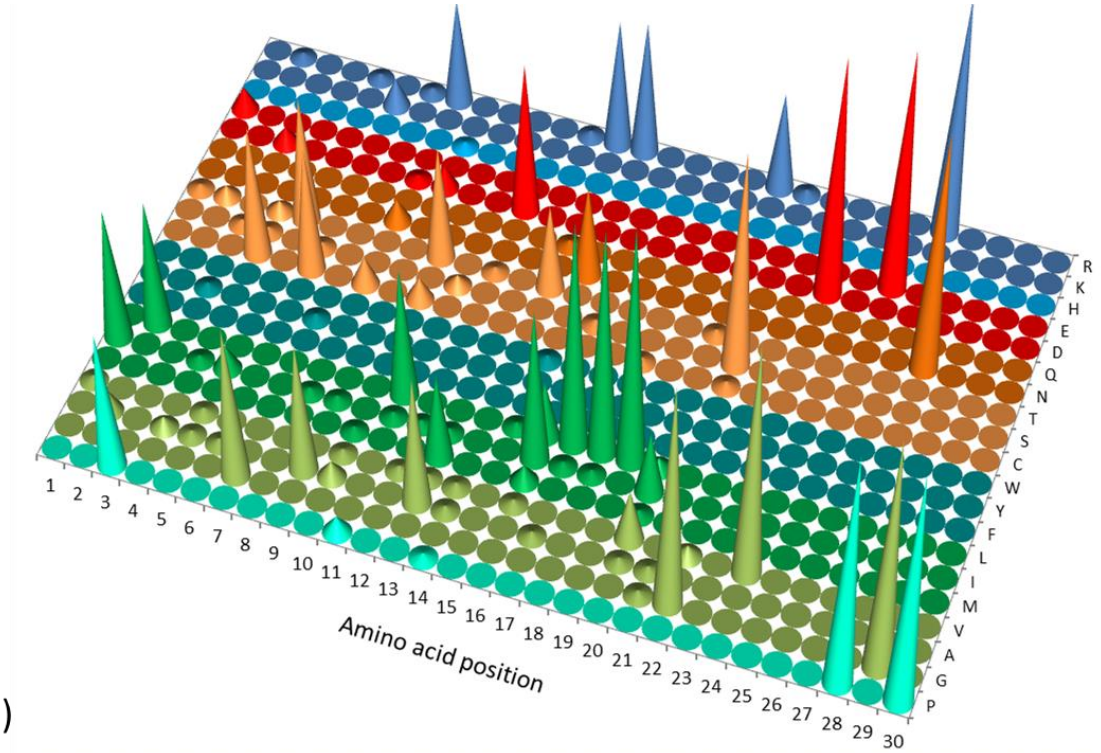
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
P	1	0	0	29	2	0	0	4	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	45	0	45
G	0	0	2	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	42	0	0	0	0	0	0	45	0
A	6	0	1	1	3	0	1	30	1	0	2	0	1	0	1	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0
V	2	0	2	0	24	3	1	2	0	0	0	2	2	1	2	1	2	0	2	0	0	0	0	45	0	45	0	0	0	0
M	0	0	0	0	1	1	0	0	0	1	0	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0
I	2	0	1	1	4	3	29	3	2	0	2	1	2	35	2	5	3	0	0	0	0	0	0	0	0	0	0	0	0	0
L	1	0	2	2	2	27	3	0	6	3	1	5	37	2	33	36	38	39	2	4	0	0	0	0	0	0	0	0	0	0
F	0	1	0	0	2	4	0	1	28	1	1	0	0	0	0	0	0	3	1	2	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	1	0	0	0	2	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
W	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	1	1	0	0	0	1	0	1	1	3	0	0	0	0	0	0	0	37	2	0	0	0	0	0	0	0	0	0	0
S	7	4	0	2	3	0	3	1	4	31	5	0	1	2	2	1	2	1	1	34	39	2	0	0	0	0	0	0	0	0
T	12	2	2	3	0	0	2	0	0	3	0	31	1	2	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
N	1	1	0	0	0	0	1	0	0	1	1	4	0	2	0	0	0	0	0	0	0	0	0	0	0	0	45	0	0	0
Q	0	0	3	1	0	0	1	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	1	1	0	2	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	45	0	0	0	0	0	0	0
E	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	45	0	0	0	0	0	0
H	5	0	0	0	0	1	0	0	0	2	4	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
K	1	3	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R	2	28	24	2	0	4	0	0	1	1	23	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Consensus Sequence:																														
	T	R	R	P	V	L	I	A	F	S	R	T	L	I	L	L	L	L	C	S	S	G	D	V	E	V	N	P	G	P

Figure 8.19 Determining the eukaryotic DVEVNPGP consensus sequence
a) Cone graph summarising the relative frequency of occurrence of the various amino acids in each position in the 2A sequence for 30 amino acids from the 2A C-terminal proline. **b)** Data-table used to draw the graph, the dataset analysed comprised the DVEVNPGP eukaryotic sequences with 2A regions longer than 13 residues (see Appendix B).

8.3.7.9 C-Terminal DVERNPGP 2A Sequences

DVERNPGP represented the third most frequent eukaryotic 2A C-terminal motif, but there were no recorded instances of this motif from viral sequences. The eukaryotic consensus sequence was `ILPCTCGRATLDARRILLLRSGDVERNPGP` (Figure 8.20); there were no direct matches to this sequence in the dataset, but the sea-urchin sequence *STR-37*: matched this consensus sequence for the C-terminal 15 amino acids. *STR-37* was found to be moderately active *in vitro* (Figure 8.5).

a)



b)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
P	0	0	24	0	0	0	0	0	0	0	3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	42	0	42
G	0	0	0	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	1	40	0	0	0	0	0	0	42	0
A	0	4	0	2	1	1	0	0	23	3	0	0	23	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0
V	1	0	0	0	1	0	0	0	0	1	0	1	1	1	0	1	0	0	0	9	0	2	0	42	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	1	0	0	15	0	0	3	0	0	0	1	0	0	0	0	0	0	0	0	0	0
I	23	0	0	1	4	0	0	1	1	0	1	1	1	0	0	27	1	1	0	11	0	0	0	0	0	0	0	0	0	0
L	0	22	0	0	0	0	0	0	0	0	23	0	0	0	1	10	40	40	42	2	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	23	0	22	0	5	0	4	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
S	0	0	1	1	0	1	0	0	0	0	2	0	0	0	0	1	0	0	0	0	38	0	0	0	0	0	0	0	0	0
T	1	2	0	2	22	0	0	0	0	20	0	1	0	15	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	4	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	42	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	3	0	0	0	0	1	4	0	0	26	0	0	0	0	0	0	0	0	0	0	42	0	0	0	0	0	0	0
E	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	42	0	0	0	0	0
H	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R	0	1	0	0	1	0	1	18	0	0	0	0	1	22	23	0	0	0	0	17	1	0	0	0	0	42	0	0	0	0
Consensus Sequence:																														
	I	L	P	C	T	C	G	R	A	T	L	D	A	R	R	I	L	L	L	R	S	G	D	V	E	R	N	P	G	P

Figure 8.20 Determining the eukaryotic DVERNPGP consensus 2A sequence
a) Cone graph summarising the relative frequency of occurrence of the various amino acids in each position in the 2A sequence for 30 amino acids from the 2A C-terminal proline. **b)** Data-table used to draw the graph, the dataset analysed comprised the DVERNPGP eukaryotic sequences with 2A regions longer than 13 residues (see Appendix B).

8.3.7.10 Consensus 2A Sequences – Alignment & Modelling

The viral and eukaryotic 2A consensus sequences for each frequently occurring DxxxNPGP motif consensus sequences were aligned using ClustalX2 in order to ascertain whether there were any generalities in 2A sequence in the upstream sequence associated with 2A (Figure 8.21). In the majority of instances position 9 upstream from the C-terminal proline was a glycine residue, followed by serine in position 10 and a hydrophobic leucine/isoleucine/tyrosine tract between residues 11 and 17. The N-terminal residues (above position 17) were more variable between sequences.

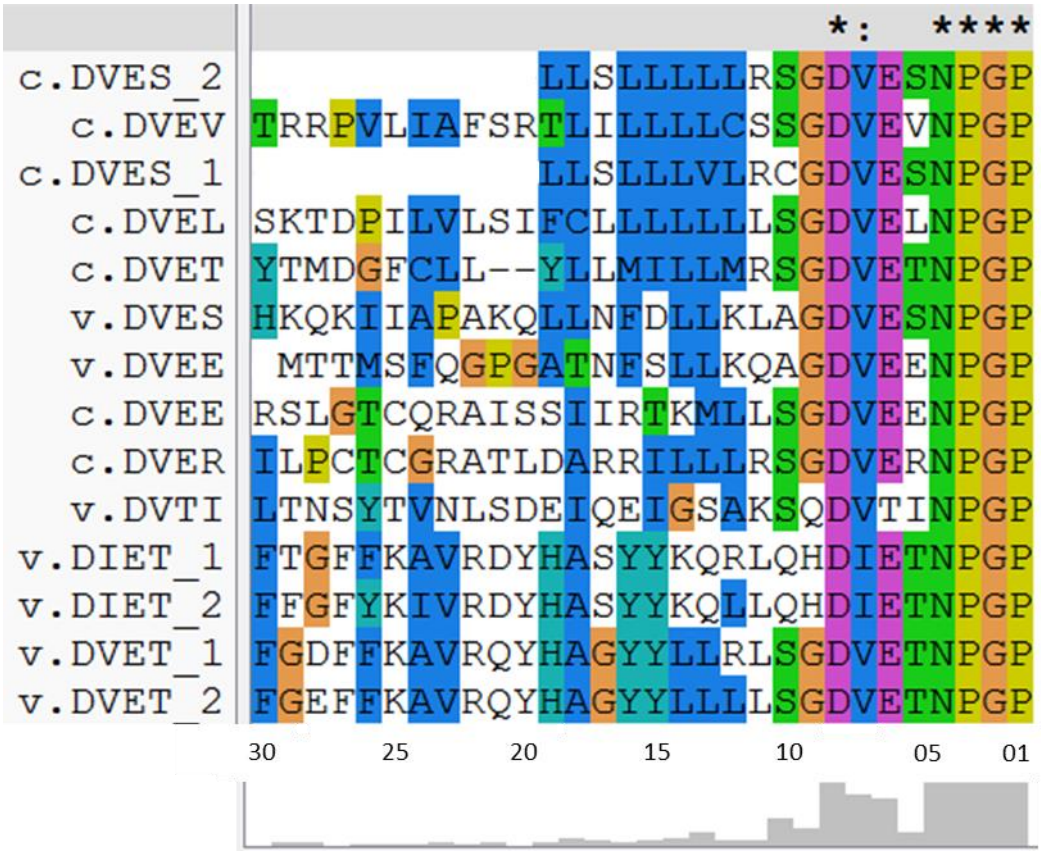


Figure 8.21 Consensus 2A sequences aligned
Alignment of the various 2A DxxxNPGP motif consensus sequences generated in Sections 8.3.7.1-9. Alignment performed using ClustalX2. *=residue completely conserved between sequences, :=residue highly conserved between sequences, v=viral consensus, c=eukaryotic (cellular) sequence, _1 and _2 represent the alternate variants of the respective consensus sequence. As the 2As peptides were various lengths, to standardise numbering, residues were now numbered from the C-terminal proline (1) to the N-terminal residue (<30).

In silico modelling was used to investigate possible 2A structure. Alignment of consensus 2A models with FMDV 2A revealed that there might be conformational similarity between active 2As. The models indicated that active 2As had the propensity to fold into structures with a central α-helical region, whereas inactive 2As were essentially unstructured (TaV 2A being a notable exception to this rule), see Figure 8.22. Interestingly, comparison of the SS7 peptide with OM-4 revealed that the active SS7 might possess this central helix, whereas the inactive OM-4 would not.

Therefore, while not an absolute requirement (see TaV 2A, Figure 8.22c); propensity to fold into a configuration with a central α-helix was apparently a good indicator that a putative 2A sequence would prove active *in vitro*. Moreover, mapping intramolecular distances onto the PEPFOLD models using PyMOL revealed that the conserved "flexible" (in terms of tolerated molecular bond angle between residues) serine-glycine pair typically occurring at positions 10 and 9 from the 2A C-terminal are likely to be positioned approximately 20 Å from the C-terminal. This serine-glycine pair would therefore be spanning the ribosome exit tunnel constriction site (refer to Figure 1.4).

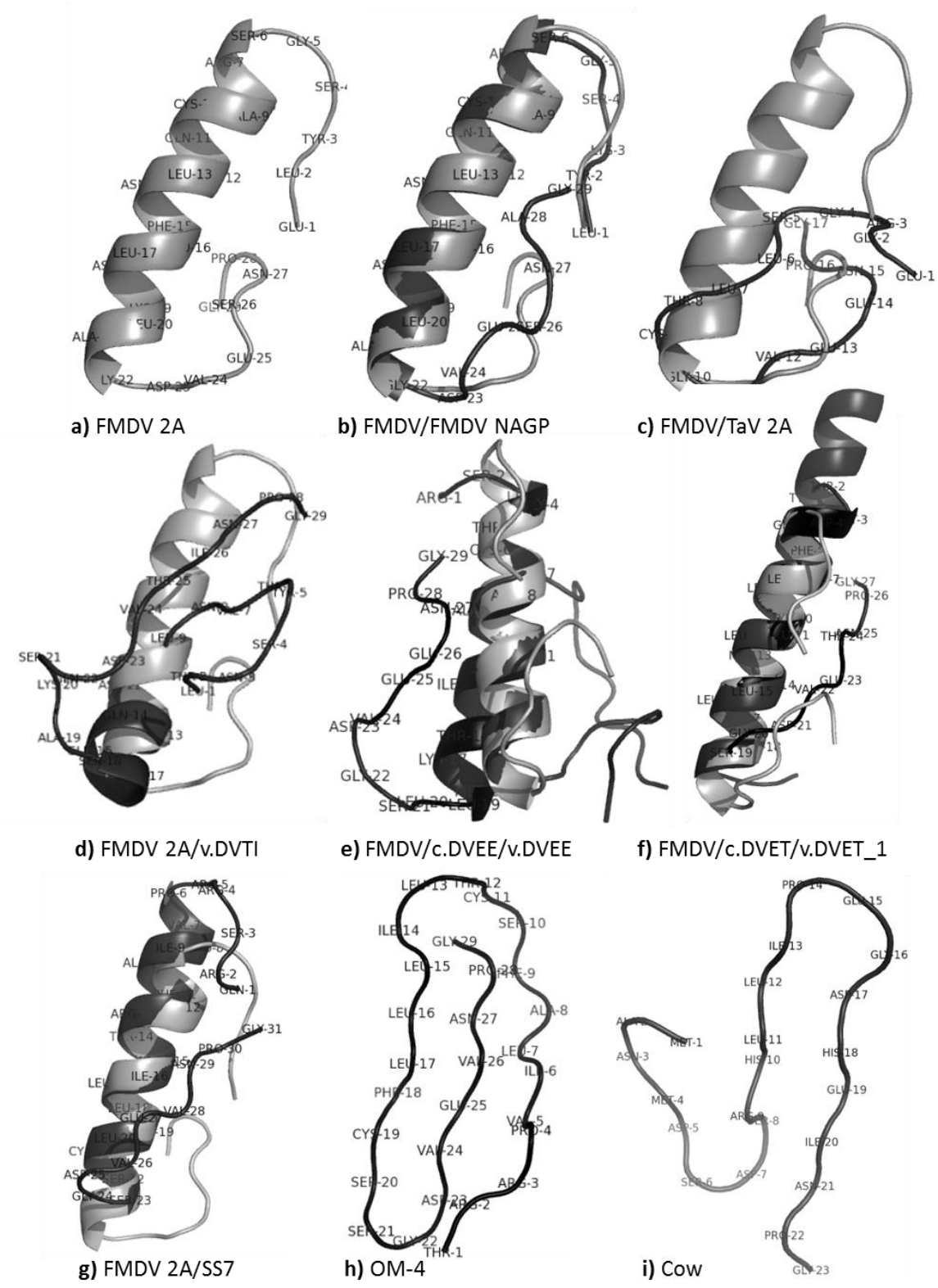


Figure 8.22 Selected 2As – peptide modelling
PEP-FOLD models of *FMDV* and selected 2A sequences – for full Legend please see overleaf.

Figure 8.22 Legend:
In silico modelling of selected 2A peptides, in each alignment, residues are numbered N to C-termini on the uppermost (dark grey or black) peptide **a)** FMDV 2A model, note the central helix and the tight turn at the C-terminal **b)** Overlay of FMDV 2A and FMDV NAGP mutant (dark grey), the central helices align but the C-termini differ in conformation (due to the NPGP to NAGP alanine substitution) **c)** FMDV 2A and TaV 2A (dark grey), TaV was highly active but did not possess a helical conformation, **d)** FMDV 2A and the viral DVTINPGP consensus sequence (*rotavirus 2A*, shown in dark grey) this sequence was largely non-helical and inactive *in vitro.* **e)** FMDV 2A and the viral (dark grey) and eukaryotic (black) DVEENPGP consensus sequences. All were helical and all were active *in vitro*. **f)** FMDV 2A plus the viral (dark grey) and cellular (black) DVETNPGP consensus sequences, again all were helical, variants of these were shown to be active *in vitro*. **g)** The active salmonid sequence SS7 (dark grey, shown here aligned with FMDV 2A (pale grey) possessed a helical architecture whereas the related inactive sequence **h)** OM-4 did not. **i)** The inactive bovine SNAT9 amino acid transporter sequence Cow was essential unstructured, this sequence architecture typifies that of the extremely low activity SNAT9 2As (report in Chapter 5). Sequences were modelled using PEP-FOLD 2011 and visualised with PyMOL v.3.1 as described in Chapter 2.1.3. The caveat of using this method to investigate likely 2A peptide architecture is that the models assume a hydrophilic environment at neutral charge and pH and without spatial constraints, not the confined space within the ribosome tunnel.

8.4 **Discussion**

8.4.1 *In Vitro* Activity Levels– Relationship to Hypothetical Peptide Architecture

In silico modelling of 2A peptide architecture revealed that most (but not all; for example, consider *TaV 2A*) active ribosome skipping 2As possessed the propensity to form an α-helix within their central region. Examination of the models suggested that residue substitution was apparently tolerated as long as this sequence architecture was maintained, thus explaining why N-terminal substitutions (upstream of the helix) were tolerated whereas those occurring nearer the C-terminal (within the helix) were not. Each C-terminal motif possessed a disparate architecture within the motif region, and it may be that each motif helped align their respective upstream helix in the position necessary for it to stall in its passage through the ribosome tunnel. This would explain why switching C-terminal motifs resulted in lower levels of ribosome skipping activity and suggests that the universally inactive -NAGP- mutation (modelled here on *FMDV 2A*, see Figure 8.22b) may function by altering the positioning of the upstream helix, negating function. The conserved serine-glycine pair at positions 10 and 9 from the 2A C-terminal was likely to be spanning the tunnel constriction site. An active 2A nascent chain is thought to transitorily halt in its passage of the exit tunnel, but the residues spanning the constriction site at this time cannot undergo molecular interactions with the constriction site proteins that would further delay the chain's progress, therefore the conserved serine-glycine doublet may result from the need to place "flexible" residues at these positions that can squeeze past the arms of the constriction without interaction. Using these models it was not possible to identify the specific residues essential for ribosome skipping.

From the PEP-FOLD models one could predict that the *Rotavirus* DVTINPGP consensus sequence would prove inactive. This was indeed the case, but this inactivity was curious insofar as DVTINPGP was the most abundant viral C-terminus motif (Figure 8.9). One might logically assume that the most widespread 2A-like sequence would also be amongst the most efficient. However, as it occurred with a structural capsid protein, here 2A-directed ribosome skipping resulting in truncated protein products would prove deleterious to the virus. Hence high occurrence did not correlate with high activity levels. It is also interesting to note that there are no records of eukaryotic DVTINPGP sequences (Figure 8.9 & Appendix B); perhaps DVTINPGP sequences were too inefficient to be sequestered for use in regulation of cellular processes. As might be expected, all other naturally occurring viral sequences investigated were active. Additionally, the *FMDV* DVELNPGP mutant was also active, showing that in this instance, the activity of the upstream context of the sequence was unaffected by a change in DxxxNPGP motif.

Despite having been identified merely on the basis of their possessing the canonical 2A C-terminal DxxxNPGP motif, the majority of tested eukaryotic 2A sequences were found to initiate translational recoding, hence validating the 2A search strategy based on cataloguing hits to the 2A C-terminal motif. However, none of the eukaryotic 2As were as efficient as the most active viral 2As (such as the *PTV 2A* sequence), instead displaying only moderate ribosome skipping abilities. Perhaps, as in some viruses (discussed in Chapter 1.10), the function of eukaryotic 2As was to regulate translation through partially effective ribosome skipping (generating both the full-length protein and the ribosome skip products). Therefore high ribosome skipping activity would not be a trait under positive evolutionary selection, and intermediate ability 2As would be the most valuable to their eukaryotic hosts.

The existence of a pair of salmonid 2A sequences, one active (*SS7*), one inactive (*OM-4*) revealed the ease with which ribosome skipping activity can be lost through mutation of any of several particular residues. These sequences suggest that as active eukaryotic 2A sequences exist today, they must have been acquired in the recent past (as a single mutation could render them ineffective) or are being retained through positive evolutionary selection as they are actively engaged in regulating both protein translation and protein trafficking (see Chapter 7.) in their eukaryotic hosts.

8.4.2 2A Peptide Phylogeny

The absence of 2A-like sequences in archaeal and bacterial entries in online databases was not unexpected as the working hypothesis of 2A activity states that 2A action is the result of interaction between the nascent peptide and the eukaryotic ribosome exit tunnel (see Chapter 1.3) Fungal and plant 2As are notable by their absence, whether this is due to a bias favouring the sequencing of “animal” genomes, or whether active 2As are not found in plants or fungi, it is too

early to say. Currently the only exceptions are a few dubious 2A-like sequences from fungi and three 2As from *Chlorella variabilis* NC64A (one of which was tested and found to be active *in vitro*, see Figure 3.5 and Figure 8.5), a green unicellular alga which is a facultative protozoan symbiont (genome published by Blanc *et al.*, 2010).

It was hoped that by recording the eukaryotic host species with 2A and 2A-like sequences it would be possible to reconstruct the likely evolutionary origins of 2A by tracing their host species phylogeny. However, the 2A sequences followed a polyphyletic distribution with either multiple losses or acquisitions during evolutionary history.

Therefore, in eukaryotes, as with viruses (Luke *et al.*, 2008), 2A may be acquired by horizontal gene transfer (HGT), perhaps through virus vectors. Certainly, there is some circumstantial evidence. Firstly, an alignment of consensus viral and cellular 2A sequences revealed some commonalities of active 2A sequences (Figure 8.21) which could point to either a common origin or to convergent evolution due to a common function. Secondly, it is interesting to consider the ecology of the host species, in particular the predominance of aquatic and parasitic species with 2A sequences (Appendix. B.). Parasitism provides an ideal setting for the transfer of genetic material between host and parasite and *vice versa*. For example, the *Chlorella* genome is known to contain suites of structural and metabolic genes obtained by viral gene transfer from fungi and its host protist (Blanc *et al.*, 2010). A shrimp virus, namely *Infectious Myonecrosis Virus* (*IMNV*), containing two 2A-like sequences is more closely related to a virus of parasitic protozoans (*Giardiavirus*) than to arthropod viruses. Did the virus acquire 2A from a protist, or did the protozoan trypanosomes acquire their 2As through viral infection? For the 2As found in marine hosts, as discussed in Chapter 3.5, certainly, the marine environment is rich in potential picornavirus vectors. However, 2A transfer vectors and the original host or hosts where translational recoding 2As evolved remain unidentified and despite the high occurrence of eukaryotic 2As, it is not possible to determine whether 2A is in essence a viral "trick" co-opted by cells, or whether it is a part of the metabolic regulatory mechanism of some eukaryotic cells that has been adopted by viruses.

Concluding Remarks

'There is nothing like looking, if you want to find something. You certainly usually find something, if you look, but it is not always quite the something you were after.' *The Hobbit* - JRR Tolkien, 1937

This thesis constitutes the first investigation into active translational recoding 2A peptides from eukaryotic proteins, and built on the previous identification and characterisation of 2As from viruses and virus-like retrotransposon elements. I identified novel eukaryotic 2A sequences, tested a representative sample for activity *in vitro*, undertook phylogenetic analyses and used *in silico* modelling to investigate their possible structural characteristics. In addition, the eukaryotic 2A sequences were screened *in silico* to ascertain their potential abilities to direct protein targeting by functioning as signal peptides. Likely signal peptide candidates were cloned into fluorescent reporter constructs and investigated using cell culture transfections.

Prior to the work documented in this thesis, there was a single report of 2A-like translational recoding sequences from a non-viral origin. Heras *et al.* found active 2A-like sequences within trypanosome non-LTR retrotransposons (Heras *et al.*, 2006) mobile genetic elements with many structural and evolutionary similarities to retroviruses (reviewed in Jurka *et al.*, 2007). Following this report, the Ryan laboratory compiled a list of approximately thirty additional putative eukaryotic 2As from sequences deposited in online public domain databases. Due to the rapid advances in genome sequencing in recent years it was felt that this list was no longer exhaustive. I repeated the search, to catalogue over four hundred putative eukaryotic 2As. I have tested a selection for translational recoding abilities *in vitro*, and determined that in general they mediate ribosome skipping similarly to viral 2As.

The eukaryotic 2As were found in a wide range of eukaryotic organisms in association with a limited number of protein types. Proteins comprising 2As include non-LTR elements, ankyrin-repeat protein-protein interaction motifs, NLR-like innate immune cascade receptor proteins, and SNAT9 sodium-dependent neutral amino acid transporter proteins.

Active 2A sequences were found in organisms from all the major eukaryotic kingdoms excepting fungi, albeit that green plants were represented by a single species of unicellular green alga. The SNAT9-associated 2As were only recorded from placental mammals (see Chapter 5.3.1) whereas the remainder of the 2As were predominantly sourced from marine and aquatic invertebrates (see Chapters 3.3.1, 4.3.1, 5.3.1 & Appendix B). Not surprisingly, considering 2A function was proposed to be unique to the eukaryotic ribosome tunnel, no putative prokaryotic 2As were identified. Mapping 2A possessing species onto a super-tree of eukaryotic life did not reveal any obvious distribution pattern (see Figure 8.3); however, as ongoing sequencing projects continue to increase the available data, the evolutionary relationships of 2A-possessing species may become

clearer. Mapping the phylogenies of eukaryotic 2A-containing proteins revealed that each protein clade in possession of 2A sequences was monophyletic; descending from a single ancestral protein with 2A, and acquisition was, in most instances, a relatively recent addition (for examples see Figure 4.5 and Figure 5.4).

The initial evolutionary origins of eukaryotic 2As and the route of transfer between proteins/organisms/phyla remain unknown. I speculated aquatic invertebrate species possessing 2A, could have acquired 2A through horizontal gene transfer from virus infections (see Chapter 3.4 and Odon *et al.*, 2013), especially considering the high viral titre in aquatic environments (reviewed in Suttle, 2005). Another possible transfer route for marine invertebrates would be through ingestion. As I have discussed in Chapter 3.4 and Odon *et al.*, 2013, the comparatively unspecialised gut architecture of many marine invertebrates predisposes them to be able to assimilate ingested material into their somatic tissues. However, genetic material encoding the 2A must be stably integrated into the germline DNA of the gametes or gamete producing tissues of its new host before the 2A can potentially be transmitted vertically to any progeny. The eukaryotic 2A sequences recorded in the databanks are highly likely to represent true instances of 2A from the hosts' germline DNA as the majority of sequencing data for these organisms was obtained from gametes and/or early stage embryos. Therefore, the 2A sequences could not have resulted from somatic cells incorporated viral DNA post-infection. If 2A is indeed spread by horizontal gene transfer, both the transfer vector(s) and the incorporation mechanisms remain to be identified. The alternate hypothesis is that the comparatively short active 2A tract could have arisen on multiple occasions through chance mutations.

The origins of the 2A peptide represent an evolutionary paradox: how did incremental evolution create a peptide sequence that apparently requires existence in its entirety to function? The answer might be that 2As are short, and their crucial eight amino acid C-terminal motif still shorter, therefore they could be generated through chance mutations. Additionally, a 2A need not have 100% functionality; a partially effective 2A generating both the full-length protein and the processing products could prove beneficial by increasing proteome diversity. In most instances, mutations leading to 2A translational recoding would result in loss of protein function, except if the 2A occurred at either the C or N-terminal of the protein, in non-coding DNA, or between active domains of a multi-domain protein. Interestingly, in viruses, successful 2A integration is apparently only possible between domains encoding for different protein types, for example in *FMDV* between the capsid and metabolic proteins (Luke *et al.*, 2008). One is reminded of the old adage that history is written only by the victors and survivors. Nowhere is this more relevant than when examining genetic evolution. Events where 2A integration resulted in a fitness disadvantage leading to host extinction will regrettably not be identifiable from the genomes of extant organisms. We can only observe causes where 2A integration is at the least neutral to host fitness.

However, does 2A acquisition confer an evolutionary fitness advantage to the host? There is presently insufficient data to address this question. However, it should be noted that the eukaryotic 2A were identified from a limited number of protein types that might be highly tolerant of 2A insertion. Namely, multiple copy number and splice variant genes (NLR-like, SNAT9, ankyrin) or non-essential movable elements (non-LTR). Indeed, in these cases, 2A could prove a positive attribute in aiding the regulation of signalling cascades or production of multiple isoforms of the protein to target multiple subcellular destinations. If this hypothesis is correct, then 2A is a perfect "selfish gene" (Dawkins, 1976, revisited by Orgel and Crick, 1980) most probably transmitted through viral-mediated horizontal gene transfer to and retained if it does not disadvantage and might on occasion advantage the host.

The phylogenetic analyses indicated that 2A insertion was probably a comparatively recent event in the evolutionary history of each eukaryotic 2A-possessing protein clade. It may be that 2A possession is a transient state, and if 2A acquisition is disadvantageous it will be quickly selected against, if neutral then it will be lost at a slower rate through build-up of mutations over evolutionary time. However, in one instance (SNAT9 proteins) were it was possible to date the arrival of 2A to approximately 60 million years ago. Therefore, it seems likely that these 2As must be under positive selection otherwise the intervening millennia of genome evolution through random mutations and genetic drift would have resulted in the loss of the short 2A sequence.

Are there inherent costs of utilising 2A? The most obvious potential cost is that mutations in the sequence could be doubly deleterious, resulting in loss of not one but two discrete functional proteins if 2A translational recoding activity were negated, and so prove lethal if the 2A were linking essential structural or metabolic proteins encoded by a single copy gene. Interestingly, as has been discussed, this does not appear to be the case, as eukaryotic 2As occur in protein clades where a number of different protein isoforms are encoded by either multiple gene copies, or multiple alternate RNA splice variants. Another factor that requires consideration is that the 2A tail will remain attached to the C-terminus of the upstream protein. In *FMDV*, the viral protein $3C^{pro}$ removes 2A from the mature 1D protein. To date, in biotechnology applications, 2A tails have rarely proved problematic, and are useful as antibody recognition tags, but they can cause loss of function through misfolding of the upstream protein (de Felipe *et al.*, 2010; reviewed in Luke *et al.*, 2010a). To date, there have been no reports from transgenic organisms of immune system activation by foreign 2A peptides (de Felipe *et al.*, 2006) but it would be imprudent to discount this risk factor. It is too early to evaluate the long-term multigenerational stability of gene suites inserted with 2A linkers.

I speculated that a major function of 2A in eukaryotic proteins might be to regulate downstream protein translation in a similar manner to that proposed for viral 2As (reviewed in Luke *et al.*,

2008). For example, I proposed that, in the case of the non-LTR 2As, the positioning of 2A at the N-terminal of ORF2 may act to regulate the expression level of the downstream APE endonuclease and reverse transcriptase domains, and so control rates of retrotransposition (Odon *et al.*, 2013). For the ankyrin-repeat and NLR-associated 2As, semi-efficient ribosome skipping 2As may generate multiple protein isoforms, a proportion of which will lack the ankyrin repeat or DEATH domain respectively, necessary for protein-protein interaction and signalling cascade initiation. In these cases, if the translational recoding capability of 2A was impaired by cellular stress then comparatively greater quantities of the full length multi-domain proteins would be generated, and hence be available to initiate or regulate signalling cascades.

In addition to their role in translational recoding, I have shown, in some instances, the 2A may "bolt-on" an additional function to the ORF, as was proposed to be the case in the *Rotaviruses* and *Totiviridae* (as discussed in Chapter 1.10). For example, I have shown that semi-efficient ribosome skipping 2As from echinoderms, sponges, and salmonids can direct a novel form of dual protein targeting, acting not as translational recoding sequences but as protein trafficking signal peptides (refer to Chapter 7.3). These direct a proportion of the newly synthesised downstream protein to a specific sub-cellular destination. This unexpected finding constitutes the first report that a translational recoding peptide can play an alternate role in directing intracellular protein trafficking. Further investigation of these dual function 2As revealed that their signalling capabilities were retained across phyla, as a sea-urchin sequence was able to direct protein trafficking in both mammalian and plant cells (Chapter 7.2.3). This lack of host specificity makes such 2A sequences potentially highly useful to future biotechnological utilities in a wide host range.

I have found that *in silico* modelling of 2A sequences indicated some commonalities regarding the predicted architecture of 2A peptides. The models revealed that active 2As typically possessed the propensity to fold into a configuration containing an α-helical central tract whereas inactive/low activity 2As were essentially unstructured. The caveat to this method was that it assumed the peptide was folding in a hydrophilic neutrally charged environment with no spatial restrictions, whereas the ribosome tunnel is known to be spatially constrained with negatively charged hydrophilic walls (Voss *et al.*, 2006; Bhushan *et al.*, 2010). The available data and *in silico* analysis techniques were insufficient to permit me to pinpoint the specific residues necessary for 2A function.

The active eukaryotic 2As described here were identified by their conserved eight residue C-terminal motif, namely D[V/I]EXNPG$^{\downarrow}$P, common also to viral 2As. At the present time, 2A is unique among the known translational recoding peptides in that the nascent peptide apparently drives the ribosome skip wherein the peptide bond between the final glycine and proline is omitted,

and this recoding is apparently driven by the nascent chain within the ribosome tunnel, without requiring either co-factors such as macrolide antibiotics or stemloop structure in the transcript mRNA (see Chapter 1.5). It seems highly improbable that 2A is unique in function, other yet to be discovered peptides could function as ribosome skipping sequences. However, if these sequences possess low amino acid homology with 2A, then probing databases with the 2A C-terminal motif will fail to reveal them. Perhaps comparable functioning sequences could be identified by repeating the discovery process that led to the characterisation of viral 2As. Examination of the sequence upstream of viral polyprotein "cleavage" sites for which the processing proteases are undetermined. It may be that some of these polypeptide "cleavages" are a result of 2A paralogues. If such sequences are identified from viruses, then their conserved features could be used to search for homologous sequences from eukaryotic genomes.

In conclusion, I hope that the new eukaryotic 2As catalogued and investigated here may prove useful for future biotechnological applications. The biotechnological community has come to rely heavily on a handful of well characterised viral 2As, but there is still widespread public mistrust of genetic modification technologies and there are legitimate if probably unjustified concerns over possible auto-immune reactions against viral derived peptides if used long-term. If eukaryotic 2As were made available for biotechnological uses, particularly 2As from species that were phylogenetically closely related to the organism being genetically modified, then these could be used in place of viral 2As. Use of such eukaryotic 2As from close genetic relatives would minimise risk of autoimmune reactions, and aid in public acceptance of genetic modification technologies as by using them researchers would not be "inserting deadly viruses". The 2As described here can provide these benefits, and more, as they represent not only ribosome skipping sequences, but also sequences that are capable of directing dual targeting of nascent proteins.

Publications Arising:

Roulston, C.; Luke, G. A.; de Felipe, P.; Cope, J.; Tilsner, J. & Ryan, M. D. (2015), 2A signal sequences that mediate translational recoding: a novel form of dual protein targeting, (submitted).

Luke, G. A.; **Roulston, C**.; Tilsner, J. & Ryan, M. D. (2015), Growing uses of 2A in plant biotechnology, in *Biotechnology*, ed. Ekinci, D. pp1-30, Intech Croatia, ISBN 978-953-51-4157-0 (accepted, in press)

Luke, G. A.; Pathania, U.S; **Roulston, C.**; de Felipe, P & Ryan, M. D (2014), DXEXNPGP - Motives for the motif, *Recent Research Developments in Virology*, **9**: 25-42.

Luke, G. A.; **Roulston C.**; Odon, V.; de Felipe, P. Sukhodub, A. & Ryan, M. D. (2013), Lost in translation: the biogenesis of non-LTR retrotransposon proteins, *Mobile Genetic Elements*, **3** (6): e27525.

Odon, V.*; Luke, G. A.*; **Roulston, C.***; de Felipe, P, Ruan, L; Escuin-Ordinas,H; Brown, J. D.; Ryan, M. D. & Sukhodub, A. (2013), APE-type non-LTR retrotransposons of multicellular organisms encode virus-like 2A oligopeptide sequences, which mediate translational recoding during protein synthesis, *Molecular Biology and Evolution*, **30** (8): 1955-1965. (*= Joint 1st Author).

References

Adams, M. J.; Lefkowitz, E. J.; King, A. M. Q. & Carstens, E. B. (2014), Ratification vote on taxonomic proposals to the International Committee on Taxonomy of Viruses (2014), *Archives of Virology*, 1-11.

Ahmadian, G.; Randhawa, J. S. & Easton, A. J. (2000), Expression of the ORF-2 protein of the human respiratory syncytial virus M2 gene is initiated by a ribosomal termination-dependent reinitiation mechanism, *EMBO Journal*, **19**: 2681-2689.

Albalat, R.; Permanyer, J.; Martinez-Mir, C.; Gonzalez-Angulo, O. & Gonzalez-Duarte, R. (2003), The first non-LTR retrotransposon characterised in the cephalochordate amphioxus, *BfCR1*, shows similarities to *CR1-like* elements, *Cellular and Molecular Life Sciences*, **60**: 803-809.

Alisch, R. S.; Garcia-Perez, J. L.; Muotri, A. R.; Gage, F. H. & Moran, J. V. (2006), Unconventional translation of mammalian *LINE-1* retrotransposons, *Genes & Development*, **20**: 210-224.

Aliyari, R.; Wu, Q.; Li, H.-W.; Wang, X.-H.; Li, F.; Green, L. D.; Han, C. S.; Li, W.-X. & Ding, S.-W. (2008), Mechanism of induction and suppression of antiviral immunity directed by virus-derived small RNAs in *Drosophila*, *Cell Host & Microbe*, **4**: 387-397.

Almen, M. S.; Nordstrom, K. J. V.; Fredriksson, R. & Schioth, H. B. (2009), Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin, *BMC Biology*, **7**:50.

Anandatheerthavarada, H. K.; Biswas, G.; Mullick, J.; Sepuri, N. B. V.; Otvos, L.; Pain, D. & Avadhani, N. G. (1999), Dual targeting of cytochrome P4502B1 to endoplasmic reticulum and mitochondria involves a novel signal activation by cyclic AMP-dependent phosphorylation at Ser128, *EMBO Journal*, **18**: 5494-5504.

Arkhipova, I. R. (2006), Distribution and phylogeny of *Penelope-like* elements in eukaryotes, *Systematic Biology*, **55**: 875-885.

Armache, J.-P.; Jarasch, A.; Anger, A. M.; Villa, E.; Becker, T.; Bhushan, S.; Jossinet, F.; Habeck, M.; Dindar, G.; Franckenberg, S.; Marquez, V.; Mielke, T.; Thomm, M.; Berninghausen, O.; Beatrix, B.; Soeding, J.; Westhof, E.; Wilson, D. N. & Beckmann, R. (2010), Cryo-EM structure and rRNA model of a translating eukaryotic 80S ribosome at 5.5-angstrom resolution, *Proceedings of the National Academy of Sciences of the United States of America*, **107**: 19748-19753.

Atkins, J. F.; Wills, N. M.; Loughran, G.; Wu, C.-Y.; Parsawar, K.; Ryan, M. D.; Wang, C.-H. & Nelson, C. C. (2007), A case for "StopGo": Reprogramming translation to augment codon meaning of GGN by promoting unconventional termination (Stop) after addition of glycine and then allowing continued translation (Go), *RNA-A, Publication of the RNA Society*, **13**: 803-810.

Ayarpadikannan, S. & Kim, H.-S. (2014), The impact of transposable elements in genome evolution and genetic instability and their implications in various diseases, *Genomics and Informatics*, **12**: 98-104.

Balaj, L.; Lessard, R.; Dai, L.; Cho, Y.-J.; Pomeroy, S. L.; Breakefield, X. O. & Skog, J. (2011), Tumour microvesicles contain retrotransposon elements and amplified oncogene sequences, *Nature Communications*, **2**: 180.

Baltimore, D. (1971), Expression of animal virus genomes, *Bacteriological Reviews*, **35**: 235-241.

Becker, T.; Bhushan, S.; Jarasch, A.; Armache, J.-P.; Funes, S.; Jossinet, F.; Gumbart, J.; Mielke, T.; Berninghausen, O.; Schulten, K.; Westhof, E.; Gilmore, R.; Mandon, E. C. & Beckmann, R. (2009), Structure of monomeric yeast and mammalian Sec61 complexes interacting with the translating ribosome, *Science*, **326**: 1369-1373.

Belsham, G. J. (1993), Distinctive features of foot-and-mouth-disease virus, a member of the picornavirus family - aspects of virus protein-synthesis, protein processing and structure, *Progress in Biophysics & Molecular Biology*, **60**: 241-260.

Bennett, V. & Chen, L. S. (2001), Ankyrins and cellular targeting of diverse membrane proteins to physiological sites, *Current Opinion in Cell Biology*, **13**: 61-67.

Bernabeu, C. & Lake, J. A. (1982), Nascent polypeptide-chains emerge from the exit domain of the large ribosomal-subunit - immune mapping of the nascent chain, *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences*, **79**: 3111-3115.

Bhushan, S.; Gartmann, M.; Halic, M.; Armache, J.-P.; Jarasch, A.; Mielke, T.; Berninghausen, O.; Wilson, D. N. & Beckmann, R. (2010), Alpha-helical nascent polypeptide chains visualized within distinct regions of the ribosomal exit tunnel, *Nature Structural & Molecular Biology*, **17**: 313-318.

Blanc, G.; Duncan, G.; Agarkova, I.; Borodovsky, M.; Gurnon, J.; Kuo, A.; Lindquist, E.; Lucas, S.; Pangilinan, J.; Polle, J.; Salamov, A.; Terry, A.; Yamada, T.; Dunigan, D. D.; Grigoriev, I. V.; Claverie, J.-M. & Van Etten, J. L. (2010), The *Chorella variabilis* NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex, *The Plant Cell*, **22**: 2943-2955.

Blobel, G. & Dobberstein, B. (1975), Transfer of proteins across membranes - Presence of proteolytically processed and unprocessed nascent immunoglobulin light-chains on membrane-bound ribosomes of murine myeloma, *Journal of Cell Biology*, **67**: 835-851.

Bos, L. (2000), 100 years of virology: from vitalism via molecular biology to genetic engineering, *Trends in Microbiology*, **8**: 82-87.

Breeden, L. & Nasmyth, K. (1987), Similarity between cell-cycle genes of budding yeast and fission yeast and the *Notch* gene of *Drosophila*, *Nature*, **329**: 651-654.

Bröer, S. (2014), The SLC38 family of sodium–amino acid co-transporters, *European Journal of Physiology*, **466**: 155-172.

Brown, J. D. & Ryan, M. D. (2010), Ribosome "Skipping": "Stop-Carry On" or "StopGo" Translation, *In:* Atkins, J. F. & Gesteland, R. F. (eds.) *Recoding: Expansion of Decoding Rules Enriches Gene Expression*. Springer-Verlag Berlin, Heidelberger Platz 3, D-14197 Berlin, Germany.

Bulgakov, V. P.; Odintsova, N. A.; Plotnikov, S. V.; Kiselev, K. V.; Zacharov, E. V. & Zhuravlev, Y. N. (2002), *Gal4*-gene-dependent alterations of embryo development and cell growth in primary culture of sea urchins, *Marine Biotechnology*, **4**: 480-486.

Charlesworth, B. & Charlesworth, D. (1983), The population-dynamics of transposable elements, *Genetical Research*, **42**: 1-27.

Choi, J.-H.; Jung, H.-Y.; Kim, H.-S. & Cho, H.-G. (2000), PhyloDraw: a phylogenetic tree drawing system, *Bioinformatics*, **16**: 1056-1058.

Christensen, H. N.; Oxender, D. L.; Liang, M. & Vatz, K. A. (1965), Use of N-methylation to direct route of mediated transport of amino acids, *Journal of Biological Chemistry*, **240**: 3609-3616.

Chung, W.-J.; Okamura, K.; Martin, R. & Lai, E. C. (2008), Endogenous RNA interference provides a somatic defense against *Drosophila* transposons, *Current Biology*, **18**: 795-802.

Colombo, S.; Longhi, R.; Alcaro, S.; Ortuso, F.; Sprocati, T.; Flora, A. & Borgese, N. (2005), N-myristoylation determines dual targeting of mammalian NADH-cytochrome b(5) reductase to ER and mitochondrial outer membranes by a mechanism of kinetic partitioning, *Journal of Cell Biology*, **168**: 735-745.

Crick, F. (1970), Central dogma of molecular biology, *Nature*, **227**: 561-563.

Cruz-Vera, L. R.; Sachs, M. S.; Squires, C. L. & Yanofsky, C. (2011), Nascent polypeptide sequences that influence ribosome function, *Current Opinion in Microbiology*, **14**: 160-166.

Davidson, A. E.; Gratsch, T. E.; Morell, M. H.; O'Shea, K. S. & Krull, C. E. (2009), Use of the sleeping beauty transposon system for stable gene expression in mouse embryonic stem cells, *Cold Spring Harbor Protocols*, **2009**: 5270.

Dawkins, R. (1976), *The Selfish Gene* Oxford, Oxford University Press.

de Felipe, P. (2004), Skipping the co-expression problem: the new 2A "CHYSEL" technology, *Genetic Vaccines and Therapy,* **2**: 13.

de Felipe, P.; Hughes, L. E.; Ryan, M. D. & Brown, J. D. (2003), Co-translational, intraribosomal cleavage of polypeptides by the foot-and-mouth disease virus 2A peptide, *Journal of Biological Chemistry,* **278**: 11441-11448.

de Felipe, P.; Luke, G. A.; Brown, J. D. & Ryan, M. D. (2010), Inhibition of 2A-mediated 'cleavage' of certain artificial polyproteins bearing N-terminal signal sequences, *Biotechnology Journal,* **5**: 213-223.

de Felipe, P.; Luke, G. A.; Hughes, L. E.; Gani, D.; Halpin, C. & Ryan, M. D. (2006), E unum pluribus: multiple proteins from a self-processing polyprotein, *Trends in Biotechnology,* **24**: 68-75.

de Felipe, P. & Ryan, M. D. (2004), Targeting of proteins derived from self-processing polyproteins containing multiple signal sequences, *Traffic,* **5**: 616-626.

Degnan, B. M.; Adamska, M.; Craigie, A.; Degnan, S. M.; Fahey, B.; Gauthier, M.; Hooper, J. N. A.; Larroux, C.; Leys, S. P.; Lovas, E. & Richards, G. S. (2008), The Demosponge *Amphimedon queenslandica*: Reconstructing the ancestral metazoan genome and deciphering the origin of animal multicellularity, *Cold Spring Harbor Protocols,* **2008**: DOI:10.1101/pdb.emo1108.

Demeshkina, N.; Jenner, L.; Yusupova, G. & Yusupov, M. (2010), Interactions of the ribosome with mRNA and tRNA, *Current Opinion in Structural Biology,* **20**: 325-332.

Denks, K.; Vogt, A.; Sachelaru, I.; Petriman, N. A.; Kudva, R. & Koch, H. G. (2014), The Sec translocon mediated protein transport in prokaryotes and eukaryotes, *Molecular Membrane Biology,* **31**: 58-84.

Ding, S. W. & Voinnet, O. (2007), Antiviral immunity directed by small RNAs, *Cell,* **130**: 413-426.

Dingwall, C. & Laskey, R. A. (1991), Nuclear targeting sequences — a consensus?, *Trends in Biochemical Sciences,* **16**: 478-481.

Donnelly, M. L. L.; Gani, D.; Flint, M.; Monaghan, S. & Ryan, M. D. (1997), The cleavage activities of aphthovirus and cardiovirus 2A proteins, *Journal of General Virology,* **78**: 13-21.

Donnelly, M. L. L.; Hughes, L. E.; Luke, G.; Mendoza, H.; ten Dam, E.; Gani, D. & Ryan, M. D. (2001a), The 'cleavage' activities of foot-and-mouth disease virus 2A site-directed mutants and naturally occurring '2A-like' sequences, *Journal of General Virology,* **82**: 1027-1041.

Donnelly, M. L. L.; Luke, G.; Mehrotra, A.; Li, X. J.; Hughes, L. E.; Gani, D. & Ryan, M. D. (2001b), Analysis of the aphthovirus 2A/2B polyprotein 'cleavage' mechanism indicates not a proteolytic reaction, but a novel translational effect: a putative ribosomal 'skip', *Journal of General Virology,* **82**: 1013-1025.

Doolittle, W. F. & Sapienza, C. (1980), Selfish genes, the phenotype paradigm and genome evolution, *Nature,* **284**: 601-603.

Doronina, V. A.; de Felipe, P.; Wu, C.; Sharma, P.; Sachs, M. S.; Ryan, M. D. & Brown, J. D. (2008a), Dissection of a co-translational nascent chain separation event, *Biochemical Society Transactions,* **36**: 712-716.

Doronina, V. A.; Wu, C.; de Felipe, P.; Sachs, M. S.; Ryan, M. D. & Brown, J. D. (2008b), Site-specific release of nascent chains from ribosomes at a sense codon, *Molecular and Cellular Biology,* **28**: 4227-4239.

Duby, G. & Boutry, M. (2002), Mitochondrial protein import machinery and targeting information, *Plant Science,* **162**: 477-490.

Ebert, T. A. & Southon, J. R. (2003), Red sea urchins (Strongylocentrotus franciscanus) can live over 100 years: confirmation with A-bomb (14) carbon, *Fishery Bulletin,* **101**: 915-922.

Fan, L.; Reynolds, D.; Liu, M.; Stark, M.; Kjelleberg, S.; Webster, N. S. & Thomas, T. (2012), Functional equivalence and evolutionary convergence in complex communities of microbial sponge symbionts, *Proceedings of the National Academy of Sciences of the United States of America*, **109**: E1878–E1887.

Fang, P.; Spevak, C. C.; Wu, C. & Sachs, M. S. (2004), A nascent polypeptide domain that can regulate translation elongation, *Proceedings of the National Academy of Sciences of the United States of America*, **10**: 4059-4064.

Firth, A. E. & Brierley, I. (2012), Non-canonical translation in RNA viruses, *Journal of General Virology*, **93**: 1385-1409.

Fredriksson, R.; Nordstroem, K. J. V.; Stephansson, O.; Haegglund, M. G. A. & Schioeth, H. B. (2008), The solute carrier (SLC) complement of the human genome: Phylogenetic classification reveals four major families, *Febs Letters*, **582**: 3811-3816.

Gao, H.; Shi, W. & Freund, L. B. (2005), Mechanics of receptor-mediated endocytosis, *Proceedings of the National Academy of Sciences of the United States of America*, **102**: 9469-9474.

Gao, Z. L.; Zhou, J. H.; Zhang, J.; Ding, Y. Z. & Liu, Y. S. (2014), The silent point mutations at the cleavage site of 2A/2B have no effect on the self-cleavage activity of 2A of foot-and-mouth disease virus, *Infection, Genetics and Evolution*, **28**:101-106

Gould, P. S. & Easton, A. J. (2005), Coupled translation of the respiratory syncytial virus M2 open reading frames requires upstream sequences, *Journal of Biological Chemistry*, **280**: 21972-21980.

Gould, P. S. & Easton, A. J. (2007), Coupled translation of the second open reading frame of M2 mRNA is sequence dependent and differs significantly within the subfamily Pneumovirinae, *Journal of Virology*, **81**: 8488-8496.

Heath, L.; van der Walt, E.; Varsani, A. & Martin, D. P. (2006), Recombination patterns in aphthoviruses mirror those found in other picornaviruses, *Journal of Virology*, **80**: 11827-11832.

Hegde, R. S. & Bernstein, H. D. (2006), The surprising complexity of signal sequences, *Trends in Biochemical Sciences*, **31**: 563-571.

Heras, S. R.; Thomas, M. C.; Garcia-Canadas, M.; de Felipe, P.; Garcia-Perez, J. L.; Ryan, M. D. & Lopez, M. C. (2006), L1Tc non-LTR retrotransposons from Trypanosoma cruzi contain a functional viral-like self-cleaving 2A sequence in frame with the active proteins they encode, *Cellular and Molecular Life Sciences*, **63**: 1449-1460.

Hibino, T.; Loza-Coll, M.; Messier, C.; Majeske, A. J.; Cohen, A. H.; Terwilliger, D. P.; Buckley, K. M.; Brockton, V.; Nair, S. V.; Berney, K.; Fugmann, S. D.; Anderson, M. K.; Pancer, Z.; Cameron, R. A.; Smith, L. C. & Rast, J. P. (2006), The immune gene repertoire encoded in the purple sea urchin genome, *Developmental Biology*, **300**: 349-365.

Horton, P. & Nakai, K. (1997), Better prediction of protein cellular localization sites with the k nearest neighbors classifier, *Intelligent Systems for Molecular Biology*, **5**: 147-152.

Horton, P.; Park, K. J.; Obayashi, T.; Fujita, N.; Harada, H.; Adams-Collier, C. J. & Nakai, K. (2007), WoLF PSORT: protein localization predictor, *Nucleic Acids Research*, **35**: W585-587.

Horvath, C. M.; Williams, M. A. & Lamb, R. A. (1990), Eukaryotic coupled translation of tandem cistrons - identification of the Influenza-B virus BM2 polypeptide, *EMBO Journal*, **9**: 2639-2647.

Hughes, P. J. & Stanway, G. (2000), The 2A proteins of three diverse picornaviruses are related to each other and to the H-rev107 family of proteins involved in the control of cell proliferation, *Journal of General Virology*, **81**: 201-207.

Hundal, H. S. & Taylor, P. M. (2009), Amino acid transceptors: gate keepers of nutrient exchange and regulators of nutrient signaling, *American Journal of Physiology-Endocrinology and Metabolism*, **296**: E603-E613.

Hutvagner, G. & Simard, M. J. (2008), Argonaute proteins: key players in RNA silencing, *Nature Reviews Molecular Cell Biology*, **9**: 22-32.

International Human Genome Sequencing Consortium (2001), Initial sequencing and analysis of the human genome, *Nature*, **409**: 860-921.

Ito, K.; Chiba, S. & Pogliano, K. (2010), Divergent stalling sequences sense and control cellular physiology, *Biochemical and Biophysical Research Communications*, **393**: 1-5.

Jacks, T.; Madhani, H. D.; Masiarz, F. R. & Varmus, H. E. (1988), Signals for ribosomal frameshifting in the Rous sarcoma virus Gag-Pol region, *Cell*, **55**: 447-458.

Jacks, T. & Varmus, H. E. (1985), Expression of the Rous sarcoma virus pol gene by ribosomal frameshifting, *Science*, **230**: 1237-1242.

Jackson, R. J.; Hellen, C. U. & Pestova, T. V. (2012), Termination and post-termination events in eukaryotic translation, *Advances in Protein Chemistry and Structural Biology*, **86**: 45-93.

Jang, S. K.; Krausslich, H. G.; Nicklin, M. J.; Duke, G. M.; Palmenberg, A. C. & Wimmer, E. (1988), A segment of the 5' nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation, *Journal of Virology*, **62**: 2636-2643.

Jenni, S. & Ban, N. (2003), The chemistry of protein synthesis and voyage through the ribosomal tunnel, *Current Opinion in Structural Biology*, **13**: 212-219.

Jurka, J.; Kapitonov, V. V.; Kohany, O. & Jurka, M. V. (2007), Repetitive sequences in complex genomes: Structure and evolution, *Annual Review of Genomics and Human Genetics*.

Kanneganti, T.-D.; Lamkanfi, M. & Nunez, G. (2007), Intracellular NOD-like receptors in host defense and disease, *Immunity*, **27**: 549-559.

Karniely, S. & Pines, O. (2005), Single translation-dual destination: mechanisms of dual protein targeting in eukaryotes, *EMBO Reports*, **6**: 420-425.

Knowles, N. J., Hovi, T, Hyypiä, T., King, A.M.Q., Lindberg, A.M., Pallansch, M.A., Palmenberg, A.C., Simmonds, P., Skern, T., Stanway, G., Yamashita, T. and Zell, R. (2012), Picornaviridae, *In:* King, A. M. Q., Adams, M.J., Carstens, E.B. and Lefkowitz, E.J. (ed.) *Virus Taxonomy: Classification and Nomenclature of Viruses: Ninth Report of the International Committee on Taxonomy of Viruses*. San Diego.

Kojima, K. K.; Matsumoto, T. & Fujiwara, H. (2005), Eukaryotic translational coupling in UAAUG stop-start codons for the bicistronic RNA translation of the non-long terminal repeat retrotransposon SART1, *Molecular and Cellular Biology*, **25**: 7675-7686.

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

Lafontaine, D. L. J. & Tollervey, D. (2001), The function and synthesis of ribosomes, *Nature Reviews Molecular Cell Biology*, **2**: 514-520.

Lander, E. S.; International Human Genome Sequencing Consortium, C.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; Funke, R.; Gage, D.; Harris, K.; Heaford, A.; Howland, J.; Kann, L.; Lehoczky, J.; LeVine, R.; McEwan, P.; McKernan, K.; Meldrim, J.; Mesirov, J. P.; Miranda, C.; Morris, W.; Naylor, J.; Raymond, C.; Rosetti, M.; Santos, R.; Sheridan, A.; Sougnez, C.; Stange-Thomann, N.; Stojanovic, N.; Subramanian, A.; Wyman, D.; Rogers, J.; Sulston, J.; Ainscough, R.; Beck, S.; Bentley, D.; Burton, J.; Clee, C.; Carter, N.; Coulson, A.; Deadman, R.; Deloukas, P.; Dunham, A.; Dunham, I.; Durbin, R.; French, L.; Grafham, D.; Gregory, S.; Hubbard, T.; Humphray, S.; Hunt, A.; Jones, M.; Lloyd, C.; McMurray, A.; Matthews, L.; Mercer, S.; Milne, S.; Mullikin, J. C.; Mungall, A.; Plumb, R.; Ross, M.; Shownkeen, R.; Sims, S.; Waterston, R. H.; Wilson, R. K.; Hillier, L. W.; McPherson, J. D.; Marra, M. A.; Mardis, E. R.; Fulton, L. A.; Chinwalla, A. T.; Pepin, K. H.; Gish, W. R.; Chissoe, S. L.; Wendl, M. C.; Delehaunty, K. D.; Miner, T. L.; Delehaunty, A.; Kramer, J. B.; Cook, L. L.; Fulton, R. S.; Johnson, D. L.; Minx, P. J.; Clifton, S. W.; Hawkins, T.; Branscomb, E.; Predki, P.;

Richardson, P.; Wenning, S.; Slezak, T.; Doggett, N.; Cheng, J. F.; Olsen, A.; Lucas, S.; Elkin, C.; Uberbacher, E., *et al.* (2001), Initial sequencing and analysis of the human genome, *Nature*, **409**: 860-921.

Langland, J. O.; Pettiford, S.; Jiang, B. M. & Jacobs, B. L. (1994), Products of the porcine group-c rotavirus NSP3 gene bind specifically to double-stranded-RNA and inhibit activation of the interferon-induced protein-kinase PKR, *Journal of Virology*, **68**: 3821-3829.

Larkin, M. A.; Blackshields, G.; Brown, N. P.; Chenna, R.; McGettigan, P. A.; McWilliam, H.; Valentin, F.; Wallace, I. M.; Wilm, A.; Lopez, R.; Thompson, J. D.; Gibson, T. J. & Higgins, D. G. (2007), Clustal W and Clustal X version 2.0, *Bioinformatics*, **23**: 2947-2948.

Lecoq, H. (2001), Découverte du premier virus, le virus de la mosaïque du tabac : 1892 ou 1898?, *Comptes Rendus de l'Académie des Sciences - Series III - Sciences de la Vie*, **324**: 929-933.

Li, H. W.; Li, W. X. & Ding, S. W. (2002), Induction and suppression of RNA silencing by an animal virus, *Science*, **296**: 1319-1321.

Li, J.; Mahajan, A. & Tsai, M.-D. (2006), Ankyrin repeat: A unique motif mediating protein-protein interactions, *Biochemistry*, **45**: 15168-15178.

Lich, J. D. & Ting, J. P. Y. (2007), CATERPILLER (NLR) family members as positive and negative regulators of inflammatory responses, *Proceedings of the American Thoracic Society*, **4**: 263-266.

Lindbo, J. A. (2007), TRBO: a high-efficiency tobacco mosaic virus RNA-based overexpression vector, *Plant Physiology*, **145**: 1232-1240.

Lu, J. & Deutsch, C. (2008), Electrostatics in the Ribosomal Tunnel Modulate Chain Elongation Rates, *Journal of Molecular Biology*, **384**: 73-86.

Luke, G. A.; de Felipe, P.; Lukashev, A.; Kallioinen, S. E.; Bruno, E. A. & Ryan, M. D. (2008), Occurrence, function and evolutionary origins of '2A-like' sequences in virus genomes, *Journal of General Virology*, **89**: 1036-1042.

Luke, G. A.; Escuin, H.; De Felipe, P. & Ryan, M. D. (2010a), 2A to the Fore - Research, Technology and Applications, *In:* Harding, S. E. T. M. P. (ed.) *Biotechnology and Genetic Engineering Reviews, Vol 26*.

Luke, G. A.; Escuin, H.; De Felipe, P. & Ryan, M. D. (2010b), 2A to the Fore - Research, Technology and Applications, *In:* Harding, S. E. T. M. P. (ed.) *Biotechnology and Genetic Engineering Reviews*.

Luke, G. A. & Ryan, M. D. (2013), The protein coexpression problem in biotechnology and biomedicine: virus 2A and 2A-like sequences provide a solution, *Future Virology*, **8**: 983-996.

Lustig, A. & Levine, A. J. (1992), One hundred years of virology, *Journal of Virology*, **66**: 4629-4631.

Luttermann, C. & Meyers, G. (2007), A bipartite sequence motif induces translation reinitiation in feline calicivirus RNA, *Journal of Biological Chemistry*, **282**: 7056-7065.

Lux, S. E.; John, K. M. & Bennett, V. (1990), Analysis of cDNA for human erythrocyte ankyrin indicates a repeated structure with homology to tissue-differentiation and cell-cycle control proteins, *Nature*, **344**: 36-42.

Machida, K.; Mikami, S.; Masutani, M.; Mishima, K.; Kobayashi, T. & Imataka, H. (2014), A translation system reconstituted with human factors proves that processing of Encephalomyocarditis Virus proteins 2A and 2B occurs in the elongation phase of translation without eukaryotic release factors, *Journal of Biological Chemistry*, **289**: 31960-31971.

Malik, H. S.; Burke, W. D. & Eickbush, T. H. (1999), The age and evolution of non-LTR retrotransposable elements, *Molecular Biology and Evolution*, **16**: 793-805.

Malkin, L. I. & Rich, A. (1967), Partial resistance of nascent polypeptide chains to proteolytic digestion due to ribosomal shielding, *Journal of Molecular Biology*, **26**: 329-346.

Martin, S. L. (2006), The ORF1 protein encoded by *LINE-1*: Structure and function during *L1* retrotransposition, *Journal of Biomedicine and Biotechnology*, **2006**: 45621.

Martinez-Salas, E. (2008), The impact of RNA structure on picornavirus IRES activity, *Trends in Microbiology*, **16**: 230-237.

Martinez-Salas, E.; Pacheco, A.; Serrano, P. & Fernandez, N. (2008), New insights into internal ribosome entry site elements relevant for viral gene expression, *Journal of General Virology*, **89**: 611-626.

Matsumoto, T.; Takahashi, H. & Fujiwara, H. (2004), Targeted nuclear import of open reading frame 1 protein is required for in vivo retrotransposition of a telomere-specific non-long terminal repeat retrotransposon, *SART1*, *Molecular and Cellular Biology*, **24**: 105-122.

Maupetit, J.; Derreumaux, P. & Tuffery, P. (2009), PEP-FOLD: an online resource for de novo peptide structure prediction, *Nucleic Acids Research*, **37**: 498-503.

McClintock, B. (1950), The origin and behavior of mutable loci in maize, *Proceedings of the National Academy of Sciences of the United States of America*, **36**: 344-355.

Meister, G. & Tuschl, T. (2004), Mechanisms of gene silencing by double-stranded RNA, *Nature*, **431**: 343-349.

Meyers, G. (2003), Translation of the minor capsid protein of a calicivirus is initiated by a novel termination-dependent reinitiation mechanism, *Journal of Biological Chemistry*, **278**: 34051-34060.

Meyers, G. (2007), Characterization of the sequence element directing translation reinitiation in RNA of the calicivirus, rabbit hemorrhagic disease virus, *Journal of Virology*, **81**: 9623-9632.

Michaely, P.; Kamal, A.; Anderson, R. G. W. & Bennett, V. (1999), Potential role of ankyrins in endocytosis, *Molecular Biology of the Cell*, **10**: 307A-307A.

Minskaia, E.; Luke, G. & Ryan, M. D. (2015), Co-expression technologies in eukaryotic cells, *In:* Zahoorullah, S. (ed.) *A Text Book of Biotechnology*. SM Group Open Access E-Books.

Minskaia, E.; Nicholson, J. & Ryan, M. D. (2013), Optimisation of the foot-and-mouth disease virus 2A co-expression system for biomedical applications, *BMC Biotechnology*, **13**: 67.

Moran, J. V.; Holmes, S. E.; Naas, T. P.; DeBerardinis, R. J.; Boeke, J. D. & Kazazian, H. H. (1996), High frequency retrotransposition in cultured mammalian cells, *Cell*, **87**: 917-927.

Morgan, D. G.; Menetret, J. F.; Radermacher, M.; Neuhof, A.; Akey, I. V.; Rapoport, T. A. & Akey, C. W. (2000), A comparison of the yeast and rabbit 80 S ribosome reveals the topology of the nascent chain exit tunnel, inter-subunit bridges and mammalian rRNA expansion segments, *Journal of Molecular Biology*, **301**: 301-321.

Mosavi, L. K.; Jr, D. L. M. & Peng, Z.-y. (2002), Consensus-derived structural determinants of the ankyrin repeat motif, *Proceedings of the National Academy of Sciences of the United States of America*, **99**: 16029-16034.

Muller, M.; Ibrahimi, I.; Chang, C. N.; Walter, P. & Blobel, G. (1982), A bacterial secretory protein requires signal recognition particle for translocation across mammalian endoplasmic-reticulum, *Journal of Biological Chemistry*, **257**: 1860-1863.

Nakai, K. & Horton, P. (1999), PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization, *Trends in Biochemical Sciences*, **24**: 34-35.

Nayak, A.; Berry, B.; Tassetto, M.; Kunitomi, M.; Acevedo, A.; Deng, C.; Krutchinsky, A.; Gross, J.; Antoniewski, C. & Andino, R. (2010), Cricket paralysis virus antagonizes Argonaute 2 to modulate antiviral defense in *Drosophila*, *Nature Structural & Molecular Biology*, **17 (5)**: 547-554.

Neerincx, A.; Lautz, K.; Menning, M.; Kremmer, E.; Zigrino, P.; Hoesel, M.; Buening, H.; Schwarzenbacher, R. & Kufer, T. A. (2010), A Role for the human nucleotide-binding domain, leucine-rich repeat-containing family member NLRC5 in antiviral responses, *Journal of Biological Chemistry*, **285**: 26223-26232.

Nguyen, M. T. H. D.; Liu, M. & Thomas, T. (2014), Ankyrin-repeat proteins from sponge symbionts modulate amoebal phagocytosis, *Molecular Ecology*, **23**: 1635-1645.

Nielsen, H.; Brunak, S. & von Heijne, G. (1999), Machine learning approaches for the prediction of signal peptides and other protein sorting signals, *Protein Engineering*, **12**: 3-9.

Nielsen, H.; Engelbrecht, J.; Brunak, S. & vonHeijne, G. (1997), Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites, *Protein Engineering*, **10**: 1-6.

Odon, V.; Luke, G. A.; Roulston, C.; de Felipe, P.; Ruan, L.; Escuin-Ordinas, H.; Brown, J. D.; Ryan, M. D. & Sukhodub, A. (2013), APE-type non-LTR retrotransposons of multicellular organisms encode virus-like 2A oligopeptide sequences, which mediate translational recoding during protein synthesis, *Molecular Biology and Evolution*, **30**: 1955-1965.

Orgel, L. E. & Crick, F. H. C. (1980), Selfish DNA - the ultimate parasite, *Nature*, **284**: 604-607.

Ortega, S.; Malumbres, M. & Barbacid, M. (2002), Cyclin D-dependent kinases, INK4 inhibitors and cancer, *Biochimica Et Biophysica Acta-Reviews on Cancer*, **1602**: 73-87.

Palmenberg (1990), Proteolytic processing of picornaviral polyprotein, *Annual Review of Microbiology*, **44**: 603-623.

Palmenberg, A. C.; Parks, G. D.; Hall, D. J.; Ingraham, R. H.; Seng, T. W. & Pallai, P. V. (1992), Proteolytic processing of the cardioviral P2 region - primary 2A/2B cleavage in clone-derived precursors, *Virology*, **190**: 754-762.

Pavlov, M. Y.; Watts, R. E.; Tan, Z.; Cornish, V. W.; Ehrenberg, M. & Forster, A. C. (2009), Slow peptide bond formation by proline and other N-alkylamino acids in translation, *Proceedings of the National Academy of Sciences of the United States of America*, **106**: 50-54.

Pearse, J. S. (2006), Perspective - Ecological role of purple sea urchins, *Science*, **314**: 940-941.

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

Petersen, T. N.; Brunak, S.; von Heijne, G. & Nielsen, H. (2011), SignalP 4.0: discriminating signal peptides from transmembrane regions, *Nature Methods*, **8**: 785-786.

Pochapin, M. B.; Sanger, J. M. & Sanger, J. W. (1983), Microinjection of Lucifer yellow CH into sea urchin eggs and embryos, *Cell & Tissue Research*, **234**: 309-318.

Powell, M. L.; Napthine, S.; Jackson, R. J.; Brierley, I. & Brown, T. D. K. (2008), Characterization of the termination-reinitiation strategy employed in the expression of influenza B virus BM2 protein, *RNA-A Publication of the RNA Society*, **14**: 2394-2406.

Rabl, J.; Leibundgut, M.; Ataide, S. F.; Haag, A. & Ban, N. (2011), Crystal structure of the eukaryotic 40S ribosomal subunit in complex with initiation factor 1, *Science*, **331**: 730-736.

Rane, N. S.; Yonkovich, J. L. & Hegde, R. S. (2004), Protection from cytosolic prion protein toxicity by modulation of protein translocation, *EMBO Journal*, **23**: 4550-4559.

Rangwala, S. H.; Zhang, L. & Kazazian, H. H., Jr. (2009), Many *LINE1* elements contribute to the transcriptome of human somatic cells, *Genome Biology*, **10**: 100.

Routh, A.; Domitrovic, T. & Johnson, J. E. (2012), Host RNAs, including transposons, are encapsidated by a eukaryotic single-stranded RNA virus, *Proceedings of the National Academy of Sciences of the United States of America*, **109**: 1907-1912.

Rueckert, R. R. & Wimmer, E. (1984), Systematic nomenclature of picornavirus proteins, *Journal of Virology*, **50**: 957-959.

Ryan, M. D.; Belsham, G. J. & King, A. M. Q. (1989), Specificity of enzyme substrate interactions in foot-and-mouth-disease virus polyprotein processing, *Virology*, **173**: 35-45.

Ryan, M. D.; Donnelly, M.; Lewis, A.; Mehrotra, A. P.; Wilkie, J. & Gani, D. (1999), A model for nonstoichiometric, cotranslational protein scission in eukaryotic ribosomes, *Bioorganic Chemistry*, **27**: 55-79.

Ryan, M. D. & Drew, J. (1994), Foot-and-mouth disease virus 2A oligopeptide mediated cleavage of an artificial polyprotein, *EMBO Journal*, **13**: 928-933.

Ryan, M. D. & Flint, M. (1997), Virus-encoded proteinases of the picornavirus super-group, *Journal of General Virology*, **78**: 699-723.

Ryan, M. D.; King, A. M. Q. & Thomas, G. P. (1991), Cleavage of foot-and-mouth disease virus polyprotein is mediated by residues located within a 19 amino acid sequence, *Journal of General Virology*, **72**: 2727-2732.

Sambrook, J.; Russell, D. W.; Sambrook, J. & Russell, D. W. (2001), *Molecular cloning: a laboratory manual*, Cold Spring Harbor Laboratory Press, 10 Skyline Drive, Plainview, New York, USA.

SanMiguel, P.; Tikhonov, A.; Jin, Y. K.; Motchoulskaia, N.; Zakharov, D.; Melake-Berhan, A.; Springer, P. S.; Edwards, K. J.; Lee, M.; Avramova, Z. & Bennetzen, J. L. (1996), Nested retrotransposons in the intergenic regions of the maize genome, *Science*, **274**: 765-768.

Sawyer, S. L.; Emerman, M. & Malik, H. S. (2004), Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G, *PLoS Biology*, **2**: 1278-1285.

Schiöth, H. B.; Roshanbin, S.; Hägglund, M. G. A. & Fredriksson, R. (2013), Evolutionary origin of amino acid transporter families SLC32, SLC36 and SLC38 and physiological, pathological and therapeutic aspects, *Molecular Aspects of Medicine*, **34**: 571-585.

Schneider, M.; Zimmermann, A. G.; Roberts, R. A.; Zhang, L.; Swanson, K. V.; Wen, H.; Davis, B. K.; Allen, I. C.; Holl, E. K.; Ye, Z.; Rahman, A. H.; Conti, B. J.; Eitas, T. K.; Koller, B. H. & Ting, J. P. (2012), The innate immune sensor NLRC3 attenuates Toll-like receptor signaling via modification of the signaling adaptor TRAF6 and transcription factor NF-kappaB, *Nature Immunology*, **13**: 823-831.

Sedgwick, S. G. & Smerdon, S. J. (1999), The ankyrin repeat: a diversity of interactions on a common structural framework, *Trends in Biochemical Sciences*, **24**: 311-316.

Shaffer, P. L.; Goehring, A.; Shankaranarayanan, A. & Gouaux, E. (2009), Structure and mechanism of a Na+ independent amino acid transporter, *Science*, **325**: 1010-1014.

Sharma, P.; Yan, F.; Doronina, V. A.; Escuin-Ordinas, H.; Ryan, M. D. & Brown, J. D. (2012), 2A peptides provide distinct solutions to driving stop-carry on translational recoding, *Nucleic Acids Research*, **40**: 3143-3151.

Shevchenko, A.; Wilm, M.; Vorm, O. & Mann, M. (1996), Mass spectrometric sequencing of proteins from silver stained polyacrylamide gels, *Analytical Chemistry*, **68**: 850-858.

Silva, J. C.; Loreto, E. L. & Clark, J. B. (2004), Factors that affect the horizontal transfer of transposable elements, *Current Issues in Molecular Biology*, **6**: 57-71.

Sodergren, E.; Weinstock, G. M.; Davidson, E. H.; Cameron, R. A.; Gibbs, R. A.; Weinstock, G. M.; Angerer, R. C.; Angerer, L. M.; Arnone, M. I.; Burgess, D. R.; Burke, R. D.; Cameron, R. A.; Coffman, J. A.; Davidson, E. H.; Dean, M.; Elphick, M. R.; Ettensohn, C. A.; Foltz, K. R.; Hamdoun, A.; Hynes, R. O.; Klein, W. H.; Marzluff, W.; McClay, D. R.; Morris, R. L.; Mushegian, A.; Rast, J. P.; Sodergren, E.; Smith, L. C.; Thorndyke, M. C.; Vacquier, V. D.; Weinstock, G. M.; Wessel, G. M.; Wray, G.; Zhang, L.; Sodergren, E.; Weinstock, G. M.; Angerer, R. C.; Angerer, L. M.; Cameron, R. A.; Davidson, E. H.; Elsik, C. G.; Ermolaeva, O.; Hlavina, W.; Hofmann, G.; Kitts, P.; Landrum, M. J.; Mackey, A. J.; Maglott, D.; Panopoulou, G.; Poustka, A. J.; Pruitt, K.; Sapojnikov, V.; Song, X.; Souvorov, A.; Solovyev, V.; Wei, Z.; Whittaker, C. A.; Worley, K.; Zhang, L.; Sodergren, E.; Weinstock, G. M.; Durbin, K. J.; Gibbs, R. A.; Shen, Y.; Song, X.; Worley, K.; Zhang, L.; Wray, G.; Fedrigo, O.; Garfield, D.; Haygood, R.; Primus, A.; Satija, R.; Severson, T.; Zhang, L.; Sodergren, E.; Weinstock, G. M.; Gonzalez-Garay, M. L.; Jackson, A. R.; Milosavljevic, A.; Song, X.; Tong, M.; Worley, K.; Ettensohn, C. A.; Cameron, R. A.; Killian, C. E.; Landrum, M. J.; Livingston, B. T.; Wilt, F. H.; Coffman, J. A.; Marzluff, W.; Mushegian, A.; Adams, N.; Belle, R.; Carbonneau, S.; Cheung, R.; Cormier, P.; Cosson, B.; Croce, J.; Fernandez-Guerra, A., *et al.* (2006), Research article - The genome of the sea urchin *Strongylocentrotus purpuratus*, *Science*, **314**: 941-952.

Srivastava, M.; Simakov, O.; Chapman, J.; Fahey, B.; Gauthier, M. E. A.; Mitros, T.; Richards, G. S.; Conaco, C.; Dacre, M.; Hellsten, U.; Larroux, C.; Putnam, N. H.; Stanke, M.; Adamska, M.; Darling, A.; Degnan, S. M.; Oakley, T. H.; Plachetzki, D. C.; Zhai, Y.; Adamski, M.; Calcino, A.; Cummins, S. F.; Goodstein, D. M.; Harris, C.; Jackson, D. J.; Leys, S. P.; Shu, S.; Woodcroft, B. J.; Vervoort, M.; Kosik, K. S.; Manning, G.; Degnan, B. M. & Rokhsar, D. S. (2010), The *Amphimedon queenslandica* genome and the evolution of animal complexity, *Nature*, **466**: 720-U723.

Stepicheva, N. A. & Song, J. L. (2014), High throughput microinjections of sea urchin zygotes, *Journal of Visualized Experiments*, e50841.

Stroud, R. M. & Walter, P. (1999), Signal sequence recognition and protein targeting, *Current Opinion in Structural Biology*, **9**: 754-759.

Suttle, C. A. (2005), Viruses in the sea, *Nature*, **437**: 356-361.

Swarts, D. C.; Makarova, K.; Wang, Y.; Nakanishi, K.; Ketting, R. F.; Koonin, E. V.; Patel, D. J. & van der Oost, J. (2014), The evolutionary journey of Argonaute proteins, *Nature Structural & Molecular Biology*, **21**: 743-753.

Talmadge, K.; Stahl, S. & Gilbert, W. (1980), Eukaryotic signal sequence transports insulin antigen in *Escherichia-coli*, *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences*, **77**: 3369-3373.

Tenson, T. & Ehrenberg, M. (2002), Regulatory nascent peptides in the ribosomal tunnel, *Cell*, **108**: 591-594.

Tu, Q.; Cameron, R. A.; Worley, K. C.; Gibbs, R. A. & Davidson, E. H. (2012), Gene structure in the sea urchin *Strongylocentrotus purpuratus* based on transcriptome analysis, *Genome Research*, **22**: 2079-2087.

Tully, D. C. & Fares, M. A. (2008), The tale of a modern animal plague: Tracing the evolutionary history and determining the time-scale for foot and mouth disease virus, *Virology*, **382**: 250-256.

van Rij, R. P. & Berezikov, E. (2009), Small RNAs and the control of transposons and viruses in *Drosophila*, *Trends in Microbiology*, **17**: 163-171.

van Rij, R. P.; Saleh, M. C.; Berry, B.; Foo, C.; Houk, A.; Antoniewski, C. & Andino, R. (2006), The RNA silencing endonuclease Argonaute 2 mediates specific antiviral immunity in *Drosophila melanogaster*, *Genes & Development*, **20**: 2985-2995.

Varoqui, H.; Zhu, H. M.; Yao, D. D.; Ming, H. & Erickson, J. D. (2000), Cloning and functional identification of a neuronal glutamine transporter, *Journal of Biological Chemistry*, **275**: 4049-4054.

von Heijne, G. (1986), A new method for predicting signal sequence cleavage sites, *Nucleic Acids Research*, **14**: 4683-4690.

Voss, N. R.; Gerstein, M.; Steitz, T. A. & Moore, P. B. (2006), The geometry of the ribosomal polypeptide exit tunnel, *Journal of Molecular Biology*, **360**: 893-906.

Walsh, A. M.; Kortschak, R. D.; Gardner, M. G.; Bertozzi, T. & Adelson, D. L. (2013), Widespread horizontal transfer of retrotransposons, *Proceedings of the National Academy of Sciences of the United States of America*, **110**: 1012-1016.

Wible, J. R.; Rougier, G. W.; Novacek, M. J. & Asher, R. J. (2007), Cretaceous eutherians and Laurasian origin for placental mammals near the K/T boundary, *Nature*, **447**: 1003-1006.

Wicker, T.; Sabot, F.; Hua-Van, A.; Bennetzen, J. L.; Capy, P.; Chalhoub, B.; Flavell, A.; Leroy, P.; Morgante, M.; Panaud, O.; Paux, E.; SanMiguel, P. & Schulman, A. H. (2007), A unified classification system for eukaryotic transposable elements, *Nature Reviews Genetics*, **8**: 973-982.

Wilhelm, S. W.; Weinbauer, M. G.; Suttle, C. A.; Pledger, R. J. & Mitchell, D. L. (1998), Measurements of DNA damage and photoreactivation imply that most viruses in marine surface waters are infective, *Aquatic Microbial Ecology*, **14**: 215-222.

Wilmanski, J. M.; Petnicki-Ocwieja, T. & Kobayashi, K. S. (2008), NLR proteins: integral members of innate immunity and mediators of inflammatory diseases, *Journal of Leukocyte Biology*, **83**: 13-30.

Wilson, D. N. & Beckmann, R. (2011), The ribosomal tunnel as a functional environment for nascent polypeptide folding and translational stalling, *Current Opinion in Structural Biology*, **21**: 274-282.

Woolhead, C. A.; McCormick, P. J. & Johnson, A. E. (2004), Nascent membrane and secretory proteins differ in FRET-detected folding far inside the ribosome and in their exposure to ribosomal proteins, *Cell*, **116**: 725-736.

Yoon, S. H.; Park, W.; King, D. P. & Kim, H. (2011), Phylogenomics and molecular evolution of foot-and-mouth disease virus, *Molecules and Cells*, **31**: 413-421.

Zhouravleva, G.; Frolova, L.; Legoff, X.; Leguellec, R.; Ingevechtomov, S.; Kisselev, L. & Philippe, M. (1995), Termination of translation in eukaryotes is governed by 2 interacting polypeptide-chain release factors, ERF1 and ERF3, *EMBO Journal*, **14**: 4065-4072.

Appendix A.

Dundee Cell Products Modified pBluescript *-with SmaI site highlighted (where 2A gene-blocks were inserted)*

CTGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCC
TAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGC
TCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGC
CATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAA
CAACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGC
TGATTTAACAAAAATTTAACGCGAATTTTAACAAAATATTAACGCTTACAATTTGCCATTCGCCATTCAGGCTGCGCAACT
GTTGGGAAGGGCGATCGGTGCGGGCCTCTTCGCTATTACGCCAGCTGGCGAAAGGGGGATGTGCTGCAAGGCGATTAAGTT
GGGTAACGCCAGGGTTTTCCCAGTCACGACGTTGTAAAACGACGGCCAGTGAATTGTAATACGACTCACTATAGGGCGA**CC**
CGGGGATATCCTCGAGGTTCCCTTTAGTGAGGGTTAATTGCGAGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGA
AATTGTTATCCGCTCACAATTCCACACAACATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGC
TAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGC
CAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGG
CTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATG
TGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGAC
GAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGA
AGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCG
CTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCC
GTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCA
GCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGC
TACACTAGAAGAACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCC
GGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAA
GATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCA
AAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCT
GACAGTTACCAATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCC
GTCGTGTAGATAACTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCACCG
GCTCCAGATTTATCAGCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATC
CAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCTACA
GGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACATGATCC
CCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTTATCACTC
ATGGTTATGGCAGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACC
AAGTCATTCTGAGAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATAATACCGCGCCACATAGC
AGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACTCTCAAGGATCTTACCGCTGTTGAGATCCAGT
TCGATGTAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGA
AGGCAAAATGCCGCAAAAAAGGGAATAAGGGCGACACGGAAATGTTGAATACTCATACTCTTCCTTTTTCAATATTATTGA
AGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGGTTCCGCGC
ACATTTCCCCGAAAAGTGCCAC

Appendix A

pSTA1 *(encodes eGFP-FMDV2A-GUS, 8023bp)*

GACGGATCGGGAGATCTCCCGATCCCCTATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATC
TGCTCCCTGCTTGTGTGTTGGAGGTCGCTGAGTAGTGCGCGAGCAAAATTTAAGCTACAACAAGGCAAGGCTTGACCGACA
ATTGCATGAAGAATCTGCTTAGGGTTAGGCGTTTTGCGCTGCTTCGCGATGTACGGGCCAGATATACGCGTTGACATTGAT
TATTGACTAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACATAACTTA
CGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGACGTATGTTCCCATAGTAACGC
CAATAGGGACTTTCCATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTATCATA
TGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTATGGGACT
TTCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACATCAATGGGCGTG
GATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAAC
GGGACTTTCCAAAATGTCGTAACAACTCCGCCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAA
GCAGAGCTCTCTGGCTAACTAGAGAACCCACTGCTTACTGGCTTATCGAAATTAATACGACTCACTATAGGGAGACCCAAG
CTGGCTAGCGTTTAAACTTAAGCTTGGTACCGAGCTCGGATCCACCATGGGGCACCACCACCACCACCACGGTAAAGGAGA
AGAACTTTTCACTGGAGTTGTCCCAATTCTTGTTGAATTAGATGGTGATGTTAATGGGCACAAATTTTCTGTCAGTGGAGA
GGGTGAAGGTGATGCAACATACGGAAAACTTACCCTTAAATTTATTTGCACTACTGGAAAACTACCTGTTCCATGGCCAAC
ACTTGTCACTACTTTCTCTTATGGTGTTCAATGCTTTTCAAGATACCCAGATCATATGAAACGGCATGACTTTTTCAAGAG
TGCCATGCCCGAAGGTTATGTACAGGAAAGAACTATATTTTTCAAAGATGACGGGAACTACAAGACACGTGCTGAAGTCAA
GTTTGAAGGTGATACCCTTGTTAATAGAATCGAGTTAAAAGGTATTGATTTTAAAGAAGATGGAAACATTCTTGGACACAA
ATTGGAATACAACTATAACTCACACAATGTATACATCATGGCAGACAAACAAAAGAATGGAATCAAAGTTAACTTCAAAAT
TAGACACAACATTGAAGATGGAAGCGTTCAACTAGCAGACCATTATCAACAAAATACTCCAATTGGCGATGGCCCTGTCCT
TTTACCAGACAACCATTACCTGTCCACACAATCTGCCCTTTCGAAAGATCCCAACGAAAAGAGAGACCACATGGTCCTTCT
TGAGTTTGTAACAGCTGCTGGGATTACACATGGCATGGACGAACTATACAAGTCCGGGTCTAGAGGAGCATGCCAGCTGTT
GAATTTTGACCTTCTTAAGCTTGCGGGAGACGTCGAGTCCAACCCCGGGCCCCACCACCACCACCACCACTTACGTCCTGT
AGAAACCCCAACCCGTGAAATCAAAAAACTCGACGGCCTGTGGGCATTCAGTCTGGATCGCGAAAACTGTGGAATTGATCA
GCGTTGGTGGGAAAGCGCGTTACAAGAAAGCCGGGCAATTGCTGTGCCAGGCAGTTTTAACGATCAGTTCGCCGATGCAGA
TATTCGTAATTATGCGGGCAACGTCTGGTATCAGCGCGAAGTCTTTATACCGAAAGGTTGGGCAGGCCAGCGTATCGTGCT
GCGTTTCGATGCGGTCACTCATTACGGCAAAGTGTGGGTCAATAATCAGGAAGTGATGGAGCATCAGGGCGGCTATACGCC
ATTTGAAGCCGATGTCACGCCGTATGTTATTGCCGGGAAAAGTGTACGTATCACCGTTTGTGTGAACAACGAACTGAACTG
GCAGACTATCCCGCCGGGAATGGTGATTACCGACGAAAACGGCAAGAAAAAGCAGTCTTACTTCCATGATTTCTTTAACTA
TGCCGGAATCCATCGCAGCGTAATGCTCTACACCACGCCGAACACCTGGGTGGACGATATCACCGTGGTGACGCATGTCGC
GCAAGACTGTAACCACGCGTCTGTTGACTGGCAGGTGGTGGCCAATGGTGATGTCAGCGTTGAACTGCGTGATGCGGATCA
ACAGGTGGTTGCAACTGGACAAGGCACTAGCGGGACTTTGCAAGTGGTGAATCCGCACCTCTGGCAACCGGGTGAAGGTTA
TCTCTATGAACTGTGCGTCACAGCCAAAAGCCAGACAGAGTGTGATATCTACCCGCTTCGCGTCGGCATCCGGTCAGTGGC
AGTGAAGGGCCAACAGTTCCTGATTAACCACAAACCGTTCTACTTTACTGGCTTTGGTCGTCATGAAGATGCGGACTTACG
TGGCAAAGGATTCGATAACGTGCTGATGGTGCACGACCACGCATTAATGGACTGGATTGGGGCCAACTCCTACCGTACCTC
GCATTACCCTTACGCTGAAGAGATGCTCGACTGGGCAGATGAACATGGCATCGTGGTGATTGATGAAACTGCTGCTGTCGG
CTTTAACCTCTCTTTAGGCATTGGTTTCGAAGCGGGCAACAAGCCGAAAGAACTGTACAGCGAAGAGGCAGTCAACGGGGA
AACTCAGCAAGCGCACTTACAGGCGATTAAAGAGCTGATAGCGCGTGACAAAAACCACCCAAGCGTGGTGATGTGGAGTAT
TGCCAACGAACCGGATACCCGTCCGCAAGTGCACGGGAATATTTCGCCACTGGCGGAAGCAACGCGTAAACTCGACCCGAC
GCGTCCGATCACCTGCGTCAATGTAATGTTCTGCGACGCTCACACCGATACCATCAGCGATCTCTTTGATGTGCTGTGCCT
GAACCGTTATTACGGATGGTATGTCCAAAGCGGCGATTTGGAAACGGCAGAGAAGGTACTGGAAAAAGAACTTCTGGCCTG
GCAGGAGAAACTGCATCAGCCGATTATCATCACCGAATACGGCGTGGATACGTTAGCCGGGCTGCACTCAATGTACACCGA
CATGTGGAGTGAAGAGTATCAGTGTGCATGGCTGGATATGTATCACCGCGTCTTTGATCGCGTCAGCGCCGTCGTCGGTGA
ACAGGTATGGAATTTCGCCGATTTTGCGACCTCGCAAGGCATATTGCGCGTTGGCGGTAACAAGAAAGGGATCTTCACTCG
CGACCGCAAACCGAAGTCGGCGGCTTTTCTGCTGCAAAAACGCTGGACTGGCATGAACTTCGGTGAAAAACCGCAGCAGGG
AGGCAAACAATCTAGTGCGGCCGCTCGAGGCCCGTTTAAACCCGCTGATCAGCCTCGACTGTGCCTTCTAGTTGCCAGCCA
TCTGTTGTTTGCCCCTCCCCCGTGCCTTCCTTGACCCTGGAAGGTGCCACTCCCACTGTCCTTTCCTAATAAAATGAGGAA
ATTGCATCGCATTGTCTGAGTAGGTGTCATTCTATTCTGGGGGGTGGGGTGGGGCAGGACAGCAAGGGGGAGGATTGGGAA
GACAATAGCAGGCATGCTGGGGATGCGGTGGGCTCTATGGCTTCTGAGGCGGAAAGAACCAGCTGGGGCTCTAGGGGGTAT
CCCCACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCC
CTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGG
CTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGG
CCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGA
ACAACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAG
CTGATTTAACAAAAATTTAACGCGAATTAATTCTGTGGAATGTGTGTCAGTTAGGGTGTGGAAAGTCCCCAGGCTCCCCAG
CAGGCAGAAGTATGCAAAGCATGCATCTCAATTAGTCAGCAACCAGGTGTGGAAAGTCCCCAGGCTCCCCAGCAGGCAGAA
GTATGCAAAGCATGCATCTCAATTAGTCAGCAACCATAGTCCCGCCCCTAACTCCGCCCATCCCGCCCCTAACTCCGCCCA
GTTCCGCCCATTCTCCGCCCCATGGCTGACTAATTTTTTTTATTTATGCAGAGGCCGAGGCCGCCTCTGCCTCTGAGCTAT
TCCAGAAGTAGTGAGGAGGCTTTTTTGGAGGCCTAGGCTTTTGCAAAAAGCTCCCGGGAGCTTGTATATCCATTTTCGGAT
CTGATCAAGAGACAGGATGAGGATCGTTTCGCATGATTGAACAAGATGGATTGCACGCAGGTTCTCCGGCCGCTTGGGTGG
AGAGGCTATTCGGCTATGACTGGGCACAACAGACAATCGGCTGCTCTGATGCCGCCGTGTTCCGGCTGTCAGCGCAGGGGC
GCCCGGTTCTTTTTGTCAAGACCGACCTGTCCGGTGCCCTGAATGAACTGCAGGACGAGGCAGCGCGGCTATCGTGGCTGG
CCACGACGGGCGTTCCTTGCGCAGCTGTGCTCGACGTTGTCACTGAAGCGGGAAGGGACTGGCTGCTATTGGGCGAAGTGC
CGGGGCAGGATCTCCTGTCATCTCACCTTGCTCCTGCCGAGAAAGTATCCATCATGGCTGATGCAATGCGGCGGCTGCATA
CGCTTGATCCGGCTACCTGCCCATTCGACCACCAAGCGAAACATCGCATCGAGCGAGCACGTACTCGGATGGAAGCCGGTC

Appendix A

TTGTCGATCAGGATGATCTGGACGAAGAGCATCAGGGGCTCGCGCCAGCCGAACTGTTCGCCAGGCTCAAGGCGCGCATGC
CCGACGGCGAGGATCTCGTCGTGACCCATGGCGATGCCTGCTTGCCGAATATCATGGTGGAAAATGGCCGCTTTTCTGGAT
TCATCGACTGTGGCCGGCTGGGTGTGGCGGACCGCTATCAGGACATAGCGTTGGCTACCCGTGATATTGCTGAAGAGCTTG
GCGGCGAATGGGCTGACCGCTTCCTCGTGCTTTACGGTATCGCCGCTCCCGATTCGCAGCGCATCGCCTTCTATCGCCTTC
TTGACGAGTTCTTCTGAGCGGGACTCTGGGGTTCGAAATGACCGACCAAGCGACGCCCAACCTGCCATCACGAGATTTCGA
TTCCACCGCCGCCTTCTATGAAAGGTTGGGCTTCGGAATCGTTTTCCGGGACGCCGGCTGGATGATCCTCCAGCGCGGGGA
TCTCATGCTGGAGTTCTTCGCCCACCCCAACTTGTTTATTGCAGCTTATAATGGTTACAAATAAAGCAATAGCATCACAAA
TTTCACAAATAAAGCATTTTTTTCACTGCATTCTAGTTGTGGTTTGTCCAAACTCATCAATGTATCTTATCATGTCTGTAT
ACCGTCGACCTCTAGCTAGAGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCC
ACACAACATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCG
CTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTT
GCGTATTGGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCA
CTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGC
CAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTC
AAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGT
TCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAG
GTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTT
ATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAG
CAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGAACAGTATTTGG
TATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAG
CGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTC
TGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTTT
AAATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACCAATGCTTAATCAGTGA
GGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCCGTCGTGTAGATAACTACGATACGGGA
GGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTATCAGCAATAAACCA
GCCAGCCGGAAGGGCCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGC
TAGAGTAAGTAGTTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCTACAGGCATCGTGGTGTCACGCTCGTCGTT
TGGTATGGCTTCATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAG
CTCCTTCGGTCCTCCGATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTC
TCTTACTGTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACCAAGTCATTCTGAGAATAGTGTATGCG
GCGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATAATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATCATTGG
AAAACGTTCTTCGGGGCGAAAACTCTCAAGGATCTTACCGCTGTTGAGATCCAGTTCGATGTAACCCACTCGTGCACCCAA
CTGATCTTCAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCGCAAAAAAGGGAAT
AAGGGCGACACGGAAATGTTGAATACTCATACTCTTCCTTTTTCAATATTATTGAAGCATTTATCAGGGTTATTGTCTCAT
GAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGA
CGTC

Appendix A

pJN1 *(encodes mCherryFP-TaV2A-eGFP, 6864bp)*

GACGGATCGGGAGATCTCCCGATCCCCTATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATC
TGCTCCCTGCTTGTGTGTTGGAGGTCGCTGAGTAGTGCGCGAGCAAAATTTAAGCTACAACAAGGCAAGGCTTGACCGACA
ATTGCATGAAGAATCTGCTTAGGGTTAGGCGTTTTGCGCTGCTTCGCGATGTACGGGCCAGATATACGCGTTGACATTGAT
TATTGACTAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACATAACTTA
CGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGACGTATGTTCCCATAGTAACGC
CAATAGGGACTTTCCATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTATCATA
TGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTATGGGACT
TTCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACATCAATGGGCGTG
GATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAAC
GGGACTTTCCAAAATGTCGTAACAACTCCGCCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAA
GCAGAGCTCTCTGGCTAACTAGAGAACCCACTGCTTACTGGCTTATCGAAATTAATACGACTCACTATAGGGAGACCCAAG
CTGGCTAGCGTTTAAACTTAAGCTTGGTACCGAGCTCGGATCCATGGTGAGCAAGGGCGAGGAGGATAACATGGCCATCAT
CAAGGAGTTCATGCGCTTCAAGGTGCACATGGAGGGCTCCGTGAACGGCCACGAGTTCGAGATCGAGGGCGAGGGCGAGGG
CCGCCCCTACGAGGGCACCCAGACCGCCAAGCTGAAGGTGACCAAGGGTGGCCCCCTGCCCTTCGCCTGGGACATCCTGTC
CCCTCAGTTCATGTACGGCTCCAAGGCCTACGTGAAGCACCCCGCCGACATCCCCGACTACTTGAAGCTGTCCTTCCCCGA
GGGCTTCAAGTGGGAGCGCGTGATGAACTTCGAGGACGGCGGCGTGGTGACCGTGACCCAGGACTCCTCCCTGCAGGACGG
CGAGTTCATCTACAAGGTGAAGCTGCGCGGCACCAACTTCCCCTCCGACGGCCCCGTAATGCAGAAGAAGACCATGGGCTG
GGAGGCCTCCTCCGAGCGGATGTACCCCGAGGACGGCGCCCTGAAGGGCGAGATCAAGCAGAGGCTGAAGCTGAAGGACGG
CGGCCACTACGACGCTGAGGTCAAGACCACCTACAAGGCCAAGAAGCCCGTGCAGCTGCCCGGCGCCTACAACGTCAACAT
CAAGTTGGACATCACCTCCCACAACGAGGACTACACCATCGTGGAACAGTACGAACGCGCCGAGGGCCGCCACTCCACCGG
CGGCATGGATGAATTGTACAAATCTAGAGCCGAGGGCAGGGGAAGTCTTCTAACATGCGGGGACGTGGAGGAAAATCCCGG
GCCCGATATCGTGTCCAAAGGGGAAGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGG
CCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGG
CAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACAT
GAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAA
CTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGA
GGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGAAGAA
CGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACAC
CCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGA
GAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTATAAGTAACT
CGAGGCCCGTTTAAACCCGCTGATCAGCCTCGACTGTGCCTTCTAGTTGCCAGCCATCTGTTGTTTGCCCCTCCCCCGTGC
CTTCCTTGACCCTGGAAGGTGCCACTCCCACTGTCCTTTCCTAATAAAATGAGGAAATTGCATCGCATTGTCTGAGTAGGT
GTCATTCTATTCTGGGGGGTGGGGTGGGGCAGGACAGCAAGGGGGAGGATTGGGAAGACAATAGCAGGCATGCTGGGGATG
CGGTGGGCTCTATGGCTTCTGAGGCGGAAAGAACCAGCTGGGGCTCTAGGGGGTATCCCCACGCGCCCTGTAGCGGCGCAT
TAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCT
TCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTG
CTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTC
GCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCT
ATTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGA
ATTAATTCTGTGGAATGTGTGTCAGTTAGGGTGTGGAAAGTCCCCAGGCTCCCCAGCAGGCAGAAGTATGCAAAGCATGCA
TCTCAATTAGTCAGCAACCAGGTGTGGAAAGTCCCCAGGCTCCCCAGCAGGCAGAAGTATGCAAAGCATGCATCTCAATTA
GTCAGCAACCATAGTCCCGCCCCTAACTCCGCCCATCCCGCCCCTAACTCCGCCCAGTTCCGCCCATTCTCCGCCCCATGG
CTGACTAATTTTTTTTATTTATGCAGAGGCCGAGGCCGCCTCTGCCTCTGAGCTATTCCAGAAGTAGTGAGGAGGCTTTTT
TGGAGGCCTAGGCTTTTGCAAAAAGCTCCCGGGAGCTTGTATATCCATTTTCGGATCTGATCAAGAGACAGGATGAGGATC
GTTTCGCATGATTGAACAAGATGGATTGCACGCAGGTTCTCCGGCCGCTTGGGTGGAGAGGCTATTCGGCTATGACTGGGC
ACAACAGACAATCGGCTGCTCTGATGCCGCCGTGTTCCGGCTGTCAGCGCAGGGGCGCCCGGTTCTTTTTGTCAAGACCGA
CCTGTCCGGTGCCCTGAATGAACTGCAGGACGAGGCAGCGCGGCTATCGTGGCTGGCCACGACGGGCGTTCCTTGCGCAGC
TGTGCTCGACGTTGTCACTGAAGCGGGAAGGGACTGGCTGCTATTGGGCGAAGTGCCGGGGCAGGATCTCCTGTCATCTCA
CCTTGCTCCTGCCGAGAAAGTATCCATCATGGCTGATGCAATGCGGCGGCTGCATACGCTTGATCCGGCTACCTGCCCATT
CGACCACCAAGCGAAACATCGCATCGAGCGAGCACGTACTCGGATGGAAGCCGGTCTTGTCGATCAGGATGATCTGGACGA
AGAGCATCAGGGGCTCGCGCCAGCCGAACTGTTCGCCAGGCTCAAGGCGCGCATGCCCGACGGCGAGGATCTCGTCGTGAC
CCATGGCGATGCCTGCTTGCCGAATATCATGGTGGAAAATGGCCGCTTTTCTGGATTCATCGACTGTGGCCGGCTGGGTGT
GGCGGACCGCTATCAGGACATAGCGTTGGCTACCCGTGATATTGCTGAAGAGCTTGGCGGCGAATGGGCTGACCGCTTCCT
CGTGCTTTACGGTATCGCCGCTCCCGATTCGCAGCGCATCGCCTTCTATCGCCTTCTTGACGAGTTCTTCTGAGCGGGACT
CTGGGGTTCGAAATGACCGACCAAGCGACGCCCAACCTGCCATCACGAGATTTCGATTCCACCGCCGCCTTCTATGAAAGG
TTGGGCTTCGGAATCGTTTTCCGGGACGCCGGCTGGATGATCCTCCAGCGCGGGGATCTCATGCTGGAGTTCTTCGCCCAC
CCCAACTTGTTTATTGCAGCTTATAATGGTTACAAATAAAGCAATAGCATCACAAATTTCACAAATAAAGCATTTTTTTCA
CTGCATTCTAGTTGTGGTTTGTCCAAACTCATCAATGTATCTTATCATGTCTGTATACCGTCGACCTCTAGCTAGAGCTTG
GCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACAACATACGAGCCGGAAGCATA
AAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGA
AACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCTCTTCCGCTTCC
TCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCC
ACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTG
CTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACA
GGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATAC
CTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTT

Appendix A

CGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCC
AACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCT
ACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGAACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTT
ACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTTTTTTTGTTTGCAAGCAGCAG
ATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCA
CGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAAGTTTTAAATCA
ATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACCAATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTA
TTTCGTTCATCCATAGTTGCCTGACTCCCCGTCGTGTAGATAACTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCT
GCAATGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTATCAGCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGA
AGTGGTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAAT
AGTTTGCGCAACGTTGTTGCCATTGCTACAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGT
TCCCAACGATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGTC
AGAAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAAGA
TGCTTTTCTGTGACTGGTGAGTACTCAACCAAGTCATTCTGAGAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCG
TCAATACGGGATAATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACTC
TCAAGGATCTTACCGCTGTTGAGATCCAGTTCGATGTAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTTTACTTTC
ACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCGCAAAAAAGGGAATAAGGGCGACACGGAAATGTTGAATA
CTCATACTCTTCCTTTTTCAATATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTATT
TAGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGACGTC

Appendix A

p132 *(encodes STR6-mCherryFP, 6152bp)*

GACGGATCGGGAGATCTCCCGATCCCCTATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATC
TGCTCCCTGCTTGTGTGTTGGAGGTCGCTGAGTAGTGCGCGAGCAAAATTTAAGCTACAACAAGGCAAGGCTTGACCGACA
ATTGCATGAAGAATCTGCTTAGGGTTAGGCGTTTTGCGCTGCTTCGCGATGTACGGGCCAGATATACGCGTTGACATTGAT
TATTGACTAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACATAACTTA
CGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGACGTATGTTCCCATAGTAACGC
CAATAGGGACTTTCCATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTATCATA
TGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTATGGGACT
TTCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACATCAATGGGCGTG
GATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAAC
GGGACTTTCCAAAATGTCGTAACAACTCCGCCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAA
GCAGAGCTCTCTGGCTAACTAGAGAACCCACTGCTTACTGGCTTATCGAAATTAATACGACTCACTATAGGGAGACCCAAG
CTGGCTAGCTCTAGAACCATGGATGGATTCTGTCTTCTCTATCTGCTCCTGATCCTCTTGATGAGATCTGGTGACGTTGAA
ACCAATCCCGGGCCCATCGATGTGAGCAAGGGCGAGGAGGATAACATGGCCATCATCAAGGAGTTCATGCGCTTCAAGGTG
CACATGGAGGGCTCCGTGAACGGCCACGAGTTCGAGATCGAGGGCGAGGGCGAGGGCCGCCCCTACGAGGGCACCCAGACC
GCCAAGCTGAAGGTGACCAAGGGTGGCCCCCTGCCCTTCGCCTGGGACATCCTGTCCCCTCAGTTCATGTACGGCTCCAAG
GCCTACGTGAAGCACCCCGCCGACATCCCCGACTACTTGAAGCTGTCCTTCCCCGAGGGCTTCAAGTGGGAGCGCGTGATG
AACTTCGAGGACGGCGGCGTGGTGACCGTGACCCAGGACTCCTCCCTGCAGGACGGCGAGTTCATCTACAAGGTGAAGCTG
CGCGGCACCAACTTCCCCTCCGACGGCCCCGTAATGCAGAAGAAGACCATGGGCTGGGAGGCCTCCTCCGAGCGGATGTAC
CCCGAGGACGGCGCCCTGAAGGGCGAGATCAAGCAGAGGCTGAAGCTGAAGGACGGCGGCCACTACGACGCTGAGGTCAAG
ACCACCTACAAGGCCAAGAAGCCCGTGCAGCTGCCCGGCGCCTACAACGTCAACATCAAGTTGGACATCACCTCCCACAAC
GAGGACTACACCATCGTGGAACAGTACGAACGCGCCGAGGGCCGCCACTCCACCGGCGGCATGGATGAATTGTACAAATAA
GCGGCCGCAATCGAATTCAAGCTTAAGTTTAAACCGCTGATCAGCCTCGACTGTGCCTTCTAGTTGCCAGCCATCTGTTGT
TTGCCCCTCCCCCGTGCCTTCCTTGACCCTGGAAGGTGCCACTCCCACTGTCCTTTCCTAATAAAATGAGGAAATTGCATC
GCATTGTCTGAGTAGGTGTCATTCTATTCTGGGGGGTGGGGTGGGGCAGGACAGCAAGGGGGAGGATTGGGAAGACAATAG
CAGGCATGCTGGGGATGCGGTGGGCTCTATGGCTTCTGAGGCGGAAAGAACCAGCTGGGGCTCTAGGGGGTATCCCCACGC
GCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCC
CGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTT
AGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCC
CTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACT
CAACCCTATCTCGGTCTATTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTA
ACAAAAATTTAACGCGAATTAATTCTGTGGAATGTGTGTCAGTTAGGGTGTGGAAAGTCCCCAGGCTCCCCAGCAGGCAGA
AGTATGCAAAGCATGCATCTCAATTAGTCAGCAACCAGGTGTGGAAAGTCCCCAGGCTCCCCAGCAGGCAGAAGTATGCAA
AGCATGCATCTCAATTAGTCAGCAACCATAGTCCCGCCCCTAACTCCGCCCATCCCGCCCCTAACTCCGCCCAGTTCCGCC
CATTCTCCGCCCCATGGCTGACTAATTTTTTTTATTTATGCAGAGGCCGAGGCCGCCTCTGCCTCTGAGCTATTCCAGAAG
TAGTGAGGAGGCTTTTTTGGAGGCCTAGGCTTTTGCAAAAAGCTCCCGGGAGCTTGTATATCCATTTTCGGATCTGATCAA
GAGACAGGATGAGGATCGTTTCGCATGATTGAACAAGATGGATTGCACGCAGGTTCTCCGGCCGCTTGGGTGGAGAGGCTA
TTCGGCTATGACTGGGCACAACAGACAATCGGCTGCTCTGATGCCGCCGTGTTCCGGCTGTCAGCGCAGGGGCGCCCGGTT
CTTTTTGTCAAGACCGACCTGTCCGGTGCCCTGAATGAACTGCAGGACGAGGCAGCGCGGCTATCGTGGCTGGCCACGACG
GGCGTTCCTTGCGCAGCTGTGCTCGACGTTGTCACTGAAGCGGGAAGGGACTGGCTGCTATTGGGCGAAGTGCCGGGGCAG
GATCTCCTGTCATCTCACCTTGCTCCTGCCGAGAAAGTATCCATCATGGCTGATGCAATGCGGCGGCTGCATACGCTTGAT
CCGGCTACCTGCCCATTCGACCACCAAGCGAAACATCGCATCGAGCGAGCACGTACTCGGATGGAAGCCGGTCTTGTCGAT
CAGGATGATCTGGACGAAGAGCATCAGGGGCTCGCGCCAGCCGAACTGTTCGCCAGGCTCAAGGCGCGCATGCCCGACGGC
GAGGATCTCGTCGTGACCCATGGCGATGCCTGCTTGCCGAATATCATGGTGGAAAATGGCCGCTTTTCTGGATTCATCGAC
TGTGGCCGGCTGGGTGTGGCGGACCGCTATCAGGACATAGCGTTGGCTACCCGTGATATTGCTGAAGAGCTTGGCGGCGAA
TGGGCTGACCGCTTCCTCGTGCTTTACGGTATCGCCGCTCCCGATTCGCAGCGCATCGCCTTCTATCGCCTTCTTGACGAG
TTCTTCTGAGCGGGACTCTGGGGTTCGAAATGACCGACCAAGCGACGCCCAACCTGCCATCACGAGATTTCGATTCCACCG
CCGCCTTCTATGAAAGGTTGGGCTTCGGAATCGTTTTCCGGGACGCCGGCTGGATGATCCTCCAGCGCGGGGATCTCATGC
TGGAGTTCTTCGCCCACCCCAACTTGTTTATTGCAGCTTATAATGGTTACAAATAAAGCAATAGCATCACAAATTTCACAA
ATAAAGCATTTTTTTCACTGCATTCTAGTTGTGGTTTGTCCAAACTCATCAATGTATCTTATCATGTCTGTATACCGTCGA
CCTCTAGCTAGAGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACAACA
TACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGC
CCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTG
GGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGG
CGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACC
GTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGA
GGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCC
TGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCA
GTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTA
ACTATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGA
GGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGAACAGTATTTGGTATCTGCG
CTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTTTTT
TTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTC
AGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAA
AATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACCAATGCTTAATCAGTGAGGCACCTA
TCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCCGTCGTGTAGATAACTACGATACGGGAGGGCTTAC
CATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTATCAGCAATAAACCAGCCAGCCG

Appendix A

GAAGGGCCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAA
GTAGTTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCTACAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGG
CTTCATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCG
GTCCTCCGATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACTG
TCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACCAAGTCATTCTGAGAATAGTGTATGCGGCGACCGA
GTTGCTCTTGCCCGGCGTCAATACGGGATAATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTT
CTTCGGGGCGAAAACTCTCAAGGATCTTACCGCTGTTGAGATCCAGTTCGATGTAACCCACTCGTGCACCCAACTGATCTT
CAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCGCAAAAAAGGGAATAAGGGCGA
CACGGAAATGTTGAATACTCATACTCTTCCTTTTTCAATATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGAT
ACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGACGTC

Appendix A

pJN132 *(STR6-mCherry-Tav2A-eGFP, 6920bp)*

GACGGATCGGGAGATCTCCCGATCCCCTATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATC
TGCTCCCTGCTTGTGTGTTGGAGGTCGCTGAGTAGTGCGCGAGCAAAATTTAAGCTACAACAAGGCAAGGCTTGACCGACA
ATTGCATGAAGAATCTGCTTAGGGTTAGGCGTTTTGCGCTGCTTCGCGATGTACGGGCCAGATATACGCGTTGACATTGAT
TATTGACTAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACATAACTTA
CGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGACGTATGTTCCCATAGTAACGC
CAATAGGGACTTTCCATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTATCATA
TGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTATGGGACT
TTCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACATCAATGGGCGTG
GATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAAC
GGGACTTTCCAAAATGTCGTAACAACTCCGCCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAA
GCAGAGCTCTCTGGCTAACTAGAGAACCCACTGCTTACTGGCTTATCGAAATTAATACGACTCACTATAGGGAGACCCAAG
CTGGCTAGCTCTAGAACCATGGATGGATTCTGTCTTCTCTATCTGCTCCTGATCCTCTTGATGAGATCTGGTGACGTTGAA
ACCAATCCCGGGCCCATCGATGTGAGCAAGGGCGAGGAGGATAACATGGCCATCATCAAGGAGTTCATGCGCTTCAAGGTG
CACATGGAGGGCTCCGTGAACGGCCACGAGTTCGAGATCGAGGGCGAGGGCGAGGGCCGCCCCTACGAGGGCACCCAGACC
GCCAAGCTGAAGGTGACCAAGGGTGGCCCCCTGCCCTTCGCCTGGGACATCCTGTCCCCTCAGTTCATGTACGGCTCCAAG
GCCTACGTGAAGCACCCCGCCGACATCCCCGACTACTTGAAGCTGTCCTTCCCCGAGGGCTTCAAGTGGGAGCGCGTGATG
AACTTCGAGGACGGCGGCGTGGTGACCGTGACCCAGGACTCCTCCCTGCAGGACGGCGAGTTCATCTACAAGGTGAAGCTG
CGCGGCACCAACTTCCCCTCCGACGGCCCCGTAATGCAGAAGAAGACCATGGGCTGGGAGGCCTCCTCCGAGCGGATGTAC
CCCGAGGACGGCGCCCTGAAGGGCGAGATCAAGCAGAGGCTGAAGCTGAAGGACGGCGGCCACTACGACGCTGAGGTCAAG
ACCACCTACAAGGCCAAGAAGCCCGTGCAGCTGCCCGGCGCCTACAACGTCAACATCAAGTTGGACATCACCTCCCACAAC
GAGGACTACACCATCGTGGAACAGTACGAACGCGCCGAGGGCCGCCACTCCACCGGCGGCATGGATGAATTGTACAAATCT
AGAGCCGAGGGCAGGGGAAGTCTTCTAACATGCGGGGACGTGGAGGAAAATCCCGGGCCCGATATCGTGTCCAAAGGGGAA
GAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAG
GGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACC
CTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCC
GCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAG
TTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAG
CTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATC
CGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTG
CTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTG
GAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTATAAGTAACTCGAGGCCCGTTTAAACCCGCTGATC
AGCCTCGACTGTGCCTTCTAGTTGCCAGCCATCTGTTGTTTGCCCCTCCCCCGTGCCTTCCTTGACCCTGGAAGGTGCCAC
TCCCACTGTCCTTTCCTAATAAAATGAGGAAATTGCATCGCATTGTCTGAGTAGGTGTCATTCTATTCTGGGGGGTGGGGT
GGGGCAGGACAGCAAGGGGGAGGATTGGGAAGACAATAGCAGGCATGCTGGGGATGCGGTGGGCTCTATGGCTTCTGAGGC
GGAAAGAACCAGCTGGGGCTCTAGGGGGTATCCCCACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTAC
GCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGC
CGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAA
ACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTT
CTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATAAGGGATTTT
GCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTAATTCTGTGGAATGTGTGTCAG
TTAGGGTGTGGAAAGTCCCCAGGCTCCCCAGCAGGCAGAAGTATGCAAAGCATGCATCTCAATTAGTCAGCAACCAGGTGT
GGAAAGTCCCCAGGCTCCCCAGCAGGCAGAAGTATGCAAAGCATGCATCTCAATTAGTCAGCAACCATAGTCCCGCCCCTA
ACTCCGCCCATCCCGCCCCTAACTCCGCCCAGTTCCGCCCATTCTCCGCCCCATGGCTGACTAATTTTTTTTATTTATGCA
GAGGCCGAGGCCGCCTCTGCCTCTGAGCTATTCCAGAAGTAGTGAGGAGGCTTTTTTGGAGGCCTAGGCTTTTGCAAAAAG
CTCCCGGGAGCTTGTATATCCATTTTCGGATCTGATCAAGAGACAGGATGAGGATCGTTTCGCATGATTGAACAAGATGGA
TTGCACGCAGGTTCTCCGGCCGCTTGGGTGGAGAGGCTATTCGGCTATGACTGGGCACAACAGACAATCGGCTGCTCTGAT
GCCGCCGTGTTCCGGCTGTCAGCGCAGGGGCGCCCGGTTCTTTTTGTCAAGACCGACCTGTCCGGTGCCCTGAATGAACTG
CAGGACGAGGCAGCGCGGCTATCGTGGCTGGCCACGACGGGCGTTCCTTGCGCAGCTGTGCTCGACGTTGTCACTGAAGCG
GGAAGGGACTGGCTGCTATTGGGCGAAGTGCCGGGGCAGGATCTCCTGTCATCTCACCTTGCTCCTGCCGAGAAAGTATCC
ATCATGGCTGATGCAATGCGGCGGCTGCATACGCTTGATCCGGCTACCTGCCCATTCGACCACCAAGCGAAACATCGCATC
GAGCGAGCACGTACTCGGATGGAAGCCGGTCTTGTCGATCAGGATGATCTGGACGAAGAGCATCAGGGGCTCGCGCCAGCC
GAACTGTTCGCCAGGCTCAAGGCGCGCATGCCCGACGGCGAGGATCTCGTCGTGACCCATGGCGATGCCTGCTTGCCGAAT
ATCATGGTGGAAAATGGCCGCTTTTCTGGATTCATCGACTGTGGCCGGCTGGGTGTGGCGGACCGCTATCAGGACATAGCG
TTGGCTACCCGTGATATTGCTGAAGAGCTTGGCGGCGAATGGGCTGACCGCTTCCTCGTGCTTTACGGTATCGCCGCTCCC
GATTCGCAGCGCATCGCCTTCTATCGCCTTCTTGACGAGTTCTTCTGAGCGGGACTCTGGGGTTCGAAATGACCGACCAAG
CGACGCCCAACCTGCCATCACGAGATTTCGATTCCACCGCCGCCTTCTATGAAAGGTTGGGCTTCGGAATCGTTTTCCGGG
ACGCCGGCTGGATGATCCTCCAGCGCGGGGATCTCATGCTGGAGTTCTTCGCCCACCCCAACTTGTTTATTGCAGCTTATA
ATGGTTACAAATAAAGCAATAGCATCACAAATTTCACAAATAAAGCATTTTTTTCACTGCATTCTAGTTGTGGTTTGTCCA
AACTCATCAATGTATCTTATCATGTCTGTATACCGTCGACCTCTAGCTAGAGCTTGGCGTAATCATGGTCATAGCTGTTTC
CTGTGTGAAATTGTTATCCGCTCACAATTCCACACAACATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAAT
GAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAAT
GAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGG
TCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAA
AGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCC
CCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTC
CCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAA

Appendix A

GCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACG
AACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATCGC
CACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTA
ACTACGGCTACACTAGAAGAACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCT
CTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTC
AAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGAT
TATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTT
GGTCTGACAGTTACCAATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGAC
TCCCCGTCGTGTAGATAACTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCT
CACCGGCTCCAGATTTATCAGCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCCT
CCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTG
CTACAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACAT
GATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTTAT
CACTCATGGTTATGGCAGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACT
CAACCAAGTCATTCTGAGAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATAATACCGCGCCAC
ATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACTCTCAAGGATCTTACCGCTGTTGAGAT
CCAGTTCGATGTAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAA
CAGGAAGGCAAAATGCCGCAAAAAAGGGAATAAGGGCGACACGGAAATGTTGAATACTCATACTCTTCCTTTTTCAATATT
ATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGGTTC
CGCGCACATTTCCCCGAAAAGTGCCACCTGACGTC

Appendix A

pSTR6-GFP *(STR6-eGFP, 6146bp)*

GACGGATCGGGAGATCTCCCGATCCCCTATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATC
TGCTCCCTGCTTGTGTGTTGGAGGTCGCTGAGTAGTGCGCGAGCAAAATTTAAGCTACAACAAGGCAAGGCTTGACCGACA
ATTGCATGAAGAATCTGCTTAGGGTTAGGCGTTTTGCGCTGCTTCGCGATGTACGGGCCAGATATACGCGTTGACATTGAT
TATTGACTAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACATAACTTA
CGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGACGTATGTTCCCATAGTAACGC
CAATAGGGACTTTCCATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTATCATA
TGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTATGGGACT
TTCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACATCAATGGGCGTG
GATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAAC
GGGACTTTCCAAAATGTCGTAACAACTCCGCCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAA
GCAGAGCTCTCTGGCTAACTAGAGAACCCACTGCTTACTGGCTTATCGAAATTAATACGACTCACTATAGGGAGACCCAAG
CTGGCTAGCTCTAGAACCATGGATGGATTCTGTCTTCTCTATCTGCTCCTGATCCTCTTGATGAGATCTGGTGACGTTGAA
ACCAATCCCGGGCCCGATATCGTGTCCAAAGGGGAAGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGC
GACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATC
TGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCTAC
CCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAG
GACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATC
GACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGAC
AAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACTAC
CAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAA
GACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTG
TATAAGTAACTCGAGGCCCGTTTAAACCCGCTGATCAGCCTCGACTGTGCCTTCTAGTTGCCAGCCATCTGTTGTTTGCCC
CTCCCCCGTGCCTTCCTTGACCCTGGAAGGTGCCACTCCCACTGTCCTTTCCTAATAAAATGAGGAAATTGCATCGCATTG
TCTGAGTAGGTGTCATTCTATTCTGGGGGGTGGGGTGGGGCAGGACAGCAAGGGGGAGGATTGGGAAGACAATAGCAGGCA
TGCTGGGGATGCGGTGGGCTCTATGGCTTCTGAGGCGGAAAGAACCAGCTGGGGCTCTAGGGGGTATCCCCACGCGCCCTG
TAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCC
TTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTT
CCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATA
GACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCC
TATCTCGGTCTATTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAA
ATTTAACGCGAATTAATTCTGTGGAATGTGTGTCAGTTAGGGTGTGGAAAGTCCCCAGGCTCCCCAGCAGGCAGAAGTATG
CAAAGCATGCATCTCAATTAGTCAGCAACCAGGTGTGGAAAGTCCCCAGGCTCCCCAGCAGGCAGAAGTATGCAAAGCATG
CATCTCAATTAGTCAGCAACCATAGTCCCGCCCCTAACTCCGCCCATCCCGCCCCTAACTCCGCCCAGTTCCGCCCATTCT
CCGCCCCATGGCTGACTAATTTTTTTTATTTATGCAGAGGCCGAGGCCGCCTCTGCCTCTGAGCTATTCCAGAAGTAGTGA
GGAGGCTTTTTTGGAGGCCTAGGCTTTTGCAAAAAGCTCCCGGGAGCTTGTATATCCATTTTCGGATCTGATCAAGAGACA
GGATGAGGATCGTTTCGCATGATTGAACAAGATGGATTGCACGCAGGTTCTCCGGCCGCTTGGGTGGAGAGGCTATTCGGC
TATGACTGGGCACAACAGACAATCGGCTGCTCTGATGCCGCCGTGTTCCGGCTGTCAGCGCAGGGGCGCCCGGTTCTTTTT
GTCAAGACCGACCTGTCCGGTGCCCTGAATGAACTGCAGGACGAGGCAGCGCGGCTATCGTGGCTGGCCACGACGGGCGTT
CCTTGCGCAGCTGTGCTCGACGTTGTCACTGAAGCGGGAAGGGACTGGCTGCTATTGGGCGAAGTGCCGGGGCAGGATCTC
CTGTCATCTCACCTTGCTCCTGCCGAGAAAGTATCCATCATGGCTGATGCAATGCGGCGGCTGCATACGCTTGATCCGGCT
ACCTGCCCATTCGACCACCAAGCGAAACATCGCATCGAGCGAGCACGTACTCGGATGGAAGCCGGTCTTGTCGATCAGGAT
GATCTGGACGAAGAGCATCAGGGGCTCGCGCCAGCCGAACTGTTCGCCAGGCTCAAGGCGCGCATGCCCGACGGCGAGGAT
CTCGTCGTGACCCATGGCGATGCCTGCTTGCCGAATATCATGGTGGAAAATGGCCGCTTTTCTGGATTCATCGACTGTGGC
CGGCTGGGTGTGGCGGACCGCTATCAGGACATAGCGTTGGCTACCCGTGATATTGCTGAAGAGCTTGGCGGCGAATGGGCT
GACCGCTTCCTCGTGCTTTACGGTATCGCCGCTCCCGATTCGCAGCGCATCGCCTTCTATCGCCTTCTTGACGAGTTCTTC
TGAGCGGGACTCTGGGGTTCGAAATGACCGACCAAGCGACGCCCAACCTGCCATCACGAGATTTCGATTCCACCGCCGCCT
TCTATGAAAGGTTGGGCTTCGGAATCGTTTTCCGGGACGCCGGCTGGATGATCCTCCAGCGCGGGGATCTCATGCTGGAGT
TCTTCGCCCACCCCAACTTGTTTATTGCAGCTTATAATGGTTACAAATAAAGCAATAGCATCACAAATTTCACAAATAAAG
CATTTTTTTCACTGCATTCTAGTTGTGGTTTGTCCAAACTCATCAATGTATCTTATCATGTCTGTATACCGTCGACCTCTA
GCTAGAGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACAACATACGAG
CCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTT
TCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCT
CTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAA
TACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAA
AGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGC
GAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGC
TTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGG
TGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATC
GTCTTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATG
TAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGAACAGTATTTGGTATCTGCGCTCTGC
TGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTTTTTTTGTTT
GCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGA
ACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAA
GTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACCAATGCTTAATCAGTGAGGCACCTATCTCAG
CGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCCGTCGTGTAGATAACTACGATACGGGAGGGCTTACCATCTG
GCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTATCAGCAATAAACCAGCCAGCCGGAAGGG

Appendix A

CCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTT
CGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCTACAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCAT
TCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTC
CGATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACTGTCATGC
CATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACCAAGTCATTCTGAGAATAGTGTATGCGGCGACCGAGTTGCT
CTTGCCCGGCGTCAATACGGGATAATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGG
GGCGAAAACTCTCAAGGATCTTACCGCTGTTGAGATCCAGTTCGATGTAACCCACTCGTGCACCCAACTGATCTTCAGCAT
CTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCGCAAAAAAGGGAATAAGGGCGACACGGA
AATGTTGAATACTCATACTCTTCCTTTTTCAATATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATAT
TTGAATGTATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGACGTC

Appendix A

pEMX *(Signal2A-mCherryFP,6183bp)*

GACGGATCGGGAGATCTCCCGATCCCCTATGGTCGACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATC
TGCTCCCTGCTTGTGTGTTGGAGGTCGCTGAGTAGTGCGCGAGCAAAATTTAAGCTACAACAAGGCAAGGCTTGACCGACA
ATTGCATGAAGAATCTGCTTAGGGTTAGGCGTTTTGCGCTGCTTCGCGATGTACGGGCCAGATATACGCGTTGACATTGAT
TATTGACTAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACATAACTTA
CGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGACGTATGTTCCCATAGTAACGC
CAATAGGGACTTTCCATTGACGTCAATGGGTGGACTATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTATCATA
TGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTATGGGACT
TTCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACATCAATGGGCGTG
GATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAAC
GGGACTTTCCAAAATGTCGTAACAACTCCGCCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAA
GCAGAGCTCTCTGGCTAACTAGAGAACCCACTGCTTACTGGCTTATCGAAATTAATACGACTCACTATAGGGAGACCCAAG
CTGGCTAGCGTTTAAACTTAAGCTTGGTACCGAGCTC**TCTAGA**ATGACTAATGCCCTTCTATTGAGATCTGGTGATGTTGA
ACTGAACCCT**GGGCCC**ATCGATGTGAGCAAGGGCGAGGAGGATAACATGGCCATCATCAAGGAGTTCATGCGCTTCAAGGT
GCACATGGAGGGCTCCGTGAACGGCCACGAGTTCGAGATCGAGGGCGAGGGCGAGGGCCGCCCCTACGAGGGCACCCAGAC
CGCCAAGCTGAAGGTGACCAAGGGTGGCCCCCTGCCCTTCGCCTGGGACATCCTGTCCCCTCAGTTCATGTACGGCTCCAA
GGCCTACGTGAAGCACCCCGCCGACATCCCCGACTACTTGAAGCTGTCCTTCCCCGAGGGCTTCAAGTGGGAGCGCGTGAT
GAACTTCGAGGACGGCGGCGTGGTGACCGTGACCCAGGACTCCTCCCTGCAGGACGGCGAGTTCATCTACAAGGTGAAGCT
GCGCGGCACCAACTTCCCCTCCGACGGCCCCGTAATGCAGAAGAAGACCATGGGCTGGGAGGCCTCCTCCGAGCGGATGTA
CCCCGAGGACGGCGCCCTGAAGGGCGAGATCAAGCAGAGGCTGAAGCTGAAGGACGGCGGCCACTACGACGCTGAGGTCAA
GACCACCTACAAGGCCAAGAAGCCCGTGCAGCTGCCCGGCGCCTACAACGTCAACATCAAGTTGGACATCACCTCCCACAA
CGAGGACTACACCATCGTGGAACAGTACGAACGCGCCGAGGGCCGCCACTCCACCGGCGGCATGGATGAATTGTACAAATA
AGCGGCCGCTATCACTAGTGGAATTCTGCAGATATCCAGCACAGTGGCGGCGGCCCGTTTAAACCCGCTGATCAGCCTCGA
CTGTGCCTTCTAGTTGCCAGCCATCTGTTGTTTGCCCCTCCCCCGTGCCTTCCTTGACCCTGGAAGGTGCCACTCCCACTG
TCCTTTCCTAATAAAATGAGGAAATTGCATCGCATTGTCTGAGTAGGTGTCATTCTATTCTGGGGGGTGGGGTGGGGCAGG
ACAGCAAGGGGGAGGATTGGGAAGACAATAGCAGGCATGCTGGGGATGCGGTGGGCTCTATGGCTTCTGAGGCGGAAAGAA
CCAGCTGGGGCTCTAGGGGGTATCCCCACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCG
TGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTC
CCCGTCAAGCTCTAAATCGGGGCATCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTGATT
AGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATA
GTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATAAGGGATTTTGGGGATTT
CGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTAATTCTGTGGAATGTGTGTCAGTTAGGGTG
TGGAAAGTCCCCAGGCTCCCCAGGCAGGCAGAAGTATGCAAAGCATGCATCTCAATTAGTCAGCAACCAGGTGTGGAAAGT
CCCCAGGCTCCCCAGCAGGCAGAAGTATGCAAAGCATGCATCTCAATTAGTCAGCAACCATAGTCCCGCCCCTAACTCCGC
CCATCCCGCCCCTAACTCCGCCCAGTTCCGCCCATTCTCCGCCCCATGGCTGACTAATTTTTTTTATTTATGCAGAGGCCG
AGGCCGCCTCTGCCTCTGAGCTATTCCAGAAGTAGTGAGGAGGCTTTTTTGGAGGCCTAGGCTTTTGCAAAAAGCTCCCGG
GAGCTTGTATATCCATTTTCGGATCTGATCAAGAGACAGGATGAGGATCGTTTCGCATGATTGAACAAGATGGATTGCACG
CAGGTTCTCCGGCCGCTTGGGTGGAGAGGCTATTCGGCTATGACTGGGCACAACAGACAATCGGCTGCTCTGATGCCGCCG
TGTTCCGGCTGTCAGCGCAGGGGCGCCCGGTTCTTTTTGTCAAGACCGACCTGTCCGGTGCCCTGAATGAACTGCACGACG
AGGCAGCGCGGCTATCGTGGCTGGCCACGACGGGCGTTCCTTGCGCAGCTGTGCTCGACGTTGTCACTGAAGCGGGAAGGG
ACTGGCTGCTATTGGGCGAAGTGCCGGGGCAGGATCTCCTGTCATCTCACCTTGCTCCTGCCGAGAAAGTATCCATCATGG
CTGATGCAATGCGGCGGCTGCATACGCTTGATCCGGCTACCTGCCCATTCGACCACCAAGCGAAACATCGCATCGAGCGAG
CACGTACTCGGATGGAAGCCGGTCTTGTCGATCAGGATGATCTGGACGAAGAGCATCAGGGGCTCGCGCCAGCCGAACTGT
TCGCCAGGCTCAAGGCGCGCATGCCCGACGGCGAGGATCTCGTCGTGACCCATGGCGATGCCTGCTTGCCGAATATCATGG
TGGAAAATGGCCGCTTTTCTGGATTCATCGACTGTGGCCGGCTGGGTGTGGCGGACCGCTATCAGGACATAGCGTTGGCTA
CCCGTGATATTGCTGAAGAGCTTGGCGGCGAATGGGCTGACCGCTTCCTCGTGCTTTACGGTATCGCCGCTCCCGATTCGC
AGCGCATCGCCTTCTATCGCCTTCTTGACGAGTTCTTCTGAGCGGGACTCTGGGGTTCGAAATGACCGACCAAGCGACGCC
CAACCTGCCATCACGAGATTTCGATTCCACCGCCGCCTTCTATGAAAGGTTGGGCTTCGGAATCGTTTTCCGGGACGCCGG
CTGGATGATCCTCCAGCGCGGGGATCTCATGCTGGAGTTCTTCGCCCACCCCAACTTGTTTATTGCAGCTTATAATGGTTA
CAAATAAAGCAATAGCATCACAAATTTCACAAATAAAGCATTTTTTTCACTGCATTCTAGTTGTGGTTTGTCCAAACTCAT
CAATGTATCTTATCATGTCTGTATACCGTCGACCTCTAGGGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAAT
TGTTATCCGCTCACAATTCCACACAACATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAA
CTCACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAA
CGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTG
CGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGA
GCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAG
CATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGC
TCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTT
TCTCAATGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTT
CAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCA
GCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTAC
ACTAGAAGGACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGC
AAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGAT
CCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCAAAA
AGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGAC
AGTTACCAATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCCGTC

Appendix A

GTGTAGATAACTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCACCGGCT
CCAGATTTATCAGCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATCCAG
TCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCTACAGGC
ATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACATGATCCCCC
ATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTTATCACTCATG
GTTATGGCAGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACCAAG
TCATTCTGAGAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATAATACCGCGCCACATAGCAGA
ACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACTCTCAAGGATCTTACCGCTGTTGAGATCCAGTTCG
ATGTAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGG
CAAAATGCCGCAAAAAAGGGAATAAGGGCGACACGGAAATGTTGAATACTCATACTCTTCCTTTTTCAATATTATTGAAGC
ATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACA
TTTCCCCGAAAAGTGCCACCTGACGTC

Appendix B.

Table A – List of Eukaryotic 2A-like Sequences (valid as of September 2014). 2A sequences listed alphabetically by host species name.

Tag	Accession no	Host species	30 amino acid 2A	protein type	SignalP D-score	Notes
AM1	EZ007780.1	*Acropora millepora*	LSLMFMMVFYNAYIPLLIRQANDVEENPGP	unknown		
AM2	EZ016056	*Acropora millepora*	SPYRFPLLSMVFLFLASLTICMDVEPNPGP	unknown	D=0.641	
AM3	EZ012739	*Acropora millepora*	RLWLTQIYLYRFYIPDLIRLANDVETNPGP	unknown		
AM4	EZ043256	*Acropora millepora*	RRGTTSLHLPGKDIYFDLTIHMDVEPNPGP	unknown		
AM5	EZ041014	*Acropora millepora*	LVCLHFGILLALFFLSCSDVEINPGP	EEP domain	D=0.508	
AM6	EZ028794	*Acropora millepora*	KKLKCSCVQXLCVAQRIKALSGDVEENPGP	unknown		
AM7	EZ015536	*Acropora millepora*	DVESNPGP	cysteine protease		
	XP_002928125.1	*Ailuropoda melanoleuca*	QRLRKVADMDSDSRHLLIPEVDHEIINPGP	amino acid transporter protein		
AF180	EO35FTK01B7DS5 (Baylor)	*Allocentrotus fragilis*	QNIDVKEADKRHITQTLLTRAGDVESNPGP	unknown		Baylor - redundant
AF74	EO35FTK01D9ZGB	*Allocentrotus fragilis*	MTNILLLRSGDVERNPGP	unknown		Baylor - redundant
AF75	EOQWS4P02IEG1N	*Allocentrotus fragilis*	MTNILLLRSGDVERNPGP	ankyrin		Baylor - redundant
AF76	EN3A38D02IKNSP	*Allocentrotus fragilis*	MTNILLLRSGDVERNPGP	unknown		Baylor - redundant
AF94	EPIY22U02HBRJJ	*Allocentrotus fragilis*	MTNILLLRSGDVERNPGP	unknown		Baylor - redundant
AF95	EOQWS4P02INTAQ	*Allocentrotus fragilis*	MTNILLLRSGDVERNPGP	unknown		Baylor - redundant
AF97	ELTDHUK03FT52O	*Allocentrotus fragilis*	MTNILLLRSGDVERNPGP	unknown		Baylor - redundant
AQ23/OFG5	XP_003388358.1	*Amphimedon queenslandica*	SRPANGHHSRSFKAAVESKSDRDVELNPGP	Ankyrin repeat		
	XP_003391287.1	*Amphimedon queenslandica*	CDTVSYAVNLLLCFMLLLLLSGDVELNPGP	Ankyrin repeat		
	XP_003382891.1	*Amphimedon queenslandica*	CHWMSNKTAAFSTNSLILLLSGDVELNPGP	Ankyrin repeat		
AQ20	XP_003385788.1	*Amphimedon queenslandica*	CDTVSYAVYLLLYFMLLLLLSGDVELNPGP	Ankyrin repeat		
	XP_003390764.1	*Amphimedon queenslandica*	CDTVLCAVYLMMYFTLLLLLSGDVELNPGP	Ankyrin repeat		

Appendix B

	XP_003390 050.1	*Amphimedon queenslandica*	QEKGPRLVLSIFCLLILLFISGDVELNPGP	Ankyrin repeat	D=0.411	
	XP_003389 053.1	*Amphimedon queenslandica*	METRPKLILSIFCLFILLFIAGDVELNPGP	Ankyrin repeat	D=0.389	
	XP_003384 088.1	*Amphimedon queenslandica*	TPETVCFLFYLFLHILLLLLSGDVELNPGP	Ankyrin repeat	D=0.439	
	XP_003391 332.1	*Amphimedon queenslandica*	DVELNPGP	Ankyrin repeat		
	XP_003390 049.1	*Amphimedon queenslandica*	QVKTPRLVLSIFCLLILLFISGDVELNPGP	Ankyrin repeat	D=0.409	
	XP_003391 290.1	*Amphimedon queenslandica*	PAAIPTVNITAVYSGPNYTSKRDVELNPGP	Ankyrin repeat, & ZipA		
AQ12/ OFG4	XP_003383 551.1	*Amphimedon queenslandica*	VSDILACFLYSVFVVKLLLLSGDVELNPGP	ATPases		
	XP_003389 880.1	*Amphimedon queenslandica*	QVKTPRLALSIFCLLILLFISGDVELNPGP	Ankyrin	D=0.427	
AQ10/ OFG1	XP_003382 893.1	*Amphimedon queenslandica*	AESRKSQHSNSHQPELIILLSGDVELNPGP	Ankyrin		
	XP_003389 882.1	*Amphimedon queenslandica*	QEKGPRLVLSIFCLLILLFISGDVELNPGP	Ankyrin	D=0.413	
	XP_003389 930.1	*Amphimedon queenslandica*	HWRQPSACTSTGNTYCGANGARDVELNPGP	Ankyrin repeat		
AQ7/ OSC6	XP_003382 892.1	*Amphimedon queenslandica*	HWMNNDETAVFSASPLILLLSGDVELNPGP	Ankyrin repeat		
AQ6/ OSC5	XP_003390 053.1	*Amphimedon queenslandica*	QEKGPRLVLSILCLLILLFLSGDVELNPGP	Ankyrin repeat	D=0.453	
	XP_003389 878.1	*Amphimedon queenslandica*	VKSQPGLILSIFCLLILLFISGDVELNPGP	Ankyrin repeat	D=0.389	
AQ4/ OFG2	XP_003387 964.1	*Amphimedon queenslandica*	GAGLRKSQHGNSHQTELILLSGDVELNPGP	Ankyrin repeat		
	XP_003389 958.1	*Amphimedon queenslandica*	LPGDTINHIFSIISHLLLFLSGDVELNPGP	Ankyrin repeat	D=0.476	
	XP_003386 731.1	*Amphimedon queenslandica*	QGKPPKLILSIFCLAILLFISGDVELNPGP	Ankyrin repeat	D=0.423	note 2nd 2A
	XP_003390 051.1	*Amphimedon queenslandica*	VKSQPGLILSIFCLLILLFISGDVELNPGP	Ankyrin repeat	D=0.389	
	XP_003390 051.1.1	*Amphimedon queenslandica*	QVKTPRLVLSIFCLLILLFISGDVELNPGP	Ankyrin repeat	D=0.409	
AQ27/ OCS2	XP_003391 203.1	*Amphimedon queenslandica*	WFFVFMVSVVFKLVSLLLLLSGDIEINPGP	Ankyrin	D=0.590	
	XP_003390 344.1	*Amphimedon queenslandica*	MLLLLSGDIELNPGP	Ankyrin		
	XP_003390 214.1	*Amphimedon queenslandica*	ASILVCIFLYFVVCRLLLFLSGDIELNPGP	P-loop NTPase		
AQ24/ OFG3	XP_003390 020.1	*Amphimedon queenslandica*	YTESNQNVCYHHFMFLLLLLAGDIELNPGP	P-loop NTPase	D=0.462	
	EPY26545. 1	*Angomonas deanei*	MWYTPVYQIHRPFLAAILLRSGDVETNPGP	EEP domain		

Appendix B

	EPY38571.1	*Angomonas deanei*	RPPTFRRMLPRSALMLMLLLGGDIERNPGP	RT_nLTR_like		
	XP_005088979.1	*Aplysia californica*	PGFFLGGQHNPAWLARLLISGGDVEQNPGP	EEP domain		note 2nd different 2A
	XP_005088979.1.1	*Aplysia californica*	PGFFLGGQHNPAWLARLLILAGDVEQNPGP	EEP domain		
	XP_005098455.1	*Aplysia californica*	MAGDVEINPGP	Calcium-binding EGF-like domain		
	XP_005103178.1	*Aplysia californica*	KIKTWKTTIYSGYRKIRLNCADDIELNPGP	unknown		
Ingi1_AC	Ingi-1_AC	*Aplysia californica*	PGFFLGGQHNPAWLARLLILAGDVEQNPGP	Ingi nonLTR	D= 0.398	From RepBase
	XM_001395993.2	*Aspergillus niger*	VVENCSWADDLGIDVILNPGP	unknown		
	AC233255.1	*Biomphalaria glabrata*	KWKFSVRDSRIKYLSLLILIAGDVESNPGP	phd superfamily, EEP domain	D=0.865 -see note	
	AC233256.1	*Biomphalaria glabrata*	KWKFSIRHSRNKYLSLLILIAGDVESNPGP	Non-LTR-like	D=0.867 full seq, D=0.366 30 aa - see note	
Cow1	DAA17954.1	*Bos taurus*	MANMDSDSRHLLIPEGDHEINPGP	Transmembrane amino acid transporter protein		
	NP_001095633.1	*Bos taurus*	MANMDSDSRHLLIPEGDHEINPGP	putative sodium-coupled neutral amino acid transporter 9		
	XP_002587815.1	*Branchiostoma floridae*	RVCSPDATATKNCAMYMLLLSGDVETNPGP	similarities to reverse transcriptase		
	EEN43826.1	*Branchiostoma floridae*	RVCSPDATATKNCAMYMLLLSGDVETNPGP	similarities to reverse transcriptase		
	XP_002591693.1	*Branchiostoma floridae*	WFEAKMCYCSYVILVLLLLMAGDVEVNPGP	Tumor necrosis factor receptor (TNFR) domain		
	XP_002614028.1	*Branchiostoma floridae*	WAEASCLLVWMVISQLMLKLAGDVEENPGP	LRR & Caspase	D=0.588	
	XP_002609384.1	*Branchiostoma floridae*	YVMMSCLLVWFMMVHKLLLQAGDIEPNPGP	Leucine-rich repeat	D=0.514	
	XP_002591878.1	*Branchiostoma floridae*	AAAPLGFKGPTGFMLAILILSGDVQKNPGP	unknown	D=0.434	
	GS005535.1	*Branchiostoma floridae*	WALFRKPKTTVFCTILLILLSGDVQPNPGP	phd zinc finger	D=0.616	
	GS010268.1	*Branchiostoma floridae*	WALFRKPKTTVFCTILLILLSGDVQPNPGP	phd zinc finger	D=0.616	
	DE019615.1	*Branchiostoma floridae*	NKLLMHNDSTKLSMILILLLSGDVEVNPGP	unknown	D=0.477	
	DE208708.1	*Branchiostoma floridae*	CMKTTDKLFLMYLCSILMAQAVDLETNPGP	phd zinc finger	D=0.453	
	Crack-10_BF	*Branchiostoma floridae*	LYHKNLLTEQCNDQVNLICLAFDIHPNPGP	Crack nonLTR		From RepBase
	Crack-11_BF	*Branchiostoma floridae*	CHVETRVNVVHLCLHTLLLLSGDVASNPGP	Crack nonLTR		From RepBase
	Crack-	*Branchiostoma*	HSVLVCDHCVTVFVVILLLLCGDIHNNPGP	Crack	D=0.492	From

Appendix B

	15_BF	*floridae*		nonLTR		RepBase
	Crack-16_BF	*Branchiostoma floridae*	DIQTNPGP	Crack nonLTR		From RepBase
	Crack-17_BF	*Branchiostoma floridae*	AVTSTSVNCVHLCFHTLLILSGDVAVNPGP	Crack nonLTR	D=0.463	From RepBase
	Crack-28_BF	*Branchiostoma floridae*	TCTERTERTLNLLVCATLLLAGDVSPNPGP	Crack nonLTR	D=0.500	From RepBase
	Crack-9_BF	*Branchiostoma floridae*	IHVKTSVNLAHLCIHTLLLLSGDVACNPGP	Crack nonLTR	D= 0.576	From RepBase
	CR1-1_BF	*Branchiostoma floridae*	KKTMIHNDSTKLSLIMILLLSGDIEINPGP	CR1 nonLTR	D=0.393	From RepBase
	CR1-2_BF	*Branchiostoma floridae*	ILRTSDRLCLLYLCSVLMSQAVDLETNPGP	CR1 nonLTR	D=0.383	From RepBase
	CR1-3_BF	*Branchiostoma floridae*	CLKTTDKLCLMYLCSILMAQAADLETNPGP	CR1 nonLTR	D=0.407	From RepBase
	CR1-11_BF	*Branchiostoma floridae*	LAPHCRPKFTLFSLTLIILLAGDVELNPGP	CR1 nonLTR	D=0.591	From RepBase
	CR1-12_BF	*Branchiostoma floridae*	PRNPLKSISVSIALLVMLTQSGDVHPNPGP	CR1 nonLTR	D=0.429	From RepBase
	CR1-18_BF	*Branchiostoma floridae*	YLRTSDRLCLLYICSVLMAQAVDLETNPGP	CR1 nonLTR	D=0.358 for 2A, but D=0.848 for start)	From RepBase - start higher sig than 2A
	CR1-10_BF	*Branchiostoma floridae*	GTDNVSAEFTQWKPAIDLTQHYDVHPNPGP	CR1 nonLTR		From RepBase
	CR1-17_BF	*Branchiostoma floridae*	TISFILSIFYSNFLLLLIVLSNDIHPNPGP	CR1 nonLTR		From RepBase
	CR1-26_BF	*Branchiostoma floridae*	NLDIFLSYTTVFISFVVILVAGDVHPNPGP	CR1 nonLTR		From RepBase
	CR1-36_BF	*Branchiostoma floridae*	DKDYGIVIQFMLPFFVLFLICGDIHPNPGP	CR1 nonLTR		From RepBase
	CR1-46_BF	*Branchiostoma floridae*	TLTICPQCILIFISLIMIILAGDIHPNPGP	CR1 nonLTR		From RepBase
	CR1-53_BF	*Branchiostoma floridae*	HFDIFLLFFPLPVLVVLSLIAGDIHPNPGP	CR1 nonLTR	D= 0.405	From RepBase
	XP_006074233.1	*Bubalus bubalis*	MANMDSDSRHLLIPEGNHEINPGP	putative sodium-coupled neutral amino acid transporter 9		
CE5	AL132860.1	*Caerorhabditis elegans*	LCETPSLPHTTFLKRKLLVRSGDVESNPGP	unknown		
CE1/3	Z49911.1 in region for protein M28.2 (in an intron)	*Caerorhabditis elegans*	LCETPSLPHTTFLKRKLLVRSGDVESNPGP	unknown, similaries to EEP domain proteins		
	XP_006187419.1	*Camelus ferus*	MANMDGDSRHLLIPEVNHEINPGP	putative sodium-coupled neutral amino acid transporter 9 isoform X1		
	XP_005617447.1	*Canis lupus familiaris*	MDSDSRHLLIPEVDHEVNPGP	putative sodium-coupled neutral amino acid transporter 9 isoform X1		
	XP_003470	*Cavia porcellus*	MANVHDRHHLISEIDDEVNPGP	putative sodium-coupled neutral amino acid		

Appendix B

	307.1			transporter 9		
	XP_004422877.1	*Ceratotherium simum simum*	MANMGSDSRHLLIPEVDHEINPGP	putative sodium-coupled neutral amino acid transporter 9 isoform 1		
	EMP28051.1	*Chelonia mydas*	RVSQPGSVRKQPKLRNALQFSLDVVTNPGP	Dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit 1		
CV1	XP_005844301.1	*Chlorella variabilis*	LRLPCSCSTTALIKRMKLLLSGDVEENPGP	unknown		
CV1	XP_005851168.1	*Chlorella variabilis*	LRLPCSCSTTALVKRMKLLLSGDVEENPGP	EEP domain		note 2nd 2A
	XP_005851168.1.1	*Chlorella variabilis*	LRLPCSCSTTALVKRMKLLLSGDVEENPGP	EEP domain		
CV1	EFN52199.1	*Chlorella variabilis*	LRLPCSCSTTALIKRMKLLLSGDVEENPGP	unknown		
	XP_006874705.1	*Chrysochloris asiatica*	MANVDNESRHLLTSEEDLEVNPGP	Putative sodium-coupled neutral amino acid transporter 9 isoform X1		
	XP_002125736.2	*Ciona intestinalis*	PTIDHNPTETHIITSTKVASNYDIALNPGP	FHA domain [Signal transduction mechanisms]		
	XP_003066996.1	*Coccidioides posadasii C735 delta SOWgp*	PKTLQLDRELSELSAPFRQGSLDVEPNPGP	fungal transcripton factor GAL4-like Zn2Cys6		
	EFW21198.1	*Coccidioides posadasii str. Silveira*	PKTLQLDRELSELSAPFRQGSLDVEPNPGP	transcrition regulator Gal4		
	XP_004678371.1	*Condylura cristata*	MANMDNSSKHLLIPDVDHEINPGP	putative sodium-coupled neutral amino acid transporter 9 isoform X1		
	JX293529.1	*Conus textile*	DGVTLGLLPSLLCYGHLLQRCGDVELNPGP	XV conotoxin	D=0.399	
	CR1-1_CGi	*Crassostrea gigas*	SRHIVVYNFYLQFFMFLLLLCGDIEVNPGP	CR1 nonLTR	D=0.409	From RepBase
	EGV96941.1	*Cricetulus griseus*	MASVDSDSRHLIPEVDLEVNPGP	Putative sodium-coupled neutral amino acid transporter 9		
	XP_003503482.1	*Cricetulus griseus*	MASVDSDSRHLIPEVDLEVNPGP	Putative sodium-coupled neutral amino acid transporter 9		
	EHJ68942.1	*Danaus plexippus*	SKRRRASSCASRGPDGDDSDDGVELSNPGP	Bromodomain, polybromo repeat I		
	EFX60676.1	*Daphnia pulex*	YNSGLWLYTLLLLAGDIERNPGP	EEP Domain		
	XP_006889620.1	*Elephantulus edwardii*	LRNLRKMANVDDEPRHRLISEVDHESNPGP	putative sodium-coupled neutral amino acid transporter 9		
	XP_005604322.1	*Equus caballus*	MDSDSRHRLIPEVDHEVNPGP	putative sodium-coupled neutral amino acid transporter 9 isoform X1		
	XP_003981016.1	*Felis catus*	MANMDSDSRHLLIPEVDHEINPGP	putative sodium-coupled neutral amino acid transporter 9 isoform X1		
	XP_001709417.1	*Giardia lamblia ATCC 50803*	NVFHILDDRWSSNAEDSRETKDDGESNPGP	unknown		

Appendix B

	XP_004058897.1	*Gorilla gorilla gorilla*	MANMNSDSRHLGTSKVDHERDPGP	amino acid transporter protein		
	EHB16430.1	*Heterocephalus glaber*	MTNVDDRHHLISEADHEVNPGP	Putative sodium-coupled neutral amino acid transporter 9		
	XP_004848734.1	*Heterocephalus glaber*	MTNVDDRHHLISEADHEVNPGP	putative sodium-coupled neutral amino acid transporter 9-like		
	XP_004906080.1	*Heterocephalus glaber*	MTNVDDRHHLISEADHEVNPGP	putative sodium-coupled neutral amino acid transporter 9-like isoform X1		
	AC079127.28	*Homo sapiens*	MLQARYLRVLAAPYSTMRNFVEDVIQNPGP	unknown		
	AAH66891.1	*Homo sapiens*	MANMNSDSRHLGTSEVDHERDPGP	amino acid transporter protein		
	NP_775785.2	*Homo sapiens*	MANMNSDSRHLGTSEVDHERDPGP	amino acid transporter protein		
	AAI01363.1	*Homo sapiens*	MANMNSDSRHLGTSEVDHERDPGP	amino acid transporter protein		
	DQ121107.1	*Hucho taimen*	TRRPVVIAFSRNLILLLLCSSGDVEVNPGP	unknown	D=0.597	
	XP_002168455.2	*Hydra vulgaris*	YNKSVYQYIHKTSFGLYAFNGIDIAPNPGP	transmembrane protein 62-like		
	XP_005319681.1	*Ictidomys tridecemlineatus*	MANMDNDSRHLLNSEVNHEINPGP	putative sodium-coupled neutral amino acid transporter 9		
IS1	DS876754 dna:scaffold scaffold:IscaW1:DS876754:1:1257:-1 (Ensembl)	*Ixodes scapularis*	MFLVLLLLLSGDVELNPGP	unknown	D=0.348	Ensembl
IS34	DS730003 dna:scaffold scaffold:IscaW1:DS730003:-26:1471:1	*Ixodes scapularis*	RQVLFLCSCERISGLKLLLSGDIELNPGP	unknown		
IS68	DS667985scaffold:iscaw1:ds667985:498:1166:1	*Ixodes scapularis*	MFSLCCQCFDVLSQVLLMSGDVELNPGP	unknown	D=0.337	
	AC192419.4-1	*Ixodes scapularis*	ATAQYVLSAPGSAWVRNGCDDNDVEANPGP	unknown		
	AC205630.1	*Ixodes scapularis*	RSLIVTCSECWGVLLLLLQLSGDVELNPGP	unknown	D=0.596	
	AC205641.1	*Ixodes scapularis*	RSLIVTCSECWGVLLLLLQLSGDVELNPGP	unknown	D=0.596	
	AC205634.1	*Ixodes scapularis*	NRRKFYYRKKVIFVLLLLLLSGDVETNPGP	Chromosome segregation ATPase		
	AC192419.4-2	*Ixodes scapularis*	NRRKLYYRKKVLFVLLLLLLSGDVETNPGP	L1 transposable element	D=0.623	

Appendix B

GU318570	*Ixodes scapularis*	LLLSGDVELNPGP	tandem repeat region	
XP_001887419.1	*Laccaria bicolor S238N-H82*	SQSSALMSSSTPFKSLKGLFSQDEESNPGP	unknown	
FR798987.1	*Leishmania braziliensis*	ITSIVKRIHVSTRRLSACLRCGDVERNPGP	chromosome 13, unknown	
FR799578.1	*Leishmania braziliensis*	IGSTLQRMFASTSSLLL*VRCGDIERNPGP	unknown	
FR799578.1.2	*Leishmania braziliensis*	IGSTLQRMFASTSSLLL*VRCGDIERNPGP	unknown	
ESO91877.1	*Lottia gigantea*	GMKTENILCLLILSSQLIAICGDVEQNPGP	unknown	D=0.589
ESP04282.1	*Lottia gigantea*	ISTGSLTISTILVISMLLVLSGDVEVNPGP	unknown	
ESO82727.1	*Lottia gigantea*	ISTGSLTISTILVIAILLVLSGDVEVNPGP	chromosome segregation protein SMC	
ESO94951.1	*Lottia gigantea*	IRLNSLTFSVFLVINVLILLSGDVETNPGP	similarity to LINE proteins	D=0.427
ESO97161.1	*Lottia gigantea*	VLHCNLMMSVLMILNVLIVLSGDVESNPGP	unknown	D= 0.461
ESO90449.1	*Lottia gigantea*	SLNLSDPFRIVVCLFLILIVAGDIEQNPGP	unknown	
ESP02481.1	*Lottia gigantea*	MTLSLVFLVNLLMLMAGDIESNPGP	Similarities to LINE	D=0.457
CR1-1_LG	*Lottia gigantea*	TLLNDTFSSILYYCFILIIRSGDIELNPGP	CR1 nonLTR	From RepBase
XP_003408087.1	*Loxodonta africana*	MANVNKESRHLLISEVDHEINPGP	putative sodium-coupled neutral amino acid transporter 9	
NP_001253166.1	*Macaca fascicularis*	MANMNSDSRHLGTSKVDHERDPGP	amino acid transporter protein	
XP_005065531.1	*Mesocricetus auratus*	MASMDSDSRPLIPEVDLEVNPGP	Putative sodium-coupled neutral amino acid transporter 9	
XP_005356860.1	*Microtus ochrogaster*	MADADSDSRHLISEVDDEVNPGP	putative sodium-coupled neutral amino acid transporter 9	
GAA99414.1	*Mixia osmundae IAM 14324*	STVAAHGQVVLKTNKQGDKYYPDVETNPGP	unknown	
GAA99414.1	*Mixia osmundae IAM 14324*	STVAAHGQVVLKTNKQGDKYYPDVETNPGP	unknown	
XP_004789844.1	*Mustela putorius furo*	MDSDSRHLLIPEVNHEIINPGP	amino acid transporter protein	
XP_004744702.1	*Mustela putorius furo*	MDSDSRHLLIPEVNHEIINPGP	amino acid transporter protein	
EPQ06479.1	*Myotis brandtii*	MPHAGDRAYNPGMCPDRESNPGP	Fermitin family like protein 3	
EPQ12632.	*Myotis brandtii*	MADLDSDSRQLLAPEADREVHPGP	amino acid transporter protein	

Appendix B

	1					
	ELK31761.1	*Myotis davidii*	MADLDSDSRQLLAPEADREVLPGP	amino acid transporter protein		
	AM878476.1	*Mytilus edulis*	YKISLLLLTNSSDIELNPGP	unknown		
	XP_001627324.1	*Nematostella vectensis*	ARVCLDILRSASIALLLVCLSNDVQLNPGP	EEP Domain	D=0.555	
	XP_001621507.1	*Nematostella vectensis*	ARVCLDLLRSASIALLLVCLLNDVQLNPGP	EEP Domain	D=0.368	
L2A-1_NVe	XP_001639586.1	*Nematostella vectensis*	GRIKRYPNSTSTFQLTRIAVSGDVSPNPGP	EEP Domain & 7 transmembrane receptor (rhodopsin family)		
Crack-3_Nve	XP_001621289.1	*Nematostella vectensis*	LRASIYMTKVGICAFSLIILSGDISLNPGP	EEP domain	D=0.337	
	XP_001630328.1	*Nematostella vectensis*	SLAGLQLSYTFLCASYLLLLAGDISTNPGP	EEP domain	D=0.361	
	XP_001622117.1	*Nematostella vectensis*	RKLIAPRSNPSSLAFRLLILSGDIPLNPGP	unknown	D=0.314	
	XP_001634001.1	*Nematostella vectensis*	RKLIAPRSNPSSLAFRLLILSGDIPLNPGP	related to pyruvate formate-lyase activating enzyme		
	XM_001628628.1	*Nematostella vectensis*	ARIHHTPASTSSFQLERLLSCGDVNPNPGP	unknown		
	XM_001627583.1	*Nematostella vectensis*	FRPGHAPASTSCFQLERLLSCGDVNPNPGP	EEP domain		
	XM_001622595.1	*Nematostella vectensis*	GRIKRYPNSTSTFQLTRIAVSGDVSPNPGP	PHD zinc finger		
	XM_001639536.1	*Nematostella vectensis*	GRIKRYPNSTSTFQLTRIAVSGDVSPNPGP	PhD zinc finger, EEP domain		
	XM_001626901.1	*Nematostella vectensis*	GRIKRYPNSTSTFQLTRIAVSGDVSPNPGP	PhD zinc finger		
	CR1-2_NV	*Nematostella vectensis*	ILRTSDRLCLLYLCSVLMSQAVDLETNPGP	CR1 nonLTR	D=0.383	From RepBase
	CR1-4_NV	*Nematostella vectensis*	FRPRRDFTRPNCYLVGLLLLCGDVASHPGP	CR1 nonLTR	D=0.362	From RepBase
	CR1-8_NV	*Nematostella vectensis*	ITYRFGRTGPSHLVMLLLILGGDVELNPGD	CR1 nonLTR	D=0.523	From RepBase
	CR1-19_NV	*Nematostella vectensis*	TSAFRKHRTFVSIIPGLLLLCGDIISQPGP	CR1 nonLTR	D=0.530	From RepBase
	CR1-20_NV	*Nematostella vectensis*	MNVGRSSSEHKHLLLCLLLLGGDIQLNPGP	CR1 nonLTR	D=0.500	From RepBase
	CR1-21_NV	*Nematostella vectensis*	RKLIAPRSNPSSLAFRLLILSGDIPLNPGP	CR1 nonLTR		From RepBase
	Neptune1_NV	*Nematostella vectensis*	TLHNRPAIFMLFVLPILIAKSMDIETNPGP	Neptune TE		From RepBase
	NV_Penelope-5_NV	*Nematostella vectensis*	TLHNRPAIFMLFVLPILIAKSMDIETNPGP	Penelope TE		From RepBase
	XP_003265979.1	*Nomascus leucogenys*	MANMNSDSRHLGTAKVDHERDPGP	amino acid transporter protein		

Appendix B

	XP_004623147.1	*Octodon degus*	MANVDDRHDLISEVDHEVNPGP	putative sodium-coupled neutral amino acid transporter 9	
	XP_004416369.1	*Odobenus rosmarus divergens*	MDSDSRHRLVPEVDHEIINPGP	amino acid transporter protein	
OM4	EZ854573.1	*Oncorhynchus mykiss*	TRRPVILAFSCTLILLLFCSSGDVEVNPGP	unknown	0.578
OM1	EZ767016	*Oncorhynchus mykiss*	RTGRPVLIAFSRTLIILLFSSGDVEVNPGP	unknown	D=0.576
OM2	EZ884838	*Oncorhynchus mykiss*	TRXPVLXAFXRTLILLLLCSSGDVEVNPGP	unknown	D=0.628
OM3	EZ888793	*Oncorhynchus mykiss*	SRRPVLIAFSRTLILLLLCSSGDVGVNPGP	unknown	D= 0.635
OM5	EZ799774.1	*Oncorhynchus mykiss*	LDDQFFLAISRNLILLLLCSSGDVEVNPGP	unknown	D=0.502
OM6	EZ796013.1	*Oncorhynchus mykiss*	PFRHTLILLLLCPSGDVEINPRP	unknown	D=0.323
SS7	CDQ99460.1	*Oncorhynchus mykiss*	SRRPVLIAFSRTLILLLLCSSGDVEVNPGP	EEP domain	D=0.629
	CDQ92808.1	*Oncorhynchus mykiss*	SRRPVLIAFSRTLILLLLCSAGDVEVNPGP	unknown	D=0.57
	CDQ57552.1	*Oncorhynchus mykiss*	SRRPVLIAFSRTLILLLLCSAGDVEVNPGP	unknown	D=0.57
	CDR01105.1	*Oncorhynchus mykiss*	SRRPVLIAFSRTLILLLLCSAGDVEVNPGP	EEP domain	D=0.57
	CDR01026.1	*Oncorhynchus mykiss*	SRRPVLIAFSRTLILLLLCSAGDVEVNPGP	EEP domain	D=0.57
	CDQ81228.1	*Oncorhynchus mykiss*	SRRPALIAFSRTLILLLLCSAGDVEVNPGP		D=0.57
	CDQ97352.1	*Oncorhynchus mykiss*	TKRPVLIAFMRTLILLLLCSSDDVEVNPGP	Exonuclease-Endonuclease-Phosphatase (EEP) domain superfamily	D=0.539
	CDQ81639.1	*Oncorhynchus mykiss*	SRRPVLIAFSRTIILLLLCSAGDVEVNPGL	Exonuclease-Endonuclease-Phosphatase (EEP) domain superfamily	D=0.460
	CDQ59560.1	*Oncorhynchus mykiss*	ARWPVCIAFSRTLILLLLCSSGDVEVNPGH	Exonuclease-Endonuclease-Phosphatase (EEP) domain superfamily	D=0.530
	CDQ84132.1	*Oncorhynchus mykiss*	SRRPVLIAFRRTLILLLLCSASDVEVNTGP	Exonuclease-Endonuclease-Phosphatase (EEP) domain superfamily	D=0.535
	XP_004275	*Orcinus orca*	MANMDSDSSHLLIPEVDNEINPGP	putative sodium-coupled neutral amino acid	

Appendix B

	195.1			transporter 9 isoform 1	
	XP_005459556.1	*Oreochromis niloticus*	MIILLAGDVERNPGP	splicing factor 3A subunit 2-like	
	XP_005461337.1	*Oreochromis niloticus*	NKPHVCLTAVLIMTSLALMDAYLVEQNPGP	transmembrane protein 121-like	D=0.319
	AFA34358.1	*Ostrea edulis*	NKDNGSNSVGKYLLILRILIAGDISTNPGP	EEP domain, partial	
	XP_003782807.1	*Otolemur garnettii*	MASVDRHHLISEVDHEINPGP	putative sodium-coupled neutral amino acid transporter 9 isoform 1	
	XP_004017029.1	*Ovis aries*	MANMDSDSRHLLIPEGDHEVNPGP	putative sodium-coupled neutral amino acid transporter 9 isoform 1	
	XP_003827419.1	*Pan paniscus*	MANMNSDSRHLGTSEVDHERDPGP	amino acid transporter protein	
	XP_001145251.1	*Pan troglodytes*	MANMNSDSRHLGTSEVDHERDPGP	amino acid transporter protein	
	XP_007079998.1	*Panthera tigris altaica*	MANMDSDSRHLLIPEVDHEINPGP	putative sodium-coupled neutral amino acid transporter 9 isoform X3	
	XP_007079996.1	*Panthera tigris altaica*	MANMDSDSRHLLIPEVDHEINPGP	putative sodium-coupled neutral amino acid transporter 9 isoform X1	
	XP_005958431.1	*Pantholops hodgsonii*	MANMDSDSRHLLIPEGDHEVNPGP	putative sodium-coupled neutral amino acid transporter 9 isoform X1	
	XP_003899730.1	*Papio anubis*	MANMNSDSRHLGTSKVDHKRDPGP	amino acid transporter protein	
	ETS82120.1	*Pestalotiopsis fici W106-1*	DDPDKKQMVQHAAKVIEEMDQDDIESNPGP	unknown	
	XP_007114962.1	*Physeter catodon*	MANMDSDSSHLLIPEVDNEINPGP	putative sodium-coupled neutral amino acid transporter 9 isoform X1	
	CCW65577.1	*Phytomonas sp. isolate EM1*	ALPNTSGSAVYFAQSGSHTESGDVTANPGP	unknown	
	XP_002815613.2	*Pongo abelii*	MANMNSDSRHLGTSQVDHERDPGP	amino acid transporter protein	
	ELK12194.1	*Pteropus alecto*	MANMDNDSRHLLIPEVDHEINPGP	Putative sodium-coupled neutral amino acid transporter 9	
	XP_006912529.1	*Pteropus alecto*	MANMDNDSRHLLIPEVDHEINPGP	putative sodium-coupled neutral amino acid transporter 9	
	XP_003305142.1	*Pyrenophora teres f. teres 0-1*	IWAFLVTWSKIILQDIPSLTNGDVSPNPGP	unknown	
	XP_006231997.1	*Rattus norvegicus*	MANVDSDSRHLISEVEHEVNPGP	putative sodium-coupled neutral amino acid transporter 9 isoform X7	
	EDM10355.1	*Rattus norvegicus*	MANVDSDSRHLISEVEHEVNPGP	similar to hypothetical protein FLJ90709, isoform CRA_b	
	NP_001030328.1	*Rattus norvegicus*	MANVDSDSRHLISEVEHEVNPGP	putative sodium-coupled neutral amino acid transporter 9	
RP1	JAA55846.1	*Rhipicephalus pulchellus*	IVLPCPEAVLAPFCSLLLLLCGDVESNPGP	Putative tick transposon	D=0.418
RP2	JAA55160.1	*Rhipicephalus pulchellus*	RCLLSILVECIDVSRTLLLLSGDIETNPGP	TATA DNA binding	D=0.337

Appendix B

				element		
RP3	JAA56189.1	*Rhipicephalus pulchellus*	HCCIAIAVDCAHIIHSLLMLSGDVEVNPGP	TATA DNA binding element	D=0.396	
RP4	JAA55762.1	*Rhipicephalus pulchellus*	PNSLVACCSRVLHYFDGLLLSGDVELNPGP	Putative tick transposon		
RP5	JAA55454.1	*Rhipicephalus pulchellus*	SFVFTSLYADIVHCLCSLLLSGDVELNPGP	L1 transposon	D=0.428	
RP5	JAA55454.1.1	*Rhipicephalus pulchellus*	SFVFTSLYADIVHCLCSLLLSGDVELNPGP	unknown	D=0.428	
RP6	JAA64478.1	*Rhipicephalus pulchellus*	RLLCPVFYDLSVSLGKLLLLAGDIETNPGP	L1 transposon	D=0.345	
RP7	JAA64408.1	*Rhipicephalus pulchellus*	NCIAVVTQLVHCLRISLLVLCGDIETNPGP	unknown	D=0.543	
RP8	JAA55744.1	*Rhipicephalus pulchellus*	HCCLSIVVQCCDVIRSLLLLAGDIETNPGP	DNA translocase from L1 transposon	D=0.439	
RP9	JAA55743.1	*Rhipicephalus pulchellus*	RRSFVIIYDCTCMLMRVLLLAGDIETNPGP	t_SNARE domain	D=0.487	
	XP_006819733.1	*Saccoglossus kowalevskii*	MCGDVESNPGP	PAT1 chromosome segregation - sporulation-specific protein 15-lik		
	XP_006821312.1	*Saccoglossus kowalevskii*	MLSGDIEVNPGP	L1 transposable element		
	XP_006820261.1	*Saccoglossus kowalevskii*	PLISMDDLSVSEGLLLAMLCDGDIEQNPGP	kinase D-interacting substrate		
SK1	XP_002731766	*Saccoglossus kowalevskii*	RYSKSSCEWMWFLVLLSFILSGDIEVNPGP	matches L1 transposon	0.425	XP file discontinued, now NW_003109978.1
SK43	gnl|ti|1870301277 name:245167785	*Saccoglossus kowalevskii*	GRIITIYCIGVSGLHFDGNYSGDTETNPGP	unknown		Baylor
SK45	Baylor Contig114660	*Saccoglossus kowalevskii*	MRSNLVLIRDIETNPGP	unknown		Baylor
SK47	Baylor	*Saccoglossus kowalevskii*	MFALYLYHCRMRSNLILIRDIETNPGP	unknown		Baylor
SK53	Baylor Contig117841	*Saccoglossus kowalevskii*	MRSNLVLIRDIETNPGP	unknown		Baylor
SK6	Baylor Contig 36012	*Saccoglossus kowalevskii*	CDQHGRIHVSLFIFGILLLLSGDIEVNPGP	unknown	D=0.411	Baylor
SK61	Baylor Contig 76190	*Saccoglossus kowalevskii*	MFSIQPLWTSKLIILCGDVESNPGP	unknown		Baylor
SS7	GU129139.1	*Salmo salar*	SRRPVLIAFSRTLILLLLCSSGDVEVNPGP	transposon within IgH locus A	D=0.629	

Appendix B

SS1	AF256957	*Salmo salar*	VSVSLFSPLHRTLILLVLCCSGDVEVNPGP	microsatellite DNA	D=0.593
SS2	EU025709.1	*Salmo salar*	HRRPILIAFSCTLILLLLCSSGDVEVNPGP	transposon	D=0.600
SS3	EU025709.2	*Salmo salar*	TRQPVLIAFSRTLTQLLLCTSGDVEVNPGP	transposon	D=0.500
SS4	AF256956	*Salmo salar*	VSVSLFSPLHRTLILLVLCCSGDVEVNPGP	transposon	D=0.593
SS5	GU817336.1	*Salmo salar*	ARRTVLIAISHTLILLLLCSSGDVEVNPGP	transposon	D=0.542
SS6	GU817336.2	*Salmo salar*	ARWPVLIAFSRTLTLLLLCSSGDVEVNPGP	transposon	D=0.643
SS8	GU129139.2	*Salmo salar*	HRRPILIAFSRTLILILLCSSGDVEVNPGP	Full length nonLTR in IgH locus A	D=0.631
SS9	EF427378.1	*Salmo salar*	ARRPVRIAFSSTLILLLLCSSGDVEVNPGP	transposon	D=0.626
SS10	EF427378.2	*Salmo salar*	IYAIVPRNLLLILLILCSSGDVEVNPGP	transposon	D=0.506
SS11	EF427378.3	*Salmo salar*	ARQPVRIAFSSTLILLILCSSGDVEVNPGP	transposon	D=0.512
SS12	EF427382.1	*Salmo salar*	ARRPVRIAFSSTLILLLLCSSGDVEVNPGP	transposon	D=0.626
SS13	EU008541.1	*Salmo salar*	TRRPFVIAFSRTLILLLHCSSGDVEVNPGP	transposon	D=0.450
SS14	EF467295	*Salmo salar*	HRRPILIAFSRTLILILLCSSGDVEVNPGP	transposon	D=0.475
SS15	AY785950.1	*Salmo salar*	RTRRPVSYFSRTLILLLLCSSGDVEVNPGP	transposon	D=0.582
	GU817337.1	*Salmo salar*	HRRPILIAFSRTLILILLCSSGDVEVNPGP	transposon	D=0.631
	GU129139.3	*Salmo salar*	TRRPVLIAFSRTLILLVLCSSGGVEVNPGP	IgH locus A - transposon	D=0.555
	GU129140.1	*Salmo salar*	TRRPVLIAFSRTLILLLLCSSGDVEVNPGP	IgH locus B - transposon	D=0.631
	EU025709.3	*Salmo salar*	SQTTNPHSFYRSLIHLLLCSSGDVEVNPEP	transposon	D=0.336
	EU025715.1	*Salmo salar*	ARQPVRIAFSRTLILLLLCSSGDVEVNPGP	transposon	D=0.612
	EU025715.2	*Salmo salar*	TRRPVLISFSRTLILLLLCSSGDVEVNSGP	transposon	D=0.620
	HM159473.1	*Salmo salar*	TRRPVLLAFSRTLILLLLCSSGDVEVNPSP	transposon	D=0.633
	EU481821.1	*Salmo salar*	PVLIAFSRTLIVLLLCFSGDVEVNPGP	transposon	D=0.618
	EU481821.2	*Salmo salar*	TRRPVLIALSHTLNQLLLCSSGDVEVNPGP	transposon	D=0.460
	EU481821.3	*Salmo salar*	TRRPVLISLSHTLNQLLLCSSGDVEVNPGP	transposon	D=0.414

Appendix B

	EU481821.4	*Salmo salar*	TRRPDLIALRHNLSLLLLCSSGDVEVNPGP	transposon	D=0.483	
	EU221179.1	*Salmo salar*	TRRPVLIAFSNNLILLLFCSSSDVEVNPGP	transposon	D=0.575	
	EU221179.1	*Salmo salar*	TRRPVLIAFSSNLILLLFCSSSDVEVNPGP	transposon	D=0.575	identical to 1st 2a in the seq
	GU552297.1	*Salmo salar*	SRRPVLIAFSRTLIILLLCSSGDVEAKPRP	transposon	D=0.483	
	HM159469.1	*Salmo salar*	TRLTVLIAFSCTLILVLL*SSGDVEVNPGP	retrotransposon		
	EF427377.1	*Salmo salar*	KHKSNIYQKLLNDLFTALLLSGDVQLNPGP	transposon	D=0.349	
	EU221180.1	*Salmo salar*	PVLIVFSRILILLLLCSSGDVEVNPGP	transposon	D=0.563	
	AF256865.1	*Salmo salar*	RARRQDLITFSRILILLLLCSSGDVEVNPG	transposon	D=0.610	
	EU221178.1	*Salmo salar*	SRRPVLIAFSRILILLLLCSSVDVEVNPCM	transposon	D=0.591	
	GQ505858.1	*Salmo salar*	PNALDDQFLLPLAVPLSYSTSGDVEVNPGP	transposon	D=0.173	
	GU817335.1	*Salmo salar*	PLAVPLSYSSSGDEEVNPGP	transposon		
	NM_001140997	*Salmo salar*	KKGTSFGPSLFTLILLLLCSSGDVEVNPGP	Salmo salar Transmembrane protein 86A (tm86a), mRNA-1 reading frame	D= 0.553	
	AC203456.8	*Salmo salar*	PLAVPSSYSSCFSGDVEVNPGP	unknown		
	EPY32652.1	*Strigomonas culicis*	VTPSRAYDITIIIVRMLLLMCGDVERNPGP	unknown, similar to retrotransposon		
SF44	EMJFY7Z01C829N	*Stronglocentrotus franciscanus*	MTNILLLRSGDVERNPGP	unknown	Baylor - redundant	
SF45	EN8UDV202I4KMY	*Stronglocentrotus franciscanus*	MTNILLLRSGDVERNPGP	unknown	Baylor - redundant	
SF46	EN8UDV202F41U5	*Stronglocentrotus franciscanus*	MTNILLLRSGDVERNPGP	unknown	Baylor - redundant	
SF47	EMWJ4TM01DDW3M	*Stronglocentrotus franciscanus*	MTNILLLRSGDVERNPGP	unknown	Baylor - redundant	
SF49	EMJFY7Z01DQGFZ	*Stronglocentrotus franciscanus*	MTNILLLRSGDVERNPGP	unknown	Baylor - redundant	
SF55	EMJFY7Z01BS5P3	*Stronglocentrotus franciscanus*	MTNILLLRSGDVERNPGP	unknown	Baylor - redundant	
STR140	Baylor GLEAN3_26442	*Stronglocentrotus purpuratus*	MTNALLLRSGDVELNPGP	NLR-like		Baylor - redundant
STR81	Baylor GLEAN3_	*Stronglocentrotus purpuratus*	SKTDLISGQIPPLSELLLLKSGDVELNPGP	NLR-like		Baylor - redundant

Appendix B

	21478					
STR81 identical2A	Baylor GLEAN3_23550	*Strongylocentrotus purpuratus*	SKTDLISGQIPPLSELLLLKSGDVELNPGP	NLR-like		Baylor - redundant
STR81 identical2A	Baylor GLEAN3_15340	*Strongylocentrotus purpuratus*	SKTDLISGQIPPLSELLLLKSGDVELNPGP	NLR-like		Baylor - redundant
STR81 identical2A	Baylor GLEAN3_11433	*Strongylocentrotus purpuratus*	SKTDLISGQIPPLSELLLLKSGDVELNPGP	NLR-like		Baylor - redundant
STR81 identical2A	Baylor AAGJ04169434.1: Contig169489	*Strongylocentrotus purpuratus*	SKTDLISGQIPPLSELLLLKSGDVELNPGP	NLR-like		Baylor - redundant
STR81 identical2A	Baylor AAGJ04083908.1 Contig83932	*Strongylocentrotus purpuratus*	SKTDLISGQIPPLSELLLLKSGDVELNPGP	NLR-like		Baylor - redundant
STR81 identical2A	Baylor AAGJ04074996.1 Contig75016	*Strongylocentrotus purpuratus*	SKTDLISGQIPPLSELLLLKSGDVELNPGP	NLR-like		Baylor - redundant
STR81 identical2A	Baylor AAGJ04035920.1 Contig35929	*Strongylocentrotus purpuratus*	SKTDLISGQIPPLSELLLLKSGDVELNPGP	NLR-like		Baylor - redundant
	XP_3723511.1	*Strongylocentrotus purpuratus*	MEVMQLFTLPLLLILCGDVELNPGP	unknown	D=0.388	
	XP_797335.3	*Strongylocentrotus purpuratus*	ISCTMDGLCLLYLLLILLMRSGDVETNPGP	NLR-like	D=0.557	
	XP_798533.3	*Strongylocentrotus purpuratus*	ICYTMDGFCLLYLLLILLMRSGDVETNPGP	Death Domain	D= 0.542	
	XP_003730835.1	*Strongylocentrotus purpuratus*	EADAIKGNDCPDVTNMLLLRSGDVERNPGP	Death Domain		
	XP_003724840.1	*Strongylocentrotus purpuratus*	MDGFCLLYLIMILLMRSGDVETNPGP	NLR-like	D=0.438	
STR6	XP_003723709.1	*Strongylocentrotus purpuratus*	MDGFCLLYLLLILLMRSGDVETNPGP	unknown, similar to ankyrin	D=0.542	
	XP_800440.1	*Strongylocentrotus purpuratus*	KADAIKGNDFPNMTYILLLRCGDVELNPGP	Death Domain		
	XP_003728571.1	*Strongylocentrotus purpuratus*	MDGFCLLYLLMILLMRSGDVETNPGP	NLR-like	D=0.473	note 2nd 2A
	XP_003730452.1	*Strongylocentrotus purpuratus*	TTDDPVKEDSACLPEMLLVKAGDVELNPGP	NLR-like		
M12V	XP_800245.3	*Strongylocentrotus purpuratus*	MDGFCLLYLLLILLVRSGDVETNPGP	NLR-like	D=0.528	
	XP_794591.3	*Strongylocentrotus purpuratus*	TSYTMDGFCLLYLLLILLMRSGDVETNPGP	ABC transporter G family member 20-	D=0.542	

Appendix B

				like		
	XP_003730118.1	*Strongylocentrotus purpuratus*	MAGFCLLYLLMILLMRSGDVETNPGP	NLR-like	D=0.531	
	XP_003730320.1	*Strongylocentrotus purpuratus*	MGGFCHLYLLMILLMRSGDVETNPGP	NLR-like	D=0.444	
SF44	XP_798371.3	*Strongylocentrotus purpuratus*	GEADTIKGNDCSDMTNILLLRSGDVERNPGP	Ankyrin repeat		
L16M	XP_003723697.1	*Strongylocentrotus purpuratus*	MDGFCLLYLLMILLMRSGDVETNPGP	NLR-like	D=0.473	
	XP_003729593.1	*Strongylocentrotus purpuratus*	MDGFCLLYLLMILLMRSGDVETNPGP	NLR-like	D=0.473	
	XP_785721.3	*Strongylocentrotus purpuratus*	MDGFCLLYLLMILLVRSGDVETNPGP	Death Domain	D=0.557	
	XP_788942.3	*Strongylocentrotus purpuratus*	EADTIKGNDCPDMANILLLRSGDVERNPGP	NLR-like		
	XP_003724839.1	*Strongylocentrotus purpuratus*	MVGFCLLYLLMILLMRSGDVETNPGP	NLR-like	D=0.534	
	XP_001198729.1	*Strongylocentrotus purpuratus*	LHPAILCSASLCFRPYLLLMAGDVEPNPGP	NLR-like		
	XP_003727629.1	*Strongylocentrotus purpuratus*	MDAFCLLYLLMILLMRSGDVETNPGP	NLR-like	D=0.483	
	XP_003729541.1	*Strongylocentrotus purpuratus*	TSYTMDGFCLLYLLLILLMRSGDVETNPGP	NLR-like	D=0.542	
	XP_001184399.2	*Strongylocentrotus purpuratus*	RSYTMDGFCLLYLLLILLMRSGDVETNPGP	NLR-like	D=0.542	
	XP_003729085.1	*Strongylocentrotus purpuratus*	TTDDPVMQESTCLPEMLLVKAGDVEQNPGP	NLR & Ankyrin		
	XP_003730724.1	*Strongylocentrotus purpuratus*	EAYAIKGNDCPDMTNILLLRSGDVERNPGP	NLR-like		
STR1	XP_797143.2	*Strongylocentrotus purpuratus*	MFVCAFILISVLLLSGDVEINPGP	endonuclease /Exonuclease & RT	0.495	redundant
STR24	XP_001196407.1	*Strongylocentrotus purpuratus*	MCAGDVQPNPGP	CR1 nonLTR	D=0.373	redundant
STR28	XP_001179204	*Strongylocentrotus purpuratus*	MGVAESTSLSHLTILLLLSGQVETNPGP	CR1 nonLTR	D=0.373	redundant
STR29	XP_791376.2	*Strongylocentrotus purpuratus*	MGVAESTSLSHLTILLLLSGQVETNPGP	CR1 nonLTR	D=0.373	redundant
STR30	XP_001199602.1	*Strongylocentrotus purpuratus*	MGVAESTSLSHLTILLLLSGQVETNPGP	CR1 nonLTR	D=0.373	redundant
STR31	XP_001200060.1	*Strongylocentrotus purpuratus*	MGVAESTSLSHLTILLLLSGQVETNPGP	CR1 nonLTR	D=0.373	redundant
STR32	XP_001185404.1	*Strongylocentrotus purpuratus*	NSSCVLNIRSTSHLAILLLLSGQVEPNPGP	CR1 nonLTR	D=0.339	redundant
STR33	XP_001184905.1	*Strongylocentrotus purpuratus*	LPVNEYRSTSLSHLTILLLLSGQVETNPGP	CR1 nonLTR	D= 0.307	redundant
STR34	XP_001196844.1	*Strongylocentrotus purpuratus*	NSTPAAMFVCVFILISVLLLSGDVEISPGP	CR1 nonLTR	D= 0.479	redundant

Appendix B

STR35	XP_001200466.1	*Strongylocentrotus purpuratus*	NSSCVLNIRSTSHLAILLLLSGQVEPNPGP	CR1 nonLTR	D= 0.339	redundant
	Rex1-1_SP	*Strongylocentrotus purpuratus*	KTPIKYYANASATFTLPIICCGDVESNPGP	REX1 nonLTR		From RepBase
STR40	GLEAN3_18025	*Strongylocentrotus purpuratus*	KSCISYYSNSTACFNIEIMCCGDVKSNPGP	REX1 nonLTR	D= 0.676	Baylor - redundant
STR55	GLEAN3_24854	*Strongylocentrotus purpuratus*	GARISYHPNTTATFQLRLLVSGDVNPNPGP	REX1 nonLTR	D=0.575	Baylor - redundant
STR61	GLEAN3_22393	*Strongylocentrotus purpuratus*	GARIRYYNNSSATFQTILMTCGDVDPNPGP	REX1 nonLTR	D=0.726	Baylor - redundant
STR89	GLEAN3_19055	*Strongylocentrotus purpuratus*	GRRIQYYNNSISTFRSELLRCGDVESNPGP	REX1 nonLTR	D= 0.496	Baylor - redundant
STR197	GLEAN3_26896	*Strongylocentrotus purpuratus*	KHPILYYTNGESSFQIELLSCGDINPNPGP	REX1 nonLTR	D=0.659	Baylor - redundant
STR51	GLEAN3_22449	*Strongylocentrotus purpuratus*	SRPILYYSNTTASFQLSTLLSGDIEPNPGP	L2 nonLTR		Baylor - redundant
STR69	GLEAN3_27016	*Strongylocentrotus purpuratus*	CRRIAYYSNSDCTFRLELLKSGDIQSNPGP	L2 nonLTR	D=0.619	Baylor - redundant
STR133	GLEAN3_00868	*Strongylocentrotus purpuratus*	KRRIPYNPNSTASFQLELLHAGDVHPNPGP	L2 nonLTR	D=0.582	Baylor - redundant
STR142	GLEAN3_14631	*Strongylocentrotus purpuratus*	KTRIPYSVNSNASFQLELLHAGDVHPNPGP	L2 nonLTR	D=0.833	Baylor - redundant
	XP_003134026.2	*Sus scrofa*	MANMDSDSRHLLIPEVDHEVNPGP	putative sodium-coupled neutral amino acid transporter 9 isoformX1		
	AF034974.1	*Sus scrofa*	NDMFAIGIQEDVIMNPGP	oestrogen receptor exon 8		
	XP_004380528.1	*Trichechus manatus latirostris*	MANVDSDSRHLLISEVDHEINPGP	putative sodium-coupled neutral amino acid transporter 9 isoform 1		
	EKC98160.1	*Trichosporon asahii var. asahii CBS 8904*	PGQQQRPARTNQASGPRQSHHGDVVLNPGP	unknown		
	EJT49640.1	*Trichosporon asahii var. asahii CBS 8904*	PGQQQRPARTNQASGPRQSHHGDVVLNPGP	unknown		
	CAA29181.1	*Trypanosoma brucei*	RSLGTCKRAISSIIRTKMLVSGDVEENPGP	ORF 1 of non_LTR		
	AFN16295.1	*Trypanosoma brucei*	RSLGTCQRAISSIIRTKMLLSGDVEENPGP	Tbingi protein, EEP domain		
	CAD21860.1	*Trypanosoma brucei*	RSLGTCQRAISSIIRTKMLLSGDVEENPGP	nonLTR reverse transcriptase		
	CAD21861.1	*Trypanosoma brucei*	RSLGTCQRAISSIIRTKMLLSGDVEENPGP	reverse transcriptase & EEP domain		
TB2	M16026.1	*Trypanosoma brucei*	RSLGTCQRAISSIIRTKMLLSGDVEENPGP	Ingi		
	M16027.1	*Trypanosoma brucei*	GPWARASVPSAVSSALRLLLSGDVEENPGP	LINE1 RT element	D=0.452	

Appendix B

CBH11175.1	*Trypanosoma brucei gambiense DAL972*	RSLGTCQRAISSIIRTKMLLSGDVEENPGP	probably non-LTR
CBH15208.1	*Trypanosoma brucei gambiense DAL972*	RSLGTCQRAISSIIRTKMLLSGDVEENPGP	unknown
CBH09339.1	*Trypanosoma brucei gambiense DAL972*	RSLGTCQHAISSIIRAKMLLSGDVEENPGP	unknown
CBH12872.1	*Trypanosoma brucei gambiense DAL972*	VPGHVPACHQQYIIRTKMLLSGDVEENPGP	unknown
CBH10019.1	*Trypanosoma brucei gambiense DAL972*	MLLSGDVEENPGP	EEP, RNase
AEL79552.1	*Trypanosoma brucei TREU927*	RSLGTCQRAISSIIRTKMLLSGDVEENPGP	probably non-LTR
AFN16296.1	*Trypanosoma congolense*	ILPCTCGRATLDARRILLLVSGDVERNPGP	Tcoingi protein, reverse transcriptase
AFN16294.1	*Trypanosoma congolense*	LRHPNRQCALQEALRQKLLLCGDVEANPGP	L1Tc retrotransposon
CCD15992.1	*Trypanosoma congolense IL3000*	ILPCTCGCATLDARRILLLVSGDVERNPGP	unknown
CCC91257.1	*Trypanosoma congolense IL3000*	IRPCTCGCATLDARRILLLVSGDVERNPGP	EEP domain
CCD15700.1	*Trypanosoma congolense IL3000*	ILPCTCGRATLDAPRILLLVSGDVERNPGP	retrotransposon
CCD11781.1	*Trypanosoma congolense IL3000*	LRHPNRQCALQEALRQKLLLCGDVEANPGP	tubulin
CCD12873.1	*Trypanosoma congolense IL3000*	ILPCTCGRATLDAQRILLLVSGDVERNPGP	EEP domain
CCD13664.1	*Trypanosoma congolense IL3000*	LRHPNRQCALQEALRQKLLLCGDVEANPGP	EEP domain
CCD12221.1	*Trypanosoma congolense IL3000*	LRHPNRQCALQEALRQKLLLCGDVEANPGP	Tubulin_FtsZ
CCD11830.1	*Trypanosoma congolense IL3000*	ILPCTCGRATLDARRILLLVSGDVERNPGP	EEP domain
CCD15356.1	*Trypanosoma congolense IL3000*	ILPCTCGCATLDARRILLLVSGDVERNPGP	EEP domain
CCD14324.	*Trypanosoma congolense*	ILPCTCGRATLDARRILLLVSGDVERNPGP	EEP domain, reverse transcriptase

Appendix B

1	*IL3000*		
CCD16992.1	*Trypanosoma congolense IL3000*	ILPCTCGRATLDARRILLLVSGDVERNPGP	EEP domain, reverse transcriptase
CCD16834.1	*Trypanosoma congolense IL3000*	ILPCMCGRATLDARRLLLLVSEDIERNPGP	unknown
CCD12673.1	*Trypanosoma congolense IL3000*	ILPCTCGRATLDARRILLLVSGDIERNPGP	unknown
CCD14270.1	*Trypanosoma congolense IL3000*	ILPCTCGRATLDARRILLLVSGDIERNPGP	unknown
CCD13868.1	*Trypanosoma congolense IL3000*	ILPCTCGRATLDARRILLLVSGDIERNPGP	EEP Domain
CCD17397.1	*Trypanosoma congolense IL3000*	ILPCTCGRATLDARRILLLVSGDIERNPGP	EEP Domain
CCD13015.1	*Trypanosoma congolense IL3000*	ILPCTCGRTTLDARRILLLVSGDIERNPGP	RT_nLTR_like EEP domain
CCD13635.1	*Trypanosoma congolense IL3000*	VLPCTCGRATLDARRILLLISGDVERNPAP	EEP domain
EKF99947.1	*Trypanosoma cruzi*	QRYTYRLRAVCDAQRQKLLLSGDIEQNPGP	trans-sialidase & EEP domain
CAB41692.1	*Trypanosoma cruzi*	QRYTYRLRAVCDAQRQKLLLSGDIEQNPGP	L1Tc EEP Domain
EKG07481.1	*Trypanosoma cruzi*	QRYTYRLRAVCDAQRQKLLLSGDIEQNPGP	EEP Domain
AFN16293.1	*Trypanosoma cruzi*	QRYTYRLRAVCDAQRQKLLLSGDIEQNPGP	RT_nLTR_like
EKF98876.1	*Trypanosoma cruzi*	QRYTYRLRAVCDAQRQKLLLSGDIEQNPGP	Tubulin_FtsZ
EKF98124.1	*Trypanosoma cruzi*	QRYTYRLRAVCDAQRQKLLLSGDIEQNPGP	Retrotransposon hot spot protein
EKG08361.1	*Trypanosoma cruzi*	QRYTYRLRAVCDAQRQKLLLSGDIEQNPGP	EEP Domain
AAA67559.1	*Trypanosoma cruzi*	QPYTYCLRALCDAQRQKLLLIGDIEQNPGP	ORF1 EEP domain
ESS60740.1	*Trypanosoma cruzi Dm28c*	QRYTYHLRAVCDAQRQKLLLSGDIEQNPGP	unknown
ESS62691.1	*Trypanosoma cruzi Dm28c*	QRYTYRLRAVCDAQRQKLLLSGDIEQNPGP	EEP Domain
EKF26873.1	*Trypanosoma cruzi marinkellei*	QRYACCLRAVCDAQRQKLLLSGDIEQNPGP	EEP Domain
EKF32178.1	*Trypanosoma cruzi marinkellei*	QRYACCMRAVCDVQRQKLLLIGDIEQNPGP	EEP Domain

Appendix B

EKF39546.1	*Trypanosoma cruzi marinkellei*	QRYACCLRAVCDAQRQKLLLSGDIEQNPGP	EEP Domain
EKF27905.1	*Trypanosoma cruzi marinkellei*	QRYACCLRAVCDAQRQKLLPSGDIEQNPGP	EEP Domain
EKF29138.1	*Trypanosoma cruzi marinkellei*	QRYACCLRAVCDAQRQKLLLSGDIEQNPGP	EEP Domain
EKF39679.1	*Trypanosoma cruzi marinkellei*	CCLRAVCDAQRQKLLLSGDIEQNPGP	EEP Domain
EKF26301.1	*Trypanosoma cruzi marinkellei*	QRHASRLRSVCDAQRQKLLLGGDIEQNPGP	EEP domain
EKF28797.1	*Trypanosoma cruzi marinkellei*	QRYACCMRAVCDAQRQKLLLSGDIEQNPGP	EEP domain
EKF30768.1	*Trypanosoma cruzi marinkellei*	MRAVCDVQRQKLLLSGDIEQNPGP	EEP domain
EKF29321.1	*Trypanosoma cruzi marinkellei*	QRYACCLRAVCDAQRQKLLRSGDIEQNPGP	EEP Domain
EKF26696.1	*Trypanosoma cruzi marinkellei*	QRYACCLRAVCDAQRQKLLLSGDIEQNPGP	EEP Domain
AFN16297.1	*Trypanosoma vivax*	ILPCTCGRATLDARRLLLLISGDVERNPGP	Tvingi, EEP domain & reverse transcriptase
CCD18325.1	*Trypanosoma vivax Y486*	ILPCTCGRATLDARRLLLLISGDVERNPGP	retrotransposon hotspot protein
CCD21457.1	*Trypanosoma vivax Y486*	ILPCTCGRAMLDARRLLLLISVDVERNPGP	EEP domain
CCD18576.1	*Trypanosoma vivax Y486*	ILPCTCGRATLDARRLLLLISGDVERNPGP	EEP domain
CCD19727.1	*Trypanosoma vivax Y486*	ILPCTCGRATLDARRLLLLISGDVERNPGP	reverse transcriptase
CCD20619.1	*Trypanosoma vivax Y486*	ILPCTCGRATLDARRLLLLASGDVERNPGP	reverse transcriptase
CCD21559.1	*Trypanosoma vivax Y486*	ILPCTCGRAALVARRLLLLTSGDVERNPGP	EEP domain
CCD20511.1	*Trypanosoma vivax Y486*	ILPCTSGRATLDARRLLLLISVDVERNPGP	EEP domain
CCD19235.1	*Trypanosoma vivax Y486*	ILPCTCGRAALDARWILLLISGDVERNPGP	reverse transcriptase
CCD17958.1	*Trypanosoma vivax Y486*	ILPCTCGRATLDARRLLLLISGDVERNPGP	EEP domain
CCD19171.1	*Trypanosoma vivax Y486*	TLPCACGCATLDARRLLLLLRGDVERNPGP	reverse transcriptase
CCD18828.1	*Trypanosoma vivax Y486*	ILPCTCGCAALDARRVLLLISGDVERNPGP	EEP domain

Appendix B

CCD20623.1	*Trypanosoma vivax Y486*	ILPCTCGRATLDARRLLLLISGDVERNPGP	RT_nLTR_like		
CCD20366.1	*Trypanosoma vivax Y486*	MLRARWLLLLISGDVERDPGP	unknown		
CCD21312.1	*Trypanosoma vivax Y486*	MLRARWLLLLISGDVERDPGP	retrotransposon hot spot protein		
CCD19869.1	*Trypanosoma vivax Y486*	ILPCTRGRATLDARRLLLLVSGGVERNPGP	EEP domain, fragment		
CCD18962.1	*Trypanosoma vivax Y486*	ILPCTCARAALDARRLLLLISGGVERNPGP	EEP Domain		
CCD20313.1	*Trypanosoma vivax Y486*	ILPCTCGRAALDAQWRLLLIFVDAERNPGP	EEP Domain		
CCD21364.1	*Trypanosoma vivax Y486*	GRATLDVLRLLLLVSGDVERNSGP	Reverse transcriptase, EEP Domain		
CCD20286.1	*Trypanosoma vivax Y486*	ILPCTCGRATLDARRLLLLISGDVERNPVP	EEP Domain		
CCD19996.1	*Trypanosoma vivax Y486*	ILPCACGRAALDARRLLLLASGDVGRNPGP	EEP Domain		
CCD20778.1	*Trypanosoma vivax Y487*	ILPCTCGRATLDARRLLLLISGDVERNPGP	reverse transcriptase of nonLTR		
XP_006149057.1	*Tupaia chinensis*	MTNMDNDSRHLIAEVDDEVNPGP	putative sodium-coupled neutral amino acid transporter 9 isoform X7		
XP_006149051.1	*Tupaia chinensis*	MTNMDNDSRHLIAEVDDEVNPGP	putative sodium-coupled neutral amino acid transporter 9 isoform X1 [Tupaia chinensis]		
XP_004322640.1	*Tursiops truncatus*	MANMDSDSSHLLIPEVDNEINPGP	putative sodium-coupled neutral amino acid transporter 9-like, partial		
XP_006206063.1	*Vicugna pacos*	MANMDSDSRHLLIPEVNHEINPGP	putative sodium-coupled neutral amino acid transporter 9 isoform X1		
CR1-L2-1_XT	*Xenopus (Silurana) tropicalis*	HNNFKSFSHLLSLSLLLLLAAGDISPNPGP	CR1-L2 nonLTR	D=0.571	From RepBase
L2-2_XT	*Xenopus (Silurana) tropicalis*	RNHFKSSAHVFSLLFLLLLAAGDVSPNPGP	L2 nonLTR	D=0.565	From RepBase
L2-3_XT	*Xenopus (Silurana) tropicalis*	KTKTYKSRSHLAFLSFLLLAAGDISPNPGP	L2 nonLTR	D=0.555	From RepBase
L2-4_XT	*Xenopus (Silurana) tropicalis*	PRAFKSRSHLLSLTLLLLLAAGDISPNPGP	L2 nonLTR	D= 0.624	From RepBase
XP_005815779.1	*Xiphophorus maculatus*	SLIPGDLLRENTDEVLQDEQLSDVENNPGP	tau-tubulin kinase 2-like isoform X2		
XP_005815778.1	*Xiphophorus maculatus*	SLIPGDLLRENTDEVLQDEQLSDVENNPGP	tau-tubulin kinase 2-like isoform X1		

Appendix B

Table B. Viral 2A Sequences – list compiled 2011.

Accession no.	Virus	2A N-terminal sequence	DXXXNPGP C-terminus motif	
NP_653077.1	polyprotein [EquinerhinitisBvirus1]	EATLSTILSEGATNFSLLKLAG	DVEL	NPGP
P27586.1	NSP3_ROTPCRecName Non-structuralprotein3; NSP3; p38; p8	GNGNPLIVANAKFQIDKILISG	DVEL	NPGP
ABR18733.1	VP1 [Foot-and-mouthdiseasevirusAsia/IRN/05]	RKQKIIAPGKQALNFDLLKLAG	DVEL	NPGP
AAW82714.1	polyprotein [Foot-and-mouthdiseasevirus-typeAsia1]	RKQKIIAPGKQVMNFDLLKLAG	DVEL	NPGP
ACD67870.1	polyprotein [Rattheilovirus1]	FSDFFKHVREYHAAYYKQRLMH	DVET	NPGP
ACF19652.1	Polyprotein [Rattheilovirus1]	FSDFFKHVREYHAAYYKQRLMH	DVET	NPGP
BAC58035.1	polyprotein [Theiler's-likevirusofrats]	FSDFFKHVREYHAAYYKQRLMH	DVET	NPGP
AEC04618.1	polyprotein [Theiler'sencephalomyelitisvirus]	FGEFFKAVRGYHADYYKQRLIH	DVET	NPGP
ACG55800.1	polyprotein [Theiler'sencephalomyelitisvirus]	FGEFFKAVRGYHADYYKQRLIY	DVET	NPGP
ACG55799.1	polyprotein [Theiler'sencephalomyelitisvirus]	FGEFFRAVRGYHADYYRQRLIH	DVET	NPGP
ABD67451.1	polyprotein [Theiler'sencephalomyelitisvirus]	FGEFFKAVRGYHADYYRQRLIH	DVET	NPGP
P08544.1	POLG_TMEV polyprotein	FGEFFKAVRGYHADYYRQRLIH	DVET	NPGP
YP_001210296.2	viralpolyprotein [Saffoldvirus]	FTDFFKAVRDYHASYYKQRLQH	DVET	NPGP
P12296.3	POLG_ENMGORecName:Full=Genomepolyprotein	VFGLYHVFETHYAGYFSDLLIH	DVET	NPGP
AAA46547.1	polyprotein [Mengovirus]	VFGLYHVFETHYAGYFSDLLIH	DVET	NPGP
YP_001686841.1	BRV2polyprotein [BovinerhinitisBvirus]	LRLTGEIVKQGATNFELLQQAG	DVET	NPGP
AAO83985.1	polyprotein [LjunganvirusM1146]	YFKIYHDKDMDYAGGKFLNQCG	DVET	NPGP
ABQ02688.1	Polyprotein [Ljunganvirus]	YFNIMHSDEMDFAGGKFLNQCG	DVET	NPGP
NP_647602.1	polyprotein [Ljunganvirus]	YFNIMHSDEMDFAGGKFLNQCG	DVET	NPGP
AAM46080.1	AF327921_1polyprotein [Ljunganvirus174F]	YFNIMHSDEMDFAGGKFLNQCG	DVET	NPGP
AEM05833.1	polyprotein [MouseMosavirus]	KVVTTDDFVVFRSAHQDVTLGG	DVET	NPGP
YP_002956072.1	polyprotein [HumancosavirusE]	MAASDGLAPRKYLSYRKIQLSG	DVET	NPGP
NP_066241.1	replicasepolyprotein [Acutebeeparalysisvirus]	TGFLNKLYHCGSWTDILLLLSG	DVET	NPGP
AAN63803.2	replicasepolyprotein [Acutebeeparalysisvirus]	RTGFLNKLYHCGSWTDILLLWSG	DVET	NPGP
AAN63804.2	AF486073_1replicasepolyprotein [Acutebeeparalysisvirus	TGFLNKLYHCGSWTDILLLLSG	DVET	NPGP
NP_044945.1	replicasepolyprotein [DrosophilaCvirus]	QGIGKKNPKQEAARQMLLLLSG	DVET	NPGP
ADF56673.2	polyprotein [DrosophilaCvirus]	QGIGKKNPKQEAARQMLLLLSG	DVET	NPGP
ADF56663.2	polyprotein [DrosophilaCvirus]	MTQGIGKKNPKQEAARQMLLLLSG	DVET	NPGP
ADF56659.2	polyprotein [DrosophilaCvirus]	MTQGIGKKNPKQEAARQMLLLLSG	DVET	NPGP
ADF56657.2	polyprotein [DrosophilaCvirus]	MTQGIGKKNPKQEAARQMLLLLSG	DVET	NPGP
ADF56653.2	polyprotein [DrosophilaCvirus]	MTQGIGKKNPKQEAARQMLLLLSG	DVET	NPGP
ADF56652.2	polyprotein [DrosophilaCvirus]	QGIGKKNPKQEAARQMLLLLSG	DVET	NPGP
ADF56671.2	polyprotein [DrosophilaCvirus]	QGIGKKNPKQEAARQMLLLLSG	DVET	NPGP
ADF56660.2	polyprotein [DrosophilaCvirus]	MTQGIGKKNPKQEAARQMLLLLSG	DVET	NPGP
ADF56684.2	polyprotein [DrosophilaCvirus]	QGIGKKNPKQEAARQMLLLLSG	DVET	NPGP
ADF56681.2	polyprotein [DrosophilaCvirus]	MTQGIGKKNPKQEAARQMLLLLSG	DVET	NPGP
ADF56676.2	polyprotein [DrosophilaCvirus]	MTQGIGKKNPKQEAARQMLLLLSG	DVET	NPGP
ADF56655.2	polyprotein [DrosophilaCvirus]	QGIGKKNPKQEAARQMLLLLSG	DVET	NPGP
ADF56672.2	polyprotein [DrosophilaCvirus]	QGIGKKNPKQEAARQMLLLLSG	DVET	NPGP

Appendix B

YP_001816886.1	protein2A [Saffoldvirus]	FTDFFKAVRDYHASYYKQRLQH	DVET	NPGP
ADF56669.2	polyprotein [DrosophilaCvirus]	ARQMLLLLSG	DVET	NPGP
YP_001686942.1	2A [BovinerhinitisBvirus]	LRLTGEIVKQGATNFELLQQAG	DVET	NPGP
ACG55802.1	polyprotein [Theiler'sencephalomyelitisvirus]	FGEFFKAGRGYHADYYKQRLIH	DVEM	NPGP
NP_040350.1	viralpolyprotein [Theilovirus]	FREFFKAVRGYHADYYKQRLIH	DVEM	NPGP
P08545.2	POLG_TMEVGRecName:Full=Genomepolyprotein	FREFFKAVRGYHADYYKQRLIH	DVEM	NPGP
P13899.1	POLG_TMEVDRecName:Full=Genomepolyprotein	FGEFFRAVRAYHADYYKQRLIH	DVEM	NPGP
ACG55801.1	polyprotein [Vilyuiskhumanencephalomyelitisvirus]	FGEFFKAVRGYHADYYKQRLIH	DVEM	NPGP
NP_740428.1	protein2A [Theilovirus]	FREFFKAVRGYHADYYKQRLIH	DVEM	NPGP
ACO92353.1	polyprotein [Saffoldvirus]	FTEFFKAVRDYHASYYKQRLQH	DIET	NPGP
ACO92355.1	polyprotein [Saffoldvirus]	FTDFFKAVRDYHASYYKQRLQH	DIET	NPGP
AEM00022.1	polyprotein [Saffoldvirus]	FTDFFKAVRDYHASYYKQRLQH	DIET	NPGP
AEK80410.1	polyprotein [Saffoldvirus]	FTDFFKAVRDYHASYYKQRLQH	DIET	NPGP
ADO20359.1	polyprotein [Saffoldvirus]	FTDFFKAVRDYHASYYKQRLQH	DIET	NPGP
ADO20358.1	polyprotein [Saffoldvirus]	FTDFFKAVRDYHASYYKQRLQH	DIET	NPGP
ADK91816.1	polyprotein [Saffoldvirus]	FSDFFKAVRDYHASYYKQRLQH	DIET	NPGP
ADK91815.1	polyprotein [Saffoldvirus]	FSDFFKAVRDYHASYYKQRLQH	DIET	NPGP
ADK91814.1	polyprotein [Saffoldvirus]	FSDFFKAVRDYHASYYKQRLQH	DIET	NPGP
ADK91813.1	polyprotein [Saffoldvirus]	FTDFSKAVRDYHASYYKQRLQH	DIET	NPGP
ACG61138.2	polyprotein [CardiovirusD/VI2223/2004]	FTDFFKAVRDYHSSYYKQRLQH	DIET	NPGP
ACG61137.2	polyprotein [CardiovirusD/VI2273/2004]	FTDFFKAVRDYHASYYKQRLQH	DIET	NPGP
CAR62533.1	polyprotein [Saffoldvirus]	FTDFFKAVRDYHASYYKQRLQH	DIET	NPGP
ADO20363.1	polyprotein [Saffoldvirus]	FTDFFKAVRDYHSSYYKQRLQH	DIET	NPGP
CBW45996.2	polyprotein [Saffoldvirus]	FTDFFKAVRDYHSSYYKQRLQH	DIET	NPGP
CBS91673.1	viralpolyprotein [Saffoldvirus]	FTDFFKAVRDYHASYYKQRLQH	DIET	NPGP
ADF28539.1	polyprotein [HumanTMEV-likecardiovirus]	FTDFFKAVRDYHASYYKQRLQH	DIET	NPGP
ACG61136.2	polyprotein [CardiovirusBR/118/2006]	FTDFFKAVRDYHASYYKQRLQH	DIET	NPGP
ACG61135.2	polyprotein [CardiovirusD/VI2229/2004]	FTGFFKAVRDYHASYYKQRLQH	DIET	NPGP
YP_001949875.1	polyprotein [HumanTMEV-likecardiovirus]	FTDFFKAVRDYHASYYKQRLQH	DIET	NPGP
ADN52625.1	polyprotein [Porcineencephalomyocarditisvirus]	VFGLYRIFNAHYAGYFADLLIH	DIET	NPGP
ACQ90253.1	polyprotein [Encephalomyocarditisvirus]	VFGLYRIFNAHYAGYFADLLIH	DIET	NPGP
ACI47518.1	polyprotein [Encephalomyocarditisvirus]	LFGLYRIFNAHYAGYFADLLIH	DIET	NPGP
ACI47517.1	polyprotein [Encephalomyocarditisvirus]	VFGLYRIFNAHYAGYFADLLIH	DIET	NPGP
ACM45091.1	polyprotein [Encephalomyocarditisvirus]	VFGLYRIFNAHYAGYFADLLIH	DIET	NPGP
ACM45090.1	polyprotein [Encephalomyocarditisvirus]	VFGLYRIFNAHYAGYFADLLIH	DIET	NPGP
ABG20637.1	polyprotein [Porcineencephalomyocarditisvirus]	VFGLYRIFNAHYAGYFADLLIH	DIET	NPGP
ABE77395.1	polyprotein [Encephalomyocarditisvirus]	VFGLYRIFNAHYAGYFADLLIH	DIET	NPGP
ABE77396.1	polyprotein [Encephalomyocarditisvirus]	VFGLYRIFNAHYAGYFADLLIH	DIET	NPGP
AAP51180.1	polyprotein [Encephalomyocarditisvirus]	IFGLYRIFSTHYAGYFSDLLIH	DIET	NPGP
CAA60776.1	completeviralprotein [Encephalomyocarditisvirus]	VFGLYSIFNAHYAGYFADLLIH	DIET	NPGP
AAL83502.1	AF356822_1polyprotein [Encephalomyocarditisvirus]	VFGLYRIFNAHYAGYFADLLIH	DIET	NPGP
CAA52361.1	unnamedproteinproduct [Encephalomyocarditisvirus]	VFGLYRIFNAHYAGYFADLLIH	DIET	NPGP
NP_056777.1	hypotheticalproteinEMCVgp1	VFGLYRIFNAHYAGYFADLLIH	DIET	NPGP

Appendix B

	[Encephalomyocarditisvirus]			
AAA43035.1	polyprotein [Encephalomyocarditisvirus]	VFGLYSIFNAHYAGYFADLLIH	DIET	NPGP
AAA43033.1	polyproteinregion [Encephalomyocarditisvirus]	VFGLYGIFNAHYAGYFADLLIH	DIET	NPGP
ABC25550.1	polyprotein [Encephalomyocarditisvirus]	VFGLYRIFNAHYAGYFADLLIH	DIET	NPGP
P17593.1	POLG_EMCVBRecName:Full=Genomepolyprotein	VFGLYGIFNAHYAGYFADLLIH	DIET	NPGP
P17594.2	POLG_EMCVDRecName:Full=Genomepolyprotein	VFGLYGIFNAHYAGYFADLLIH	DIET	NPGP
P03304.1	POLG_EMCVRecName:Full=Genomepolyprotein	VFGLYRIFNAHYAGYFADLLIH	DIET	NPGP
ABI15777.2	polyprotein [Encephalomyocarditisvirus]	IFGLYHIFETHYAGYFADLLIH	DIET	NPGP
ACO92357.1	polyprotein [Saffoldvirus]	FTDFFKAVRDYHASYYKQRLQH	DIET	NPGP
YP_002268402.1	polyprotein [Senecavalleyvirus]	RAWCPSMLPFRSYKQKMLMQSG	DIET	NPGP
YP_002956076.1	polyprotein [HumancosavirusD]	IIARPYIRESSNVSRLKLLLSG	DIET	NPGP
YP_002956075.1	polyprotein [HumancosavirusB]	PPRPLSTSIRSRAAYLRQKLMH	DIET	NPGP
YP_001950226.1	protein2A [HumanTMEV-likecardiovirus]	FTDFFKAVRDYHASYYKQRLQH	DIET	NPGP
NP_919029.1	polyprotein [Ectropisobliquapicorna-likevirus]	GQRTTEQIVTAQGWAPDLTQDG	DVES	NPGP
NP_277061.1	polyprotein [Perinanudavirus]	GQRTTEQIVTAQGWVPDLTVDG	DVES	NPGP
AAQ17044.1	polyprotein [Ectropisobliquapicorna-likevirus]	GQRTTEQIVTAQGWAPDLTQDG	DVES	NPGP
NP_201566.1	polyprotein [EquinerhinitisBvirus2]	VADWENLLSQGATNFDLLKLAG	DVES	NPGP
AAT01719.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
AAT01787.1	polyprotein [Foot-and-mouthdiseasevirus-typeSAT1]	YKVSLVAPEKQMANFALLKLAG	DVES	NPGP
ADI24382.1	polyprotein [Foot-and-mouthdiseasevirus-typeSAT1]	YKVALVAPAKQLSNFDLLKLAG	DVES	NPGP
AAT01782.1	polyprotein [Foot-and-mouthdiseasevirus-typeSAT1]	YKTAITKPAKQMCSFDLLKLAG	DVES	NPGP
AAT01786.1	polyprotein [Foot-and-mouthdiseasevirus-typeSAT1]	YKTALVKPAKQLCNFDLLKLAG	DVES	NPGP
AAT01783.1	polyprotein [Foot-and-mouthdiseasevirus-typeSAT1]	YQTALTKPAKQLCNFDLLKLAG	DVES	NPGP
YP_022777.1	polyprotein [Foot-and-mouthdiseasevirus-typeSAT1]	YKTSIVRPAKQLCNFDLLKLAG	DVES	NPGP
AAT01788.1	polyprotein [Foot-and-mouthdiseasevirus-typeSAT1]	YKTTLVKPAKQLSNFDLLKLAG	DVES	NPGP
AAT01784.1	polyprotein [Foot-and-mouthdiseasevirus-typeSAT1]	YQTALVRPAKQLCNFDLLMLAG	DVES	NPGP
AAT01785.1	polyprotein [Foot-and-mouthdiseasevirus-typeSAT1]	HKTALVKPAKQLCNFDLLKLAG	DVES	NPGP
AAT01789.1	polyprotein [Foot-and-mouthdiseasevirus-typeSAT1]	YKTAITKPVKQLCNFDLLKLAG	DVES	NPGP
ACM79369.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
ACL52159.1	polyprotein [Foot-and-mouthdiseasevirus-typeSAT2]	RFDAPIGVERQTLNFDLLKLAG	DVES	NPGP
ABN05227.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
AAT01746.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQSIIAPAKQLLNFDLLKLAG	DVES	NPGP
AAT01705.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
P49303.1	POLG_FMDVZRecName:Full=Genomepolyprotein	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
ADI24381.1	polyprotein [Foot-and-mouthdiseasevirus-typeSAT2]	RFDAPIGVEKQLCNFDLLKLAG	DVES	NPGP
ADI24380.1	polyprotein [Foot-and-mouthdiseasevirus-typeSAT2]	RFDAPIGVEKQLCNFDLLKMAG	DVES	NPGP
ABU87557.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
ABU87556.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP

Appendix B

ABU87555.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPTKQLLNFDLLKLAG	DVES	NPGP
AAQ11227.1	polyprotein [Foot-and-mouthdiseasevirus-typeSAT2]	RFDSPIGVKKQLCNFDLLKLAG	DVES	NPGP
CAB62902.1	polyprotein [Foot-and-mouthdiseasevirus-typeSAT2]	RFDAPIGVAKQLLNFDLLKLAG	DVES	NPGP
AAT01795.1	polyprotein [Foot-and-mouthdiseasevirus-typeSAT3]	YKTPLVKPEKQLCNFDLLKLAG	DVES	NPGP
AAT01794.1	polyprotein [Foot-and-mouthdiseasevirus-typeSAT3]	YKTPLVKPEKQLCNFDLLKLAG	DVES	NPGP
AAT01796.1	polyprotein [Foot-and-mouthdiseasevirus-typeSAT3]	YKTPLVKPDKQMCNFDLLKLAG	DVES	NPGP
YP_003328989.1	polyprotein [Foot-and-mouthdiseasevirus-typeSAT3]	YKIKLVAPDKQLCNFDLLKLAG	DVES	NPGP
YP_024314.1	polyprotein [Foot-and-mouthdiseasevirus-typeSAT2]	RFDAPIGVEKQLLNFDLLKLAG	DVES	NPGP
AAT01790.1	polyprotein [Foot-and-mouthdiseasevirus-typeSAT2]	RFDAPIGVEKQLFNFDLLKLAG	DVES	NPGP
AAT01791.1	polyprotein [Foot-and-mouthdiseasevirus-typeSAT2]	RFDAPIGVEKQLCNCDLLKLAG	DVES	NPGP
AAT01737.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	YKQQIIAPAKQLLNFDLLKLAG	DVES	NPGP
AAT01704.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
AEE65049.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
AEE65048.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
AEE65047.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
AEE65046.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
AEE65045.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
AEE65044.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
AEE65037.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
AEE65036.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
ADN28046.1	polyprotein [Foot-and-mouthdiseasevirus-typeO]	HKQKIVAPVKQSLNFDLLKLAG	DVES	NPGP
ADM16570.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
ADM16569.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
ADM16568.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
ADM16567.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
ADM16566.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
ABV03522.1	polyprotein [Foot-and-mouthdiseasevirus-typeAsia1]	RKQEIIAPEKQTLNFDLLKLAG	DVES	NPGP
AAT01728.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	RKQKIIAPEKQLLNFDLLKLAG	DVES	NPGP
AAT01725.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQSLNFDLLKLAG	DVES	NPGP
AAT01756.1	polyprotein [Foot-and-mouthdiseasevirus-typeO]	HKQKIVAPVKQTLNFDLLKLAG	DVES	NPGP
AAT01711.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
AAT01699.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	YKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
AAT01700.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	YKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
AAT01736.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPEKQLLNFDLLKLAG	DVES	NPGP
AAT01709.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
AAT01696.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	YKQKIIAPEKQLLNFDLLKLAG	DVES	NPGP
AAT01708.1	polyprotein [Foot-and-mouthdiseasevirus-	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP

Appendix B

	typeA]			
AAT01695.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPGKQLLNFDLLKLAG	DVES	NPGP
AAT01744.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPEKQLLNFDLLKLAG	DVES	NPGP
AAT01731.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
AAT01729.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPEKQLLNFDLLKLAG	DVES	NPGP
AAT01733.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPEKQLLNFDLLKLAG	DVES	NPGP
AAT01698.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQRIIAPAKQLLNFDLLKLAG	DVES	NPGP
AAT01726.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	RKQKIIAPEKQLLNFDLLKLAG	DVES	NPGP
AAT01734.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIITPVKQLLNFDLLKLAG	DVES	NPGP
AAT01713.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPTKQLLNFDLLKLAG	DVES	NPGP
AAT01715.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPSKQLLNFDLLKLAG	DVES	NPGP
AAT01716.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
AAT01701.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	YKQKIIAPEKQLLNFDLLKLAG	DVES	NPGP
AEE65052.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
AEE65051.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKTIAPAKQLLNFDLLKLAG	DVES	NPGP
AEE65050.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
AEE65043.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
AEE65042.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
AEE65041.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQFLNFDLLKLAG	DVES	NPGP
AEE65039.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
AEE65038.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
ADX97244.1	polyprotein [Foot-and-mouthdiseasevirus-typeO]	HKQKIVAPVKQILNFDLLKLAG	DVES	NPGP
ADV52248.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVES	NPGP
ADV52245.1	polyprotein [Foot-and-mouthdiseasevirus-typeO]	HKQKIVAPVKQLLNFDLLKLAG	DVES	NPGP
ADV52243.1	polyprotein [Foot-and-mouthdiseasevirus-typeO]	HKQKIVAPVKQLLNFDLLKLAG	DVES	NPGP
ADV52247.1	polyprotein [Foot-and-mouthdiseasevirus-typeO]	HKQKIVAPVKQSWKFDLLKLAG	DVES	NPGP
ADU05399.1	polyprotein [Foot-and-mouthdiseasevirus-typeO]	HKQKIVAPVKQLLNFDLLKLAG	DVES	NPGP
ADR83528.1	polyprotein [Foot-and-mouthdiseasevirus-typeO]	HKQKIVAPVKQSLNFDLLKLAG	DVES	NPGP
ADR66170.1	polyprotein [Foot-and-mouthdiseasevirus-typeO]	HKQKIVAPVKQLLNFDLLKLAG	DVES	NPGP
ADR66169.1	polyprotein [Foot-and-mouthdiseasevirus-typeO]	HKQKIVAPVKQLLNFDLLKLAG	DVES	NPGP
ADR51745.1	polyprotein [Foot-and-mouthdiseasevirus-typeO]	HKQKIVAPVKQLLNFDLLKLAG	DVES	NPGP
ADR51742.1	polyprotein [Foot-and-mouthdiseasevirus-typeO]	HKQKIVAPVKQLLNFDLLKLAG	DVES	NPGP
ADR51741.1	polyprotein [Foot-and-mouthdiseasevirus-typeO]	HKQKIVAPVKQLLNFDLLKLAG	DVES	NPGP
ADM36039.1	polyprotein [Foot-and-mouthdiseasevirus-typeO]	HKQKIVAPVKQLLNFDLLKLAG	DVES	NPGP
ADH32285.1	polyprotein [Foot-and-mouthdiseasevirus-typeO]	HKQKIVAPVKQSLNFDLLKLAG	DVES	NPGP
ADH32284.1	polyprotein [Foot-and-mouthdiseasevirus-typeO]	HKQKIVAPVKQSWKFDLLKLAG	DVES	NPGP
ADH32283.1	polyprotein [Foot-and-mouthdiseasevirus-typeO]	HKQKIVAPVKQSWKFDLLKLAG	DVES	NPGP

Appendix B

ADC92548.1	polyprotein [Foot-and-mouthdiseasevirus-typeO]	YKQKIVAPVKQSLNFDLLKLAG	DVES	NPGP
ADC92547.1	polyprotein [Foot-and-mouthdiseasevirus-typeO]	HKQKIVAPVKQSLNFDLLKLAG	DVES	NPGP
NP_851403.1	non-structuralpolyprotein [Kashmirbeevirus]	IGFLNKLYKCGTWESVLNLLAG	DIEL	NPGP
AEJ21074.1	nonstructuralproteinNSP3 [HumanrotavirusC]	GTGYPLIVANSKFQIDKILISG	DIEL	NPGP
ADP76634.1	nonstructuralproteinNSP3 [HumanrotavirusC]	GVGYPLIVANSKFQIDKILISG	DIEL	NPGP
ADP76633.1	nonstructuralproteinNSP3 [HumanrotavirusC]	GVGYPLIVANSKFQIDKILISG	DIEL	NPGP
ADP76632.1	nonstructuralproteinNSP3 [HumanrotavirusC]	GTGYPLIVANSKFQIDKILISG	DIEL	NPGP
AAX21225.1	nonstructuralprotein3 [HumanrotavirusC]	GVGYPLIVANSKFQIDKILISG	DIEL	NPGP
AAX21224.1	nonstructuralprotein3 [HumanrotavirusC]	GTGYPLIVANSKFQIDKILISG	DIEL	NPGP
AAX21226.1	nonstructuralprotein3 [HumanrotavirusC]	GVGYPLIVANSKFQIDKILISG	DIEL	NPGP
P34717.1	NSP3_ROTBSRecName:Full=Non-structuralprotein3	GIGNPLIVANSKFQIDRILISG	DIEL	NPGP
YP_392486.1	nonstructuralprotein3 [RotavirusC]	GAGYPLIVANSKFQIDKILISG	DIEL	NPGP
NP_919029.1	polyprotein [Ectropisobliquapicorna-likevirus]	TRGGLQRQNIIGGGQRDLTQDG	DIES	NPGP
NP_277061.1	polyprotein [Perinanudavirus]	TRGGLRRQNIIGGGQKDLTQDG	DIES	NPGP
AAQ17044.1	polyprotein [Ectropisobliquapicorna-likevirus]	TRGGLQRQNIIGGGQRDLTQDG	DIES	NPGP
ADR51746.1	polyprotein [Foot-and-mouthdiseasevirus-typeO]	HKQKIVAPVKQLLNLDLLKLAG	DIES	NPGP
ADR51744.1	polyprotein [Foot-and-mouthdiseasevirus-typeO]	HKQKIVAPVKQLLNFDLLKLAG	DIES	NPGP
ADR51743.1	polyprotein [Foot-and-mouthdiseasevirus-typeO]	HKQKIVAPVKQLLNSDLLKLAG	DIES	NPGP
YP_002956074.1	polyprotein [HumancosavirusA]	VLPRPLTRAERDVARDLLLIAG	DIES	NPGP
ACT78456.1	p101 [Bombyxmoricypovirus1]	RTAFDFQQDVFRSNYDLLKLCG	DIES	NPGP
AAL27618.1	AF433660_1unknown [Bombyxmoricypovirus1]	RTAFDFQQDVFRSNYDLLKLCG	DIES	NPGP
BAB20438.1	BmCPV-1p101 [Bombyxmoricypovirus1]	RTAFDFQQDVFRSNYDLLKLCG	DIES	NPGP
BAB20437.1	BmCPV-1p101 [Bombyxmoricypovirus1]	RTAFDFQQDVFRSNYDLLKLCG	DIES	NPGP
ACJ48052.1	polyprotein [Ljunganvirus64-7855]	YFKVYHDIEMDHSTKCFLNQCG	DVEE	NPGP
NP_653143.1	polyprotein [Porcineteschovirus]	MTVMAFQGPGATNFSLLKQAG	DVEE	NPGP
ACT66681.1	polyprotein [Porcineteschovirus5]	MVALSFQGPGATNFSLLKQAG	DVEE	NPGP
AAK12409.1	AF296115_1polyprotein [Porcineteschovirus6]	MMLQGPGATNFSLLKQAG	DVEE	NPGP
AAK12385.1	AF296091_1polyprotein [Porcineteschovirus6]	MTTMSFQGPGATNFSLLKQAG	DVEE	NPGP
AAK12381.1	AF296087_1polyprotein [Porcineteschovirus2]	MTTMMLQGPGATNFSLLKQAG	DVEE	NPGP
AAK12401.1	AF296107_1polyprotein [Porcineteschovirus2]	MTTMSLQGPGATNFSLLKQAG	DVEE	NPGP
AAK12402.1	AF296108_1polyprotein [Porcineteschovirus2]	MTTLSLQGPGATNFSLLRQAG	DVEE	NPGP
AAK12411.1	AF296117_1polyprotein [Porcineteschovirus6]	MTTMMLQGPGATNFSLLKQAG	DVEE	NPGP
AAK12403.1	AF296109_1polyprotein [Porcineteschovirus2]	MTTMSLQGPGATNFSLLKQAG	DVEE	NPGP
ADP65803.1	polyprotein [Porcineteschovirus2]	MTTMTLQGPGATNFSLLKQAG	DVEE	NPGP
AAK12386.1	AF296092_1polyprotein [Porcineteschovirus7]	MTVVSFQGPGATNFSLLKQAG	DVEE	NPGP
AAK12407.1	AF296113_1polyprotein [Porcineteschovirus4]	MTTLTLQGPGATNFSLLKQAG	DVEE	NPGP
AAK12406.1	AF296112_1polyprotein [Porcineteschovirus4]	MTALTFQGPGATNFSLLKQAG	DVEE	NPGP
AAK12405.1	AF296111_1polyprotein [Porcineteschovirus4]	MTTMSLQGPGATNFSLLKQAG	DVEE	NPGP
AAK12383.1	AF296089_1polyprotein [Porcineteschovirus4]	MTTMMLQGPGASNFSLLKQAG	DVEE	NPGP
AAK12398.1	AF296104_1polyprotein [Porcineteschovirus1]	MTRMSFQGPGATNFSLLKQAG	DVEE	NPGP
AAK12382.1	AF296088_1polyprotein [Porcineteschovirus3]	MTTMSFQGPGASSFSLLKQAG	DVEE	NPGP
AAK12397.1	AF296103_1polyprotein [Porcineteschovirus1]	MTRMSFQGPGATNFSLLKQAG	DVEE	NPGP
AAK12390.1	AF296096_1polyprotein [Porcineteschovirus11]	MTRMSFQGPGATNFSLLKRAG	DVEE	NPGP

Appendix B

AAK12394.1	AF296100_1polyprotein [Porcineteschovirus1]	MTTMTLQGPGATNFSLLKQAG	DVEE	NPGP
AAK12396.1	AF296102_1polyprotein [Porcineteschovirus1]	MTAMAFQGPGATNFSLLKQAG	DVEE	NPGP
AAK12418.1	AF231769_1polyprotein [Porcineteschovirus1]	MTTLSYQGPGATNFSLLKQAG	DVEE	NPGP
AAK12384.1	AF296090_1polyprotein [Porcineteschovirus5]	MTTMLFQGPGAANFSLLRQAG	DVEE	NPGP
AAK12417.1	AF231768_1polyprotein [Porcineteschovirus1]	MTTMSYQGPGATNFSLLKQAG	DVEE	NPGP
AAK12416.1	AF231767_1polyprotein [Porcineteschovirus1]	MTTISYQGPGATNFSLLKQAG	DVEE	NPGP
ABC84373.1	polyprotein [Porcineteschovirus1]	MTTLSYQGPGATNFSLLKQAG	DVEE	NPGP
AAK12413.1	AF296119_1polyprotein [Porcineteschovirus10]	MTTLSFQGPGATNFSLLKQAG	DVEE	NPGP
AAK12389.1	AF296095_1polyprotein [Porcineteschovirus10]	MTTLSFQGPGATNFSLLKQAG	DVEE	NPGP
BAB32828.1	polyprotein [Porcineteschovirus1]	MTTLSYQGPGATNFSLLKQAG	DVEE	NPGP
AAK12388.1	AF296094_1polyprotein [Porcineteschovirus9]	MTTMAFQGPGATNFSLLKQAG	DVEE	NPGP
ABY71756.1	non-structuralpolyprotein [Israelacuteparalysisvirusofbees]	IGFLNKLYRCGDWESILLLLSG	DVEE	NPGP
ABY71754.1	non-structuralpolyprotein [Israelacuteparalysisvirusofbees]	IGFLNKLYRCGDWESILLLLSG	DVEE	NPGP
ACD01399.1	polymerasepolyprotein [Israelacuteparalysisvirusofbees]	IGFLNKLYRCGDWESILLLLSG	DVEE	NPGP
ACT66680.1	polyprotein [Porcineteschovirus4]	MTAMFLQGPGATNFSLLKQAG	DVEE	NPGP
AAK12410.1	AF296116_1polyprotein [Porcineteschovirus6]	MTTMMLQGPGATNFSLLKQAG	DVEE	NPGP
AAR31762.1	polyprotein [Porcineteschovirus]	MTRMSLQGPGASNFSLLKQAG	DVEE	NPGP
AAR31753.1	polyprotein [Porcineteschovirus2]	MTTMSLQGPGATNFSLLKQAG	DVEE	NPGP
AAR31752.1	polyprotein [Porcineteschovirus3]	MTTMTFQGRGATNFSLLKQAG	DVEE	NPGP
ACT66676.1	polyprotein [Porcineteschovirus1]	MTVMTFQGPGATNFSLLKQAG	DVEE	NPGP
AAR31744.1	polyprotein [Porcineteschovirus1]	MTVMTFQGPGATNFSLLKQAG	DVEE	NPGP
AAR31747.1	polyprotein [Porcineteschovirus1]	MTVMTFQGPGATNFSLLKQAG	DVEE	NPGP
AAK12404.1	AF296110_1polyprotein [Porcineteschovirus2]	MTTMSFQGPGATNFSLLKQAG	DVEE	NPGP
AEL29704.1	polyprotein [Porcineteschovirus]	MTTMTLQGPGATNFSLLKQAG	DVEE	NPGP
AAK12391.1	AF296097_1polyprotein [Porcineteschovirus1]	MTTMSYQGPGATNFSLLKQAG	DVEE	NPGP
AAK12399.1	AF296105_1polyprotein [Porcineteschovirus1]	MTVMTFQGPGATNFSLLKQAG	DVEE	NPGP
AAR31760.1	polyprotein [Porcineteschovirus]	MTTLSFQGPGATNFSLLKQAG	DVEE	NPGP
AAR31759.1	polyprotein [Porcineteschovirus]	MTTLSFQGPGATNFSLLKQAG	DVEE	NPGP
AAR31761.1	polyprotein [Porcineteschovirus]	MTTLSFQGPGATNFSLLKQAG	DVEE	NPGP
AAK12395.1	AF296101_1polyprotein [Porcineteschovirus1]	MTVMTFQGPGATNFSLLKQAG	DVEE	NPGP
AAK12393.1	AF296099_1polyprotein [Porcineteschovirus1]	MTRLSFQGPGATNFSLLKQAG	DVEE	NPGP
AAR31749.1	polyprotein [Porcineteschovirus2]	MMLQGPGATNFSLLKQAG	DVEE	NPGP
AEL29706.1	polyprotein [Porcineteschovirus]	MTTMSFQGPGATNFSLLRQAG	DVEE	NPGP
AAR31765.1	polyprotein [Porcineteschovirus1]	MTVMTFQGPGATNFSLLKQAG	DVEE	NPGP
ACT66678.1	polyprotein [Porcineteschovirus3]	MTTMSFQGPGATNFSLLKQAG	DVEE	NPGP
ACT66683.1	polyprotein [Porcineteschovirus5]	MTTMSFQGPGATNFSLLKQAG	DVEE	NPGP
AAR31754.1	polyprotein [Porcineteschovirus2]	MTTMSLQGPGATNFSLLKQAG	DVEE	NPGP
AAR31764.1	polyprotein [Porcineteschovirus1]	MTVMAFQGPGATNFSLLKQAG	DVEE	NPGP
ACT66684.1	polyprotein [Porcineteschovirus5]	MTTMSFQGPGATNFSLLKQAG	DVEE	NPGP
AAK12408.1	AF296114_1polyprotein [Porcineteschovirus5]	MTTMSFQGPGATNFSLLKQAG	DVEE	NPGP
AAR31751.1	polyprotein [Porcineteschovirus]	MTTLSFQGPGATNFSLLKQAG	DVEE	NPGP
AAK12414.1	AF296120_1polyprotein [Porcineteschovirus11]	MTTLSFQGPGATNFSLLKQAG	DVEE	NPGP
AAR31748.1	polyprotein [Porcineteschovirus11]	MARMSFQGPGATNFSLLKQAG	DVEE	NPGP
AEL29705.1	polyprotein [Porcineteschovirus]	MTTMSLQGPGATNFSLLKQAG	DVEE	NPGP

aa

Appendix B

AAK12400.1	AF296106_1polyprotein [Porcineteschovirus1]	MTVMTFQGPGATNFSLLKQAG	DVEE	NPGP
AAR31767.1	polyprotein [Porcineteschovirus1]	MTVMAFQGPGATNFSLLKQAG	DVEE	NPGP
ACT66686.1	polyprotein [Porcineteschovirus11]	MTRVSFQGSGATNFSLLKQAG	DVEE	NPGP
AAK12415.1	AF296121_1polyprotein [Porcineteschovirus11]	MTAMALQGPGATNFSLLKQAG	DVEE	NPGP
AAC97195.1	capsidproteinprecursor [Thoseaasignavirus]	RGPRPQNLGVRAEGRGSLLTCG	DVEE	NPGP
NP_573542.1	capsidproteinprecursor [Euprosternaelaeasavirus]	RRLPESAQLPQGAGRGSLVTCG	DVEE	NPGP
AAQ14330.1	capsidprotein [Thoseaasignavirus]	RGPRPQNLGVRAEGRGSLLTCG	DVEE	NPGP
AEE65040.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNSDLLKLAG	DVEP	NPGP
ABV03521.1	polyprotein [Foot-and-mouthdiseasevirus-typeO]	HKQKIVAPEKQLLNFDLLKLAG	DVEP	NPGP
AAP79123.1	polyprotein [Foot-and-mouthdiseasevirus-typeO]	HKQKIVAPVKQLLNFDLLKLAG	DVEP	NPGP
AAT01702.1	polyprotein [Foot-and-mouthdiseasevirus-typeA]	HKQKIIAPAKQLLNFDLLKLAG	DVEP	NPGP
ABI93984.1	polyprotein [Foot-and-mouthdiseasevirus-typeAsia1]	RKQEIIAPEKQMMNFDLLKLAG	DVEP	NPGP
AEL29703.1	polyprotein [Porcineteschovirus]	ALTTMSLQGPGATNFSLLKQAG	DIEE	NPGP
ACT65996.2	polyprotein [Porcineteschovirus8]	ALTTMSLQGPGATNFSLLKQAG	DIEE	NPGP
AAK12412.1	AF296118_1polyprotein [Porcineteschovirus8]	ALTTMSLQGPGATNFSLLKQAG	DIEE	NPGP
AAK12387.1	AF296093_1polyprotein [Porcineteschovirus8]	ALTTMSLQGPGATNFSLLKQAG	DIEE	NPGP
AEL12438.1	polymerasepolyprotein [Israelacuteparalysisvirusofbees]	IGFLNKLYRCGDWDSILLLLSG	DIEE	NPGP
ABY57949.1	non-structuralpolyprotein [Israelacuteparalysisvirusofbees]	IGFLNRLYRCGDWDGILLLLSG	DIEE	NPGP
ACD01403.1	polymerasepolyprotein [Israelacuteparalysisvirusofbees]	MGFLNKLYRCGDWDSILLLLSG	DIEE	NPGP
ACD01401.1	polymerasepolyprotein [Israelacuteparalysisvirusofbees]	IGFLNKLCQCGDWDSILLLLSG	DIEE	NPGP
YP_001040002.1	polymerasepolyprotein [Israelacuteparalysisvirusofbees]	IGFLNKLYRCGDWDSILLLLSG	DIEE	NPGP
ACT66677.1	polyprotein [Porcineteschovirus2]	MTTMTLQGPGATNFSLLKQAG	DIEE	NPGP
ACT66679.1	polyprotein [Porcineteschovirus4]	MTAMFLQGPGATNFSLLKQAG	DIEE	NPGP
AAR31746.1	polyprotein [Porcineteschovirus2]	MTTMTLQGPGATNFSLLKQAG	DIEE	NPGP
AAR31745.1	polyprotein [Porcineteschovirus2]	MTTMTLQGPGATNFSLLKQAG	DIEE	NPGP
AAK12392.1	AF296098_1polyprotein [Porcineteschovirus1]	MTVVTYQGPGATNFSLLKQAG	DIEE	NPGP
AAR31763.1	polyprotein [Porcineteschovirus1]	MTVMTFQGPGATNFSLLKQAG	DIEE	NPGP
AAR31766.1	polyprotein [Porcineteschovirus1]	MTVMTFQGPGATNFSLLKQAG	DIEE	NPGP
ACT66682.1	polyprotein [Porcineteschovirus5]	MTTLTFQGPGATNFSLLRQAG	DIEE	NPGP
CAP58274.1	polyprotein [Saffoldvirus]	FTDFFKAVRDYHASYYKQRLQH	DIEA	NPGP
AEE69077.1	polyprotein [Saffoldvirus]	FTDFFKAVRDYHASYYKQRLQH	DIEA	NPGP
YP_003620399.1	p81 [Providence virus]	TLMGNIMTLAGSGGRGSLLTAG	DVEK	NPGP
YP_003620399.1	p81 [Providence virus]	NSDDEEPEYPRGDPIEDLTDDG	DIEK	NPGP
AAF09193.1	polyprotein [Foot-and-mouth disease virus (strain O1)]	HKHKIVAPVKQLLNFDLLKLAG	DMES	NPGP
BAJ14095.1	outer capsid protein VP4 [Porcine rotavirus A]	LTNSYTTDLSDEIKEIGSSKSQ	DVTI	NPGP
BAJ14093.1	outer capsid protein VP4 [Porcine rotavirus A]	LTNSYTTDLSDEIKEIGSSKSQ	DVTI	NPGP
BAJ14091.1	outer capsid protein VP4 [Porcine rotavirus A]	LTNSYMTDLSDEIEEIGSSKSQ	DVTI	NPGP
AAY46581.1	outer capsid protein VP4 [Porcine rotavirus]	LTNSYTTDLFDEIKEIGSSKSQ	DVTI	NPGP
AAY46580.1	outer capsid protein VP4 [Porcine rotavirus]	LTNSYTTDLFDEIKEIGSSKSQ	DVTI	NPGP
ABF83605.1	outer capsid protein VP4 [Porcine rotavirus]	LTNSYTTDLSEEIKEIGSSKSQ	DVTI	NPGP
ABH03539.1	outer capsid protein VP4 [Porcine rotavirus]	LTNSYTTDLSEEIKEIGSSKSQ	DVTI	NPGP

Appendix B

Q8JNB1.1	VP4_ROT46 Outer capsid protein	LTNSYTTDLSDEIEEIGSSKSQ	DVTI	NPGP
AEL20323.1	structural protein VP4 [Porcine rotavirus A]	LTNSYTVNLSDEIQDIGSAKSQ	DVTI	NPGP
AEF01496.1	VP4 [Giant panda rotavirus A]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ADN93327.1	VP4 [Bovine rotavirus A]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ADN93326.1	VP4 [Bovine rotavirus A]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ADN93325.1	VP4 [Bovine rotavirus A]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACS27082.1	VP4 outer capsid protein [Porcine rotavirus A]	LNNSYTTDLSDEINEIGSLKSQ	DVTI	NPGP
ACG50676.1	outer capsid protein VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
AAS02091.1	outer capsid spike VP4 protein [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
AAZ38156.1	VP3 [Porcine rotavirus A]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
P11114.1	VP4_ROTP5	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
BAA03845.1	VP4 protein [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACN22281.1	VP4 [Rotavirus A]	LANSYTSELQDTIDDISAQKSQ	DVTI	NPGP
ACC91713.1	VP4 [Rotavirus A]	LANSYTSELQDTIDDISAQKSQ	DVTI	NPGP
ACC91712.1	VP4 [Rotavirus A]	LANSYTSELQDTIDDISAQKSQ	DVTI	NPGP
ACC91711.1	VP4 [Rotavirus A]	LANSYTSELQDTIDDISAQKSQ	DVTI	NPGP
ACC91708.1	VP4 [Rotavirus A]	LANSYTSELQDTIDDISAQKSQ	DVTI	NPGP
ACI31966.1	VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACI31940.1	VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56214.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56212.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56211.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56210.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56208.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56207.1	VP4 [Porcine rotavirus]	LPNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56204.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56203.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56202.1	VP4 [Porcine rotavirus]	LSNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56201.1	VP4 [Porcine rotavirus]	LPNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56199.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56188.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQDIGSAKSQ	DVTI	NPGP
ACO56186.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56185.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56184.1	VP4 [Porcine rotavirus]	LTNSTQVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56171.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56168.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56163.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56154.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56153.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56152.1	VP4 [Porcine rotavirus]	LTNSHTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56151.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56149.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56148.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56147.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP

Appendix B

ACO56141.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56139.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56138.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56137.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56135.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56134.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQDIGSAKSQ	DVTI	NPGP
ACO56133.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56132.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56131.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56129.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQDIGSAKSQ	DVTI	NPGP
ACO56128.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQDIGSVKSQ	DVTI	NPGP
ACO56127.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56126.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56125.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56124.1	VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACB71152.1	VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACB71147.1	VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACO56195.1	truncated VP4 [Porcine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACI31965.1	VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACI31964.1	VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACI31963.1	VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACI31962.1	VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACI31959.1	VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACI31957.1	VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACI31956.1	VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACI31955.1	VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACI31953.1	VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACI31951.1	VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACI31950.1	VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACI31948.1	VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACI31947.1	VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACI31946.1	VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACI31945.1	VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACI31944.1	VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACI31943.1	VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACI31938.1	VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACI31937.1	VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ACI31936.1	VP4 [Bovine rotavirus]	LTNSYTVNLSDEIQEIGSAKSQ	DVTI	NPGP
ADN03191.1	VP4 [Rotavirus pig/2B/IRL/2005/P [13]/ [22]]	LTNSYTTDLSDEIEEIGSSKSQ	DVTI	NPGP
ADN03190.1	VP4 [Rotavirus pig/48/IRL/2006/P [13]]	LTNSYTTDLYDEIEEIGSQKSQ	DVTI	NPGP
ADN03189.1	VP4 [Rotavirus pig/1/IRL/2007/P [13]]	LTNSYTTDLYDEIEEIGSQKSQ	DVTI	NPGP
CAQ76897.1	outer capsid protein [Rotavirus A Hu/6784/2000/ARN/Cameroon]	NTSNHTPVNLLTKSRDWIDKSQ	DVTI	NPGP
ACZ06085.1	outer capsid protein VP4 [Porcine rotavirus]	LTNSYTVNLPDEIQEIGSAKSQ	DVTI	NPGP
AAM95775.1	outer capsid protein VP4 [Porcine rotavirus A]	FYTTDLSDEIEEIGSSKSQ	DVTI	NPGP

dd

Appendix B

AAL31530.1	AF427125_1 outer capsid protein VP4 [Porcine rotavirus A]	LTNSYTVNLSDEIQDIGSAKSQ	DVTI	NPGP
BAD90838.1	outer capsid protein VP4 [Porcine rotavirus A]	TDLFDEIEEIGSSKSQ	DVTI	NPGP
ABY60445.1	outer capsid protein [Rotavirus A pig/Slovenia/P83/G5P [6]]	VELSDEIKTIGSKKNQ	DVTI	NPGP
AAA79027.1	outer capsid protein [Bovine rotavirus 993/83]	LANSYTSELQDTIDDISVQKSQ	DVTI	NPGP
BAD22593.1	capsid protein [Porcine rotavirus]	NLSDEIQDIGSAKSQ	DVTI	NPGP
BAD22592.1	capsid protein [Porcine rotavirus]	LSDEIQEIGSAKSQ	DVTI	NPGP
BAK08662.1	outer capsid protein VP4 [Porcine rotavirus A]	EIEEIGSSKSQ	DVTI	NPGP
BAK08661.1	outer capsid protein VP4 [Porcine rotavirus A]	EIEEIGSLKSQ	DVTI	NPGP
ADQ89898.1	outer capsid protein [Porcine rotavirus]	LTNSYTTDLSEEIKEIGSSKSQ	DVTI	NPGP
AAM95774.1	outer capsid protein VP4 [Porcine rotavirus A]	EIEEIGSSKSQ	DVTI	NPGP
BAK08660.1	outer capsid protein VP4 [Porcine rotavirus A]	EIQEIGSAKSQ	DVTI	NPGP
BAK08659.1	outer capsid protein VP4 [Porcine rotavirus A]	EIQEIGSAKSQ	DVTI	NPGP
AAX13501.1	outer capsid structural protein VP4 [Porcine rotavirus A]	SLFDEIKEIGSSKSQ	DVTI	NPGP
ADM15521.1	VP4 [Rotavirus A Po/CE-M-06-0007/Canada/2006/G11P [13]I5E9]	EIGSLKSQ	DVTI	NPGP
ABH03540.1	outer capsid protein VP4 [Porcine rotavirus A]	SR	DVTI	NPGP
BAJ13370.1	outer capsid protein VP4 [Rotavirus A]	LTNSYTTDLSDEIDEIGSSKSQ	DVTV	NPGP
ACT53279.1	outer capsid spike VP4 protein [Bovine rotavirus]	LTNSYTVELSDEIQEIGSTKTQ	DVTV	NPGP
AAZ04534.1	VP4 outer capsid protein [Porcine rotavirus A strain 134/04-15]	LTNSYTTDLSDEIEEIGSSKSQ	DVTV	NPGP
Q96802.1	VP4_ROTRF capsid proteins Hemagglutinin	LTNSYTVELSDEIQEIGSTKTQ	DVTV	NPGP
BAC85485.1	structural protein 4 [Bovine rotavirus]	LTNSYTVELSDEIQEIGSTKTQ	DVTV	NPGP
AEO45636.1	VP4 [Human rotavirus A]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ADV52474.1	outer capsid protein VP4 [Human rotavirus A]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ADO78483.1	outer capsid protein VP4 [Rotavirus A]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ADI59479.1	outer capsid protein [Human rotavirus A]	LTNSYSVDLYDEIEQIGSEKNQ	DVTV	NPGP
ADI59478.1	outer capsid protein [Human rotavirus A]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ADI59477.1	outer capsid protein [Human rotavirus A]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ACL80633.1	outer capsid protein [Human rotavirus A]	LTNSYSVDLYEEIEQIGSEKTQ	DVTV	NPGP
P17465.1	VP4_ROTBN outer capsid protein Hemagglutinin	LTNSYTVELSDEIQEIGSTKTQ	DVTV	NPGP
BAJ19470.2	outer capsid protein VP4 [Rotavirus A Hu/NhaTrang/V20/2006/VNM]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
AEE69102.1	outer capsid protein VP4 [Human rotavirus A]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ADV52473.1	outer capsid protein VP4 [Human rotavirus A]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ADT82720.1	outer capsid protein VP4 [Human rotavirus A]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ADT82719.1	outer capsid protein VP4 [Human rotavirus A]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ADT82718.1	outer capsid protein VP4 [Human rotavirus A]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ADB81101.2	outer capsid protein VP4 [Human rotavirus A]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ACV83293.3	outer capsid protein VP4 [Human rotavirus A]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ADT82721.1	outer capsid protein VP4 [Human rotavirus A]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ABD60996.1	capsid protein [Human rotavirus A]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ABD60997.1	capsid protein [Human rotavirus A]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ADE43130.1	major outer capsid protein VP4 [Rotavirus human/BJ-CR5317/China/2008/G9P [8]b]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
CAC21839.1	outer capsid protein [Rotavirus A]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
AEE69096.1	outer capsid protein VP4 [Human rotavirus A]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ACR22893.1	outer capsid protein VP4 [Human rotavirus A]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP

Appendix B

ACJ65724.1	outer capsid protein VP8* subunit [Human rotavirus A]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ADE44947.1	outer capsid protein [Human rotavirus P [8]]	LPNSYSGDLYAEIEQIGSEKTQ	DVTV	NPGP
ADN03188.1	VP4 [Rotavirus pig/2F/IRL/2005/P [26]]	LTNSYTTDLSDEIEEIGSSKSQ	DVTV	NPGP
ADD84302.1	protease-sensitive outer capsid protein [Human rotavirus A]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ADE44949.1	outer capsid protein [Human rotavirus P [8]]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
BAJ19472.1	outer capsid protein VP4 [Rotavirus A Hu/NhaTrang/V30/2006/VNM]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ABW03042.1	non-glycosylated outer capsid protein VP4 [Human rotavirus A]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ABW03039.1	non-glycosylated outer capsid protein VP4 [Human rotavirus A]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ABB84940.1	VP4 [Rotavirus A]	LSSSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ABB84927.1	VP4 [Rotavirus A]	SVDLYDEIEQIGSEKTQ	DVTV	NPGP
AEE69100.1	outer capsid protein VP4 [Human rotavirus A]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ABW03043.1	non-glycosylated outer capsid protein VP4 [Human rotavirus A]	LPNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ABS29578.1	VP4 [Human rotavirus A]	Q	DVTV	NPGP
ADR30152.1	protease-sensitive outer capsid protein [Human rotavirus A]	LTNSYSVDLYDEIEQIGSEKTQ	DVTV	NPGP
ABS29577.1	VP4 [Human rotavirus A]	EQIGSQKTQ	DVTV	NPGP
ACU82759.1	major outer capsid protein [Human rotavirus A]	NSYSVDLHDEIEQIGSEKTQ	DVTV	NPGP

‘Whenever I have found out that I have blundered, or that my work has been imperfect, and when I have been contemptuously criticised, and even when I have been overpraised, so that I have felt mortified, it has been my greatest comfort to say hundreds of times to myself that *"I have worked as hard and as well as I could, and no man can do more than this"*.’

Autobiography - Charles Darwin, 1876

‘With magic, you can turn a frog into a prince. With science, you can turn a frog into a PhD and you still have the frog that you started with.’

The Science of Discworld - Sir Terry Pratchett, Ian Stewart and Jack Cohen, 1999

~ THE END ~