Structure-function studies of the ORF1 protein from the insertion-site specific retroposon M5 found in Indo-Pakistan urban malarial vector *Anopheles stephensi*.

Daniel Oluwatosin Akinbosede

Supervised by
Dr Pryank Patel and Dr Colin Malcolm

Submitted to the University of Hertfordshire in partial fulfilment of the requirements of the degree of Master of Science in Biotechnology

School of Life and Medical Sciences

University of Hertfordshire

United Kingdom

June 2018

DECLARATION

I declare that:

(a) All the work described in this report have carried out by me – and all the results (including any survey findings, etc.) given herein were first obtained by me – except where I may have given due acknowledgement to others;

(b) All the prose in this report have written by me in my own words, except where I may have given due acknowledgement to others and used quotation marks, and except also for occasional brief phrases of no special significance which may be taken from other people's work without such acknowledgement and use of quotation marks;

(c) All the figures and diagrams in this report have been devised and produced by me, except where I may have given due acknowledgement to others.

I understand that if I have not complied with the above statements, I may be deemed to have failed the project assessment, and/or I may have some other penalty imposed up-on me by the Board of Examiners.

Signed …… Date……...

Name: Daniel Oluwatosin Akinbosede.

ABSTRACT

In the 1950s Barbara McClintock inferred the occurrence of transposition: the movement of small segments of DNA - entities known as transposable elements from one position of the genome to another (McClintock, 1950). Classification of transposable elements in regards to mechanism of transposition distinguishes them into two groups; transposons (Class II) and retroposons (Class I). The term retrotransposon was coined as it illustrates the transposition of these elements is dependent on the reverse transcription of RNA to DNA through a reverse transcriptase, also known as the 'copy and paste' transposition. The M5 retroposon has been found in numerous mosquito species such as *Anopheles stephensi.* M5 present in these *Anopheles* is a class 1, non-LTR transposable element of the jockey clade family with two open reading frames (ORF). Due to its APE like endonuclease, M5 should transpose to random sites of the genome. However, in *A. stephensi* the element has been reported to transpose with site specificity. The aim of the project was to gain structural and functional information on the role of the ORF1 protein of M5 in order to understand the element's site specificity.

To perform functional and structural studies, an *Escherichia coli* expression vector was designed with a synthetic AsM5 ORF1 insert. Heterologous expression and purification of ORF1p in *E.coli* produced signs degradation or very low yield of the unfolded protein possibly due to the host's inability to process some eukaryotic features required for the protein. *Saccharomyces cerevisiae* was then chosen as an expression system for AsM5 ORF1p production. The AsM5 ORF1 gene was cloned from the *E.coli* expression vector into the pYES2/CT *S.cerevisiae* expression vector and sequencing verified that the AsM5 ORF1 insert was successfully cloned into the

pYES2/CT vector. Optimization of the lysis and expression protocol in *S.cerevisiae* had slowed progress but a highly effective method of cell lysis was developed. Expression of the full length ORF1p in *S.cerevisiae* was not confirmed and difficulties in expression could be attributed to the fact that the original synthetic ORF1 sequence which was cloned is codon optimised for expression in *E.coli* hindering expression in *S.cerevisiae*.

CPSF100_C was one of the conserved domains identified in the AsM5 ORF1 amino acid sequence using conserved domain web tools. For further analysis the CPSF100_C domain was cloned into an *E.coli* vector and successfully expressed in rich media, the protein was the purified using immobilized metal ion affinity chromatography (IMAC) and ion exchange chromatography (IEC). In order to progress to NMR studies of the domain, ^{15}N labelled expression of CPSF100_Cp was performed. Usually, several expression showed the non-peptide fusion partner glutathione *S*-transferase (GST) being expressed without the protein of interest possibly due to mRNA instability when the gene is expressed in minimal media. The identification of protein domains such as CPSF100_C and their interactions with nucleic acids and other proteins will likely be the key to understanding AsM5's site specific retrotransposition.

Contents

LIST OF ABREVIATIONS

APE - Apurinic/apyrimidinic endonuclease

AsM5 – *Anopheles stephensi* M5

Cam – Chloramphenicol

CHO – Chinese hamster ovary cells

DNA – Deoxyribonucleic acid

DTT – Dithiothreitol

ECL – Enhanced Chemiluminescence

GAG – Group specific antigen

GST – Glutathione *S*-transferase

HMM – Hidden markov model

IEX – Ion exchange chromatography

IMAC – Ion affinity chromatography

IPTG - Isopropyl β-D-1-thiogalactopyranoside

Kan – Kanamycin

LB – Luria Bertani

LINE – Long Interspersed Nuclear Elements

LTR – Long terminal repeat

MES – Morpholinoethansulfonic acid monohydrate

NEB - New England Biolabs

NMR – Nuclear magnetic resonance

OD – Optical density

ORF – Open reading frame

ORFp – Open reading frame protein

PCR – Polymerase chain reaction

POI – Protein of interest

PPM – Parts per million

RNA – Ribonucleic acid

SDS - Sodium dodecyl sulfate

SDS PAGE - Sodium dodecyl sulphate polyacrylamide gel electrophoresis

SINE – Short Interspersed Nuclear Elements

SMART – Simple modular architecture research tool

SOC - Super Optimal broth with Catabolite repression

TIR- Terminal inverted Repeats

Y-PER™ - Yeast protein extraction reagent

LIST OF FIGURES

Chapter 1

GENERAL INTRODUCTION

1.1 - TRANSPOSABLE ELEMENTS

In the 1950s Barbara McClintock inferred the occurrence of transposition (McClintock, 1950): the movement of small segments of DNA - entities known as transposable elements from one position of the genome to another (Hartwell, Hood, Goldberg, Reynolds, & Silver, 2010). Transposable elements range from 50bp to 10kb in size. They are present in both prokaryotic and eukaryotic organisms and make up 50% of the human genome (Flutre, Permal, & Quesneville, 2012). Often branded as 'selfish DNA' due to their ability to replicate themselves whilst making no notable contribution to their host; transposable elements have been studied in a large number of model organisms and it is well understood that they have had a profound effect in the shaping of eukaryotic genomes (Malik & Eickbush, 2000). The literature shows that there are several ways in which the activity of transposable elements can have an impact on a genome in both positive and negative ways. For example, the movement of a transposable element can inactivate genes, alter the expression levels of genes or induce potentially dangerous illegitimate recombination (Muñoz-López & García-Pérez, 2010). This is all made possible by the different specific mechanisms of transposition.

The transfer of transposable elements is usually vertical which describes the transfer of genetic material from parent to progeny. However, many have speculated that the detection of horizontal transfer which is the transfer of genetic material between unrelated individuals, is important to understanding the origin and spread of transposable elements and in assessing their impact on genetic diversity (Crainey, Garvey, & Malcolm, 2005). The speculation arose because horizontal transfer has long been recognized as a crucial mechanism driving bacterial evolution. Though it is well characterised in bacteria, the evolutionary importance of horizontal transfer in multicellular eukaryotes is still poorly understood (Schaack, Gilbert, & Feschotte, 2010). A vast amount of research such as the work done in 'Massive horizontal transfer of

transposable elements in insects' by Peccoud, Loiseau, Cordaux & Gilbert, 2017 has been done to study the horizontal transfer of transposable elements in eukaryotes and the consensus on the topic is ever changing.

1.2 - CLASSIFICATION OF TRANSPOSABLE ELEMENTS AND MECHANISMS OF TRANSPOSITION

Classification of transposable elements in regards to mechanism of transposition distinguishes them into two groups; transposons (Class II) and retrotransposons (Class I). The mechanism utilised by transposons is referred to as 'cut and paste transposition'; where the transposable element is cut from one site in a chromosome and pasted into a new site (Snustad & Simmons, 2003). Transposons consist of a transposase gene that is flanked by two Terminal Inverted Repeats (TIRs). The transposase recognises these TIRs to perform the excision of the transposon, which is inserted into a new genomic location such as a different chromosome (Muñoz-López & García-Pérez, 2010). There have been several reported instances of horizontal transfer of transposons across phyla such as PiggyBac, a well-known transposon that utilises this method by recognising TTAA chromosomal sites. (Schaack, Gilbert, & Feschotte, 2010). The element transposase recognises specific inverted terminal repeat sequences (ITRs) on both ends of the element before cutting the sequence ready integrate into TTAA chromosomal sites.

The term retrotransposon was coined as it illustrates the point that transposition of these elements is dependent on the reverse transcription of RNA to DNA. The transposition of a retrotransposon, also known as the 'copy and paste' transposition begins with its transcription by RNA polymerase into an RNA that encodes a reverse transcriptase – like enzyme (Hartwell, Hood, Goldberg, Reynolds, & Silver, 2010) which copies RNA into single stranded cDNA; this is then used as a template to produce double stranded cDNA. Figure 1 shows that in addition to the Class I and Class II groups, there are further subdivisions in the world of transposable elements. Class I elements consist of Short Interspersed Nuclear Elements (SINEs), Long

Interspersed Nuclear Elements (LINEs) and Long Terminal Repeat (LTR) Retrotransposons. LTR retrotransposons have LTRs that vary from 100 base pairs to 5 kb and are similar in structure and life cycle to retroviruses (Muñoz-López & García-Pérez, 2010). They transpose by synthesising a double-stranded DNA intermediate, using the element's reverse transcriptase and RNA as a template. The completed complementary DNA is then introduced into the host chromosome via a recombination event involving an associated integrase or recombinase (Beauregard, Curcio, & Belfort, 2008). In plant genomes, LTR retrotransposons are the most populous repetitive sequence, for example there are able to make up 75% of the maize genome (Baucom et al., 2009).

SINEs are a sub-category by virtue of their reliance on LINE encoded proteins. LINEs, retroposons and non LTR Retrotransposons are interchangeable terms and they can be found in high copy number and are widespread in eukaryotic genomes (Yadav, Mandal, Rao, & Bhattacharya, 2009). LINEs are often several kilobases long and contain two open reading frames (ORFs) encoding a group specific antigen (gag) protein (ORF1), an endonuclease and reverse transcriptase domains (ORF2) together enabling the element for autonomous retrotransposition (Schmidt, 1999), which is illustrated in figure 1. This thesis refers to protein domains as distinct structural or potentially functional units of a protein (Oh & Yi, 2016).

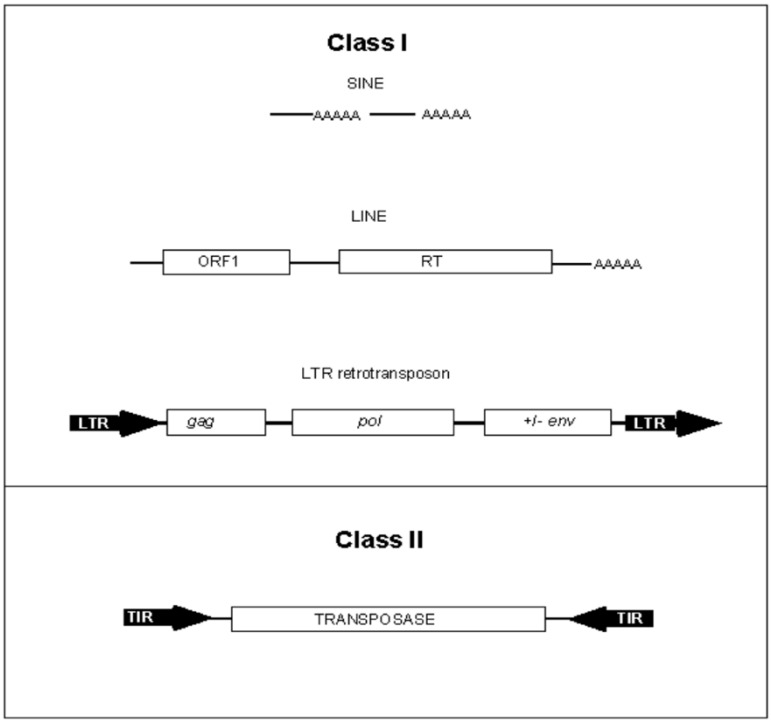

Figure 1: An Image depicting the subdivisions of transposable elements and their composition (Bowen & Jordan, 2002). The figure shows the two main classes (I & II) of transposable elements and their subdivisions. Particular interest goes to the structure of the LINE class I element which is shown to in encode an open reading frame and reverse transcriptase domain followed by poly A tail.

SINEs are similar to LINEs, but are shorter (<500 bases), simpler, and almost undoubtedly dependent on LINE reverse transcriptase and endonuclease functions for retrotransposition (Weiner, 2002). In some cases, they have been seen to have their own endonucleases that would allow them to cleave their way into a genome, however the majority of SINEs integrate at chromosomal breaks (Muñoz-López & García-Pérez, 2010). *Alu* elements are examples of SINEs and are the most abundant in the human genome; present in more than one million copies, which altogether represent 10% of the whole genome mass (Häsler & Strub, 2006). As they are often found in non-coding DNA, mutations in Alu elements are usually of no consequence. However, some Alu elements are involved in translational regulation. BRCA1, a DNA repair protein whose mutation is associated with breast cancer is likely the best-

characterised example of translation regulation by an Alu element. The 80 kb genomic sequence of this gene is composed at 40% of *Alu* elements' (Häsler & Strub, 2006), resulting in a high risk of mutation. In addition to this *Alu* elements have been shown to be involved in RNA editing and alternative splicing (Häsler & Strub, 2006) and more recently suggested as novel regulators of gene expression in type 1 diabetes susceptibility genes (Kaur & Pociot, 2015).

In depth research into the origin and phylogeny of transposable elements has resulted in very refined classification of some groups of transposable elements; this is especially true for non-LTR retrotransposons. Malik, Burke, & Eickbush (1999) conducted analysis of non-LTR retrotransposons based on an extended sequence alignment of their reverse transcriptase domain. They found that all identified non-LTR elements could be grouped into 11 distinct clades that each date back to before the divergence of the major animal phyla (Malik, Burke, & Eickbush, 1999). Clades are generally known as a group of organisms that are identifiable with similar structural features that possess a single ancestor (Malik, Burke, & Eickbush, 1999). The clades were named after the earliest determined member within the family (Lovsin, Gubensek, & Kordis, 2001). The resulting names were as follows L1, RTE, Tad1, R1, R2, R4, LOA, I, CR1, CRE and Jockey. Figure 2 shows how these families are related and outlines their basic structure.

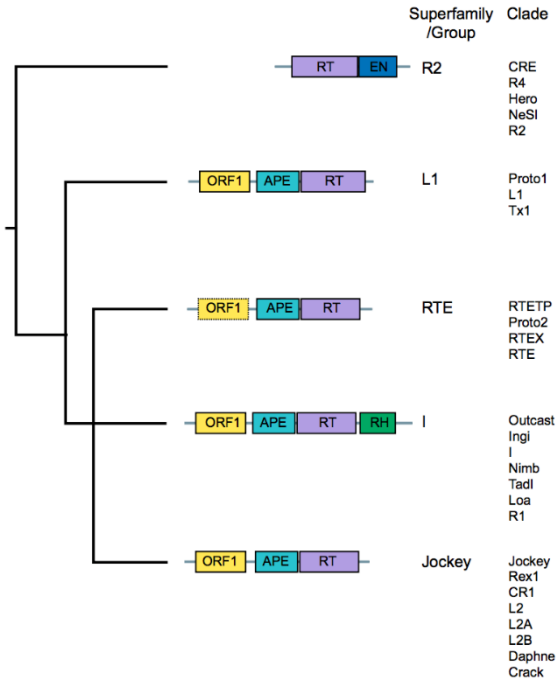

Figure 2: A phylogenetic tree of the R2, L1, RTE, I and jockey families, their basic outline and their subdivisions (Metcalfe & Casane, 2014). The diagram shows elements from the L1, RTE and jockey clades to have identical structures. The ORF2 of these elements are depicted as a apurinic/apyrimidic endonucleases and reverse transcriptase coding sequence

Elements of the jockey clade are examples of LINEs due to their reverse transcriptase domain and based on their structure they can be divided into two groups. The first group has a single open reading frame (ORF) that encodes RT in the middle and a restriction endonuclease near its C-terminus. The second group has two ORFs: ORF1 and ORF2; the latter encodes two domains responsible for retrotransposition: apurinic/apyrimidinic endonuclease (APE-like endonuclease) domain at the N-terminus and reverse transcriptase domain in the middle (Novikova et al., 2007). The APE like endonucleases are known to be involved in the general mechanisms of DNA repair in both eukaryotes and prokaryotes (Malik, Burke, & Eickbush, 1999). A study showed the phylogeny of the AP-endonucleases (APE) agrees with the phylogeny of the reverse transcriptase in the elements. The APE like endonuclease phylogeny indicates that among the eight non-LTR clades containing this endonuclease, the L1 clade is

the oldest, followed by the RTE clade. The acquisition of these endonuclease by the non-LTR lineage from a host repair machinery appears to be of ancient origin because it is not possible to resolve whether this domain was obtained from a prokaryotic or a eukaryotic source (Malik, Burke, & Eickbush, 1999). This particular study concluded that the acquisition of AP endonucleases by non-LTR retrotransposons was a single event that occurred early in the evolution of eukaryotes.

1.3 - JOCKEY CLADE AND THE M5 RETROPOSON

The jockey clade is a very large bracket with several smaller more specific element groups, their structure is similar to that of the R1 and L1 clades. Movement of elements such as these can significantly increase the size of an organism's genome. The jockey family is represented by several subfamilies of elements in *Drosophila* but also in mosquitos like *Anopheles gambiae*. Transposable elements in the jockey clade, representing eight mosquito species, were examined by Crainey,*et al* (2005) and were found to be made up of three monophyletic groups of sub-elements JM1, JM2, and JM3. Horizontal transfer of retrotransposons including jockey clade elements is not fully understood. Crainey, et al concluded that there was no evidence for horizontal transfer events after analysing a large data set.

The M5 retrotransposon has been found in numerous mosquito species such as *A. gambae*, *Anopheles sinensis, Anopheles stephensi and Anopheles maculatus.* Due to its APE like endonuclease, M5 should transpose to random sites of the genome as only two of more than 20 APE-encoding clades such as Tx1 and R1 have been shown to have site specificity (Fujiwara, 2015). However, in *A. stephensi* work performed by Adams (2015) has reported the M5 element to transpose with site specificity (Adams, 2015). This is made even more unusual by the fact that M2, a close relative of M5 also has an APE like domain but does not exhibit site specificity. As can be seen in figure 2, the jockey clade element is very similar in structure to L1 and RTE elements. As such the ORF2 of M5 is similar to those mentioned above and is

extremely well characterised, boasting two functional proteins RT and APE, which are both well conserved across many of the non-LTR elements. ORF1 on the other hand is not as well conserved hence the inability to label the role or function the resulting protein would have on the element. This alone makes these ORF1s worth studying in detail as their characterisation would offer answers to many of the unanswered questions surrounding the retrotransposition of the element.

1.4 – ORF1 PROTEIN

To date, there is no published literature regarding the structure of AsM5. As with all similar elements, the protein encoded by ORF1 in AsM5 is particularly poorly understood, though figure 3 does present a schematic that identifies three conserved zinc finger motifs. In this thesis, motifs are discussed as small conserved portions of a protein that can be used to determine structure but can have a variety of functions. Motifs can either be of conserved sequence or structure (Chiang, Gelfand, Kister & Gelfand, 2007). L1 is arguably the best understood LINE element and, as can be seen in figure 2, has a very similar organisation to that of jockey clade elements. The two known functions of the protein encoded by ORF2, the APE and RT, were readily predicted based upon sequence homology and general phylogenetic analysis.

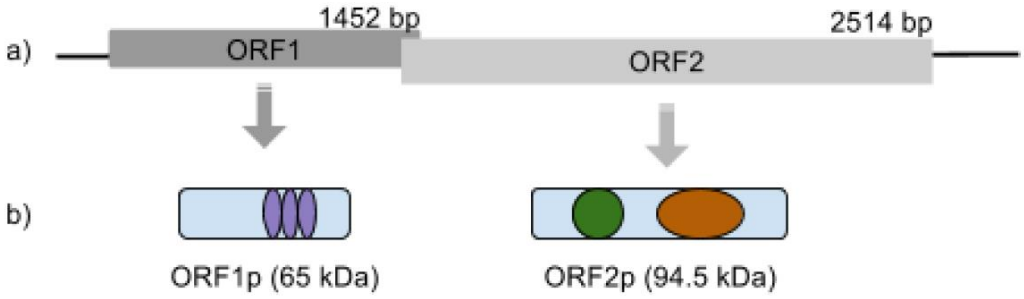

Figure 3: A diagram of the AsM5 retroposon ORFs depicting its encoded proteins. In its two ORFs AsM5 has some highly conserved protein domains including three Zinc fingers (purple) in ORF1p and the RT (green) and APE (orange) in ORF2 (Ramón 2016).

Sequence/function comparisons have so far provided very little regarding the function of the ORF1 protein, even in an element as well studied as L1. Several expression studies along with *in-vitro* and *in-vivo* experiments have added more knowledge and L1 ORF1p has been reported to be essential for the retrotransposition of the element. It has been demonstrated that mutations in ORF1p halted retrotranspotition in both mouse and human L1 (Martin, 2006), a study yet to be replicated in M5. The experiments showed that there was no retrotransposition reported in human L1 mutants when two amino acid residues in the protein were replaced with a stop codon in separate experiments. In both of these cases, the frequency of retrotransposition was less than 0.06% of the original wild-type element (Martin, 2006). Research also shows that L1 binds nucleic acids and the coiled coil domain on the protein is largely responsible for this, though this domain is not often identified when the AsM5 amino sequence is put through conserved protein domain searches. Nonetheless the same review that highlights this also warns that the powerful affinities observed with ORF1p purified from the insoluble fraction of *E.coli* heterologous expression were not as strong when the same assay was done using protein purified from the soluble fraction, possibly due to the fact that soluble protein does not need refolding and is less susceptible to denaturation and thus more likely to function as normal (Kolosha and Martin, 1997).

In addition to the documented interaction of L1 ORF1p with nucleic acids, there is also the matter of its interaction with other proteins most notably, chaperone proteins. Chaperones are generally multifunctional proteins found in all eukaryotic organisms. In this case the specific chaperones discussed are proteins that prevent incorrect interactions between histones and nucleic acids and in turn prevent the formation of nucleosomes (Ransom, Dennehey and Tyler, 2010). The L1 OF1p contains zinc-finger motifs resembling those of AsM5, a group of protein motifs that bear some similarity to retroviral gag proteins, which play a crucial role in retroviral replication, a possible reason for their evolutionary retention in retrotransposons. It is a certainty that the chaperone activity of ORF1p in L1 is necessary for retrotransposition. L1 ORF1p's chaperone activity is vital, it was demonstrated that single point mutations which remove ORF1p's ability to interact with chaperone proteins also destroys retrotransposition activity, even if the point mutations had no effect on RNA or single stranded DNA binding affinity (Martin et al., 2005). Both *in vitro* and *in vivo* work has confirmed that the protein binds both RNA and DNA, with a higher affinity for single-stranded than double-stranded nucleic acids. In addition to this, nucleic acid chaperone activity of the protein likely contributes more directly to reverse transcription of the entire element than previously thought (Martin, 2006).

1.5 – AIMS AND OBJECTIVES

It is worth noting that site specific elements usually target the same sequences in different repetitive units within the genome (Fujiwara, 2015) and unpublished work carried out by Adams 2015 on AsM5 suggests that the element might be targeting histone gene clusters in transposition (Adams, 2015). It has often been implied that transposable elements could be used as a target to control a population. A large proportion of research on the topic is on ways to exploit their presence, by using them to manipulate genomes of organisms that transmit tropical diseases such a malaria (Muñoz-López & García-Pérez, 2010). In some site-specific

elements, evidence has been presented that suggests ORF1p could be involved in gaining access to the target genomic site into which its element is pasted (Fujiwara, 2015).

The aim of this project was to gain better understanding the role of AsM5 ORF1 protein in the retrotransposition of the element. This aim was approached by using protein prediction software and conserved domain databases to identify and study any functional domains within AsM5 ORF1p that makes it unique when compared with its closest relatives. Prokaryotic and eukaryotic heterologous systems were used to express AsM5 ORF1 protein by creating constructs containing the gene of interest. The protein would then be purified for further structural and functional studies.

The specific objectives for the project were as listed:

- An alignment of ORF1p sequences closely related to M5 ORF1p derived from available whole genome sequencing projects and supplemented by PCR and DNA sequencing.
- Locations of domain boundaries predicted and confirmed by computational and sequence analysis of M5 related ORF1ps.
- Structural predictions of the protein region of ORF1ps, to identify as yet unknown functional elements in low-complexity regions.
- Expression and purification of ORF1 with confirmed functionality using E. coli or S. cerevisiae expression vectors.
- Using experimental work to support structural studies of the M5 retroposon ORF1p.

Chapter 2

BIOINFORMATIC ANALYSIS

2.1 – INTRODUCTION

Studying proteins using bioinformatic analysis of their sequences has become an indispensable part of the biotechnology field. When performing analysis of a gene and its function, there are several reasons to choose amino acid sequences rather than DNA sequences. These include the much larger alphabet of amino acids (20 amino acids versus 4 bases) and the lower signal-to-noise ratio in protein sequence searches. Arguably the most important feature is the closeness between a protein sequence and function. In addition to this, the availability of good, well annotated databases of protein sequences and protein sequence signatures are constantly improving the field (Derbyshire et al., 2015). Conventionally, the first step in protein analysis is to search databases for similar sequences. This usually indicates how well characterised the sequence or similar sequences are, though it is difficult to infer much about the protein from a single sequence. To better study the sequence, alignments are usually built to create a consensus for a protein family, or to identify conserved domains and motifs or highly conserved residues that may be important for function, for example in an active site (Mulder & Apweiler, 2001). Once similar sequences are identified it becomes possible to put small pieces of the story surrounding the protein together. High identity hits from a 'BLASTP' or protein blast search will usually help identify orthologous and paralogous sequences.

Unfortunately, jockey clade elements of *Anopheles* are not particularly well characterised so BLASTP searches show very little in this case. To characterise the ORF1 protein a different approach was taken, one studying the conserved domains and motifs within the protein rather than studies based on the entire protein. Protein domains are evolutionarily conserved sequences in proteins that frequently match structural and functional units of other proteins

across species(Fong & Marchler-Bauer, 2008). Protein domains come in families. A grouping of functional diversity, and a large number of clusters assembled by obvious sequence similarity, can be reduced to between several hundred and a few thousand domain superfamilies. The classification of a superfamily usually depends on how aggressively one group clusters together based on 3D-structural and/or functional similarities determined by structural and functional analysis of other proteins containing the same domains (Marchler-Bauer, 2004). The specific function and sometimes structure of a protein and its homologue usually depends on its combination of domains; two-thirds of prokaryotic proteins and 80% of eukaryotic proteins have more than one domain (Fong & Marchler-Bauer, 2008). This makes the identification of legitimate conserved domains in a protein one of the most important steps in determining its function.

One of the most popular and effective tools used to identify conserved domains is Simple Modular Architecture Research Tool or 'SMART' (Letunic, Doerks & Bork, 2015). SMART was originally a tool for identifying signalling domains but has since expanded. It works by performing multiple sequence alignments of representative family members. On this web based tool, there are more than '400 domain families found in signalling, extracellular and chromatin-associated proteins'. Phylogenetic origins, functional class, tertiary structures and functionally important residues are all taken into consideration when comprehensively annotating these domains (Schultz, 2000).

Pfam is another such web tool, a database of curated protein families. Each of these families is defined by a profile hidden Markov model (HMM) and at least two alignments. Profile HMMs are models used for the statistical searching of homologous sequences built from an aligned set of family-representative sequences. The current release of Pfam, version 27.0, contains 14 831 Pfam-A protein families (Finn et al., 2013). Pfam is arguably the most robust web tool of its kind due to its use of information from internationally established sources such as UniProt, SwissProt and CATH. Open-source web software has opened the field to developers

approaching the issues of identifying conserved domains from different perspectives. PROSITE consists of entries that describe protein domains, families and functional sites, as well as associated patterns to identify the domains. It is complemented by a collection rule based patterns, which allows more stringency of the patterns by offering additional information about functionally and/or structurally critical amino acids, though PROSITE heavily relies on the annotation of domains from UniProt and SwissProt database entries (Sigrist et al., 2009).

The literature suggests that each of these web based tools for identifying conserved domains have different strengths and weaknesses when considering the size of their databases and the accuracy at which they identify the domains. To make the most of all the best features from these tools, other web based tools have been created that work by combining several of the most popular tools in the sector such as Pfam and SMART. 'MOTIF Search' is such a tool (Kanehisa, 2002). Functioning using data from the database of Clusters of Orthologous Groups of proteins (COGs), PROSITE, SMART, Pfram and NCBI. Its ability to combine all the well-developed tools listed makes it transformative in the field.

In this chapter, these web tools were used to identify conserved domains within AsM5 ORF1p and compare them to those identified in the most closely related sequences in other *Anopheles* retroposons.

2.2 – METHODS

2.2.1 – AsM5, FROM DNA SEQUENCE TO AMINO ACID SEQUENCE.

The DNA nucleotide sequence for AsM5 obtained from Adams (2015) was translated to an amino acid sequence using the online ExPASy translate tool at https://web.expasy.org/translate/. Both the DNA and the amino acid sequence can be found in appendices 6.1 and 6.2 respectively (Adams 2015).

2.2.2 – BUILDING A PHYLOGENTIC TREE USING AsM5 RELATED SEQUENCES.

A series of ORF1 amino acid sequences from closely related retroposons found in *Anopheles, Aedes* and *Culex* mosquitos were used to create a phylogenetic tree on both http://www.trex.uqam.ca/ and https://www.ebi.ac.uk/Tools/msa/clustalo/ to allow for comparison. These sequences were obtained via a combination personal communication with Colin Malcom (Dr Colin Malcolm 2016, pers.comm) and the PhD thesis of Taif Adams (Adams 2015). It should also be noted that clustal omega (Sievers et al., 2014) is mainly a multiple sequence alignment tool, the tree is based on a simple neighbour joining method without distance correction. The tree built at http://www.trex.uqam.ca/ was created using a distance matrix method. The tree is obtained by using the method named 'Weighted least-squares method' and the distance-based methods is then 'polished using the procedure of quadratic approximation of its branch lengths' (Boc, Diallo & Makarenkov, 2012). The large phylogenetic tree of 38 sequences was then used to identify the most closely related sequences to AsM5 ORF1p. The sequences used for the creation of this tree can be found in appendix 6.3.

2.2.3 – ALIGNMENT OF THE MOST CLOSELY RELATED ORF1p SEQUENCES TO AsM5 ORF1p

Following the identification of the most closely related sequences to AsM5 ORF1p, the three sequences identified were aligned with AsM5 ORF1p to identify sites of conservation using the multiple sequence alignment tool on Clustal Omega at

https://www.ebi.ac.uk/Tools/msa/clustalo/. The sequence alignment was done using the default setting on the web link above. The most notable parameters are follows: Max Guide Tree Iterations = 1, Max HMM Iterations = 1 and number of (combined guide-tree/HMM) iterations = 0. The sequences used for this alignment can be found in the appendix 6.3.

2.2.4 – IDENTIFICATION OF CONSERVED DOMIANS USING WEBTOOLS.

SMART was used to identify conserved domains and motifs in AsM5 and the closely related sequences identified above at http://smart.embl-heidelberg.de/. This was used to compare the conserved domains and motifs across these sequences. In addition to the use of SMART, MOTIF Search was used also to identify conserved domains in these sequences at http://www.genome.jp/tools/motif/. When using MOTIF search the database parameters selected are shown in figure 4.

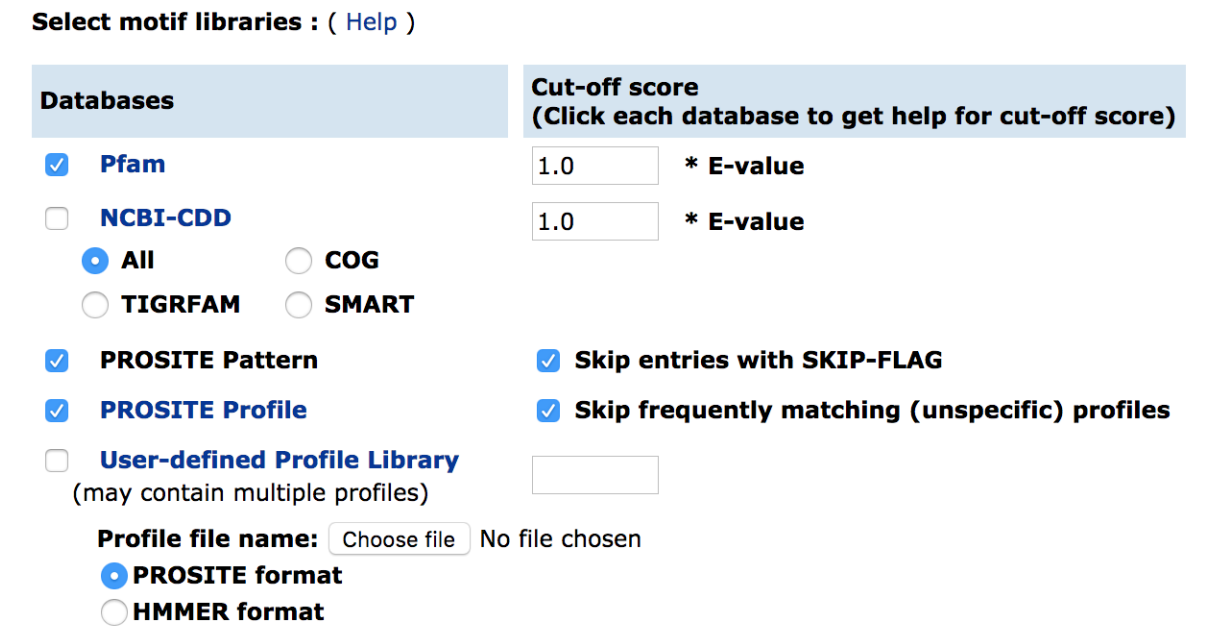

Figure 4: The selected search parameters when using 'MOTIF search' for conserved domain searches.

2.3 – RESULTS

The phylogenetic tree constructed using the 38 sequences mentioned in section 2.2.3 can be seen in figure 5 and figure 6. The focus of the task was to identify the closes relatives of AsM5 ORF1. Whilst there are some discrepancies between other phylogeny branches, both figures 5 and 6 agree that the most similar sequences are from *Anopheles dirus, Anopheles farauti* and *A.maculatus.*

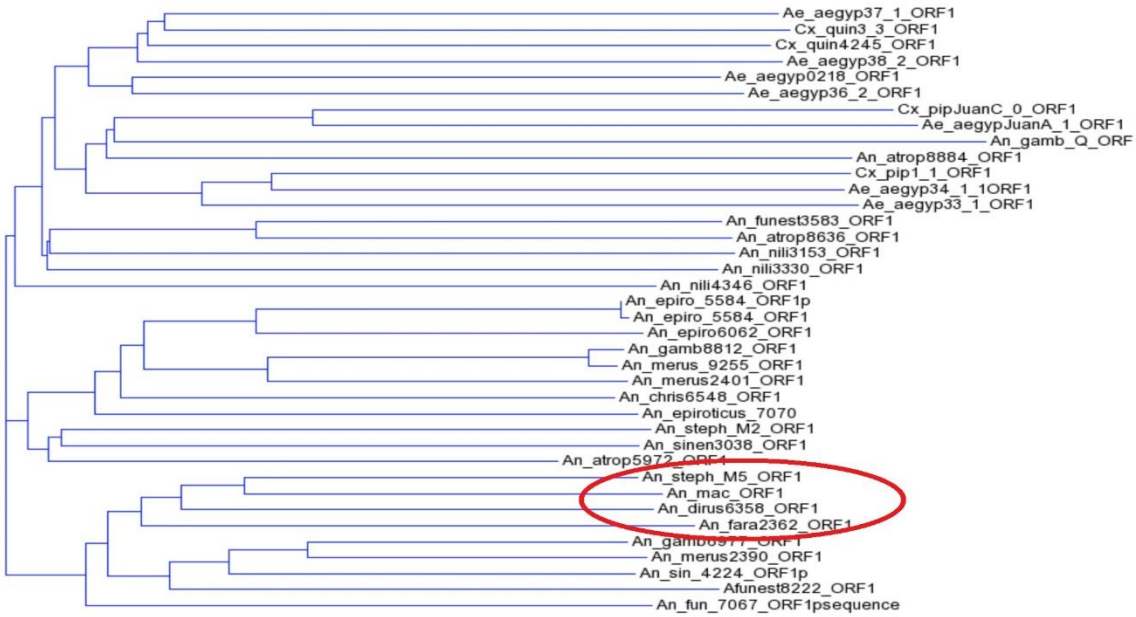

Figure 5: A phylogenetic tree constructed with the 37 other ORF1 sequences closely related to AsM5 ORF1 using a web tool at www.trex.uqam.ca/. The red ring highlights the sequences found to be of the closet relation to AsM5 ORF1

Figure 6: A phylogenetic tree constructed with the 37 other ORF1 sequences closely related to AsM5 ORF1 using a web tool at www.abi.ac.uk/Tools/msa/clustalo. The red ring highlights the sequences found to be the closest relatives of AsM5 ORF1.

The protein sequences of the closest AsM5 relatives identified from the generated phylogenetic trees were aligned (figure 7). The alignment shows that there are both highly and poorly conserved sections of the protein. There are some clear gaps in the alignment due to the proteins, difference in length of amino acid chain. Though there are a few poorly conserved positions within it, there is a large generally well conserved section in the alignment which in includes all four sequences. The area of continuously high conservation start position 136, 134, 63 and 110 for the related sequences in *A.farauti*, *A.dirus A.maculatus* and *A.stephensi* respectively. The area of high conservation ends in positions 353, 420, 286 and 337 in the same order. The most populous residues in the conserved area were small hydrophobic residues and residues with hydroxyl, sulfhydryl and amine functional groups.

Figure 7: A multiple sequence alignment of the jockey ORF1 sequences identified to be AsM5's closest relatives. The alignment shows areas of both high and low conservation. The area of continuously high conservation start position 136, 134, 63 and 110 for the related sequences in A.farauti, A.dirus A.maculatus and A.stephensi respectively. The area of high conservation ends in positions 353, 420, 286 and 337 in the same order. * = Fully conserved residue, : = Conservation between groups of strongly similar properties, . = conservation between groups of weakly similar properties (EMBL-EBI, 2018).

The next test in analysis of these sequences was the identification of conserved domains and motifs using web tools. The AsM5 ORF1 sequences was put through both the ‘SMART’ and ‘MOTIF search’ tools to identify conserved domains and motifs, the results can be seen in figure 8. The SMART search identified four regions of low complexity as well as three zinc figures. The Zinc fingers are identified in both web tool searches, however, they do slightly disagree on the exact location of the fingers. MOTIF search identifies the domain at positions 302-322 while SMART identifies it at position 256-313. Towards the N terminus of the protein, there is disagreement between the web tools. MOTIF search identifies a domain known as CPSF100_C from Pfram whilst SMART identifies area of low complexity in roughly the same position. These same tests were carried out with the closely related sequences from *A.farauti*, *A.dirus A.maculatus and can be seen in figures 9, 10 and 11 respectfully.*

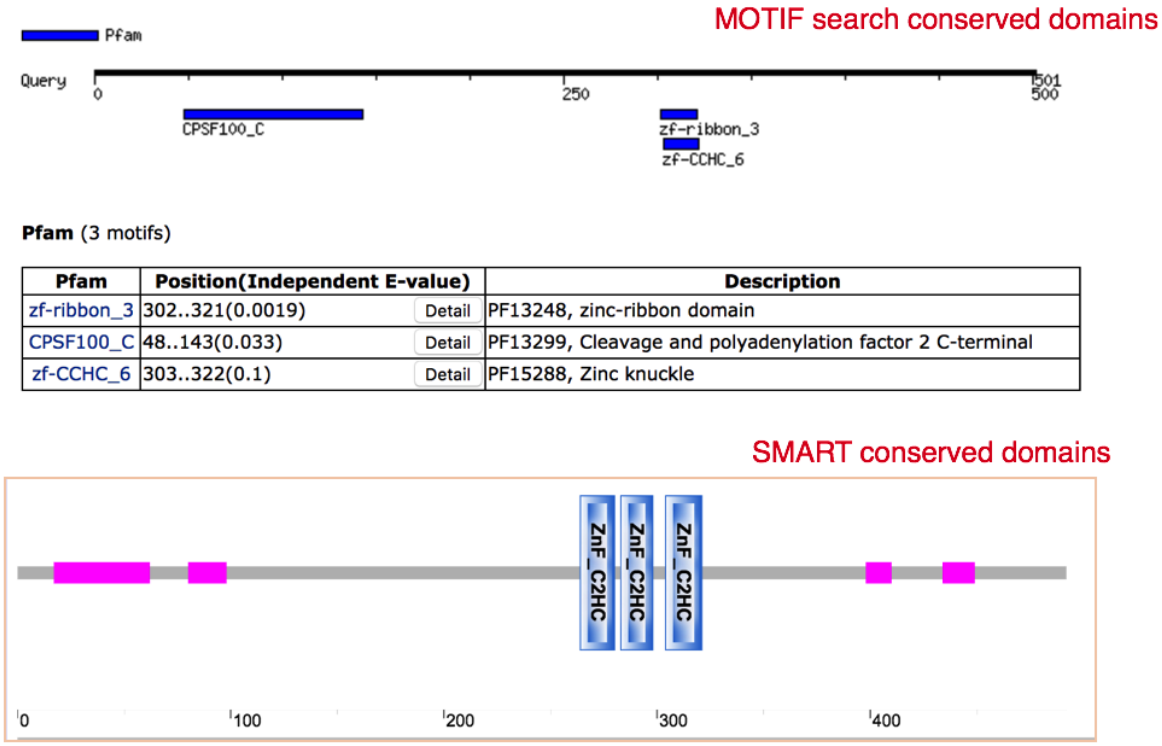

Figure 8: The products of conserved domain searches of AsM5 in both the MOTIF search and SMART web tools. The MOTIF search identified three domains and motifs, the details are on the schematic. The SMART search identified seven domains that include three zinc finger motifs and four areas of low complexity (in pink).

The identification of conserved domains from the AsM5 ORF1 related sequence in *A. farauti* is shown in figure 9. In the MOTIF search result there is only one identified conserved domain 'TFIIF_alpha' towards the N-terminus in position 27 to 120. In the diagram from the SMART search, there are again four areas of low complexity and one zinc finger domain in a central position from 288-304. Unlike the searches in AsM5 ORF1, there is no similarities between the SMART and MOTIF search results, most notably the lack of a zinc finger domain the MOTIF search result.

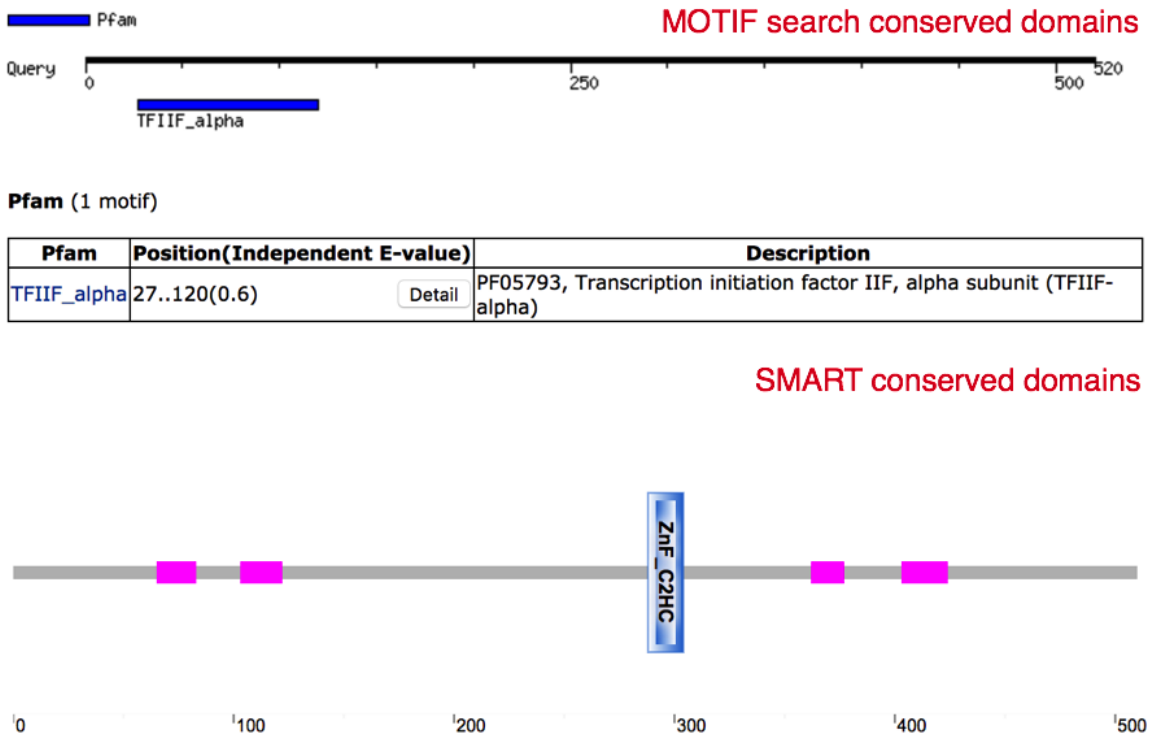

Figure 9: The products of conserved domain searches of the AsM5 ORF1 related sequence identified in A.farauti using both the MOTIF search and SMART web tools. TFIIF_alpha is the only domain featured in the MOTIF search diagram. The SMART domain search yielded five domains including a zinc finger and four areas of low complexity shown in pink.

Figure 10 outlines the results also from web tool tests on the AsM5 related sequence in *A.dirus*. The figure again shows that there are no similarities between the domains and motifs identified in the MOTIF search results when compared with the SMART domain results. The MOTIF search structure shows a domain towards the N terminus of the protein named FAM104 from position 16 – 86. The figure also shows seven domains and motifs identified from the SMART search, all previous seen in analysis of other sequences. There are three areas of low complexity towards the N terminus of the protein in positions 1-8, 42-73 and 103-130. The other three areas of low complexity are towards the C terminus at positions 380-392, 413-450 and 473-488. In addition to these areas of low complexity the figure also shows a zinc finger domain at position 131-284.

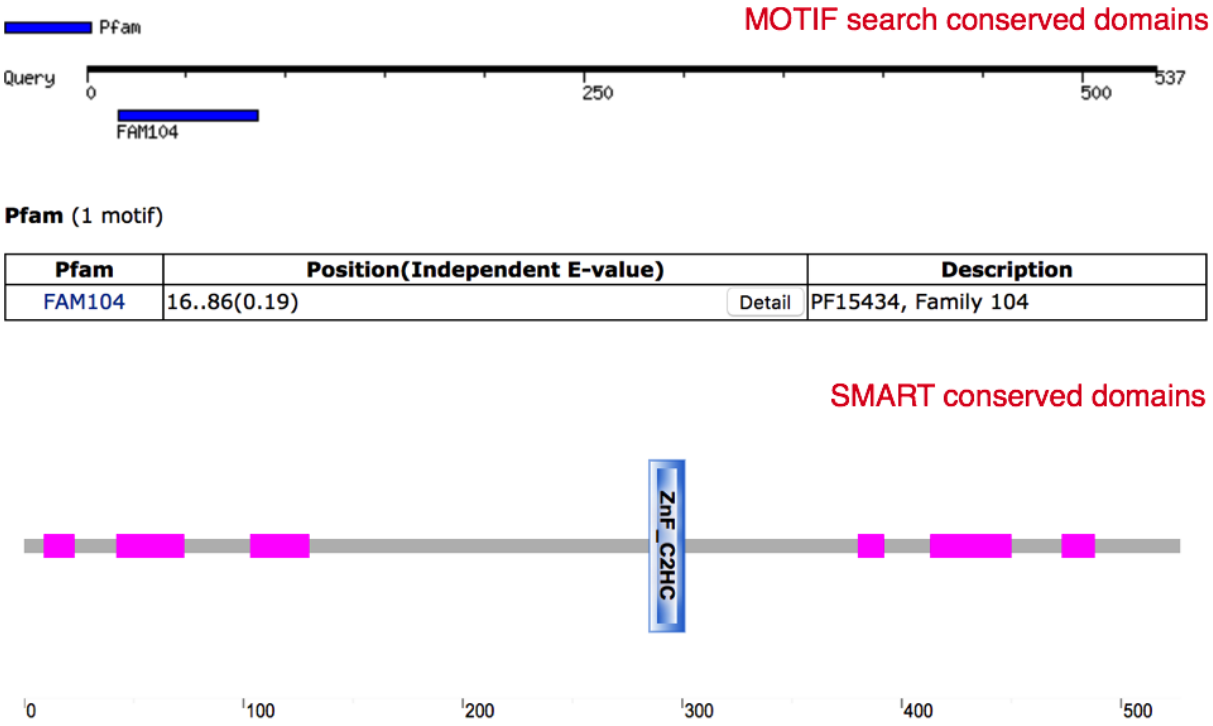

Figure 10: The products of conserved domain searches of the AsM5 ORF1 related sequence identified in A.dirus using both the MOTIF search and SMART web tools. The regions in pink on the SMART conserved domains represent areas of low complexity.

The final AsM5 related sequence from *A.maculatus* was also analysed using MOTIF search and SMART and is displayed in figure 11. The figure highlights a zinc finger ribbon from positions 254-271 in the MOTIF search diagram, this was the only domain predicted with confidence by this web tool. The SMART domains and motifs predicted with confidence in the same sequence are three areas of low complexity at positions 28-46, 309-331 and 346-358. Though not shown on the schematic, there were other interesting motifs identified by the SMART web tool. Three zinc finger motifs were below the threshold feature on the diagram and were positioned at residues 233-248 and 253-275.

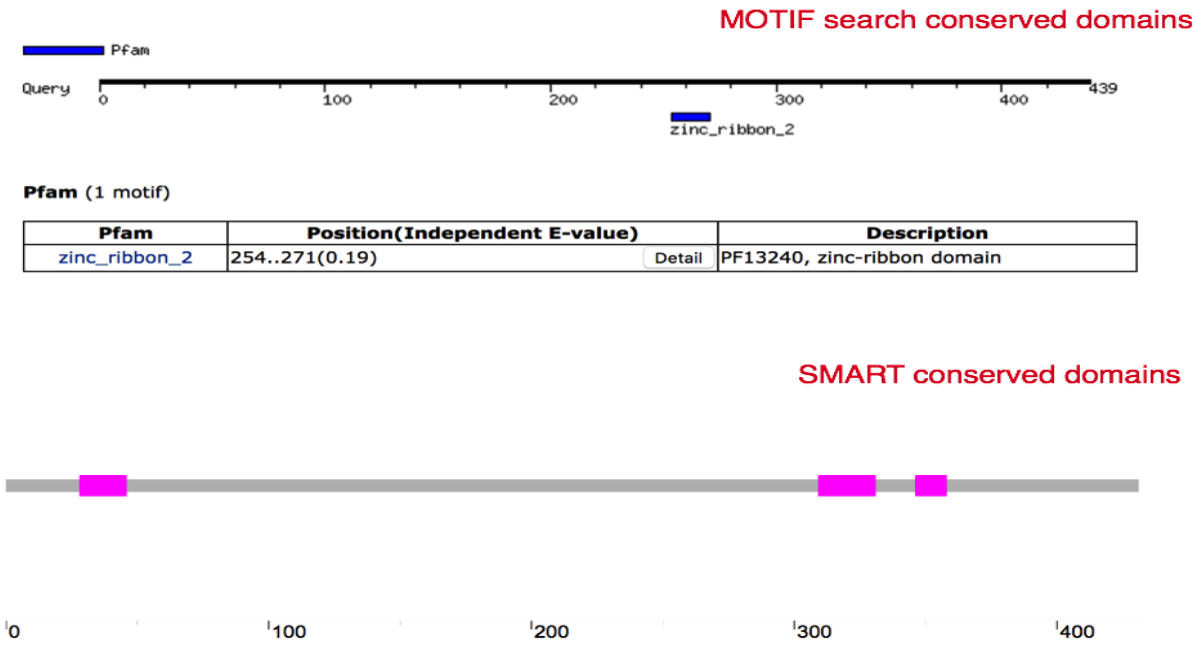

Figure 11: The products of conserved domain searches of the AsM5 ORF1 related sequence in A.maculatus using both MOTIF search and SMART web tools. The regions in pink on the SMART conserved domains represent areas of low complexity.

In three out of the four AsM5 related sequences, conserved domains were identified towards the N terminus of each ORF1 protein. They were CPSF100_C, TFIIF_alpha and FAM104 in *A.stephensi. A.farauti* and *A.dirus* respectively. The positions of these domains are listed in table 2. These sequences were aligned using https://www.ebi.ac.uk/Tools/msa/clustalo/ and the results are displayed in figure 12.

Table1: A table detailing the position of N terminus conserved domains identified by searches in web tools.

Identity of amino acid sequence and name of domain	Position of domain
TFIIF_alpha in *Anopholes farauti*	27-120
FAM104 in *Anopholes dirus*	16-86
CPSF100_C in *Anopholes stephinsi*	48-148

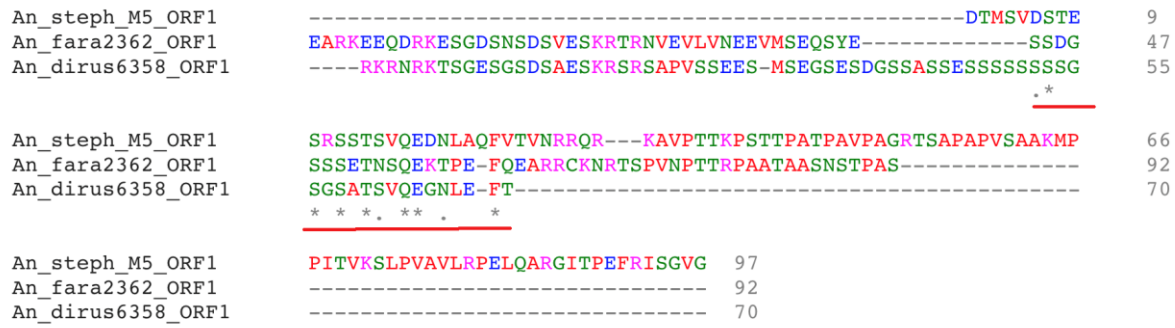

*Figure 12: A multiple sequence alignment of the CPSF100_C domain in A.stephensi, TFIIF_alpha in A.farauti and FAM104 in A.dirus. The alignment generally shows a lack of conservation between the three domains. The red line highlights an area that shows some residue similarity as well as points of sequence identity. * = Fully conserved residue, : = Conservation between groups of strongly similar properties, . = conservation between groups of weakly similar properties (EMBL-EBI, 2018).*

The alignment in figure 12 shows a general lack of conservation between the three proteins. There is the obvious issue of their different sizes with the FAM104 being significantly shorter than the other two. Nonetheless, there is a small area of similarity between the three domains underlined in red.

2.4 - SUMMARY

Phylogenetic analysis was successfully used to identify the closet relatives of AsM5 ORF1 from a vast list of elements found in several mosquito species. The selection of the elements in *A.maculatus*, *A.farauti and A.dirus* allowed for comparative analysis with AsM5's closest relatives. Conserved domain searches using web tools showed that the greatest similarity between these ORF1 proteins was the presence of one or more zinc fingers. The stand out features from conserved domain and motif searches on the AsM5 ORF1 amino acid sequence were CPSF100_C and the zinc finger knuckle. The identification of conserved domains and motifs in AsM5 ORF1p not only added a wealth of knowledge to what is known about the proteins functionality but also guided much of the laboratory work carried out later. Both the CPSF100_C and the zinc knuckle boast features that give tangible clues regarding the protein's function. Each identified structure is documented to interact with other proteins as well as bind nucleic acids such as DNA and RNA.

Chapter 3

EXPRESSION OF FULL LENGTH ORF1p AND ORF1p CONSERVED DOMAINS/MOTIFS IN *Escherichia coli*

3.1 – INTRODUCTION

In order to tackle the aim of this project, structural and functional assays would require protein samples encoded by the ORF1 sequence. *E.coli* is one of the organisms of choice for the production of recombinant proteins, its use as a cell factory is well-established and it has become the most popular expression platform (Rosano and Ceccarelli, 2014). *E.coli* offers a rapidly growing host that can be used for all parts of the process; from bulking up the vector to expression of the protein of interest (POI). Though expression of a recombinant protein may impart a metabolic burden on the microorganism, causing a considerable decrease in generation time, high cell density cultures are usually easily achieved. Choosing the most appropriate *E.coli* strain and expression vector are decisions that must not be taken lightly; the successful expression or even the potential yield of the POI can vary drastically between strains and vectors. Expression vectors have been improved and optimised over the years. The most common expression plasmids in use today are the result of multiple combinations of replicons, promoters, selection markers, multiple cloning sites, and fusion protein/fusion protein removal strategies (Rosano and Ceccarelli, 2014).

Many expression vectors, used in the production of heterologous proteins, function in much the same way. The use of the *lac* operon has been well characterised in prokaryotic organisms and the *lac* promoter plays a vital role in its function. Though the *lac* promoter is only induced in the presence of lactose, *E.coli* cells prefer to use glucose as their carbon source. In tandem this makes induction of the *lac* operon difficult in the presence of readily metabolisable carbon sources such as glucose. If lactose and glucose are both present, expression from the *lac* promoter is not fully induced until all the glucose in the media has been used up

(Muraoka et al., 1991). To overcome such obstacles, mutants of the *lac* operon were developed, for example, the mutant *lac*UV5 reduces but does not eliminate sensitivity to glucose regulation in rich media. Though this mutant revolutionised the field, it came with novel problems such as the disadvantage of sometimes having unacceptably high levels of expression in the absence of inducer. The vector suite chosen for expression in this project (pOPIN) contained pET28a vectors. In this system, the DNA sequence encoding the POI is cloned behind a promoter recognised by the phage T7 RNA polymerase. The specificity of T7 RNA polymerase for its own promoters in addition with its ability to inhibit the host RNA polymerase with rifampicin allows it to exclusively express of genes under the control of a T7 RNA polymerase promoter (Tabor & Richardson, 1985). Thus, the system can be induced by lactose or in this case its non-hydrolysable analogue isopropyl β-D-1-thiogalactopyranoside (IPTG) (Rosano and Ceccarelli, 2014).

Discouraging growth of untransformed cells is crucial when growing cultures for protein expression. Because of this, selection genes are included in nearly all vectors on the market. Selection of the pOPIN vectors used a kanamycin (Kan) antibiotic resistance gene. In addition to being able to select for successful transformants, it is vital to have an ability to detect the POI during the expression and purification process. Small peptide tags as well as non-peptide fusion partners are often used in combination throughout the field. Small peptide tags are arguably the most important of the two as they are often used as the tool for purification by chromatography (Rosano and Ceccarelli, 2014). For example, in this case, vectors are designed to produce 'His-tagged' proteins which can be recovered by immobilised metal ion affinity chromatography (IMAC) using Ni^{2+} beads. The non-peptide fusion partners often work as solubility enhancers (Hammarström et al., 2009) and for this project vectors possessing glutathione *S*-transferase (GST) and small ubiquitin-like modifier (SUMO) tags were available.

IMAC is just one of many chromatographic techniques used to isolate recombinant proteins. IMAC is based on the interactions between a transition metal ion (Co^{2+}, Ni^{2+}, Cu^{2+}, Zn^{2+}) immobilised on an agarose matrix and specific small peptide tags. The amino acid Histidine shows the strongest interaction, which is why 6x His tags are usually coded into expression vectors. Electron donor atoms on the histidine imidazole ring readily form coordination bonds with the immobilised metal and those bonds can be dissociated by adding a buffer with a high concentration of imidazole (Norouzi, Hojati and Badr, 2016). Though IMAC is an extremely effective purification tool, it is often used in combination with other chromatography methods to achieve a truly clean sample. Ion Exchange Chromatography (IEX) is widely used because the buffer conditions can be adapted to suit a comprehensive range of proteins rather than only being applicable to a specific functional group of proteins (Cutler, 2004).

For biochemical functional and structural analysis to be performed the fusion protein partner must be removed from the POI. Due to their functionality and size, these fusion partners can adversely affect any data collected from functional and structural studies. The should also be removed because they also can interfere with protein activity and structure (Rosano and Ceccarelli, 2014). In this case, the vectors in the suite were designed with a protease recognition site for removal via enzymatic cleavage.

In this chapter, vectors containing the full length AsM5 ORF1 sequence are used to express the ORF1 protein. The knowledge gained from bioinformatics analysis is used to clone the CPSF100_C conserved domain into and expression vector and the conditions for optimal expression and purification are investigated.

3.2 - METHODS

3.2.1 – TRANSFORMATION

The vector used was a DNA plasmid called pOPINB with an AsM5 ORF1 insert. This vector was a gift from Dr. Paul Elliott (Laboratory of Molecular Biology, Cambridge, UK) containing a 6 x histidine tag and a kanamycin resistance gene. Transformation of the DNA plasmid into *E. coli* was performed using a heat shock method. Competent *E.coli* Rosetta cells were deforested from -80°C to room temperature and incubated on ice. The DNA construct was added into the cells and incubated on ice for 30 minutes. The mixture was heat shocked in a 42°C water bath for 30 seconds and then placed back on ice (See appendices for DNA construct sequence). SOC media is added and the transformed cells were incubated at 37°C for 1 hour with agitation. 50μl, 100μl and 150μl of the transformed cells were then inoculated on separate Luria-Bertani (LB) agar with added kanamycin 34ug/ml (Kan) and chloramphenicol 34ugml(Cam) for antibiotic selection of positive transformants. The agar plates were then incubated for at 37°C for 16 hours. The details of the media components can be found in the appendices.

3.2.2 – EXPRESSION OF FULL LENGTH ORF1 PROTEIN IN *Escherichia coli*.

A single isolated colony of transformed cells was used to inoculate 100ml LB/Kan/Cam broth in a 250 ml conical flask and grown for 16hours at 37°C shaking at 150 RPM. This starter culture was then used to inoculate 500ml/Kan/Cam broth in a two-litre flask to an optical density (OD) of 0.100 at 600nm. The flask was then grown at 37°C with shaking to an OD of 0.600 – 0.800 OD at 600nm. A sample of un-induced culture was taken and IPTG was added to the remaining culture at a final concentration of 1mM, then incubated at 37°C for 3 hours. Both induced and un-induced samples were centrifuged and pelleted (6000 x g for 30 minutes

at 4°C) then re-suspended in lysis buffer (50mM Tris, 100mM NaCl, 1mM EDTA, pH 7.5). After re-suspension the culture was centrifuged again (6000 x g for 30 minutes at 4°C) and they supernatant discarded before freezing the pellet at -20°C)

3.2.3 – PREPARATION OF INCLUSION BODIES (SONICATION)

The pellet was defrosted and re-suspended in lysis buffer and lysozyme protease inhibitors and DNase were added and kept on ice for 5 minutes. The mixture was sonicated 10 minutes with short bursts of 30 seconds followed by intervals of 30 seconds for cooling whilst kept on ice at all times. The sonicated mixture was centrifuged at 15 000 x g for 50 minutes to separate the soluble and insoluble fractions. The resulting pellet was harvested and the supernatant discarded. The pellet was then re-suspended in lysis buffer and centrifuged at high speed to wash the pellet. This was repeated twice. The pellet was then re-suspended in urea buffer (8M urea, 50mM HEPES, pH 7.4) and left to solubilise on a roller for 16 hours. The purification process was IMAC, performed on a 'His-Trap' ÄKTA start system. Two buffers were used in the purification process (A = 50mM HEPES, 8M urea pH 7.4. and B = 50mM HEPES, 8M urea, 500mM Imidazole, pH 7.4). The program performed several wash steps after loading the cell lysate on to the column before eluting the POI by increasing the concentration of imidazole in the buffers by a gradient.

3.2.4 - PREPARATION OF INCLUSION BODIES (TRITON-X).

The pellets were defrosted and re-suspended in 30ml of lysis buffer per litre of culture, then incubated at 37°C for 30 minutes with agitation. Nonidet P-40 was added to a concentration of 1% and incubated at 4°C for 50 minutes with mild agitation. $MgSO_4$ was added to a concentration of 15μM along with DNase, the mixture was then incubated at 37°C with agitation for 30 minutes. To separate the soluble and insoluble fractions the mixture was centrifuged at 15,000 x g for 40 minutes at 4°C. Triton-X100 was added to the lysis buffer at 0.5% and it was used to re-suspend the resulting pellet and then centrifuged again at high speed. Phosphate-buffered saline (PBS) was used to wash the resulting pellet (Inclusion

bodies) and repeated 3 times. The inclusion bodies were then re-suspended in urea buffer (50mM NaH_2PO_4, 300mM NaCl, 8M urea) on a roller for 12 hours to solubilise.

3.2.5 – PURIFICATION OF FULL LENGTH ORF1 PROTEIN.

The first step in the protein purification was IMAC performed on a 'His-Trap' ÄKTA start program. Two buffers were used in the purification process (A = 50mM NaH_2PO_4, 300mM NaCl, 8M urea pH 8.0. B = 50mM NaH_2PO_4, 300mM NaCl, 8M urea, 500mM Imidazole, pH 8.0). The solubilised inclusion bodies were centrifuged at high speed (15 000 x g) and the supernatant was filtered through 0.22um pores. The sample was loaded on to the AKTÄ and run through a 'His-Trap' program with a Ni^{2+} agarose column. The program included an equilibration of the column, application of the sample, wash steps of unbound protein and elution of the POI with a gradient of increasing imidazole concentration. Following the purification, the fractions including the flow through and wash steps were loaded onto a 12.5% acrylamide gel for analysis.

3.2.6 – 'In-Fusion®' CLONING OF CPSF100_C DOMAIN.

The CPSF100 domain was cloned into a pOPIINK vector using the 'In-Fusion®' cloning method. The CPSF100 sequence was first amplified using a novel primer pair (one cycle at 95°C for 2 minutes; 95°C for 40 seconds, 50.1°C for 30 seconds and 72°C for 2 minutes repeated for 30 cycles; final extension at 72°C for 7 minutes and held at 4°C after completion) and run on a 0.7% agarose gel. After visualising the gel, the correct bands were cut from the gel and the DNA was extracted using a Monarch DNA gel extraction kit (New England Biolabs. Following gel extraction, the manufacturers protocol 'In-Fusion Cloning Procedure for Spin-Column Purified PCR Fragments' was followed to clone the CPSF_100 sequence into the pOPINK vector (GST tag). After the ligation reaction, the new construct was transformed into TOP10 *E.coli* cells and grown on LB/Kan/IPTG/X-Gal plates. Blue/white screening and colony

PCR (a single colony of transformants diluted in sterile distilled water and used as the DNA in a PCR to check for amplification of specific DNA) was used to identify positive transformants. These were then used to inoculate 5ml LB/Kan broth in order to bulk the cells up for DNA plasmid extraction. A Sigma-Aldrich mini prep plasmid extraction kit was used according to the manufacturers' protocol.

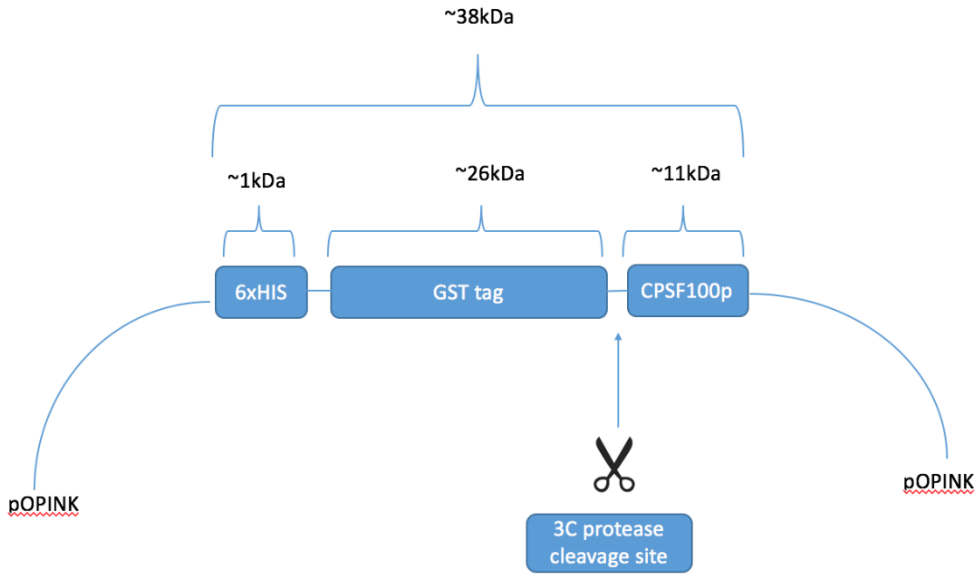

Figure 13: An image outlining the components of the pOPINK/CPSF100_C construct. To the far left is the 6xHis tag (~1kDa) which is followed by the ~26kDa GST tag and then the protein of interest CPSF100 ~11kDa. All these components together make a ~38kDa recombinant protein.

3.2.7 – EXPRESSION TRIALS OF THE CPSF100_C DOMAIN.

Expression trials of CSPF100_Cp began with small scale cultures. 15ml starter cultures and 50ml expression cultures at varying conditions in order to optimise the expression protocol. Expression was trialled with varying lengths of incubation, varying temperatures and OD600 at induction. Trials of CPSF100_Cp expression were carried out at 37°C for 3 hours and 18°C for 16 hours before scaling up to large expression cultures. Expression at 18°C for 16 hours was scaled up and trials continued by inducing expression at an OD600 of 0.4, 0.5 and 0.6.

3.2.8 – EXPRESSION OF CPSF100_C DOMAIN.

The pOPINK/CPSF100_C construct was transformed into BL21* expression cells following the transformation method outlined in section 3.2.1 using LB/Kan plates to select for positive transformants. A single colony of the cells was used to prepare a starter culture to ultimately inoculate a 500ml/Kan broth to an optical density (OD) of 0.1 at 600nm. The culture was then incubated at 37°C with shaking to near saturation. A sample of un-induced culture was taken and 1mM IPTG to the remaining culture and incubated at 18°C for 16 hours. Though this was the final optimised method, several expression trials were performed, at both small and large scales. Both induced and un-induced samples were centrifuged and pelleted (6000 x g for 30 minutes at 4°C) then re-suspended in buffer (500mM NaCl, 20Mm Tris, 2mM DTT, pH 6.8) and frozen at -20°C.

3.2.9 – HIS STAIN.

In order to confirm the presence of the recombinant POI a ‘his-tag specific’ assay was performed. InVision™ His-Tag Stain was used on acrylamide gels after SDS page following the manufacturers protocol.

3.2.10 – ^{15}N LABELLED EXPRESSION OF CSPF_100 DOMAIN.

The pOPINK/CPSF100_C construct was transformed into BL21* expression cells following the transformation method outlined in section 3.2.1 using LB/Kan plates to select for positive transformants; a single colony of the cells was used to prepare the starter culture. The expression media was a minimal media made by adding 50ml of 10x solution A (0.88 M Na_2HPO_4 & 0.55 M KH_2PO_4), $CaCl_2$ to 0.05mM, MgSO4 to 0.001M and Thiamine to 10 mg/ml with sterile distilled water in a two litre conical flask. The starter culture cells were then used to inoculate the minimal media to an optical density (OD) of 0.15 at 600nm and then expressed as outlined in section 3.2.7.

3.2.11 – PURIFICATION OF CPSF100 PROTEIN.

The cell lysate sample was centrifuged at high speed (15 000 x g) and the supernatant was filtered through 0.22um pores for purification. The IMAC purification of CPSF100 protein was performed using the method highlighted in section 3.2.5 using different buffers. Buffer A (500mM NaCl, 20Mm Tris, 2mM DTT, pH 6.8) and Buffer B (500mM NaCl, 20mM Tris, 2mM DTT, 500mM Imidazole, pH 6.8). Following IMAC, the sample was further purified using Ion Exchange Chromatography. The sample was first dialysed into a low salt buffer A (50mM NaCl, 20mM Tris, 2mM DTT, pH 8.0) and bound to a cellulose column through anion exchange. Following this, a salt gradient will be employed to elute the protein using buffer B (500mM NaCl, 20mM Tris, 2mM DTT, pH 8.0).

To further confirm the protein was a fusion protein and to prepare a sample for further analysis, 3C protease was used to cleave the GST tag from the CPSF100_C protein by incubating the protease and the fusion protein overnight at 4°C. In order to separate the 3C protease and GST tag from the CSPF100_C protein, a ‘reverse his trap’ was performed by applying the

protein sample to a Ni^{2+} agarose column and collecting it; buffer B was used to elute the GST and protease which had bound the column as a result of their 6x-his tag. The protein sample was dialysed in to buffer A then into MES buffer using SnakeSkin™dialysis tubing before being concentrated to a concentration of 70mM. The concentrated sample was then used for one-dimensional nuclear magnetic resonance (NMR) studies. A sample from each step of dialysis and reverse his trap were run on 18% acrylamide gels for SDS-PAGE.

3.2.12 – PURIFICATION OF 15N LABELLED CPSF100 PROTEIN.

The purification of ^{15}N labelled CPSF100_C was performed following the method outlined in section 3.2.9. However, in this case the IEX step was excluded and the protein sample was dialysed in to buffer A using SnakeSkin™dialysis tubing before the a reverse 'his-trap' was performed.

3.3 – RESULTS

The transformation of the ORF1/pOPINB was successful and confirmed using a colony PCR. It had already been confirmed from previous unpublished work performed by Akinbosede (2016) showed that the pOPINB vector expressed ORF1p in inclusion bodies. The SDS PAGE gels run from the solubilised inclusion bodies yielded very little evidence to show the overexpression of ORF1p. Attempts to purify the solubilised inclusion bodies from both the method of sonication and detergent (Triton-X100) via IMAC were unsuccessful.

Infusion® cloning of the CPSF100_C (322bp) sequence into the pOPINK vector was successful as confirmed by colony PCR (Figure 14). DNA sequencing of the constructs showed that not only was the insert present, but the sequence was in the correct frame for protein expression. As can be seen in figure three of the constructs sequenced were aligned with the CPSF100_C DNA sequence with 100% identify.

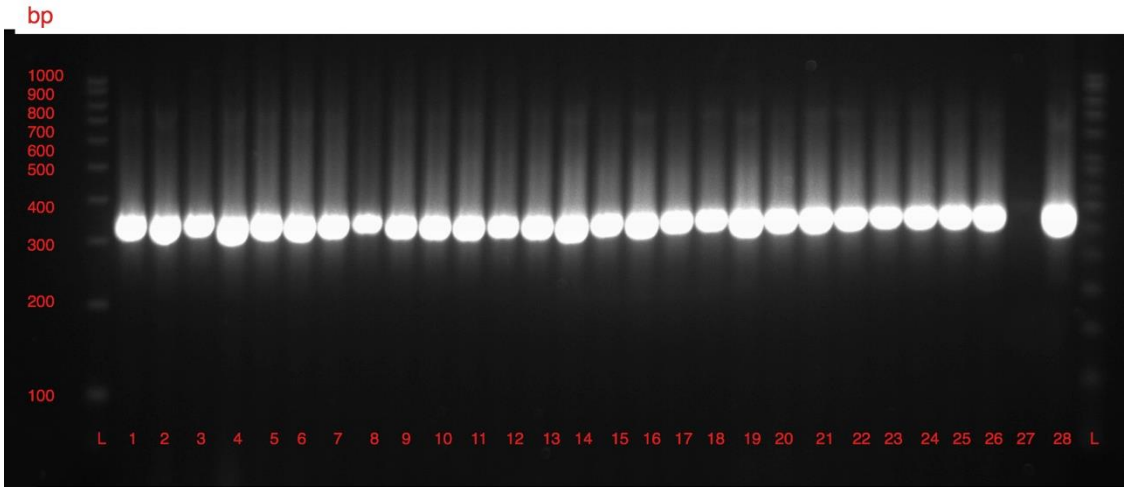

Figure 14: 1.8% agarose gel showing the products of a colony PCR using colonies transformed with the CPSF100_C/pOPINk construct. L (left) = 100bp ladder, 1-26 = PCR products, 27 = negative control, 28 = positive control, L (right) = 50bp ladder

When expressing ORF1p, several expression trials were performed to check for the presence of recombinant protein. Initial expression trials presented very little sign of the POI but the expression at 18°C for 16 hours with normal induction did present a promising band (~37

kDa) in the insoluble fraction as can be seen in figure 15. This band seen in the insoluble induced pellet can also be seen in the insoluble pellet of the un-induced sample.

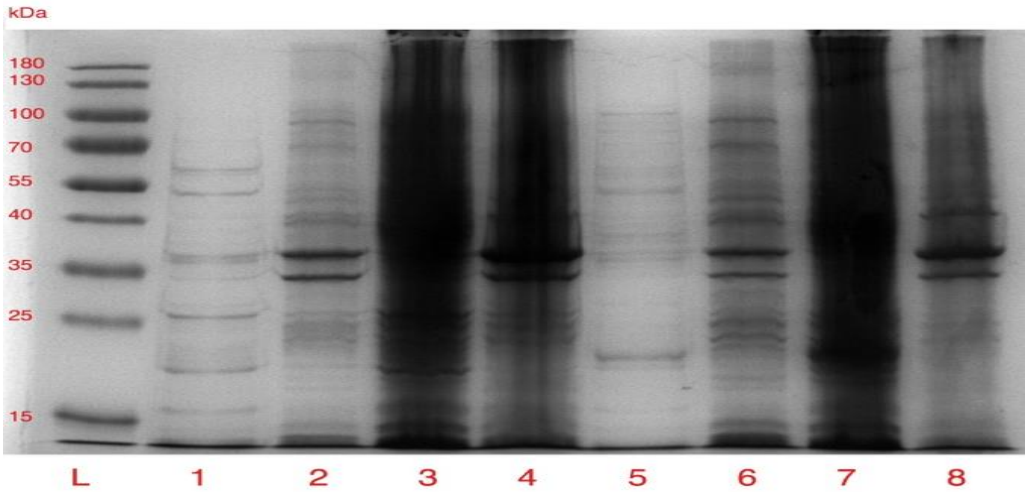

Figure 15: A 12.5% acrylamide gel with samples from a trial expression at 18□C/16 hours. L = Ladder, 1 = un-induced soluble, 2 = un-induced insoluble, 3-4 are concentrated versions of 1-2 respectively, 5 = induced soluble, 6 = induced soluble, 7-8 are concentrated versions of 5-6 respectively.

Figure 16 shows the results of a IMAC purification of CPSF100_Cp after SDS PAGE. The gels show that the POI is successfully expressed and eluted with increasing concentrations of imidazole. The band at ~38 kDa in lanes 2-7 on gel 2 match that of the POI which is predicted to be ~11kDa and the GST/His tag (~26 kDa) bound together. Though the POI is successfully bound to the column and eluted with high concentrations of imidazole, the gel shows that there is some unspecific binding to the column, possibly due to degraded forms of the POI.

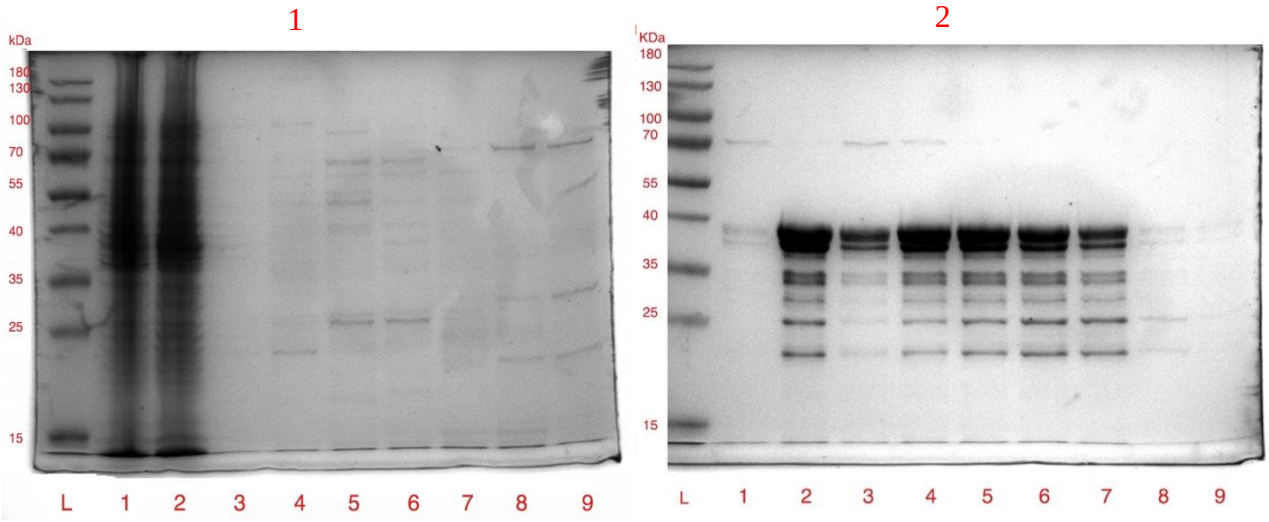

Figure 16: Two 12.5% SDS PAGE gels using the samples collected from the IMAC purification of cell lysate after CPSF100_C expression. Lanes 1 and 2 in gel 1 show the flow through collected as the lysate was applied to the column. All the subsequent lanes in both gels the samples produced by the wash steps and elutions with an increasing concentration of imidazole (0-500mM).

To further purify the sample, IEX was performed and the results can be seen in figure 17. It is clear that the POI was eluted with increasing concentrations of NaCl (Gel 2, fractions 4 – 7). However, it is also clear that purification via IEX did not make the sample much cleaner. Nonetheless, the results make it clearer that the protein present is indeed the POI.

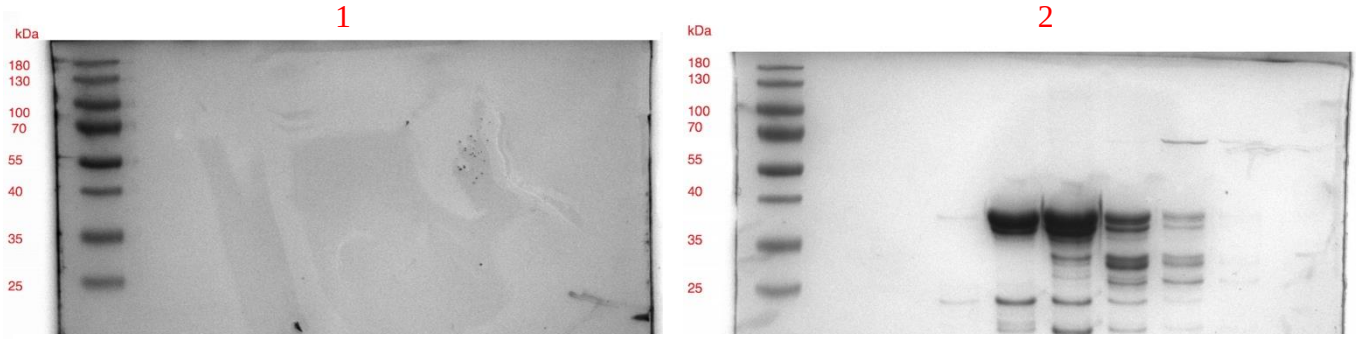

Figure 17: Two 12.5% SDS PAGE gels using the samples collected from the IEX purification of the protein sample collected from IMAC. Lanes 1 – 4 in gel 1 show the early wash steps. All the subsequent lanes in both gels the samples produced by the elutions of the POI with an increasing concentration of NaCl (0-500mM).

After IEX, the protein sample was prepared for cleavage by the 3C protease in order to cleave the GST and His-tag from CPSF100p. Figure 18 shows that the sample from IEX in lane 1 is cleaved in lane 2, which shows the presence of the GST/His-tag, the 3C protease and the CPSF100_Cp. In lane 3 the cleaved sample is run through the Ni^{2+} agarose column separating CPSF100_Cp from the tags which are bound to the column and eluted in lane 4, though a small amount of CPSF100_Cp can still be seen in the lane. Though there is still a little CPSF100_Cp in lane 4, it is clear that in lane 3 there is a clean sample of CPSF100_Cp without the presence of any other protein or tag.

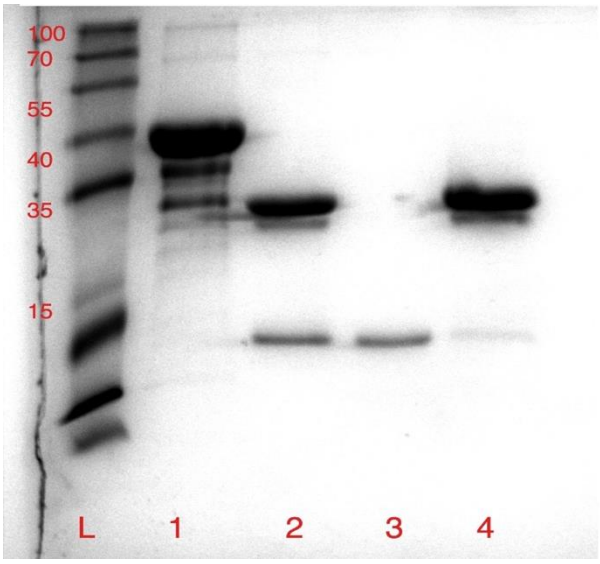

Figure 18: A 15% acrylamide gel showing samples obtained from cleavage of the 3C protease of the POI. L = Ladder, 1 = un-cleaved sample, 2 = cleaved sample before reverse his-trap, 3 = flow through of cleaved sample, 4 = elution of bound protein via IMAC buffer B (500mM Imidazole).

After a pure sample of the CPSF100_Cp was obtained through several purification techniques and enzymatic cleavage; the protein was then dialysed into a pH buffer containing 2-Morpholinoethanesulfonic acid monohydrate (MES) and concentrated in preparation for NMR studies, the result of which is shown in figure 19. On initial appearances, the 1D spectrum of

the protein is accurate based on knowledge of the sequence as demonstrated by the lack of peaks in the aromatic region of the spectrum; correlating to the absence of aromatic amino acids in the sequence. The spectrum shown is indicative of that of an unstructured protein. The presence of tall, sharp and undispersed peaks suggests that the protons in CPSF100_Cp are subjected to very similar chemical shifts due to the lack of shielding that would be present if the protein was folded. This confirms the outcome of the protein disorder prediction that was performed using a web tool named DISOPRED (Ward, Sodhi, McGuffin, Buxton & Jones, 2004) at http://bioinf.cs.ucl.ac.uk/psipred/. Very large peaks such as those seen at 3.7 ppm are confirmed to be caused by the MES buffer the protein was stored in.

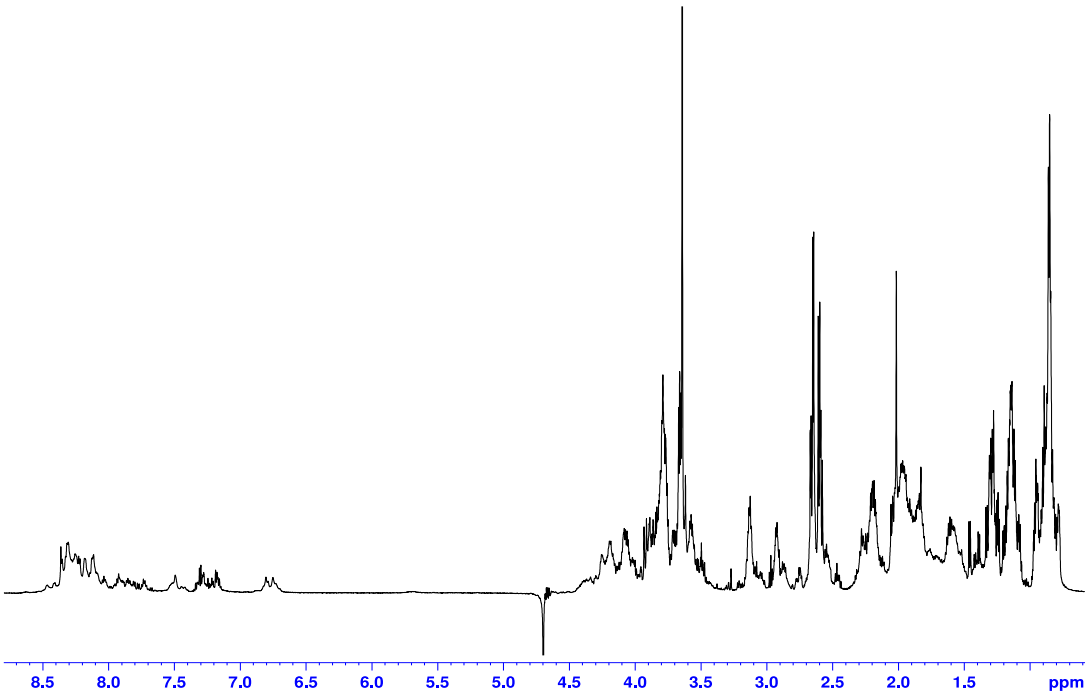

Figure 19: A 1H 1D NMR spectrum of unlabelled CPSF100_Cp. The spectrum is typical of an unfolded or disordered protein, confirmed by the presence of tall, sharp and undispersed peaks suggests that the protons in CPSF100_Cp are subjected to very similar chemical shifts due to the lack of shielding. The water peak can be seen at 4.7 ppm and the large peak at 3.7ppm is attributed to the MES buffer of which the protein is stored. 1H spectrum was acquired on a Bruker 700MHz spectrometer with cryoprobe at Francis Crick Institute by Alain Oregioni. Number of scans = 128; number of dummy scans = 16; spectral width =

15.9406 ppm; Water suppression was achieved using excitation sculpting with gradients. Spectra was processed using TopSpin v3.5 pl 7.

In figure 20 there are 2 gels showing the samples collected from an IMAC purification of the ^{15}N labelled expression of CPSF100_Cp. It was expected that the gel samples would resemble that of the unlabelled expression but that does not appear to be the case when comparing the two sets of results. Figure 8 as well as figure 6 shows a 'his' labelled protein being eluted with increasing concentrations of imidazole. However, in the case of figure 20 the eluted protein is ~ 27 kDa and not the expected ~ 38kDa that is seen in figure 6. This procedure was repeated from transformation to IMAC purification to find the same result, a protein ~10 kDa smaller than expected.

Nonetheless, the purification process was continued and the reverse his trap was performed as can be seen in figure 21. It would appear that there was no cleavage of the eluted protein as no band can be seen in the flow through (Lane 2) and the sample in the post cleavage sample is identical in size to that of the control sample form IMAC.

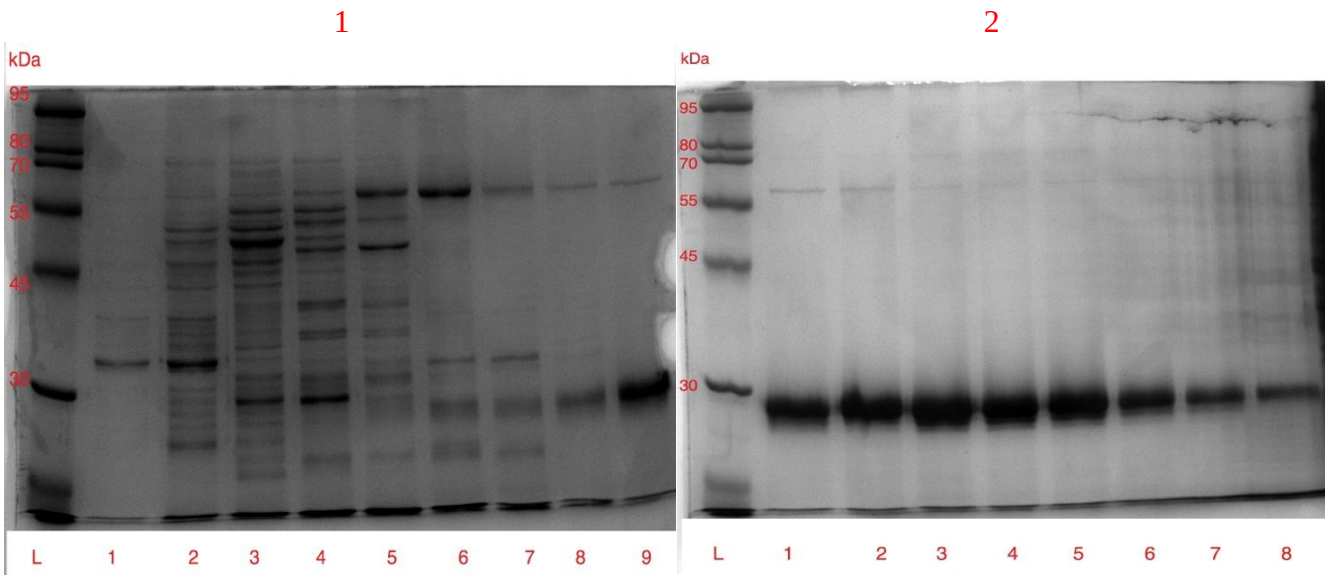

Figure 20: Two 12.5% SDS PAGE gels using the samples collected from the IMAC purification of cell lysate after 15N labelled CPSF100_Cp expression. Lanes labelled 'L' contain a protein standard ladder. All the subsequent lanes in both gels the samples produced by the wash steps and elutions with an increasing concentration of imidazole (0-500mM).

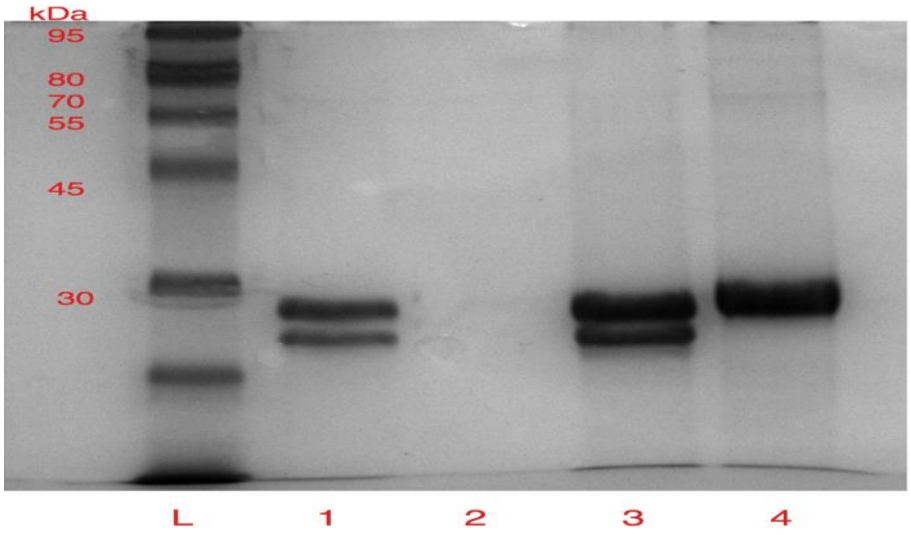

Figure 21: A 15% acrylamide gel showing samples obtained from cleavage of the 3C protease of the POI. L = Ladder, 1 = cleaved sample before reverse his-trap, 2 = flow through of cleaved sample, 3 = elution of bound protein via IMAC buffer B (500mM Imidazole) and 4 = Control sample from IMAC purification

3.4 - SUMMARY

In conclusion, the expression of full length ORF1p in *E.coli* was not productive enough for structural or functional studies. The process of solubilising the inclusion bodies was time consuming and unreliable. Identification of conserved domains and motifs in chapter 2 offered a life line to further studies of the protein. The CPSF100_C domain and the zinc finger motifs (Zinc finger knuckle) sparked immediate interest when matching the literature regarding them to current knowledge regarding the element as a whole. The decision was made to clone both the CPSF100_C domain and the zinc finger motifs into appropriate pOPIN vectors. In-fusion® cloning of CPSF100_C was successful but the same was not true for the zinc finger knuckle. Initial PCRs of the zinc finger knuckle sequence did not present the desired result to proceed with the cloning protocol leaving CPSF100_C as the primary focus. After several expression optimisation trials, this chapter shows the successful expression of the CPSF100_C domain in *E.coli* using the pOPINK vector (GST tag). Following purification and dialysis, a sample of this protein was used for 1D 1H NMR which showed that the protein is disordered or unfolded in its native state.

A phenomenon was observed when attempting to express ^{15}N CPSF100_Cp in minimal media. The repeated expression of the GST tag without ^{15}N CPSF100_Cp was as unexpected as it was difficult to explain.

Chapter 4

EXPRESSION OF FULL LENGTH ORF1p IN *Saccharomyces cerevisiae*.

4.1 – INTRODUCTION.

'There is no universally applicable solution for the production of all recombinant proteins' (Bill, 2014). And there is still no effective way to accurately predict which host system is most suitable to any particular protein, especially when trying to reach the highest functional yields. In biotechnology, recombinant proteins can be produced using a variety of different cell factories. This includes bacteria such as *E.coli*, yeast such as *Saccharomyces cerevisiae*, insect cells infected with vectors such as baculovirus and mammalian cells (Hou, Tyo, Liu, Petranovic & Nielsen, 2012).

S.cerevisiae, commonly referred to as baker's yeast is a single celled eukaryotic fungal organism with seemingly endless applications to biotechnology, farming and food technology. As arguably the best studied and researched eukaryotic model organisms, there is an enormous wealth of knowledge encompassing its genomics, biochemistry and physiology. There is also decades of work detailing its large-scale fermentation performance and how that can be utilised to enable this organism as an industrial powerhouse (Nielsen & Jewett, 2008).

Usually, the limiting factor of recombinant protein expression is often the ability to obtain sufficient quantities of the protein for clinical studies or for production at suitably low cost to allow for its availability in the wider market (Werner 2004). As previously mentioned, different host factories have been described and single celled microbes are often preferred due of their quick growth, high biomass potential and well-characterised biological and modification mechanisms (Porro, Sauer, Branduardi & Mattanovich, 2005). In the biotechnology industry there is almost always a decision to be made when choosing the appropriate cell factory as

most of the recombinant protein production is achieved in Escherichia coli, *Pichia pastoris*, *S.cerevisiae*, and Chinese hamster ovary cells (CHO cells) (Hou, Tyo, Liu, Petranovic & Nielsen, 2012).

Model organisms have been paramount in research over several decades because they provide a framework on which it is possible to develop and optimize methods can then be used in higher organism (Karathia, Vilaprinyo, Sorribas & Alves, 2011). *S.cerevisiae* serves as an important model for all eukaryotes and many of the genes that have had the greatest impact in human medicine were first discovered as homologs in the yeast. It was also the first eukaryotic organism to have its genome sequenced and many breakthroughs in all the biosciences have been pioneered using *S.cerevisiae* as a model organism (Nielsen & Jewett, 2008). There are many reasons to perform recombinant protein expression in *S.cerevisiae* rather than some of the other systems available, particularly prokaryotic hosts. Yeast cells such as S.cerevisiae can offer the better of two worlds. They provide many of the advantages of recombinant expression in microbes such as fast growth speed, easy genetic manipulation using expression vectors and most importantly, low cost media. Whilst still offering some of the features of higher eukaryotic organism, most notably post translational modifications and secretory expression.

As stated in section 3.1, expression vectors are at the heart of any expression system. The vector chosen was based on a *GAL1* promoter. The *GAL1* promoter in *S.cerevisiae* is induced by the presence of galactose in the media and strongly repressed by the presence of glucose. There are two sites within the GAL1 promoter that mediate glucose repression. This functions by glucose first inhibiting transcription activation by the GAL4 protein. Secondly, a promoter element actually confers glucose repression independently of GAL4 to ensure the regulation of the promoter (Flick & Johnston, 1990).

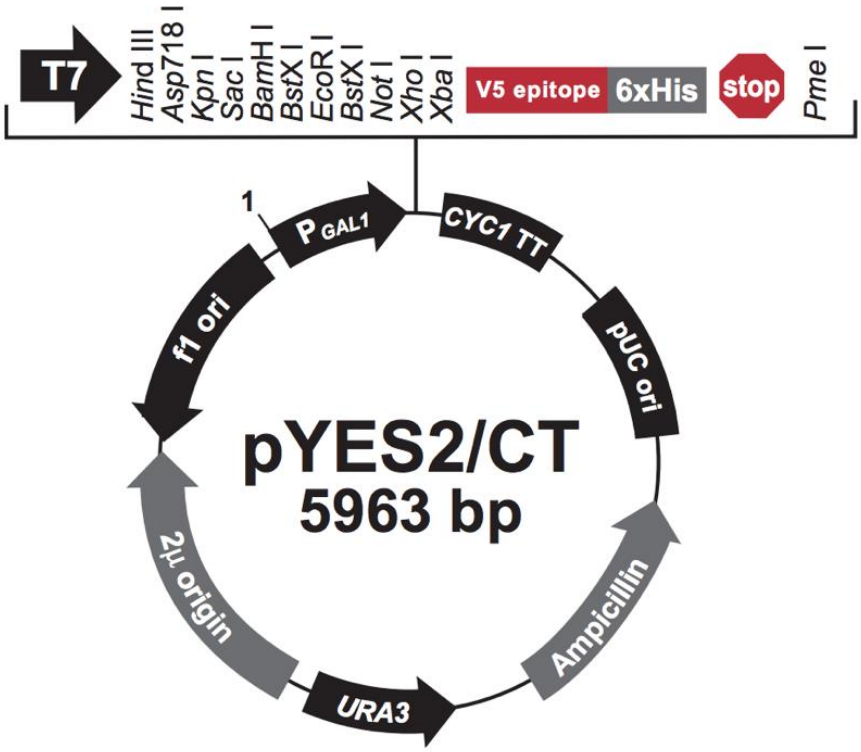

Figure 22: A map of the pYES2/CT S.cerevisiae expression vector from Invitrogen™ (Fisher, 2018).

In this case the vector chosen for expression the protein of interest (POI) was pYES2/CT from the suite of Invitrogen™ pYES vectors. The pYES2 suite of plasmids have been constructed for varying purposes, such as recombinant protein expression, several cloning strategies with subsequent expression, expression with both N-terminus and C-terminus tags for detection and affinity purification (Section 3.1) (Porat, 2018). This vector has a *GAL1* promoter followed by restriction sites for several restriction enzymes such *EcoR1*, *Not1* and *Xba1* to allow for restriction cloning of a gene for expression. In addition to the restriction sites, there is also an ampicillin resistance gene for selection when bulking up the new construct in *E.coli.* Upstream of the *GAL1* promoter and the ampicillin resistance gene there is a gene named '*URA3*' and a pUC and 2µ origin to create and maintain a high copy number. *URA3* encodes an enzyme known as orotidine 5-phosphate decarboxylase (ODCase) which is involved in the synthesis of pyrimidine ribonucleotides such as uracil, without which a cell cannot transcribe any genes and thus cannot survive (François, Chapeland-Leclerc, Villard & Noël, 2004). Reminiscent of

the vectors discussed in section 3.1, there is a 6 x histidine tag as well as a V5 epitope included in the pYES2/CT vector to allow for purification of the recombinant protein via immobilised metal ion affinity chromatography (IMAC). The V5 epitope is based on the P and V proteins from the paramyxovirus of simian virus 5 (SV5) with a peptide sequence of GKPIPNPLLGLDST (Sivagnanam et al., 2010).

In this chapter the synthetic ORF1 gene inserted in the pOPIN *E.coli* vector suite was cloned into the pYES2/CT expression vector using restriction cloning. The new construct was then transformed into *S.cerevisiae* competent cells and induced for expression by manipulation of the *GAL1* promoter.

4.2 – METHODS.

4.2.1 - RESTRICTION CLONING 0F ORF1 INTO THE pYES2 YEAST VECTOR.

The ORF1 DNA sequence was first amplified via PCR using novel primer pair *M5AspORF1* (forward and reverse) from the *E.coli* vector poPINb . The primers introduced an EcoR1 restriction site to the 5' end and a Xba1 restriction site to the 3' end of the sequence. The amplification conditions were one cycle at 95°C for 2 minutes; 95°C for 40 seconds, 52.4°C for 30 seconds and 72°C for 2 minutes repeated for 30 cycles; final extension at 72°C for 7 minutes and held at 4°C after completion (DA ORF1). The PCR products were purified using a 'QIAquick PCR purification kit (250)'. The purified DNA was cloned into the Promega pGEM-T easy vector followed by a transformation into high efficiency JM109 competent cells according to the manufacture's specifications and protocols. Blue/white screening was used to select colonies that had been transformed with the correct insert. The white colonies were checked via colony PCR (DA ORF1) using a Sigma-aldrich 'REDTaq ReadyMix PCR Reaction mix' and *M5AspORF1* primers. A colony with the correct insert size was used to inoculate a 5ml LB broth and grown at 37°C and 150 RPM for 16 hours. The plasmid DNA was extracted using a Sigma-Aldrich mini prep plasmid extraction kit to the manufacture's protocol.

Restriction digests were performed using EcoR1 and Xba1 enzymes from New England Biolabs (NEB). Following the manufacture's protocol, a double digest of both the pYES2 vector and the pGEM-T/ORF1 was performed then run on a 0.7% agarose gel. After visualising the gel, the correct bands were cut from the gel and the DNA was extracted using a Monarch DNA gel extraction kit from NEB.

A Ligation reaction combined 5μl 2X ligation buffer, 1μl T4 DNA ligase, 50ng of vector DNA and the insert DNA added at a ratio of 3:1, the volume was then made up to 10μl with nuclease free water. The ligation reaction was incubated at room temperature for 1 hour. 2μl of ligated

DNA was added to 50μl of JM109 competent cells to be incubated on ice for 20 minutes. The mixture then underwent heat shock by placement into a 42°C water-bath for 50 seconds; then immediately returned to ice for 2 minutes. 800μl of SOC media was added to the mixture and incubated at 37°C for 1 hour with agitation. The transformation was then spread on LB plates with 50μg/ml ampicillin and incubated at 37°C for 16 hours. Colony PCR (DA ORF1) using *M5AspORF1* primers was employed to check that the insert was present. The colonies that were positive for the ORF1 insert were used to inoculate 5ml LB broths and grown for 16 hours at 37°C and 150 RPM. The plasmid DNA was extracted using a Sigma-Aldrich mini prep plasmid extraction kit to the manufacture's protocol. Figure 22 is a schematic outlining the structure of insert for expression in the vector.

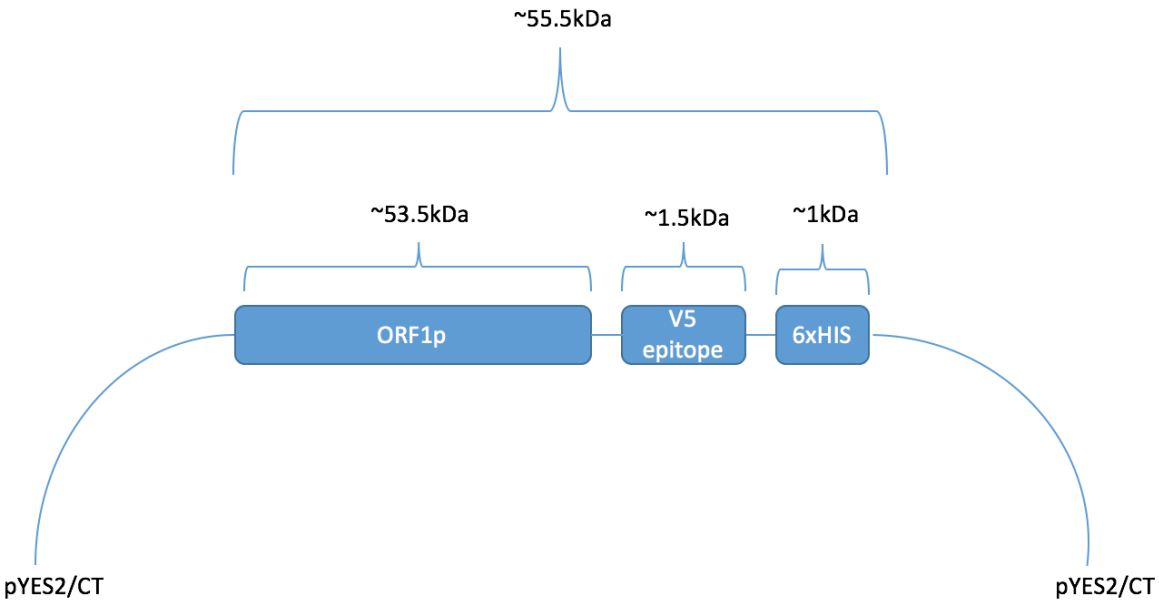

Figure 23: An image outlining the components of the pYES2/CT/ORF1 construct. To the far left is the protein of interest ORF1 ~53.5 is followed by the V5 epitope (~1.5kDa) and 6xHis (~1kDa).

4.2.3 – TRANSFORMATION OF *Saccharomyces cerevisiae* WITH THE pYES2CT/ORF1 CONSTRUCT.

The transformation was done using the 'S. c. EasyComp™ Transformation Kit' from thermo fisher scientific. A tube containing 50 µl of competent S.cerevisiae cells were equilibrated to room temperature and 1 µg of the pYES2CT/ORF1 construct was added. 500 µl of solution three was added to the cell mixture and vortexed vigorously. The transformation reaction was then incubated at 30°C for one hour whist mixing the reaction vigorously every 15 minutes. 100 µl of the reaction was then plated on to SC minimal media minus uracil (selective media) and incubated at 30°C for three days. The components of the media can be found in the appendices.

4.2.3 – EXPRESSION TRIALS OF FULL LENGTH ORF1p IN *Saccharomyces cerevisiae.*

For expression of ORF1p in *S.cerevisiae*, a single colony of cells transformed with pYES2CT/ORF1 were used to inoculate 15 ml of SC selective media with 2% raffinose (Starter media) and grown overnight at 30°C and 200 RPM for 19 hours. Following overnight growth, the OD600 of the sample was taken and the appropriate amount of overnight culture necessary to obtain an OD600 of 0.4 in 50 ml of media was calculated and removed. The equation used is as follows; 0.4 OD600 * Expression media volume (ml)/ 0D600 of starter culture. The culture was then pelleted at 6000 x g for five minutes before being re-suspended in 50ml of SC selective media with 2% galactose (Expression media) and grown overnight for 30°C and 200 RPM for 24 hours removing a sample after 3, 6, 12 and 24 hours of expression.

4.2.4 – LARGE SCALE EXPRESSION TRIALs OF ORF1p IN *Saccharomyces cerevisiae.*

The methods in section 4.2.3 were scaled up when performing large scale expression trials of ORF1p in *S.cerevisiae.* Rather than a 15ml starter culture, a 100ml starter culture in a 250ml conical flask was used to inoculate 500ml of expression culture in a two litre conical flask. After inoculation the culture was incubated at 30°C and 200 RPM. In order to optimise the

expression and later the lysis method, several trials were performed using varying lengths of expression including 16 hours and 24 hours. Following expression, the cultures were pelleted at 6000 RPM for 10 minutes, and re-suspended in breaking buffer (50 mM sodium phosphate, pH 7.4, 1 mM EDTA, and 5% glycerol, with added protease inhibitor cocktail). They were then pelleted again following the same conditions and frozen at -20°C.

4.2.5 – PREPARATION OF CELL LYSATE AFTER *Saccharomyces cerevisiae* EXPRESSION.

In order to lyse *S.cerevisiae* cells after expression, several methods were used in order to optimise the procedure. The first method of lysis was performed using acid washed glass beads. The frozen pellet was then defrosted and re-suspended in equal volume of breaking buffer followed by an addition of acid washed glass beads also to an equal volume. The mixture shaken vigorously in cycles for 10 minutes (30 seconds on, 30 seconds off) whilst keeping it as cold as possible in order to prevent damage of the proteins. Following the lysis step, the cell lysate was centrifuged at 15 000 x g for 45 minutes to separate the soluble proteins from the cell debris and acid washed glass beads.

The second lysis method was performed using 'Y-PER™ Yeast Protein Extraction Reagent' from Thermo Fisher Scientific. Cells were re-suspend in an appropriate amount of Y-PER reagent as indicated by the manufacturers' protocol before the mixture was vortexed gently until homogeneous. Thermo Scientific Protease Inhibitor Cocktail was then added to preserve the protein mixture before agitating at room temperature for 20 minutes. The cell debris was pelleted by centrifuging the mixture at 15 000 x g for 10 minutes (small expression) or 45 minutes (large expression).

The final method for lysis was a result of several optimisation trials. In this case the expression culture cells were removed from the incubator whilst till in logarithmic phase, pelleted at 6000 x g for 10 minutes, then frozen. After the cells were defrosted, they were re-suspended in Y-

PER reagent as indicated by the manufacturers' protocol. This was followed by agitation at room temperature for 30 minutes and sonication for 10 minutes with short bursts of 30 seconds followed by intervals of 30 seconds for cooling whilst kept on ice at all times. The cell debris was pelleted by centrifuging the mixture at 15 000 x g for 10 minutes (small expression) or 45 minutes (large expression).

4.2.6 - ANALYSIS of *Saccharomyces cerevisiae* EXPRESSION SAMPLES

After expression of the ORF1p in *S.cerevisiae,* the samples were run on SDS-PAGE gels to check for overexpression of the recombinant protein. After lysis, the soluble samples were mixed with loading buffer and boiled before being loaded and run on 12.5% acrylamide gels the stained with coomassie blue. To confirm the presence of ORF1p, western blots were carried out. The Western blot was performed by first repeating the SDS-PAGE gel then wet blotting the gel to transfer the proteins onto a PVDF membrane.

Following the transfer, the PVDF membrane was agitated in blocking buffer (2% Tween 80 in PBS) overnight at 4°C. The membrane was then washed in the primary antibody which was 6x-His Tag Monoclonal Antibody from Thermo Fisher Scientific diluted 1:3000 with PBS buffer. The membrane was incubated with the primary antibody for two hours at room temperature with agitation and then washed with PBS for 10 minutes five times. Once the primary antibody was completely washed, the membrane was incubated in secondary antibody (BSA antibody produced in rabbit) at a dilution of 1:300 for one hour at room temperature. Once the incubation was completed, the membrane was washed in PBS for 10 minutes five times. For Enhanced Chemiluminescence (ECL) detection 2ml of a mixture of equal volumes solution A (0.2mM coumaric acid, 1.25mM Luminol) and solution B (0.3% v/v H_2O_2) were added to the membrane and incubated at room temperature for five minutes. The membrane was the exposed to film in a cassette for five minutes before the film was developed to show areas of specific binding.

Due to unacceptable amounts of unspecific binding the Western blot procedure was repeated several times to optimize the process with varying dilutions of both the primary and secondary antibodies. Table 2 shows all the dilutions used in order to optimise the procedure.

Table 2: A table outlining the different antibody dilutions performed in order to optimize the western blot procedure.

	Primary antibody dilution.	Secondary antibody dilution
Original dilutions	1:3000	1:300
Optimisation 1	1:5000	1:500
Optimisation 2	1:6000	1:500
Optimisation 3	1:8000	1:500

4.2.7 - GLUCOSE INHIBITION EXPERIMENT.

An experiment was designed to check the activity of the GAL1 promoter by inhibiting it with glucose. The methods in section 4.2.3 were repeated but with two starter cultures and two expression flasks. Whilst one of the expression flasks continued with 2% galactose as the metabolisable sugar, the other was made with 2% glucose and they were grown for 16 hours at 30°C and 200 RPM. Samples were taken and cell lysates were made using the optimised lysis method before running on acrylamide gels for SDS-PAGE. To check for overexpression of the recombinant protein the ‘His stain’ method outlined in section 2.3.9 was performed on the samples from this experiment.

4.2.8 – CODON USAGE ANALYSIS

In order to investigate the effects of codon usage on the expression of AsM5 ORF1p in *S.cerevisiae*, a bioinformatics web tool was used to analyse the frequency of rare *S.cerevisiae* present in the DNA sequence. The web tool used was titled Codon Usage from the Sequence Manipulation Suite found at http://www.bioinformatics.org/sms2/codon_usage.html. The AsM5 ORF1 DNA sequence was put through the database which gave the number of time each codon was used to code for an amino acid. These results were then compared with a list of the eight least used codons in *S.cerevisiae* as outlined by the European Molecular Biology Laboratory (EMBL) at https://www.embl.de/pepcore/pepcore_services/cloning/choice_expression_systems/codons8/

4.3 – RESULTS

The restriction cloning of the ORF1 sequence from *E.coli* vector pOPINb to *S.cerevisiae* vector pYES2/CT was successful as demonstrated by colony PCR and DNA sequencing of the construct of the insert using sequencing primers. Once the insert was confirmed to be intact and in the correct orientation, expression trails were carried out. Expression trials initially yielded very little. SDS PAGE gels did not present adequate evidence that the recombinant protein was being produced in the expression media.

Western blots were performed to confirm the presence of the recombinant protein. Initial attempts of this procedure produced very unclear images. There was a large amount of unspecific binding of the antibodies, so much so that it was impossible to determine the expression of the recombinant proteins. The procedure was repeated and the antibodies were diluted as seen in table two. Even with the heavy dilutions of the antibodies there were many failures with the Western blot protocol including a high amount of unspecific interaction and binding to the membrane.

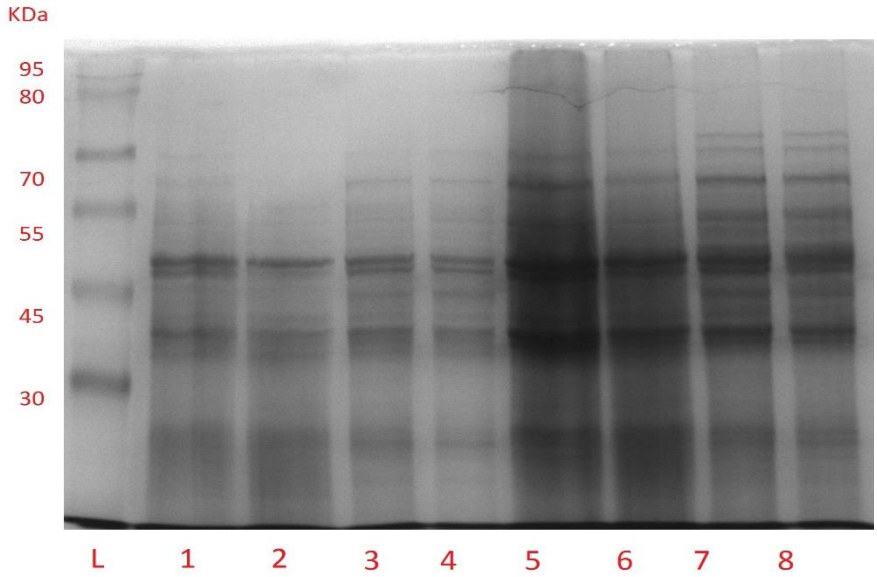

Figure 24: SDS-PAGE gel of the samples collected from the glucose inhibition experiment. Lane 1 = growth in glucose A, Lane 2 = growth in glucose B, Lane 3 = growth in galactose A, Lane 4 = growth in galactose B. Lanes 5 - 8 are concentrated versions of lanes 1-4 in the same order.

When troubleshooting the expression of ORF1p or lack thereof in *S.cerevisiae,* several experimental parameters were investigated one of which was the use of rare codons in the organism. Table 3 outlines the eight least used codons in *S.cerevisiae* along with the number of times that codon appears in the ORF1sequence cloned from the *E.coli* vector. The table shows that 63 of the amino acid residues in the sequence are coded for by one of these rare codons (16%) with CCG coding for proline the most frequent.

Table 3: A table presenting the 8 least used codons in S.cerevisiae and their frequency of appearance in the AsM5 ORF1 sequence.

Codon and respective amino acid of rare codon.	Number of times identified in AsM5 ORF1
AGG - Arginine	**5**
CGA – Arginine	**8**
CGG – Arginine	**6**
CGC – Arginine	**10**
CCG – Proline	**14**
CUC – Leucine	**0**
GCG – Alanine	**12**
ACG – Threonine	**8**

4.4 – SUMMARY

Structural and functional studies of the full length ORF1p expressed *S.cerevisiae* was hindered by the time needed to optimise the expression and purification protocol. Unexpectedly, the lysis of the cells after expression was one of the largest hurdles to overcome. The realisation that *S.cerevisiae* cells in logarithmic phase are much easier to break than cells in stationary phase informed the design of the final lysis method. Even after optimisation of lysis, expression of ORF1p was not confirmed by Coomassie blue staining or by Western blot. Western blots of the *S.cerevisiae* cell lysate were undesirably sensitive and repeated optimizations did not seem to confirm the presence of the His-tagged ORF1p. Another attempt was made to detect ORF1 using the 'His-tag' stain outlined in section 2.3.9.

This method also resulted in too much background information to confirm expression of ORF1p with any confidence.

Naturally, the idea that ORF1p was not being expressed at all was explored. When troubleshooting heterologous protein expression in *S.cerevisiae,* one of the most frequent discussions was the consideration of rare codons in the mRNA sequence of the desired protein. The methods used in section 4.2.8 yielded a result that showed 13% of the codons present in the ORF1 sequence cloned into the pYES2/CT vector were in a list of the eight rarest codons used by *S.cerevisiae.* The expression of the ORF1p is likely to be reliant on the resolution of this issue.

Chapter 5

DISCUSSION

5.1 – CONSERVED MOTIFS & DOMAINS.

5.1.1 – ZINC FINGERS

One of the main aims of this project was to use bioinformatics to identify conserved domains in the AsM5 ORF1; the background, methods and results of that analysis can be seen in chapter two. Chapter two presents the results of several conserved domain and motif searches in AsM5 ORF1 and its most closely related sequences from similar jockey clade elements in other *Anopholes* mosquitos. The conserved domain and motif searches were performed using two web tools; 'SMART' and 'MOTIF search'. Figures 8, 9,10 and 11 display the results for these searches, giving an insight into the conservation of the ORF1p function across species. The MOTIF search web tool (Figure 8) confirmed the three zinc fingers already identified in the protein as described in the general introduction (Figure 3). In combination, SMART and MOTIF search results showed that a zinc finger protein or a similar zinc ribbon was present in all of the AsM5 related sequences. Zinc finger proteins are present in an enormous variety of organisms and literature regarding their function is constantly expanding. The structure of a zinc finger is quite well conserved over time and the literature now classifies variations of the protein, the classical zinc finger motifs has a short beta hairpin and an alpha helix. First identified in *Xenopus* oocytes as a zinc-binding motif, zinc fingers were shown to be responsible for binding DNA in the transcription factor IIIA (McDowall, 2018). In the past few years, there have also been reports on zinc finger proteins that show RNA binding activity such as the HIV-1 nucleocapsid (CCHC) and reovirus s3 (C2H2) (Brown, 2005).

As outlined in section 1.4, work was carried out on the retroposon L1 showed that the ORF1p contained zinc-finger motifs. In L1, zinc fingers are associated with a nucleic acid chaperone,

which is critical for retroviral replication (Martin, 2006). Martin suggests that the zinc fingers play a role in 'copy' part of the 'copy and paste' mechanism of the L1 element. Research into site-specific non-LTR retroposons by Fujiwara 2015 using SART1 from the Tx1 clade showed that zinc knuckles do play a role in the mechanism for site specificity. The term zinc knuckle in this case refers to a group of more than two zinc fingers. Fujiwara inferred that as in retroviruses, zinc knuckle motifs participate in interactions between retroviral RNA and Gag proteins. Similar to the mutation experiments performed in L1, detailed mutation analysis showed that three zinc knuckle motifs from the ORF1p in SART1 are involved with its mRNA in a site specific manner, suggesting that the motifs may play an important role in all site specific non-LTR elements (Fujiwara, 2015). Though, this involvement may be due to the ORF1p-ORF1p or ORF1p-ORF2p interactions reported by Matsumoto et al (Matsumoto, Hamada, Osanai & Fujiwara, 2006).

5.1.2 – CPSF100

Another notable result from the conserved domain and motif searches was the identification of the CPSF100_C domain in AsM5 ORF1p. CPSF100_C was not predicted in any of the other closely related sequences and has been reported to be involved in the formation of a complex that interacts with histone-specific processing factors (Sullivan, Steiniger & Marzluff, 2009). The CPSF100_C domain is described by the MOTIF search web tool as the C terminus of a polyadenylation and cleavage factor. Studies of the complete CPSF100 protein in *Drosophila* have shown that it forms a core heterodimeric complex with the proteins CPSF73, Symplekin and interacts with histone specific processing factors (Sullivan, Steiniger & Marzluff, 2009),

though this domain identified in AsM5 is probably just a sequence that is extremely similar to the C terminus of full length CPSF100p. The role that this protein plays with nucleic acids and histone genes in particular was certainly worth investigating. *In vivo* studies have demonstrated that the knock down of CPSF100 in the complex outlined above caused histone

pre-mRNA mis-processing (Sullivan, Steiniger & Marzluff, 2009) adding some endorsement to the idea that AsM5 targets histone genes for site specific transposition. Though the complex described by Sullivan, Steiniger & Marzluff is also involved in the processing of poly(A) RNAs which of course could suggest that the activity of CSPF100 might actually more broad as oppose to histone gene specific. A conclusion supported by Kolev, Yario, Benson & Steitz when they stated that CPSF100 acts in a complex with other proteins in the process of maturation of most eukaryotic pre-messenger RNAs (Kolev, Yario, Benson & Steitz, 2008). After considering the literature regarding CPSF100, the association of the protein to histone genes made the identification of CPSF100_C as a conserved domain in AsM5 a notable discovery worth investigating. This was rationale behind the cloning and subsequent expression of the CPSF100_Cp domain in *E.coli*.

5.1.3 – 1D NMR ANALYSIS OF CPSF100_Cp

1D NMR analysis of CPSF100_Cp suggests that the protein is mostly disordered. NMR signals of individual residues are often variable depending on their chemical environment but are usually in the vicinity of the random coil shift value (Guo & Tugarinov, 2009). Chapter 3.3 explains that this 1D NMR spectrum is typical of an unstructured protein due to the presence of tall, sharp and undispersed peaks which suggests that the protons in residues CPSF100_Cp are not variable due to a lack of shielding. When looking at a folded protein, dispersed peaks are clearly visible as a result of such shielding and the disorder of this protein could be attributed to several factors. It is well documented that some proteins remain in a disordered form until they interact with other proteins or ligands. It is conventionally understood that many proteins are intrinsically disordered in native form and fold upon binding, though this is not true for all proteins, as disorder can also be found in the bound state (Fong et al., 2009). Another reason for the disorder in this protein is the repeated presence of the residue proline. Proline is an unusual residue because its side chain folds back on to its amino terminus and forms a ring with the backbone of the amino acid. This structure causes changes

the bond angle of the peptide bonds that proline forms with other amino acids which in turn affects it ability to form the hydrogen bonds required for alpha helices in secondary structure. This hindrance to the formation of alpha helices makes proteins with several prolines prone to disorder. 10% of the CPSF100_Cp sequence is made up of prolines and over half of those prolines are predicted to be a part of disordered protein binding.

5.1.4– THE OTHER C-TERMINAL CONSERVED DOMAINS

The web tool conserved domain searches using SMART and MOTIF search identified CPSF100_C as the C-terminal conserved domain in only AsM5, though other conserved domains were identified in two of the three AsM5 closely related sequences. TFIIF_alpha in *Anopholes farauti* and FAM104 in *Anopholes dirus* were also indentified in figure 11. A multiple sequence alignment of CPSF100_C, TFIIF_alpha and FAM104 showed very little similarity between the sequences suggesting different associated functions which were then confirmed by literature. TFIIF_alpha is described as a subunit or associating protein of RNA polymerase II involved in stimulation elongation of nucleic acid sequences (Funk, Nedialkov, Xu & Burton, 2002). FAM104 is a very under reported domain with all online database providing minimal information and only describing it to a part of a family of proteins found in eukaryotes.

5.2 – FAULTY CPSF100_C EXPRESSION IN MINIMAL MEDIA

In an effort to study the folding of CPSF100_Cp through nuclear magnetic resonance (NMR), the protein was prepared for ^{15}N labelled expression. After expression of CPSF100_Cp in LB media was confirmed and the fusion protein was purified, the pOPINK/CPSF100_C construct was used to express ^{15}N CPSF100p in minimal media containing ^{15}N ammonium chloride. The results for the expression are presented in figure 19, where SDS page analysis shows the expression of a protein approximately 27 kDa in size. As outlined in the results section, the ^{15}N labelled protein purified was ~11kDa smaller than expected. After further analysis, it

became clear that this this was the expression of the GST and His tag without CPSF100_Cp. Even after the integrity of the construct was checked via PCR and DNA sequencing, the construct continued to express normally in LB but the fusion tags without the CPSF100_Cp in minimal media. Expression of recombinant proteins in minimal media can often require re-optimisation of the expression procedure as the availability of nutrients does sometimes affect the cells ability to not only produce complex recombinant proteins, but to thrive and survive. *E.coli* cells cultured in rich media such as LB grow much faster and are consistent with well-known patterns of protein synthesis in rapidly growing cells. In contrast, *E.coli* cells cultured on minimal median grow much slower and show a different pattern of gene expression and regulation. Cells grown on minimal medium display elevated gene expression of sequences involved in biosynthesis of building blocks. Most notably, almost half of known RpoS related genes are expressed at higher levels in minimal media than they are in rich media (Dong & Schellhorn, 2008). The RpoS gene encodes sigma factor S, which is essential for the transcription of a range of stationary phase and stress resistance genes (Hengge-Aronis, Lange, Henneberg & Fischer, 1993). Essentially, when growing in minimal media *E.coli* cells become far more concerned with survival than growth and division. The issues regarding fusion proteins expressed in minimal media are poorly documented in the literature; especially true for this particular issue of free-tag expression.

The lack of directly relatable troubleshooting information in the literature made it clear that free-tag overexpression is a rare experimental problem. As outlined above, *E.coli* cells growing in minimal media are more invested in survival than thriving. GST is a strongly expressed non-peptide fusion tag and it's expression in *E.coli* is consistent and robust unlike the newly cloned CPSF100_Cp. It is possible that whilst dealing with the stresses of switching from rich media to minimal media, the host cells do not prioritise maintenance of the infrastructure required to express complex non-essential proteins. Another possible explanation for the free GST expression is simply that the GST-CPSF100_Cp protein might be cleaved and degraded by cellular proteases. This behaviour is observed in expression

using non peptide tag maltose binding protein (MBP) as a solubility enhancer (Korepanova et al., 2007).

5.3 – OPTIMISATION OF *Saccharomyces cerevisiae* LYSIS PROTOCOL

Optimisation of the *S.cerevisiae* lysis method was an unusually time consuming portion of this project. The *S.cerevisiae* cell wall is a strong and sturdy structure that provides physical protection amongst other features. The position of the cell wall is in two parts, the inner wall and the outer wall. The inner layer is mainly responsible for the rigid strength of the wall, consisting of β1,3-glucan and chitin which represent approximately 50–60% of the wall dry weight. The outer layer's duty is mainly rooted in performing cell to cell interactions and consists of glycosylated mannoproteins (Klis, 2002). The pYES2/CT vector manual recommended a procedure for the lysis of *S.cerevisiae* cells but did not give sufficient warning to the consequences of growing the culture to saturation. The recommended lysis method was extremely ineffective for disruption of the cell wall when the culture was allowed to grow into stationary phase. The literature makes it clear that cells in the stationary-phase have thick, less porous cell walls making it very difficult to lyse them using the glass bead method outlined in section 4.2.5. The thickened cell wall of stationary phase cells are not only tough enough to withstand shaking with acid washed beads but they are also resistant to digestion by the enzymatic activities of some enzymes such as zymolase (Werner-Washburne, Braun, Johnson & Singer, 1993). This resistance to enzymatic activity and the risk of protein degradation are the reasons enzymatic lysis of cell wall was not chosen as an alternative. As stated in section 4.2.5, the final lysis protocol was mainly devised from the realisation that cells in the logarithmic phrase are far easier to break open than cells in the stationary phase.

5.4 – CODON USAGE

After the optimisation of the lysis protocol, there were still some experimental hurdles to overcome in the expression of the full length AsM5 ORF1p in *S.cerevisiae.* After several expression trails, there was still no promising evidence that AsM5 ORF1p was being expressed. One the issues considered when troubleshooting the failure to express the protein was codon usage. Living organisms are subject to a degenerate genetic code, as several codons are known to code for the same amino acid (Sharp et al., 1988). Nucleotide, codon and amino acid preferences are subject to variation among genes and organisms. Codon usage preferences occur because there are 64 codons but only 20 amino acids to code for (Hamady, Wilson, Zaneveld, Sueoka & Knight, 2009) and there is on-going speculation on the evolutionary powers that drive these codon preferences context (Gustafsson, Govindarajan & Minshull, 2004). It is now well known that synonymous codons are generally not used with equal frequency (Sharp et al., 1988). Because of this, codon optimisation has been used as a technique to improve fusion protein expression. The consensus is that highly expressed proteins are typically encoded by genes with optimal codons due translation efficiency and mRNA stability. Zhou et al., 2016 also stipulated that in general, the overall translation efficiency of an mRNA sequence is mainly determined by the efficiency of translation initiation (Zhou et al., 2016).

The AsM5 ORF1 sequence inserted into the pYES2CT yeast vector was cloned from the synthetic AsM5 ORF1 sequence originally codon optimised for expression in *E.coli* and the pOPIN vectors because proteins are often challenging to express outside their original context. One of the reasons why codon bias affects heterologous expression is reported to be 'because preferred codons correlate with the abundance of cognate tRNAs' (Gustafsson, Govindarajan & Minshull, 2004). An illustrative example in *E.coli* is the tRNA that infrequently reads the codons AGG and AGA for Arginine ($tRNA_4^{Arg}$), is found in very low concentrations within the cell (Bulmer, 1987). Table 4 portrays the eight rarest codons used in *S.cerevisae* as

outlined by the EMBL and the frequency in they which they appear in the AsM5 ORF1 sequence. In total, 63 of the amino acid residues in AsM5 ORF1p are coded for by these rare codons and make up 13% of the protein's amino acid residues. Presnyak *et al* demonstrated that codon usage is the main factor for RNA stability in *S.cerevisiae* due to its effects on gene translation. This leads to the a possible conclusion that these 63 amino acid residues coded for by the any of the eight least used codons in *S.cerevisae* lead to dire consequences in the expression of this fusion protein.

5.5 – FURTHER WORK

Successful protein expression experiments are heavily reliant on the preliminary work of cloning and expression trials. The smoothness and speed of these stages usually set the pace for the work that is to come. Unfortunately, work in this project was limited by the time needed to perform cloning experiments and expression trials. The inability to express ^{15}N labelled CPSF100_Cp was particularly limiting as it denied an NMR spectra with which real progress could be made when working to solve the structure of the domain. Nonetheless, 1DNMR of the CSPF100_Cp did inform that the protein is very disordered in its native state and experimental work should be done to investigate any change in formulation after interaction with other proteins or nucleic acids. The other future work required for this project would include overcoming the inability to express ^{15}N labelled CPSF100_Cp and the re-cloning of the AsM5 ORF1 sequence in the pYES2/CT vector. Re-cloning the vector to remove those rare codons might make it easier for the cell to express the protein. Expression of the full length AsM5 ORF1p could be used to investigate its interaction with histone gene mRNAs via pull down assays and even protein to protein interactions with the chaperone proteins that guide them.

5.6 – CONCLUSION.

In conclusion, AsM5 ORF1 is identified to possess interesting conserved domains and motifs, which could lead to a better understanding of the element's site specific retrotransposition. The Zinc fingers identified are present in other retrotransposons both site and non-site specific. The CPSF100_C domain identified towards the N terminus adds more evidence to the idea that AsM5 targets histone genes during transposition. Progress was made in the heterologous expression of full length ORF1p after cloning the sequence from a pOPIN vector into pYES2/CT. Though there was no conclusive evidence that the protein was expressed, information from troubleshooting certainly brings the work closer to completion. CPSF100_Cp was expressed in *E.coli* and purified with the aim of performing functional analysis using nucleic acids or other proteins. 1D NMR analysis of the CPSF100_Cp supported data obtained from protein disorder prediction software that showed the protein is highly disordered in it native form.

CHAPTER 6

APPENDIX

6.1 – AsM5 ORF1 DNA SEQUENCE

ATGCCGAAGCGTGGTAGAAGGAGGAGTAGGAAGCCAGACTCGGTAGGATCGAGCTCCGATTCGGCCGA

GTCCAAGCGCTCGAGGAGCGTGGTTTCTTCCTCTTCGGAGGAATCTGACGATACGATGAGCGTGGACA

GTACGGAGTCACGTTCATCCACCAGCGTGCAGGAGGATAACCTGGCACAATTTGTCACCGTAAACCGG

CGACAACGGAAGGCAGTCCCGACAACGAAGCCTTCCACAACACCAGCTACGCCCGCTGTTCCTGCAGG

GCGTACATCGGCTCCGGCTCCCGTATCGGCGGCAAAAATGCCTCCGATCACGGTGAAGTCACTCCCAG

TAGCTGTCCTGCGTCCGGAACTGCAGGCTCGTGGAATCACACCAGAGTTCCGTATCTCCGGCGTAGGC

ACGTCAATCACCGTTCGATCTCCTGCTGAACAGCAGGAGGTCCTTAACTACCTGCAGCAGCGGAATGC

GGAATATTTTTCGCATGACGCTAAAAACATGCGTCCCTTCAAGGCGGTGCTTCGTGGGCTTCCGGAAA

CGGACCTCGCGGAGATCGTTTGTGAACTGAAGGAAATCCACCAGCTCGACGTTTTGGAGGCGTTCGAG

ATCAAGCGCCGCGCAGAGGGCATTCAAACCAGGTTGTACCTGGTTCATTTCAAGCGAGGAACATGCTC

GCTAAAAAAGCTGGAGGCAGTACGGTCAATCCAGCAAGTCATCGTGCGATGGGAGCCGTACCGCGGAG

GGAAGAAAGGCCCGACGCAATGCCATCGATGTCAGGCTTTTGGGCATGGTACTCGCCATTGCCAAATT

AAACCTCGGTGTGCCATCTGCGCGGCGGAGCATCTCTCGGAGCAGTGTCCATCGAGTTCGGGCACAGT

AAAGTGCTCGAACTGTGGTGCTGCTCATCGCGCCGATGATCCGTCGTGTCCAAAGCGGGCCAAATACA

TTGAGATTCGTCAGCGCGCCAACGGTCGAAACTCTGCTCCTCCACCAGCCAAAGCTAACGTGTGGCAC

GCGCTTCCACCGTTAGCCACCATCCAAACCACACTCCCACACTCCATTCCTCCTCCGGTGTTGCACAC

TGCTCCCAAGGCCAAAAGCTTTGCGCAGATTGTGTCAGCACCAACCACTCCAAGCGTCCGACCTGCGG

CAGCGCGCATCCCACAACCTAACCCAACAGTACCACAACCTAAACCAACCTCTTCTTCTTCCTTGCAA

TCCACAGCACCGAGATACAACCTTGCGAAGCGACTGCAGGACATCAAGAACGCTCCAGACACACCAGC

TACAACACCAACTACAACTCCAGCCACAACTTCATCGGAAGACCTGTTTAGCCCGGAAGAGCTATTCG

CTATATTTAGCAGAATGCTCCCGAAGATCCGCCTTTGCCGCAACAAGGGAGAACAAATCGCCGTTATC

GGAGAACTATTGATGCTCCTTCACTGA

6.2 – AsM5 ORF1p AMINO ACID SEQUENCE.

>An_steph_M5_ORF1

MPKRGRRRSGKPDSVGSSSDSAESKRSRSVVSSSSEESDDTM

SVDSTESRSSTSVQEDNLAQFVTVNRRQRKAVPTTKPSTTPATPAVPAGR

TSAPAPVSAAKMPPITVKSLPVAVLRPELQARGITPEFRISGVGTSITVR

SPAEQQEVLNYLQQRNAEYFSHDAKNMRPFKAVLRGLPETDLAEIVCELK

EIHQLDVLEAFEIKRRAEGIQTRLYLVHFKRGTCSLKKLEAVRSIQQVIV

RWEPYRGGEKGLTQCHRCQAFGHGTRHCQIKPRCAICAAEHLSEQCPSSS

GTVKCSNCGAAHRADDPSCPKRAKYIEIRQRANGRNSAPPPAKANVWHAL

PPLATIQTTLPHSIPPPVLHTAPKAKSFAQIVSAPTTPSVRPAAARIPQP

NPTVPQPKPTSSFSLQSTAPRYNLAKRLQDIKNAPDTPATTPTTTPATTS

SEDLFSPEELFAIFSRMLPKIRLCRNKGEQIAVIGELLMLLH

6.3 – ORF1p AMINO ACID SEQUENCES

>Ae_aegypJuanA_1_ORF1

LNMVSTTNKRKGESLNSLLPSKKVGFKTVTTRGKNGRKDASPECEVSSKG

EMNNCIEMSNQFDALDKFSEHQIEAASSPGSLIQVRKQRVPPIVVSCSEF

GGFRQEILNSIRGIKVSFQIAKKGDCRVLPETLKDRELLLKHLEEKKHKF

FTYDDKTERLFKVVLKGLSSDYKSPEEIKNGINDLLGFSPVQVIIMKKRT

QSGIVRKGLSQEFYLVHFNKKELNNIKALEKAKLLFDVRVTWEHFQKPGG

NYQNPTQCRRCQKWGHGTKNCRMDAKCMICGGSSHAKDVCPVKEDTTKFI

CCNCGANHKSNFWNCPSRKKVIEARARQMKDNIRYDNGRFRNLPGRVSNN

AHFSVNDRLIMNHTHQEDHNHAHSQTNFIPSGSRSNLSISNVSTHGKSFA

DIVAGNSNSSPVRSMGTHSTCFKSNGKNPTATGNSASSSTGNSNGKSHDM

SASDFNFLTEQLNLMIDAMFKATTMTEAVQVGVKFTNQIVIGLRFSNGSK

>Cx_pipJuanC_0_ORF1

ENFACKMRQNKGKRKSSEDLVVTSVKRLNRKPANANRKRKQPLLRSDSDS

ECEVNPPIPLTNSFGVLSETDDKEPSPRTEPSAVEKRVKAPPIVVTSVSD

LASFRTQLKNCKETCNLKVSFQLGRRGECRLLTESLQDHQTFVGYLKNHK

HNFYTYETKNARPFKAVLKGLSNDLSVDEIKNELKVLLGFAPSQVIPMKK

KSNGNISRFGLTSQFYLIHFNRNEINNLKILDKVQFLFHVRVKWEHFKKH

GGNGQNLTQCRGCQAFGHGTDHCAMVPKCMVCGDSSHDKDNCPVKEVTQF

KCANCGGNHKSNFWDCPIRKKVLDSRAKHQPKSKPKFSQSQVVPASLNQT

FVLSHSNNSRNTPTVEKLGNNNGISYANVVSGSSTNFKSSTNLSEIGQVP

QISFENFSAGNALGSSDLGDVTFEKMTFLQNSLFGLIQTMSNATSMMEAI

QIGLKFANDVVLTLKFNHGSK

>An_nili4346_ORF1

AEVESEKRSYLLRSACDARSALQSAINDHMGRGPGKRLRPSSSSSEEISL

SETDSSSEDXXSCSETSSSSEDDSSIRSVMEFDTQEQPFIEVTAKKPPKP

KAKPAAAPVSAAARPAAPTQTTTSAPPTAAKSAKIPPVVVRSPAPHELRK

IFASFRGIQFKITGAGTQILPPNIEVHRAVTEHLALLKHEYYTHDFVGDK

PYKVVLRGLPITNEEEILSELREIHGLTPTAAYRIKRRHEVEGSHHSCLY

LIHFKKGTCTLQTLRAIRAVGSIIVRWEEYRGGRPSVTQCFRCQGFGHGT

KHCHMRPKCAKCSKEHLTDQCDQEAVSPKCANCGGTHHGRDLTCPQRAKF

KEIRAAASNKQLKKQLRASQPVPAPPQLASNKAFPPLAPPSKAPTAKTPK

PAIPPGLEYAFMAKQSGAAPLPSAAEPVETTEGAEPHDLASLFKIFIAMK

ARLLQCRTRMDQMTIIAELLLTHG

>An_nili3153_ORF1

PFGSLILYQTKILYSYSECSTVSAASAMGRKKKKIAKKEVVSSDAASKPP

KVPTNTSLEVVDLVLDALHHRPEASSPVAVATALPVASAPKRRHLSSSNS

SSDSDTCVPCSSEMNTSSDSDATEGNDDEFMVVNSRRKPQKPKLPTPAPA

SIPAAAKAAPVAASSSSARKIPPIYVKSPPFSQLRGELNRNIGSGFDMAM

RGVGVRITTKRIEVHRSIRQYLDSIKAEYFSHALVEEKPFKVVLRGLPRD

CEGEIAQEMKAVHQLEVSAVHRIGRPGDDQNRHHSVLYLVLFKKGATTVP

QLQNIRKLCDLLVKWEAHRGGPRTVLQCRRCLRFGHGAANCKLPMQCANC

SKEHNESVCVAVPPEAPKCSNCGQNHKATDPACIIRVRTLQQRVQPPAAA

AHAARPPPPPITSASFPPLAPRRNPVPAAPAPVPKPRFTANSRLVAAAAS

PDVVAVPKAIIPKHVPSMPVSAKPFAAVVKSPALPVAPTPPAPSVDDVFY

NEALTLINNTLFKIHSQMVALLRSCSSRADQLAAITEFANRYG

>An_merus_9255_ORF1

SRVDCGAEQTCFLSVRQIAKRRQLAYDSVCAPRCSAVMKRMSGRKENQQE

RSRSNSQNSNESKRARIKTQDAYDETVSTENDEFTQVWAKGRRQASNVLM

DVNVEASTSAPTKLTSKPNGKLPPIVVKSMPLASLRPELQSRKLYVEYQL

SGIGTKIFAKSLADHRAIISLLEGKKVEFFTHDLKEDRPFKAVIRGLPLI

ETEDIVDELKVNYNLEVTEVFRIKRKNEENQSYHQQLYLAHFKRGSCSMK

KLETVRTIQSVIVKWESYRGGHKGPTQCLRCQNFGHGTRNCRIQPHCAVC

AESHHTDSCNAKNNVDATVKCANCGENHRARDVTCPQRTKYQQIQILANQ

KIRRNHSSAAKGNQRAPPPPLSSTEHFPLTGMPSSAPSSSAFPRKNTNTQ

VPPGFQYNLAQRLINAQTTIPEPISTQQENLYDATTLMQIFKEMSTKLRS

CRTKADQITVLGELIITYG

>An_gamb8812_ORF1

IDVILRALGWDEPSCTCIRGPCYRFAILESRVDCGAEQTCFLSVRQIEKR

RQLAYDSVCAPRCSVMERMSGRKENQQERSRSNSLNSNESKRARMETQDA

YDETVSTENDEFTQVWAKGRRQASNVLMDVNVEASTSAPTKHTSKPNAKL

PPIVVKSMPLASLRPDLQSRKLYVEYQLSGIGIKIFAKSLADHRAIISLL

EGKKVEFFTHDLKEDRPFKAVIRGLPLIEIEDIVDELKVNYNLEVTEVFR

IKRKNEENQSYHQQLYLAHFKRGSCSMKKLETVRTIQSVIVKWESYRGGH

KGPTQCLRCQNFGHGTRNCRIQPRCAVCAESHHTDSCNAKNNVDATVKCA

NCGENHRARDVTCPQRTKYQQIQILANQKIRRNHSSAAKNNQRAPPPPLS

STEHFPLTGMPSSAPSSSAFPRKNTNTQVPPGFQYNLAQRLINAQTTIPE

PISTQQENLYDATTLMQIFKEMSTKLRSCRTKADQITVLGELIITYG

>An_merus2401_ORF1

RVHAALKTLGSYHSVRYCTSDIFERRGTVAYELSNRSDVFFGCCARTTNT

YNSNFEIVCVCAIWPGAMKRRDRGKENEQNRSRSNSESSRDSKRSKINVV

NSVEEREALTTTSMDTGEFFEVHRKGQKRVQNAQPIVNDGASTSASPKQT

TQTNTRLPPIVVKSLSLASLRPELQARKLYAEFQLSGIGTKIFAKNLADH

FTIINMLESKKAEFFTHDLKENRPFKAVIRGLPLMEIDDIIDELKISYKL

EVTEVHRIKRRDETNQNYHQQLYLAHFKRGSCSMNKLQAVRTIQSVIIKW

ESYRGGHKGPTQCLRCQGFGHGTCNCRILPRCAICAEPHLTDTCNVNNPQ

STVAKCANCGANHRARDVECPQRAKYQQIRKLANERGHRHHTAAEKPRAP

PPNLSSQVHFPSAGMPSSAPSSSALPQKNHTMQVPPGFQYNLAQRLIDAQ

KVASEPIPTENENLHDTTTLLQIFKEMSSKLRACKTKADQIAVLGELIIT

YG

>An_epiro_5584_ORF1

LCEKFSVEMGRKKKRDRNKDRSDSSSCESIASKVSCVTEAFEADMEDQNI

DTEVEEFIEVLPRKIKGKTSGAEDLNNFGASTSKPSNSADNLPRKLPPMV

VKSLPLSIIKPQLSSRRIQAEYQLCGIGTKIFVHTKENRSEVINFLKQHG

VEFFTHDLKEERPFKAVIRGLPLMEIQELKDELVHLYQLDVLEVHRIKRR

NEETTNYHHQIYLVHFKRGTCTMNKLQEVRTIQSVIIQWESYRGGHKGPT

QCLCCQGFGHGTRNCNVKPNCANCAENHLTSECPTSNVEGTVAKCVNRGQ

NHLVRDAQCPQRRKYQEIRKLAADRSRKQPAVPRAPVPAISSMQHFPAMP

MGMPPATSSSTAWRKKPAAPAVPPPPGFQYNLQQRLIDAQNINTEPESTE

ELHDASTLMNIFRLMSSKLRKCRSKIDQITVLGEMIIQYG

>An_epiro_5584_ORF1p

LCEKFSVEMGRKKKRDRNKDRSDSSSCESIASKVSCVTEAFEADMEDQNI

DTEVEEFIEVLPRKIKGKTSGAEDLNNFGASTSKPSNSADNLPRKLPPMV

VKSLPLSIIKPQLSSRRIQAEYQLCGIGTKIFVHTKENRSEVINFLKQHG

VEFFTHDLKEERPFKAVIRGLPLMEIQELKDELVHLYQLDVLEVHRIKRR

NEETTNYHHQIYLVHFKRGTCTMNKLQEVRTIQSVIIQWESYRGGHKGPT

QCLRCQGFGHGTRNCNVKPNCANCAENHLTSECPTSNVEGTVAKCVNCGQ

NHLVRDAQCPQRRKYQEIRKLAADRSRKQPAVPRAPVPAISSMQHFPAMP

MGMPPATSSSTAWRKKPAAPAVPPPPGFQYNLQQRLIDAQNINTEPESTE

ELHDASTLMNIFRLMSSKLRKCRSKIDQITVLGEMIIQYG

>An_epiro6062_ORF1

MVKKRKERHRNSSESSNDSIASKVSRVMEESAGDMDIGPYDQDADTKVDD

FIEVVSRKQKRKASPAEQAVNYGTSTSGGHYGASTPSATKRYGASTSSEN

RSASATTKKFPPIVVKSLPLFVLRPQLNLRGLRVEYQLSGMGTKVFVHSK

DDRCAVLNFLKENKVEFFTHDLKEERPFKAVIRGLPLMETEDVKAELVQE

YQLDVLEVHRIKRCNEETTNFHHMVHFKRGTGTLNKLQAVRTIQSIIVRW

EPYKGGRKGPTQCLRCQGFGHGTRNCYTVPKCANCAKEHPNEECPTNNIE

GSVVKGINCEESHKARDVECPQRTKYLQIRKLAADRTKQKSTTSAPRPPP

PKPSSVEHFPPMVKTMPTATPSSTAWSKKPAAPTVPPSFHYNLQQRLVDA

QRTEPEPEPVDEIHDAITLMKIFKEMSFKLRKCRTKIDQITVLGELIIQY

G

>An_chris6548_ORF1

KMPTIEVKTMHLASLKPALQSRNINAMYQLTGIGTKVFVKTKQEHEAVIG

FLETSCVEFFSHEMKEDKPYNVVIRGLPLIELDDIIAELTEQHQLQVLEM

FRNKRRNEENQPYHNQLYLEHLQRGSCTLAGLQTIKSVQSVIVRWEVYQN

SHKGPVQCGRCQGFGHGTRNCRLKPNCATCALDHLTDMCPTKEEPQTMKC

KNCMGPHWANSGTCHLRTKYIEICQQASTRSRKQPVQPRAQLPPMTLQNF

PHLPIITPPCAYRSTMRSIVAPVILPGIKYNQIQRLAAAQRNEPIQTVED

SLYDASTLMVIFKKMAIKLKGCRTKHHQIAVLVEL

>An_epiroticus_7070

TKVSKPETKRSGVKPKRKRSPSTSSNSSSSSHFVTEYEVSSSEDTLTSAM

ETDEEGFQRVTAKKGEKKSIKQMKNNPASTINTAVANTLPPSCAIPATPS

TSRAMNPTPSATKTPSAANVSDKSFQRKLPPIVVKNLHIATLKPELTKRN

INAIYQLSGIGTKVFVKTKADFDTVKSFLAENQVEFFSHEIKQEKPFKAV

IRGLPLLELDDIKNELVEEHQLQIVEIFRIKRRNEEVQAYHNQLYLVHFK

RGTCTLAGLQTIKSIGSIIIRWEAYRNGHKGPVQCGRCQSFGHGTRNCRL

KPKCAICSLEHLTEVCATAEEPASTKCTNCNGPHRANDTSCPQRTK

>An_atrop5972_ORF1

ARKINAEFQLTGIGTKVYVRTRTEYVAVLALMENSKAEFYTHEIREERPF

KVVIRGLPYMDTNDIADELRVYHGLETREIFVIKRRNEGKRTFHHQLYLV

HFKRGSCTLASLQAIRSIQSVIVRWEPYRGGRKGPTQCLRCQDFGHGTRH

CRLQPRCANCAGNHLTNDCSANTEEVNKCANCQGNHRANNVECPQRAKYQ

EVRKLASSRGQTRKPLPSSATSTSKSPAPPALQPTEVLPAPVPTVLPPVT

SENPWTKVKVNTPKIPPGFQSNVAHRLSQPERPKSVPAAAGDLPPETEEL

HDAATLMLIFQEMTTQLRHCRTKLEQVTVLGRISIRYG

>An_sinen3038_ORF1

CQNYSSGLKWSVPPFGIMPTEKDENKKRTARSGSSGSEEGEAKRKVTGTG

VTPVPDANMAASSDEAQEGFQPVRSRSKRGSIGEASSSANPAAAPTVGKK

AAPGGSTVPIVTTRRPPPVVVQNISYAALRQKLHARNIDAEYQFGSIGTK

IFVKTRQEHTALIKLLEHTKTEFFTHDLRDDRPFKAVIRGLPLLETDEIP

DEPRDSYALEVEEVFRIKRRAEDKISYHNQLYLVHLKRASCNLTTLQTVR

SIFSVRVKWEPYRRGPRGPIQCHRCQAFGHGARNCHLPPKCVNCSLPHFT

ANCQQPVPPKCANCGEEHEAKSPECPHRTKYLEIRQKAMAGKSKKRSTNR

PATPPPPPVTVAAFPSLPNKPQLPSSAPSVALSGLANPSAIPPGFEYAAK

AKNVPPVSDHNLNPGQPEAPLYDSATLMQIFMDMTERLSSCRSRRKQIAV

LGEIIIRYG

>An_merus2390_ORF1

PLTGIHRIFCAPEKSLTKPVLFEFDFILTSSSIMSGKGKKHTKRVRDDSS

ESDSEESKCSVRRVSSLDAASKRMAVDLPENNKSVIETCTSDTMDQEAIR

SEFTVVTRRKKNNSIRMSTDKTSTGLSVNPPAVPTAPLAGQNVSTSPGPT

YKPPPIVVKTISITELRPELQSRGFKPKFRLSGIGTSIIACSKTEYDDII

KYLHERKAEFFTHDAKQDRPFKAVLRGLPEMEIHEIEEELRGVYQLDVIE

TFEIKRRINTIHSRLYLVHFKRGTCSLKKLEMVRTIQQVIIRWEPYRGNK

KGPTQCHRCQDFGHGTRHCNINPRCALCAGQHITDICPTKDQQEALKCSN

CAGPHRADDQTCPRRTKYVEIRQQASRRQTQHKNLPNLGIRPINLKPLAS

APPAITSIPVTPSPTIRPSPPQSSIVLPSNRTVPEVTRSPPGFITLAQRL

ENARNAPDVSTSAADGELFSMEELFNIFKKMISKIRLCRNKTEQLAVIGE

LLMLNG

>An_gamb6977_ORF1

NRFSFCHARFSASSSIMSGEGGKHAKRGRNSCSDSGSDSGMSKRSVRRIC

PTADNDALAQRLNNENNSINDTSSSDTMEQEEISQEFTLVNRKKGSARRR

STQKTSAGPSANPPAASATPLAGRKVNPTSESNFKPPPIVVKTIPVAELR

PELQSRGFTPQFRLSGIGTSIVTRSRSEYDGVVKYLQERKAEFFSHDAKQ

DRPFKAVLRGLPEMDIPEIVEELQGRYQLEVLEAFEIKRRIDTIHSRLYL

VHFKRGTCSLKKLEAVRTIQQVIIRWEAYRGGKKGPTQCHRCQDFGHGTR

HCNIQPRCARCAGQHITEACPTKDQNEALKCSNCSGPHGADDPTCPRRAK

YVEVRQQASRRQAKHQHPPNPVHRQIPPRPLITAPPMVRPPPATSTHTAT

QPSSSVMQTNRTASIEPNTPPGFITLAQRLENARNAPDTPTPVSEANLFS

MQELFSIFTKMMSKLRLCRNKAEQLAVIGELLMLNG

>Afunest8222_ORF1

TSRFSCSRSVTRKTDFLLFSCALPSLSTMSERGEKQVKRGRKNASDNDSG

SDSGESKRSMRRVFSPTLENNNMTIEHSFYETSTSGTMEQEEVHDEFRVV

KHRKKMKNTTKNTTTAKISAGPSANTPAAPAVLQANRNARSSPSSICKPP

PIVVKTIALSELRPELQSRGLTPEFRLSGIGTSILTRCKADFDGVLKYLR

ERKAEFFSHDPKEDRPFKVVLRGLPKMEVHEIVEELRECYQLSVVEAFEI

KRRAQNIHSMIYLVHFKRGTCSLKKLEAVRTIQQVIVRWEPYRGGKKGPT

QCHRCQDFGHGTRHCNIQPRCALCAKEHLTDKCPTKEHATLKCSNCAGAH

RADDPTCPRRAKFLEVRQLXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXTSTSKGNLFSMNEVTI

RTNSILSKIRLCRNKANQLVVIGELLMLNR

>An_sin_4224_ORF1p

VGSPLSSDCASGLYLTRRKAAYISYAAKDQRPFKAVLRGLPAMPLDEIRS

RYQLDVTEAFEIKRRAEGIHSRLYLVHFRRGTCTLKTLEAARSIQQVIVR

WEAYRGGKKGPTQCHRCQEFGHGTRHCNVKPRCVLCAGQHTSETCPSADG

HQAVKCSNCAGPKGARNHRQKHQHPQPPAQQQNKNWPQVSAFPPLASTKT

PPSALSSTASPCGPTAVTEPATPPATISLAQRLENAKKASDTPAPVPEGD

LFSMEELFSIFSKMLSRIRQCRNKADQLAVIGELLLCY

>An_mac_ORF1

NYCKYSSTSIQEENLDKSVEVNCRRKKTPAASASTATPTPAAPAAPLAGR

KSAPTTASAVKMPPIVVKTIPVAVLRPELQARGITPEFRLSVVGTSIIVR

SPAEQQEVYSYLQQRNAEFFSHDAKDMRPFKAMLRGLPEMELDDIVAELK

GKHHLDVLEAFEIKRRAEGIHSRLYLVHFKRGTCSLKKLEAVRSLYXVII

RXKQYRGSKKGSTQYDRCQAFVHSTRHCRIKPRCVICAQEHLSDQCPTND

RIAKCTNCGAVHRADDPACTHRAKYQELRRHMNSRKSSHQQMQTNVWKAF

PPLTTTVFVVPQAMPVSSVVPAAAPSLKPVARKVASSVTSSVPIGEPPNP

SELPVPEPAANLQVSSGHKTNLAQRLEKARNTPGPSASAPEDLFTPEELY

QICITMLSKFRLCRTKAEQLAVIVELIKLHG

>An_dirus6358_ORF1

FSERSAMPKRSKKSGRKRNRKTSGESGSDSAESKRSRSAPVSSEESMSEG

SESDGSSASSESSSSSSSSGSGSATSVQEGNLEFTVVRRSPGKAPAPAGV

VLTRSPTTKTTPSIPSIPATPQPTPVPATAGKFPPIVVRTVPVSVLRPEL

QARGFTPAFRLSSVGTSILVRSCAEQQGVLTYLQQRNAEFFTHDAKDQRP

FKAVLRGLPATEIPEIVEELRNQHQLDVLEAFEIKRRAEGIHSRLYLVHF

KRGTCSLNKLKEVRSINQVIVRWEAYHGGKKGPTQCHRCQEFGHGTRHCR

LLPKCVMCAAQHLSENCPHSGTVVPAKCTNCSAAHRADDPACPRRASYIN

LRQLANNNRKPTLQQPQPAHRVTPQSAHRVTPPTVPAPIPAVSAWATTSL

HNSTPAAAVGSRPSAVPVTTPAPSATPPINPPPAPTTITPAPAPKSSKPA

GTSTNNSDRKVNLVQRLENARNAEPTPPAPTPEDAEDDLYTMEELLQIFK

TTLAKIRLCRNKYDQLAVIAEMMSLHG

>An_fara2362_ORF1

WKPQVYFYELPVQCRAIQIVFARLYYDEARKEEQDRKESGDSNSDSVESK

RTRNVEVLVNEEVMSEQSYESSDGSSSETNSQEKTPEFQEARRCKNRTSP

VNPTTRPAATAASNSTPASPAAGRTSASAPISAGKLPPVVVRSLPISTLR

PQLQSRGLVPAFKISSVGTSIYTRSQAEYQGVIGYLRRRSAEFFTHDTKD

QRPLKAVVRGLPAMELNDIVTELREEHQLDVLEAFEIKRRAEGIQSRLYL

VYFRRASCSLKKLEQVRSIQQVMVRWESCVGGKKGPTQCHRCQEFGHGTR

HCQLKVRCVICAGQHTSDTCPSMGQTTPAQCANCNAAHRADDPSCPRQAK

YIEGRRLANERKQAPQPKQPPRQKTPQYLPAPHPAVNVWKTSNMKLSFAP

TKTSDTIPSQTPTPPPTQLTKLPTNIPAQPVSTATLPTNIPVQPAPTAIA

AVETSPTSAFSKNLSQRLENDRNAPDTRHPSKLQQSRFFLHAGAVHHFPN

LPSKASALQK

>An_fun_7067_ORF1psequence

VFKYLDAIKAEYYTHAPRDERSYKAVIRGLPAMDVEEIADELWNQHNLEV

LGVFHMKRRDESIESKLYLVIFKRGTCSLAKLSAVRSIRQVIVRWEAYRG

GKQGLTQCFRCQSFGHGTRHCHMKPRWALCAEEHVSDSCPAASQTVQTQQ

FKCVNCNGDHRASDPSCPRSAQYKKMRQQVFNRFRNLKQQPXRRKTAETA

PASTAASSRTVCXDLRHGKYIGSCTDPSALCSTVTSQRSSTPATSPGFQV

NLAQRLKIAREAVTTPDISTEDDAASFMQLMNIVKKYVPLIRAWRSIEVK

LVVLVDLIA

>An_steph_M2_ORF1

CVLPMGKRKKREVLRPSLSNPSELTPKKPKETLVCVSDGTPDDADEEIGE

FAEVSYKGGARRKSVNAAPSDNNISXAASKEVRPPPVVVKQPSLFQLKHE

LKDFSGVEFQCIGIGVKVYVKSLEDHQRLISLLEGRKDALFYTHDIPQPK

PFKAVLRGLPLTDANEVMAELRDRYSMKPIEVFRIKRRNEDTNTYSSHLY

LVHFEKGTCSQEALKEVRTIQSIRVRWEAYRGGRRRLTQCLRCQAFGHGS

RNCRMKPKCPNCSLEHVLAECKAASDTLRCANCNGGHKANDPQCPQLAKY

RQIRERASAQQRVNFRKAMQTKVVPASSAFPPLNPRAERPIPTVSAADTH

RKEPPTVVIPPGMQYAMVAKMTSPFRCNTAEAPLPSDTGAPLHDAATLMG

IFTEMVERLSTCRSRRDQITVIGEFAIRYG

>An_nili3330_ORF1

MGKRKRKSVKKVGRASASSSEETCEAANGVSTTVSPPAQPSGPKRTCLPD

PDQGSTTVPKMCESSLAPSLSDSDIESLGEFREVRSRRRSSILASAVSAM

APNTSACSVTTTSSAASPMVSSGPTVAVAPLSESTKFTAGPSSRRRPPPI

TVMQPHVDIVRKELQNHAVDLKLCGPGVRVLAKDKAAFDVAHATLIRMNV

NFYTHAYPTEKPYKVVVRGLPLLDPNEIVTELQDKYHLKASSAYHIKRRA

EDTRKYYDCMYLVCFPKGTVDLAGLKVVRKLCDLLVTWEAYRGGPRSVTQ

CLRCQRFGHGDRNCFLQPICGNCAQEHLQSACTWTPQAAPKCANCGENHR

ANDPTCPSRIRYLATRQKYPVVNQPNTKPTTTSSSALPAPLASLSAFPPL

KPCTGKVPANSSTITPHSSSVKNPVNSARAPAVAHAPLDKAPAVDNRVSA

QTRSIVSGSTVAPSYQYAMAAKGMQPQFTSPQAEPEEELLVMDAVIKLII

EFGPRLKLCRTRIEQISVIGEILFRYG

>An_atrop8636_ORF1

GIPRLMGKRGRKKRGKEDTVSPPQESARLADVPPTDKRFKRSSESIDPTI

DDTHLDDFVEVSSRRSSCTKSAKSAASGGKQPNSSLSQESLLSLEGAMDV

DHRSPGSGRAKTTNKEAPRVPASSIPCSANTPAIRSARIPPIVVNAPYHQ

LRAELSGIPGIVYQFSGAHVKLLTSVPETRDRVLALLKTTKREFFTHEAR

SERPFKAVIRGLPELPESDILSALRAQSLEPIAVHKISKDPEQSQRRQAC

LFLVHFVKGTINLAALKSIRTIDCIRVSWEAHRGGKGRIVQCHRCQAFGH

GTRNCSMRQRCENCSQEHDVASCPIVPAEAAKCANCNGNHCSSDKFCPSR

QHYESIRQQALDKRQKKPKPLTTPQSIRPPPLVEASTFPPIKPSDLSAPA

TTAYAAPAALTHSTGEKAAVPALASLFKNTDSKIEQSAAHTTTMKAPSVH

TTNKQAYPVPAPTTRTDGSPGDGNDDFNIQEWIDILYVMTQRLRLCRSRP

EKFAVIAELAIRYGC

>An_funest3583_ORF1

RSANSVNGSFCSRMTRRGRKKKREDEVSPSVVSTGAPPPKIMLTTEPESE

KFIEVCSRRSRHSKQTSSPSMQTDAVPTRAINHYQEDILSQEGTIDTEDC

SHGSGRAKTTNKAGPPLLASSAASNRTIRPARIPPIVVNAPYHQLRADLA

GIPGIVYQFAGPYVKVLTSLVETRDRVLTLLKASCVEFFTHEIRSEKPLK

VVIRGLPDLPEEEIIAALREQSLEPLAVHKIHKQHEERQHRQACLYLAHF

TKGTITLAALKCIRTIDCIRVSWEAHRGGKGRTVQCHRCQAFGHGTRNCS

MKQRCENCSKEHATEACSILSPEAPKCANCQGSHRSNDPDCPSRHQYHQM

RQKVSFSNQRQLKPSRANRAAMPPPPAPTNSSFPPLRRSGPPTHATVANA

GPVTTFASVVNDSLPTNTAVPLASPHRAVAESITRTHTMRANTMQAPAIR

ASTPHPASSPEMLPGNEDDFSIEEWVEILRVMTQRFRLCRTRQEKFAVIA

ELAIRYGC

>Ae_aegyp36_2_ORF1

VRKVIASAMGGKRKKKSLSPNKSQSSPLSKKDKRSSASSGVDFGRELNAS

QSNMYALLDDISDHDDDQCSIHTSATPNERVPRSQVEARAIVSGTVKANS

PPMKKKLPPLVVKSLPLEKLKKVMQAINVRAEFQLTGMGIKLIVKSDQEF

SKAKTYLSKSGAEFFTHDVAAEKPFKAVVRGLSQMDTNEILCELKDVYKL

QPLAVFNINRRAATSSTKYRDCLYLVHFAKGSATLGALKAVRTLNDCIVK

WEAYRGPNRSVTQCMRCLNFGHGTRNCNLKPRCNFCSQEHWTENCVLEGA

CEFRCANCSGQHMSTDKRCPKLEEYQRIRKQATTRNQPNQQKKKKPNLIN

LDEFPELPPPMSSSGWQRSRSPPRAPPGGIPPGFRWGNLNGGSEPVNQGF

PQPEALPTSVSSTIAHLAALVAEMQKMMMQMMQMFLSFNVQRQGC

>Ae_aegyp0218_ORF1

LSKCFDYVLPSMGKRGRRRSSNGSKQNSPQSALKKPKEGSPTGHRKRAKS

SNIVAGTQSVITEAGTAENRSEMGGFGDIGQTDDFITPFIHQNGNQPSKI

PPLVVKSIPLGQLKQDLRANGIDAQFKLTRIGIKIVVHTKEAMEATKAYL

QRKKAEYFTHDAPEEKPFKAVIRGLPITEKSLIEAELIQHYKLQPVAIHV

IARKFSEGDNRDCLYHVHFRKGSTTLNALKAVRTLNDMIVTWEAYRGSHR

DVTQCMRCLNFGHGTRNCNLKPRCNICAHPHITADCPHEDVAAFKCVNCG

SGHKASDKICPKREAYKQIRKNAATRNLPGRRAPENQQLFRQDEFPALQQ

NSKQRQQPNVTPSWPRQPRTTTATPSSSQHPVPQVSANAASFPDDECESV

PQSGSLYAPEELVRIFLDMSDKLKRCRSRHEQVETLGVFLIQYGR

>Cx_quin4245_ORF1

LVTLSLVVDVRSSILVTQVEKLRKLLAVTNEVAMGKRGGGAAPGQGSAKV

IKGDRNSLLNANPYAPLAGGSGGTTVEKRIKLPPIFTPVKEIAKLMEAMN

KAKLHPNYKLCSTGTKILCCTEELFNGVKSFFKQAKIEFYTHDVAAAKPM

KVVIRGLPAREKPENIMDELVKVHKLKPVAVFEMTRQNKEINYRDSLYLI

HLERGSATLAELKKIKAIAHIVVEWEMYRPQHREVTQCKNCQAFGHGTKN

CAMAPKCPKCAGPHCEVDCEAEMDDETAVKCVNCGNNHPASDKACPKRAE

FMLIRKKASTRNQPNRSNRVKSLTNDDENFPEIPRRPIPVLEPLPLPGKN

PAGKPPGTPKPPPPGWGNPGGSKQQPLQQQPEEKLFSKDELLDIFDVMIE

KMCRCRTRVEQLRTLGRFIIQYGH

>Cx_quin3_3_ORF1

INMTRRVPAQNAGNIALPAAQAGKKVGISKRKVEASTDGQKPGISKRKVL

LEVTSNSKKTKKSDGTVPMDEEEENSSSHEEQLLKNNKFAGLPDEDQVAE

AKENEVKQRKEKLPPFYVRQSAATIDFRAGLVELIKSGKVLGNIRLCQDG

FKVLVQSRQHYQLVKDYLTENEAEYFTHDVVMDKPYKIVVRGLYDMPVEE

LAAELKVLKLDVLAVHKMSRRNKDIKYRDQLYLLHLAKGSTTLPELKAIR

AVFNIIVSWERYRPVHRDVTQCFNCLGFGHGGKNCHLKRRCAKCGTDAHI

TSQCIQDSLVKCLNCNGEHSSTDRKCPKRAEFVKIRQQASTKNQPQRRRT

PPALVEENFPPLQPRRQVPNLAPLPLDPRKRAEVNHPRPGSSQEPRPPPP

GFSQEPRPTQEPAVEENGNDLYTSTELLNIFKQMSATLRGCKTKTQQIEV

LTSFVIQYGS

>Ae_aegyp37_1_ORF1

EELERFSPGLAMGDTAAGVAASEISSRSASGSEMRCKGSEKRPASGNTSD

TVAPKKFANNMYSVLTDGDAGNSPVAVKKRKPKQQCVATEPERKCKCPPI

FVKGDPPNLRASIRDCIRNGYFRGSFRLCSEGVKLMLESKESFDNAKDFL

TKRKWEFFTHDMPGTKPLKVLLRGLDDMAVDELVEELEFHDLKPVKVDKI

ARHDRTRKYRDQLYLVHLEHGSTTLKDLRAIKIINSTVVEWQKYKPVHRE

VTQCMNCLRFGHGTRNCSMASRCSTCGGNHQNEACDQMDESQPKCANCGE

KHRATDKNCPKRAEFLVIRQRASTKNQPRKTVAPPPLTSAHFPQIPKPQR

SIPVLPPLQPQQRLVAAAASVPSKAQCSQAPPINQWHQPPPGFRRQDNTP

LPPEDAAPLYSSEQLAPIFSDLVARLRSCKSRFDQNYTLGLFVIENGY

>Ae_aegyp38_2_ORF1

VARQVRILPVRVRSALFAISRSRDCSPALTYLLGVASQKASLQAEQPRPI

AMGRNRKQKADSASILAPLADSGQTSAPKRARNEDANPAAYSRLLANNQF

ASLPVDQAPPGAKVPPLFTASKDLSALRSELAANNIRPLFKLCHTGTKIM

CASGADYDKAGKLLKAKGVEFYTHDAPGSKPLKVLVRGLPELTPEAILDE

LKAAGLKPTNVFPIRRAQGGRHRDQLYLAHLEKGSTTMAGLTRVRALFNI

VVEWERYRPKKRGVTQCGNCLAFGHGTRNCHMKPRCGKCAGAHATITCQP

MEEGIEPKCANCGANHEGSSRNCPKRAEFLAIRQQASAKKLGRQRQRQPP

PPLTEEHFPTPRYQVPNLPPLPPTHRQASRQPAPSVQHRLAAAAAAPPVQ

NAPPPGWGNPGRSAPGTPPSDDGSLYTPEQMLEYTRDLFQRLRACRSKSE

QINAANSVVFAFLAKYGP

>Ae_aegyp34_1_1ORF1

FLKIQSEKSKTLGKRTPGVSSNVASSECSPNGGPCSIMIRKNYDKSKVRI

TSTQTEISTMVDNDLLTANVQQRRRHNSTDENSMRPRNQSSDSGHQFSSQ

PIAGCSNANNVLIAVPNVPTENPFDTLMDNEELQERVTPQQSASKIHCPP

IFVQNGTVKDINKLMSSLEVGEKNYAQKIIKGGIRLHVKEKTKFTVVVAA

LKSENVKFFTHGTSDEVPIRIVLGGLPVLDLEEVREELKQANVLPVEVKL

LYSSKDEDSALYLLKFPKGAVKLKELQKIKMLFNVVVSWRFFSRRIGEVI

QCYRCQKFGHGMRNCNMDAKCVKCGELHLTKDCTLPARRATDDRSKIRCA

NCSQNHTSSYKGCPARKNHIQENEEKKKMQSSRRKDAPALSHAPGGRSFR

STFVTPSKSFADAIKDGSSATVVAAAAVAVDGAAGGGGYAGPDQSELFSL

HEFMNLASDLFTRLSSCKTKAQQFLALSELMIKYVYNG

>Cx_pip1_1_ORF1

FITAPGVFASTLGKRKQPKPPPLPGEGSPARDSSSQILQRVNYKKVRKQQ

QEIAVTTPLSRRHRRNSLGGISSTSDQLNNHRPGTQPGGGTGGFPTFNQY

EALDFDISGDDEENNNNGGDGETAAAGNGAAVGSVVKNPVPVPTKVRCPP

IFVYGSSVPALNRLLSTTQLGIDDYHLRVNKGHIQIRVSTKIHFTAVVSK

LKNSDVQFYTHGTSDETPVKIVLSGLPVFPVEDVKLELESVFLRPTSVRQ

MGKSKHGDYALYLLQFEKGTVKLQELQQIKALFNVIVRWRHYSKKKSDVV

QCFRCQQYGHGMRNCHLEAKCVKCGERHQTTACVLPARADVVVNDDRSQI

RCANCSQNHTANYKGCPTRLKYLQDLKAKKKTSPASRSNAPKVSAVPAPA

PRPLGGDLSQLLGSIANPGVSYSQAVQGQPESSTLFTVEEFMCLASELFT

RLSNCQSKAMQFLALSELIIKFVYNGQP

>Ae_aegyp33_1_ORF1

FSFGQWLPQSFMGKTIKADGVPSGPSDGGSRSGVRKISSILERRSYDKGK

TKAQLTSTLQTDSSDQIVVPVVEPHLNRSRSASMSDFPVLESENSGGPAQ

VCPSIPLRNSFELLVQQNNIDDVENEMTNVSQNIQTNSARCPPITVWKMS

VQDINKLLYQLNGDGKFVLKNSKGAVQIRTKCSSLFVDIQEALKQLNAEF

YTHATRGDASVKIVLSGLPVYNIEEIKTELAKNNISPREVKLLYKTRDSS

SALYVLNFAKGTVKLNKLREVQYLFNVVVSWRHWTRRVNDILQCFRCQRF

GHGSRHCNMQLRCVKCGKQHTSGDCTIPKKASGGSISKTHKDIKCANCGQ

NHAASFRECPYRLEFIKRQVSSVSRQNPNGGPTINPPRKFTSSWVTQNRS

FAEVVSRPTTSEMRPEAANTTTERNNTNLFTLSEFLGLAREMFNRFRGCT

SREEQFFALQELMAKYLYIH

>An_atrop8884_ORF1

HFTDNCQKPEPPKCANCGASHRANHPECPKREQFRELQKRSRNRARQKTQ

YQQVPLAQPPSQLSARLDDLRRHQDDGQKRQKQQTAPVQYKQSLPQQQNV

QQKEANSSTPSYSNVLSPNSSVKATIRPAPAHQQIEKNKELFDPEKLVEI

FNEIMDAVRSCTTKHEQLACLAKLIIRYA

>An_gamb_Q_ORF1_432429

QCCAVITRDFAMAAICFSCAEPLEATGCIISCAYCDATFHRGCCKLPPEL

IDAVLSNVDLHWSCIGCTNMLKNPRCRSVKEIGAQVGFQAALNSAVAAIG

KLVEPIVAEVRSGFTLLQTASTPHNRNSDPRPATGRKRRRIIEDSASPGV

NKIVNSRGNTLCAASSPNAYTNTTIAVQPAPTQPHELVGTDPLSSPLQAA

PREPFTDRIWIRLSAYQRPSLWNKWSLSVKRRLATDDVIAYCLLRRGVSV

DSMNWLSFKVRVPAILRDAALTPSTWPVGIGVREFFQSRQHDHQTSSPIA

TRNRFTTRTPATSTEHRYTTRTPTTTHRLAARTSTPPDPETTSSQQCHPP

VNDTLEAPNSTLVSGPPQNHRASSPHLHQSTIDRFFLN

6.4 – In-Fusion® CLONING PRIMERS

Primers used for cloning of the CPSF100_C domain into pOPINk

Forward,

5” AAGTTCTGTTTCAGGGCCCGTCGGAGGAATCTGACGATAC 3”

Reverse,

5” ATGGTCTAGAAAGCTTTAAACGGTGATTGACGTGCCTA 3”

6.5 – RESTRICTION CLONING PRIMERS

Primers used for cloning ORF1 into pYES2/CT

Forward,
5” TAAGCAGAATTCATGCCGAAGCGTGGTAGAAG 3”

Reverse,

5” TCCTGCTCTAGATAGGTGAAGGAGCATCAATAGTTCTC 3”

6.6 – SC MINIMAL MEDIA

SC is synthetic minimal defined media for *S.cerevisiae*.

0.67% yeast nitrogen base (without amino acids but with ammonium sulfate)

2% carbon source (Raffinose in starter media & Galactose in expression media)

0.01% (adenine, arginine, cysteine, leucine, lysine, threonine, tryptophan)

0.005% (aspartic acid, histidine, isoleucine, methionine, phenylalanine, proline, serine, tyrosine, valine)

2% agar (for plates)

CHAPTER 7

REFERNCES

Adams, T. (2015). Survival, demise, and propagation of retroposons in the tandem arrays of histone gene clusters in Anopheles mosquitoes (PhD). Queen Mary, University of London.

Akinbosede, D. (2016). Heterologous expression of open reading frames (ORF) 1 & 2 encoded in the novel M5 retroposon of Anopheles stephensi (Undergraduate). University of Hertfordshire.

Bateman, A. (2004). The Pfam protein families database. Nucleic Acids Research, 32(90001), 138D-141. doi: 10.1093/nar/gkh121

Baucom, R., Estill, J., Chaparro, C., Upshaw, N., Jogi, A., & Deragon, J. et al. (2009). Exceptional diversity, non-random distribution, and rapid evolution of Retroelements in the B73 Maize genome. Plos Genetics, 5(11).

Beauregard, A., Curcio, J., & Belfort, M. (2008). The take and give between retrotransposable elements and their hosts. Annual Review Of Genetics, 42, 587-617.

Bill, R. (2014). Playing catch-up with Escherichia coli: using yeast to increase success rates in recombinant protein production experiments. Frontiers In Microbiology, 5. doi: 10.3389/fmicb.2014.00085

Boc, A., Diallo, A., & Makarenkov, V. (2012). T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. Nucleic Acids Research, 40(W1), W573-W579. doi: 10.1093/nar/gks485

Brown, R. (2005). Zinc finger proteins: getting a grip on RNA. Current Opinion In Structural Biology, 15(1), 94-98. doi: 10.1016/j.sbi.2005.01.006

Bulmer, M. (1987). Coevolution of codon usage and transfer RNA abundance. Nature, 325(6106), 728-730. doi: 10.1038/325728a0

Chiang, Y., Gelfand, T., Kister, A., & Gelfand, I. (2007). New classification of supersecondary structures of sandwich-like proteins uncovers strict patterns of strand assemblage. Proteins: Structure, Function, And Bioinformatics, 68(4), 915-921. doi: 10.1002/prot.21473

Crainey, J., Garvey, C., & Malcolm, C. (2005). The origin and evolution of mosquito APE retroposons. Molecular Biology And Evolution., 22(11), 2190-7. Retrieved from

Cutler, P. (2004). Protein Purification Protocols: Second Edition. Totowa, NJ: Humana Press Inc.

Derbyshire, M., Gonzales, N., Lu, S., He, J., Marchler, G., Wang, Z., & Marchler-Bauer, A. (2015). Improving the consistency of domain annotation within the Conserved Domain Database. Database, 2015(0), bav012-bav012. doi: 10.1093/database/bav012

Dong, T., & Schellhorn, H. (2008). Control of RpoS in global gene expression of Escherichia coli in minimal media. Molecular Genetics And Genomics, 281(1), 19-33. doi: 10.1007/s00438-008-0389-3

EMBL. (2017). Protein expression and purification core facility - cloning - choice of expression systems. Retrieved from https://www.embl.de/pepcore/pepcore_services/cloning/choice_expression_systems/

EMBL-EBI. (2018). EMBL-EBI < Help < Tools < ClustalW2 FAQ. Retrieved from https://www.ebi.ac.uk/Tools/msa/clustalw2/help/faq.html

Finn, R., Bateman, A., Clements, J., Coggill, P., Eberhardt, R., & Eddy, S. et al. (2013). Pfam: the protein families database. Nucleic Acids Research, 42(D1), D222-D230. doi: 10.1093/nar/gkt1223

Fisher, T. (2018). pYES2/CT Yeast Expression Vector. Retrieved from https://www.thermofisher.com/order/catalog/product/V825120

Flick, J., & Johnston, M. (1990). Two systems of glucose repression of the GAL1 promoter in Saccharomyces cerevisiae. Molecular And Cellular Biology, 10(9), 4757-4769. doi: 10.1128/mcb.10.9.4757

Flutre, T., Permal, E., & Quesneville, H. (2012). Transposable element Annotation in completely sequenced Eukaryote Genomes (pp. 17-39). Springer Science + Business Media.

Fong, J., & Marchler-Bauer, A. (2008). Protein subfamily assignment using the Conserved Domain Database. BMC Research Notes, 1(1), 114. doi: 10.1186/1756-0500-1-114

Fong, J., Shoemaker, B., Garbuzynskiy, S., Lobanov, M., Galzitskaya, O., & Panchenko, A. (2009). Intrinsic Disorder in Protein Interactions: Insights From a Comprehensive Structural Analysis. Plos Computational Biology, 5(3), e1000316. doi: 10.1371/journal.pcbi.1000316

François, F., Chapeland-Leclerc, F., Villard, J., & Noël, T. (2004). Development of an integrative transformation system for the opportunistic pathogenic yeastCandida lusitaniae usingURA3 as a selection marker. Yeast, 21(2), 95-106. doi: 10.1002/yea.1059

Fujiwara, H. (2015). Site-specific non-LTR retrotransposons. Microbiology Spectrum, 3(2). doi: 10.1128/microbiolspec.mdna3-0001-2014

Funk, J., Nedialkov, Y., Xu, D., & Burton, Z. (2002). A Key Role for the α1 Helix of Human RAP74 in the Initiation and Elongation of RNA Chains. Journal Of Biological Chemistry, 277(49), 46998-47003. doi: 10.1074/jbc.m206249200

Guo, C., & Tugarinov, V. (2009). Selective 1H-13C NMR spectroscopy of methyl groups in residually protonated samples of large proteins. Journal Of Biomolecular NMR, 46(2), 127-133. doi: 10.1007/s10858-009-9393-0

Gustafsson, C., Govindarajan, S., & Minshull, J. (2004). Codon bias and heterologous protein expression. Trends In Biotechnology, 22(7), 346-353. doi: 10.1016/j.tibtech.2004.04.006

Hamady, M., Wilson, S., Zaneveld, J., Sueoka, N., & Knight, R. (2009). CodonExplorer: an online tool for analyzing codon usage and sequence composition, scaling from genes to genomes. Bioinformatics, 25(10), 1331-1332. doi: 10.1093/bioinformatics/btp141

Hammarström, M., Hellgren, N., van den Berg, S., Berglund, H., & Härd, T. (2009). Rapid screening for improved solubility of small human proteins produced as fusion proteins in Escherichia coli. Protein Science, 11(2), 313-321. doi: 10.1110/ps.22102

Hartwell, L., Hood, L., Goldberg, M., Goldberg, A., & Silver, L. (2010). Genetics: From genes to genomes (4th ed.). New York: McGraw-Hill Higher Education.

Häsler, J., & Strub, K. (2006). Alu elements as regulators of gene expression. Nucleic Acids Research, 34(19), 5491–5497. Retrieved from

Häsler, J., & Strub, K. (2006). Alu elements as regulators of gene expression. Nucleic Acids Research, 34(19), 5491-5497. doi: 10.1093/nar/gkl706

Hengge-Aronis, R., Lange, R., Henneberg, N., & Fischer, D. (1993). Osmotic regulation of rpoS-dependent genes in Escherichia coli. Journal Of Bacteriology, 175(1), 259-265. doi: 10.1128/jb.175.1.259-265.1993

Hou, J., Tyo, K., Liu, Z., Petranovic, D., & Nielsen, J. (2012). Metabolic engineering of recombinant protein secretion by Saccharomyces cerevisiae. FEMS Yeast Research, 12(5), 491-510. doi: 10.1111/j.1567-1364.2012.00810.x

Kanehisa, M. (2002). The KEGG databases at GenomeNet. Nucleic Acids Research, 30(1), 42-46. doi: 10.1093/nar/30.1.42

Karathia, H., Vilaprinyo, E., Sorribas, A., & Alves, R. (2011). Saccharomyces cerevisiae as a Model Organism: A Comparative Study. Plos ONE, 6(2), e16015. doi: 10.1371/journal.pone.0016015

Kaur, S., & Pociot, F. (2015). Alu Elements as Novel Regulators of Gene Expression in Type 1 Diabetes Susceptibility Genes?. Genes, 6(3), 577-591. doi: 10.3390/genes6030577

Klis, F. (2002). Dynamics of cell wall structure in Saccharomyces cerevisiae. FEMS Microbiology Reviews, 26(3), 239-256. doi: 10.1016/s0168-6445(02)00087-6

Kolev, N., Yario, T., Benson, E., & Steitz, J. (2008). Conserved motifs in both CPSF73 and CPSF100 are required to assemble the active endonuclease for histone mRNA 3′-end maturation. EMBO Reports, 9(10), 1013-1018. doi: 10.1038/embor.2008.146

Kolosha, V., & Martin, S. (1997). In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. Proceedings Of The National Academy Of Sciences, 94(19), 10155-10160. doi: 10.1073/pnas.94.19.10155

Korepanova, A., Moore, J., Nguyen, H., Hua, Y., Cross, T., & Gao, F. (2007). Expression of membrane proteins from Mycobacterium tuberculosis in Escherichia coli as fusions with maltose binding protein. Protein Expression And Purification, 53(1), 24-30. doi: 10.1016/j.pep.2006.11.022

Letunic, I., Doerks, T., & Bork, P. (2015). SMART: recent updates, new developments and status in 2015. Nucleic Acids Research, 43(D1), D257-D260. doi: 10.1093/nar/gku949

Lovšin, N., Gubenšek, F., & Kordi, D. (2001). Evolutionary dynamics in a novel L2 Clade of Non-LTR Retrotransposons in Deuterostomia. Molecular Biology And Evolution, 18(12), 2213-2224. Retrieved from

Malik, H., & Eickbush, T. (2000). NeSL-1, an ancient lineage of site-specific non-lTR retrotransposons from Caenorhabditis elegans. Genetics., 154(1), 193-203. Retrieved from

Malik, H., Burke, W., & Eickbush, T. (1999). The age and evolution of non-lTR retrotransposable elements. Molecular Biology And Evolution, 16(6), 793-805. Retrieved from

Marchler-Bauer, A. (2004). CDD: a Conserved Domain Database for protein classification. Nucleic Acids Research, 33(Database issue), D192-D196. doi: 10.1093/nar/gki069

Martin, S. (2006). The ORF1 protein encoded by LINE-1: Structure and function during L1 Retrotransposition. Journal Of Biomedicine And Biotechnology, 2006. Retrieved from

Martin, S., Cruceanu, M., Branciforte, D., Wai-lun Li, P., Kwok, S., Hodges, R., & Williams, M. (2005). LINE-1 Retrotransposition Requires the Nucleic Acid Chaperone Activity of the ORF1 Protein. Journal Of Molecular Biology, 348(3), 549-561. doi: 10.1016/j.jmb.2005.03.003

Matsumoto, T., Hamada, M., Osanai, M., & Fujiwara, H. (2006). Essential Domains for Ribonucleoprotein Complex Formation Required for Retrotransposition of Telomere-Specific Non-Long Terminal Repeat Retrotransposon SART1. Molecular And Cellular Biology, 26(13), 5168-5179. doi: 10.1128/mcb.00096-06

McClintock, B. (1950). The origin and behavior of mutable loci in maize. Proceedings Of The National Academy Of Sciences, 36(6), 344-355. doi: 10.1073/pnas.36.6.344

McDowall, J. (2007). Protein of the month. Retrieved from https://www.ebi.ac.uk/interpro/potm/2007_3/Page1.htm

Metcalfe, C., & Casane, D. (2014). Modular organization and reticulate evolution of the ORF1 of Jockey superfamily transposable elements. Mobile DNA, 5(1), 19. doi: 10.1186/1759-8753-5-19

Mulder , N., & Apweiler, R. (2001). Tools and resources for identifying protein families, domains and motifs. Genome Biology, 3(1). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC150457/pdf/gb-2001-3-1-reviews2001.pdf

Muñoz-López, M., & García-Pérez, J. (2010). DNA Transposons: Nature and applications in Genomics. Current Genomics, 11(2). Retrieved from

MURAOKA, A., ITO, K., NAGASAKI, H., & TANAKA, S. (1991). Phosphoenolpyruvate:carbohydrate phosphotransferase systems in Enterococcus faecalis. Nippon Saikingaku Zasshi, 46(2), 515-522. doi: 10.3412/jsb.46.515

Nielsen, J., & Jewett, M. (2008). Impact of systems biology on metabolic engineering ofSaccharomyces cerevisiae. FEMS Yeast Research, 8(1), 122-131. doi: 10.1111/j.1567-1364.2007.00302.x

Norouzi, R., Hojati, Z., & Badr, Z. (2016). Overview of the recombinant proteins purification by affinity tags and tags exploit systems. Journal Of Fundamental And Applied Sciences, 8(3), 90. doi: 10.4314/jfas.v8i3s.168

Novikova, O., Śliwińska, E., Fet, V., Settele, J., Blinov, A., & Woyciechowski, M. (2007). CR1 clade of non-lTR retrotransposons from Maculinea butterflies (Lepidoptera: Lycaenidae): Evidence for recent horizontal transmission. BMC Evolutionary Biology, 7(1), 93.

Oh, K., & Yi, G. (2016). Prediction of scaffold proteins based on protein interaction and domain architectures. BMC Bioinformatics, 17(S6). doi: 10.1186/s12859-016-1079-5

Peccoud, J., Loiseau, V., Cordaux, R., & Gilbert, C. (2017). Massive horizontal transfer of transposable elements in insects. Proceedings Of The National Academy Of Sciences, 114(18), 4721-4726. doi: 10.1073/pnas.1621178114

Porat, A. (2018). Protein Expression In Yeast Using INVSc1 And pYES2/CT From Invitrogen. Retrieved from https://www.biocompare.com/Product-Reviews/40593-Protein-Expression-In-Yeast-Using-INVSc1-And-pYES2-CT-From-Invitrogen/

Porro, D., Sauer, M., Branduardi, P., & Mattanovich, D. (2005). Recombinant Protein Production in Yeasts. Molecular Biotechnology, 31(3), 245-260. doi: 10.1385/mb:31:3:245

Presnyak, V., Alhusaini, N., Chen, Y., Martin, S., Morris, N., & Kline, N. et al. (2015). Codon Optimality Is a Major Determinant of mRNA Stability. Cell, 160(6), 1111-1124. doi: 10.1016/j.cell.2015.02.029

Ransom, M., Dennehey, B., & Tyler, J. (2010). Chaperoning Histones during DNA Replication and Repair. Cell, 140(2), 183-195. doi: 10.1016/j.cell.2010.01.004

Ried, T., Difilippantonio, M., Lim, S., & Hummon, A. (2007). Isolation and solubilization of proteins after TRIzol® extraction of RNA and DNA from patient material following prolonged storage. Biotechniques, 42(4), 467-472. doi: 10.2144/000112401

Rosano, G., & Ceccarelli, E. (2014). Recombinant protein expression in Escherichia coli: advances and challenges. Frontiers In Microbiology, 5. doi: 10.3389/fmicb.2014.00172

Schaack, S., Gilbert, C., & Feschotte, C. (2010). Promiscuous DNA: Horizontal transfer of transposable elements and why it matters for eukaryotic evolution: Trends in ecology & evolution. Retrieved from

Schultz, J. (2000). SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Research, 28(1), 231-234. doi: 10.1093/nar/28.1.231

Sharp, P., Cowe, E., Higgins, D., Shields, D., Wolfe, K., & Wright, F. (1988). Codon usage patterns inEscherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogasterandHomo sapiens; a review of the considerable within-species diversity. Nucleic Acids Research, 16(17), 8207-8211. doi: 10.1093/nar/16.17.8207

Sievers, F., Wilm, A., Dineen, D., Gibson, T., Karplus, K., & Li, W. et al. (2014). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology, 7(1), 539-539. doi: 10.1038/msb.2011.75

Sigrist, C., Cerutti, L., de Castro, E., Langendijk-Genevaux, P., Bulliard, V., Bairoch, A., & Hulo, N. (2009). PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Research, 38(suppl_1), D161-D166. doi: 10.1093/nar/gkp885

Sivagnanam, M., Janecke, A., Müller, T., Heinz-Erian, P., Taylor, S., & Bird, L. (2010). Case of syndromic tufting enteropathy harbors SPINT2 mutation seen in congenital sodium diarrhea. Clinical Dysmorphology, 19(1), 48. doi: 10.1097/mcd.0b013e328331de38

Snustad, P., Gardner, E., & Simmons, M. (2003). Principles of genetics (3rd ed.). New York, NY: John Wiley and Sons (WIE).

Sullivan, K., Steiniger, M., & Marzluff, W. (2009). A Core Complex of CPSF73, CPSF100, and Symplekin May Form Two Different Cleavage Factors for Processing of Poly(A) and Histone mRNAs. Molecular Cell, 34(3), 322-332. doi: 10.1016/j.molcel.2009.04.024

Sullivan, K., Steiniger, M., & Marzluff, W. (2009). A Core Complex of CPSF73, CPSF100, and Symplekin May Form Two Different Cleavage Factors for Processing of Poly(A) and Histone mRNAs. Molecular Cell, 34(3), 322-332. doi: 10.1016/j.molcel.2009.04.024

Ward, J., Sodhi, J., McGuffin, L., Buxton, B., & Jones, D. (2004). Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. Journal Of Molecular Biology, 337(3), 635-645. doi: 10.1016/j.jmb.2004.02.002

Weiner, A. (2002). SINEs and lINEs: The art of biting the hand that feeds you. Current Opinion In Cell Biology, 14(3), 343–350.

Werner, R. (2004). Economic aspects of commercial manufacture of biopharmaceuticals. Journal Of Biotechnology, 113(1-3), 171-182. doi: 10.1016/j.jbiotec.2004.04.036

Werner-Washburne, M., Braun, E., Johnson, G., & Singer, R. (1993). Stationary Phase in the Yeast Saccharomyces cerevisiae. MICROBIOLOGICAL REVIEWS, 57(2), 383-401.

Yadav, V., Mandal, P., Rao, D., & Bhattacharya, S. (2009). Characterization of the restriction enzyme-like endonuclease encoded by the Entamoeba histolytica non-long terminal repeat retrotransposon EhLINE1. FEBS Journal, 276(23), 7070-7082.

Zhou, Z., Dang, Y., Zhou, M., Li, L., Yu, C., & Fu, J. et al. (2016). Codon usage is an important determinant of gene expression levels largely through its effects on transcription. Proceedings Of The National Academy Of Sciences, 113(41), E6117-E6125. doi: 10.1073/pnas.1606724113