Recombinant Spidroins from Infinite circRNA Translation

Maksim G. Ivanov, BSc.

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

September 2022

Abstract

Spidroins are a diverse family of peptides and the main components of spider silk. They can be used to produce sustainable, lightweight and durable materials for a large variety of medical and engineering applications. Spiders' territorial behaviour and cannibalism precludes farming them for silk. Recombinant protein synthesis is the most promising way of producing these peptides. However, many approaches have been unsuccessful in obtaining large titres of recombinant spidroins or ones of sufficient molecular weight. The work described here is focused on expressing high molecular weight spidroins from short circular RNA molecules. Mammalian host cells were transfected with designed circular-RNA-producing plasmid vectors. A backsplicing approach was implemented to successfully circularise RNA in a variety of mammalian cell types. This approach could not express any recombinant spidroins based on a variety of qualitative protein assays. Further experiments investigated the reasons behind this.

Additionally, due to the diversity of spidroins in a large number of spider lineages, there are potentially many spidroin sequences left to be discovered. A bioinformatic pipeline was developed that accepts transcriptome datasets from RNA sequencing and uses tandem repeat detection and profile HMM annotation to identify novel sequences. This pipeline was specifically designed for the identification of repeat domains in expressed sequences. 21 transcriptomes from 17 different species, encompassing a wide selection of basal and derived spider lineages, were investigated using this pipeline. Six previously undescribed spidroin sequences were discovered. This pipeline was additionally tested in the context of the suckerin protein family. These proteins have recently been investigated for their potential properties in medicine and engineering including adhesion in wet environments. The computational pipeline was able to double the number of suckerins known to date. Further phylogenetic analysis was implemented to expand on the knowledge of suckerins. This pipeline enables the identification of transcripts that may have been overlooked by more mainstream analysis methods such as pairwise homology searches. The spidroins and suckerins discovered by this pipeline may contribute to the large repertoire of potentially useful properties characteristic of this diverse peptide family.

Acknowledgements

Throughout my work, I have been supported most by my principal supervisor Keith Spriggs. He has always provided me with freedom to choose the direction of my research. His suggestions never limited my freedom to choose and his advice has always served to help me. Through his unique perspective, I had the opportunity to expand my work into bioinformatics and transcriptomics. This was at first an intimidating prospect but Keith's advice and support helped me thrive and enjoy this.

My secondary supervisor Sara Goodacre has always helped me throughout my work, especially in her area of expertise – spider biology and evolutionary biology. Additionally, she has always been optimistic and supportive which has helped me remain calm even throughout the most frustrating and hopeless days of my project.

My group member Declan Grewcock has been a mentor and a dear friend throughout my doctoral education. He was the one who trained me in many lab techniques at the start of my work. He always assisted me when I needed it and crucially he advised me on different aspects of circRNA biology. We have had many discussions about circRNA developments throughout my time in the GRRB but also about films, food and our personal lives.

Arnelle Edwards was another group member who shared a lot of the same experiences as me and helped me throughout my work. Her help with western blotting and qPCRs is greatly appreciated. Leah Ashley taught me the fundamentals of transcriptomics and provided me with a valuable transcriptome that was the basis for part of my research. Her help came at a critical time when I was struggling with bioinformatics and I was clueless about transcriptomics.

Paulyna Magaña, Angela Downie, Fahad Almulhim and Ryan Morris are members of the GRRB who have assisted me in the lab at critical moments of my research. More importantly, they were my close friends whose support and advice helped me maintain a positive outlook throughout my education. AJ Alamoudi, Steven Lawrence, Christopher Roberts, Peter Wellham and Abdul Hafeez have likewise helped me throughout my lab work and their contribution is much appreciated.

Charlotte Deall advised me on the properties of different spidroin sequences and her long-read sequencing *Argyroneta* transcriptome was valuable in my research. Similarly, Ella Deutsch kindly provided me with a *Cyrtophora* transcriptome for my work.

Natalya Pavlova from the Memorial Sloan Kettering Cancer Center, USA was kind enough to advise me on the design of tRNA qPCR primers. Without her help, my experiment would have been unsuccessful. Likewise, Alan Costello from the Scripps Research

Institute, USA advised me on the construction of my RNA circularisation plasmids which I appreciate greatly.

My internal assessor Anna Piccinini has always been kind and considerate throughout my time in the GRRB. Her advice about several experiments has been invaluable. I have taken advantage of many opportunities to develop my skills and improve my career prospects through courses planned and administrated by Anna.

Jolanta Beinarovica, Neil Thomas and David Harvey from the Biodiscovery Institute, University of Nottingham have likewise provided me with valuable advice throughout my research. Joe Tomlinson and Olubusola Olaleye were my group members whose friendliness and insightfulness I appreciate. The other members of the GRRB throughout the years have always made my time there pleasant. Spending time with them has been very enjoyable.

My research was supported by the BBSRC's Doctoral Training Partnership. The supervisors of this programme have always done their best to support me with my mental health, career planning and research activities.

My family and my girlfriend have supported me during every pleasant and difficult moment of my education. The positive mentality that they have instilled in me has been invaluable in every moment of my work. Their love and support have ensured that I reached this stage. I cannot thank them enough.

Table of Contents

List of Figures

List of Tables

Abbreviations

AcSp – Aciniform spidroin

ACTB – Actin-β

AgSp – Aggregate spidroin

ANOVA – Analysis of variance

ARS – Aminoacyl-tRNA synthetase

BLAST – Basic local alignment search tool

bp, bps – Basepair, basepairs

COS-1 – *Cercopithecus aetiops* cell line-1

circRNA – Circular RNA

CMV - Cytomegalovirus

CrSp – Cribellate spidroin

CTC – Continuously translating circular RNA

CySp – Cylindriform spidroin

DMEM - Dulbecco's Modified Eagle Medium

dNTP – Deoxynucleoside Triphosphate

DTT – dithiothreitol

eIF – Eukaryotic (translation) initiation factor

FBS – Foetal bovine serum

GA – Glycine-rich and poly-alanine repeat

GAPDH - Glyceraldehyde-3-phosphate dehydrogenase

GFP – Green fluorescent protein

HEK293 – Human embryonic kidney 293 cell line

HMM – Hidden Markov model

HRP – Horseradish peroxidase

HSF – Human splice site finder

IGV – Integrated Genomics Viewer

IORF – Infinite open reading frame

IRES – Internal ribosome entry site

IVA – *In vivo* assembly

LB – Lysogeny broth

m^{6}A – N^6-methyladenosine

MaSp – Major ampullate spidroin

MCF-7 - Michigan Cancer Foundation-7 cell line

miRNA – MicroRNA

MiSp – Minor ampullate spidroin

NIH/3T3 – National Institutes of Health / 3-day transfer, inoculum 3 x 10^5 cells

PAGE – Polyacrylamide gel electrophoresis

PBS – Phosphate-buffered saline

PCR – Polymerase chain reaction

Pfam – Protein families database

PySp – Pyriform spidroin

qPCR – Quantitative polymerase chain reaction

RCA – Rolling circle amplification

RCI – Reverse-complementary intron

RCT – Rolling circle translation

RIPA - Radioimmunoprecipitation assay

RNase - Ribonuclease

Rotl – R Open Tree of Life

RPL13 – Ribosomal protein L13

RPM – Rotations per minute

RT-PCR – Reverse transcription polymerase chain reaction

RTA – Retrolateral tibial apophysis

SDS – Sodium dodecyl sulphate

SMRT - Single-molecule real-time sequencing

SP – Signal peptide

SRA – Sequence Read Archive

TBST – Tris-buffered saline and tween

TFB1,2 – Transformation buffers 1 and 2

TPM – Transcripts per million

tRNA – Transfer RNA

TSA – Transcriptome Shotgun Assembly database

TUBB – Tubulin-β

TuSp – Tubiliform spidroin

YTHDC1 - YTH domain-containing protein 1

YTHDF3 - YTH domain protein F3

1 Introduction

1.1 Spider silk

Spidroins are insoluble proteins that are the main structural component of spider silk, the material that spider webs are built from. There is a single diverse protein family that contains all spidroins from all spider species (Gatesy et al., 2001; Guerette et al., 1996). While spider webs are the most well-known application of spider silk, spiders produce a wide variety of structures that are used for prey capture, courtship, self-preservation, underwater breathing and many other applications. This variety is enabled by a large diversity of spidroin proteins. A wide variety of spidroin homologs and orthologs is known from a large number of spiders (Rising and Johansson, 2015). The majority of spidroin proteins undergo very complex protein folding to form fibres that are the building material for their webs. These proteins have been of particular interest to materials scientists and biologists for the past three decades due to their extraordinary fibre strength, elasticity and toughness (Gosline et al., 1999). While spiders' behaviour precludes farming them for natural silk synthesis, numerous competing approaches have been developed to produce recombinant spidroins (Spiess et al., 2010). However, many of these approaches have struggled with replicating and translating the spidroins' particularly long and repetitive nucleotide sequences (Zhang et al., 2019b).

1.1.1 Properties

Some spider silk fibres possess a mechanical strength (the ability of a fibre to withstand pulling forces without breaking) of around 0.5-1.2 GPa which is comparable to steel's (1.5 GPa). However, spider silk achieves this property at only a fifth of steel's weight (Gosline et al., 1999; Ko et al., 2001; Zemlin, 1968). Mechanical toughness is the amount of deformation that a fibre can withstand without breaking (measured as energy per volume). Spider silk (150-160 MJ/m^3) exceeds Kevlar three times in this regard (50 MJ/m^3) (Gosline et al., 1999; Humenik et al., 2011; Vollrath and Knight, 2001).

These silk fibres are also very elastic (the ability of a fibre to stretch and return to its original length). Some types of silk can stretch up to five times as long before having to return to their prior length (Eisoldt et al., 2011). Spider silk is very resistant to fibre fatigue. It can retain its elastic properties even after 1000 cycles of strain without breaking or losing its elasticity (Hennecke et al., 2013).

Windlass is the ability of spider silk to maintain its strain without sagging. This phenomenon can be observed in a subset of the fibres in a capture web where glue droplets on the fibres gradually aggregate parts of the fibre into a ball by sticking to it (Elettro et al., 2016). Spidroins are also known for their biocompatibility since they are non-immunogenic and non-pyrogenic (Fredriksson et al., 2009; Hedhammar et al., 2010).

While the degradation of most spidroins has not been studied in detail before, these proteins are not as long-lived as plastics and can be degraded naturally, thus avoiding potential pollution on the scale seen with plastics (Coekin, 2021). Particularly a subset of silks has been investigated as a biodegradable alternative to polymer glues (Roberts et al., 2020).

In spiders, these proteins are produced at ambient temperatures without the use of harsh detergents and solvents. Potentially, large-scale production of these could similarly be performed under mild conditions with a low carbon footprint (Vollrath et al., 2013).

1.1.2 Spider Silk Diversity

Spiders use their silk for a wide variety of structures that are important for their survival and reproduction. There are several types of webs that are used in prey capture and some of the most complex of these are the orb webs (Figure 1.1 below). These types of webs are usually composed of five different silks. Major ampullate

(MaSp) is used for the spokes (also used for the lifeline when escaping danger) and is very elastic and tough (Kelly et al., 2020b). Minor ampullate (MiSp) silk is used as a reinforcement fibre and usually forms an auxiliary spiral that supports the orb web's integrity. It is non-elastic but very tough and also used in prey wrapping by some spiders (Colgin and Lewis, 1998; La Mattina et al., 2008). Pyriform silk (PySp) is used as a cement material to anchor the web to various surfaces (Simmons et al., 2019). Aggregate silk (AgSp) uniquely does not form fibres and is instead a type of glue that spiders use on their webs so that prey would stick to it (Moon, 2018). Flagelliform silk is remarkable for its elasticity which often results in prey becoming entangled in the web (Hayashi and Lewis, 2001). A combination of some of these silk types can be used by certain spiders to produce simpler structures called cobwebs which are far less ordered (Boutry and Blackledge, 2008).

Additionally, aciniform silk (AcSp) is used to wrap and immobilise prey after it has been captured. This silk type is very tough which prevents captured animals from escaping (Hayashi et al., 2004).

Female spiders can also produce egg sac structures that are important in protecting their offspring. These are primarily composed of cylindriform silk (CySp also known as tubiliform silk (TuSp)) which is very high in stiffness (very resistant to pulling forces) (Lin et al., 2009). A rough visualisation of these structures can be seen in Figure 1.1.

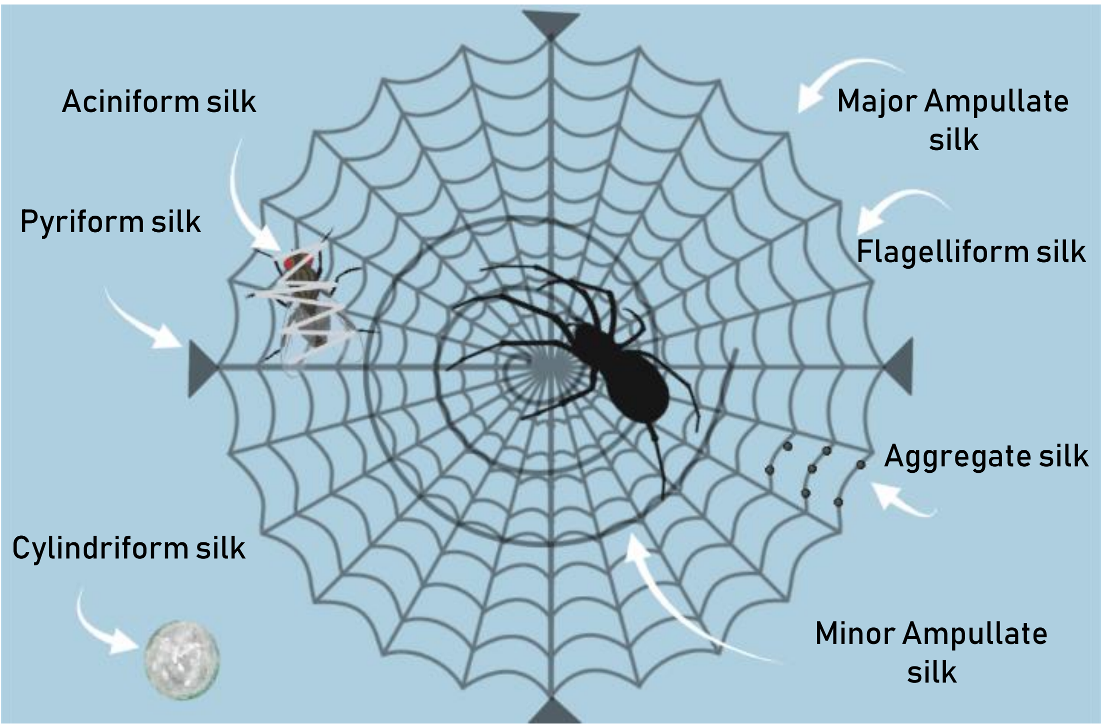


Figure 1.1 – **Schematic representation of spider web structures from the different silk types.** An orb web is presented here which can be produced from up to six different types of silk. The cylindriform egg sacs are formed separately from the prey-capture webs. This figure does not reflect the entire diversity of possible spider web structures and serves only as a representation.

Cribellate silk (CrSp) is another form of silk that is remarkable for its adhesive properties and elasticity. This takes the form of fine threads whose microscopic size enables attachment through van der Waal's forces (Michalik et al., 2019). Many other types of structures have been reported previously (Buchli 2015; Schutz et al., 2007; Yip et al., 2019).

This large diversity of spidroin types and applications is a result of the hundreds of millions of years of evolution in an order of animals which inhabits the majority of Earth's habitats. Spider silk has most likely existed since at least 380 million years ago based on the diversity that exists within modern spiders (Selden et al., 2008). Lineages of spiders that diverged early in history such as tarantulas and tube-dwelling spiders possess only a single silk gland. However, the number of silk glands present in a spider is not always indicative of the maximum number of producible silk types. Silk in these spiders has different uses in their lifetimes and may be composed of multiple spidroins (Kono et al., 2019; Starrett et al., 2012). On the other hand, more recently-evolved lineages like the orb-weavers may have up to seven different silk glands which produce even more types of spidroins (Hormiga and Griswold, 2014; Kono et al., 2019; Rising and Johansson, 2015). It is not known what percentage of spidroins in existence is known to researchers.

1.1.3 Spidroin Structure and Organisation

Much research has been directed at understanding the relationship between spidroin properties and structure. The spidroins are broadly composed of conserved non-repetitive terminal domains which flank a repetitive central domain (Figure 1.2). While the terminal domains are short (~100 amino acids each) the central domain is very long and repetitive (Garb et al., 2010). This has led to average overall spidroin lengths of around 3500 amino acids (Kono et al., 2019; Xia et al., 2010). Additionally, the members of the spidroin family vary greatly in their intron sequences. Some spidroins contain no introns while others possess hundreds (Wen et al., 2017).

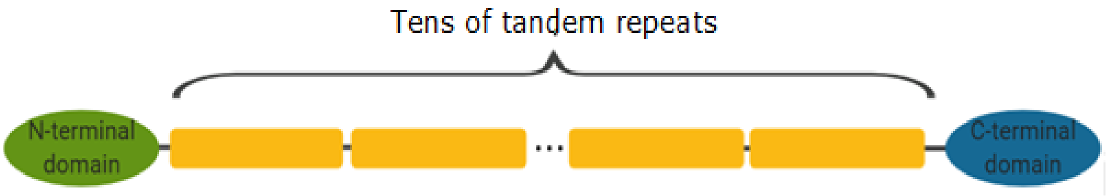


Figure 1.2 - **Schematic representation of spidroin domain architecture**. The terminal domains (green and blue) have unique sequences (within the peptide) and strong sequence conservation within the protein family. These domains enable spidroin-spidroin interactions (Xu and Lewis, 1990). The majority of the peptide’s sequence is composed of the central domain (yellow). This is a very

repetitive region composed of numerous tandem repeats that vary largely in size and composition (Bratzel and Buehler, 2012; Gosline et al., 1999).

Individual peptide repeats are usually iterated 25-30 times and complete spidroins' sizes are between 250-350 kDa on average. This length enables the structures that are assembled from these proteins to rely on fewer intermolecular interactions which overall contributes to fibres that do not break easily (Bratzel and Buehler, 2012; Xu and Lewis, 1990; Zheng and Ling, 2019).

During fibre formation, the native N-terminus is indispensable. Terminus-terminus interactions are directly involved in bringing two spidroins together (Ries et al., 2014). While the terminal domains are largely invariable in the spidroin gene family, the central domains are quite variable from homolog to homolog. It is this variability which gives different spidroin types their variable properties (Gatesy et al., 2001).

A silk fibre may be composed of several paralogous spidroins. For example in dragline silk (used by spiders for the outer rims and spokes in their orb webs) up to three major ampullate spidroins (MaSp1, 2 and 3) can be found. Non-spidroin structural proteins have also been identified but they have not been studied in detail (Hinman and Lewis, 1992; Kono et al., 2019).

Each repetitive unit may be composed in part of smaller homorepeats like the poly-alanine tracts (common in major ampullate spidroins) that assemble into β-sheet structures and give the protein its strength (Hayashi et al., 1999). Several other motifs can also make up the repeat units. GGX motifs (glycine-glycine then usually either alanine, leucine, glutamine, tyrosine or serine) enable some spidroin regions to fold into 3^1-helical secondary structures (Figure 1.3B). Such helices form parts of an amorphous region within the assembled silk fibre. They function as springs, by uncoiling and coiling as a fibre is pulled on or relaxed thus conferring elastic properties to dragline and flagelliform silk fibres (Becker et al., 2003; Hayashi et al., 1999; Kummerlen et al., 1996). The poly-alanine and GGX-rich tracts alternate within a spidroin's peptide sequence (Figure 1.3A).

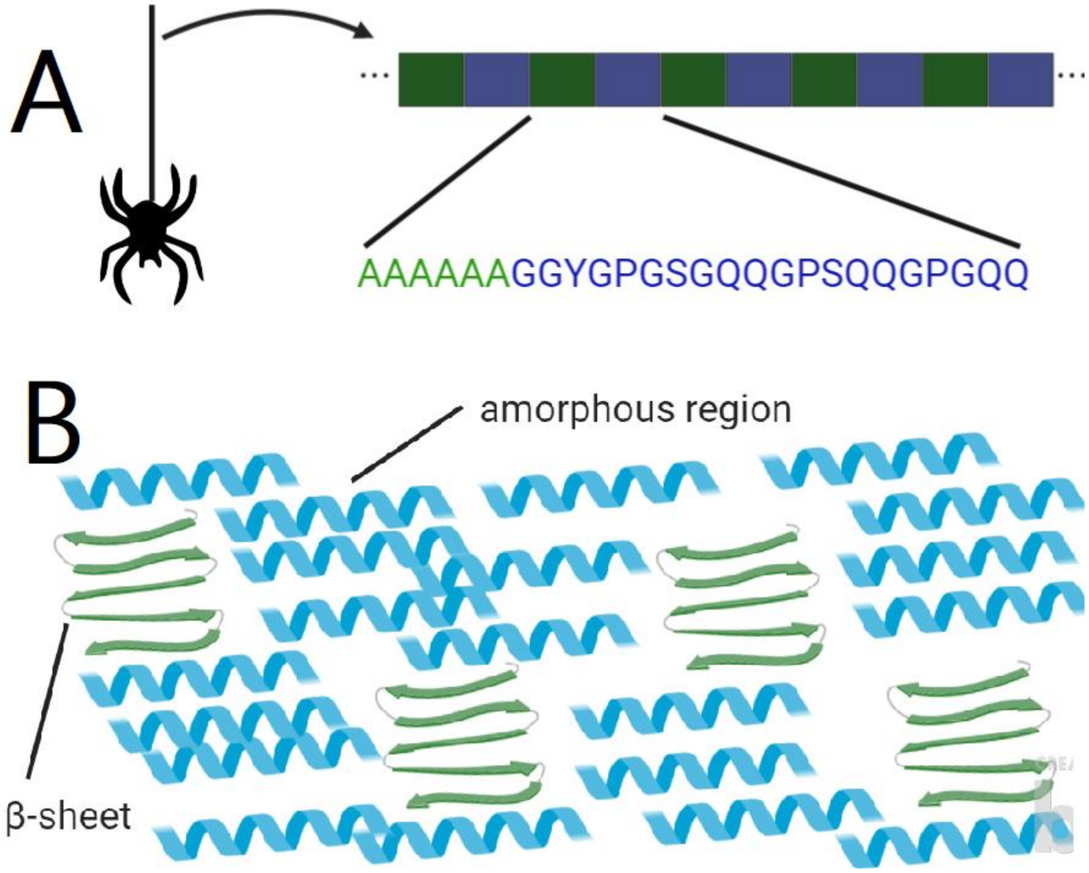


Figure 1.3 – **Assembly of MaSp into higher order structures**. (A) A representation of the alternating MaSp tandem repeat regions from the primary structure is shown in green (poly-alanine tracts) and blue (glycine-rich). (B) The MaSp silk assembles into fibres which are composed of amorphous, 3^1-helix-containing (blue) and β-sheet secondary structures (green).

Other amino acid motifs are also common throughout the spidroins. The Gly-Pro-Gly-X-X (GPGXX) motif folds into an elastin-like β-turn which also contributes to spidroin elastic properties and is similarly found in MaSp and FLAG spidroins (Jenkins et al., 2010). Gly-Gly-

Tyr (GGY) and Gln-Gln (QQ) motifs are less common and predominantly located in MaSp2. These motifs are most likely involved in noncovalent intermolecular binding between MaSp2 molecules and regional exclusion of MaSp1 resulting in a pattern within a fibre that is favourable for elasticity. This knowledge can be exploited by varying the relative quantities of MaSp1 and 2 thereby producing adjustable elasticity of the silk (Malay et al., 2017).

In the case of aciniform spidroins, these short motifs are very uncommon but large repeat regions are ubiquitous (200-375 amino acids). These regions confer slightly different mechanical properties to the aciniform fibres but their overall secondary structure again is mostly β-sheet/helical (Tremblay et al., 2015).

Pyriform spidroins for example possess repeat units that have a more complex structure and a length exceeding 200 amino acids. These are enriched in glutamines and prolines (Chaw et al., 2017).

Due to the presence of short repeat units in some of the spidroins, indel events have generated a large amount of heterogeneity between repeats. This is potentially in part due to polymerase slippage and recombination events and is most likely responsible for the large variability in protein lengths that have evolved in the superfamily (Garb and Hayashi, 2005; Zhou et al., 2021). Conversely, it has been observed in aciniform spidroins that selection pressures drive a homogenisation of these repeat units, thus reducing the potential for faulty silk production in spiders (Ayoub et al., 2013).

Interactions between regions of the central domain overall contribute to the proper assembly of spidroin fibres in the spider's silk gland (Bosia et al., 2010). The amino acid motifs discussed here are summarised in Table 1.1 below.

Table 1.1 – **Summary of spidroin properties and their repeat motifs.**

Silk Name	Property	Amino Acid Motifs	Reference
Aciniform (AcSp)	High toughness (resistance to deformation)	TTX, SSX, XQQ, GGX	(Hayashi et al., 2004)
Aggregate (AgSp)	Aqueous, viscous adhesion	GGX, GPGGX, SSX, TTX	(Sahni et al., 2010; Stellwagen and Renberg, 2019)
Cribellate (CrSp)	Dry adhesion through van der Waals forces	SSX	(Correa-Garhwal et al., 2019; Hawthorn and Opell, 2002)
Cylindrical (CySp)	High toughness	TTX, A_n, XQQ, SSX	(Zhao et al., 2006)
Flagelliform (FLAG)	High elasticity	$(GA)_n$, GGX, GPGGX, SSX	(Hayashi and Lewis, 1998; Hayashi et al., 1999)
Major Ampullate 1 (MaSp1)	High ultimate strength (resistance to pulling)	$(GA)_n$, A_n	(Lawrence et al., 2004)
Major Ampullate 2 (MaSp2)	High elasticity	TTX, SSX, XQQ, GGX, A_n, GPGGX, GPGQQ	(Lawrence et al., 2004); (Sponner et al., 2005)
Minor Ampullate (MiSp)	Elasticity and toughness	$(GA)_n$, A_n, GPGQQ	(Colgin and Lewis, 1998; Guinea et al., 2012; Vienneau-Hathaway et al., 2017)
Pyriform (PySp)	High adhesive strength and flaw tolerance	XQQ, SSX	(Wang et al., 2019; Wolff et al., 2015)

1.1.4 Natural Spidroin Expression and Protein Folding

The natural process of spidroin expression and fibre formation in spiders is remarkably complex. This complexity is very important in successful silk fibre formation and is difficult to mimic artificially in industrial settings. This is partly why natural silk fibres' properties are difficult to replicate. There are dedicated silk gland organs in spider abdomens that are responsible for silk formation. Each type of silk is synthesised by a separate gland. Only a small number of spidroins are expressed in each gland (Jorge et al., 2022).

In general, the silk gland is composed of four distinct regions (Figure 1.4). Spidroins are expressed and secreted into a lumen known as the tail. They are then stored in a concentrated aqueous solution known as dope (typically 30-50% spidroin) in a sac termed the ampulla (Rising and Johansson, 2015; Yamaura et al., 1985). Here the spidroins aggregate in micelles with their hydrophobic domains oriented inward and their more hydrophilic termini oriented toward the aqueous solution (Eisoldt et al., 2012)https://doi.org/10.1002/bip.22006. Next, the solution passes through a narrowing S-shaped duct (Heidebrecht et al., 2015). Here an ion gradient and a decreasing pH gradient result in dimerization of the N- and C-terminal domains. Water is gradually reabsorbed as the solution travels through the gland. The spherical micelles are gradually stretched into longer fibres and the shearing forces stimulate the correct folding of the beta-sheets within the hydrophobic central domains (Andersson et al., 2014; Heidebrecht et al., 2015; Vollrath and Knight, 2001). Finally, the spider can control the thickness of the fibre through a special valve located near the end of the silk gland (Vollrath and Knight, 1999).

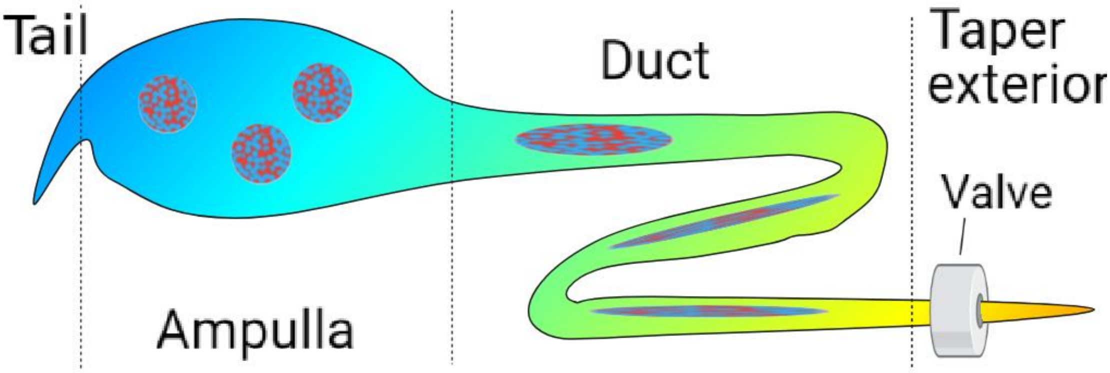


Figure 1.4 – **Schematic representation of a spider silk gland**. There is a pH gradient inside the gland from slightly basic (dark blue) to acidic (orange). This ensures the proper folding of the spidroins. Cells in the tail region produce spidroins which assemble into micellar structures in the ampulla. Under decreasing pH and gradual physical extrusion, the spidroins start to form slightly elongated shapes in the duct region. Finally, a muscle-controlled valve is involved in constraining the space of the spidroins to form fibre structures to the exterior of the spider.

Normally, several emerging silk threads assemble into a single thicker silk fibre (Alfaro et al., 2018). However, cribellate silk is produced from a specialised organ termed a cribellum. This sieve-like structure covers the exit of silk from the glands and enables this silk to form numerous microscopic threads. The fine structure and large surface area give cribellate silk its specific adhesive properties through intermolecular van der Waal's forces (Michalik et al., 2019).

Additionally, there are non-spidroin proteins that are thought to have a role in fibre formation and stability but these have not been studied in as much detail (Kono et al., 2019).

1.2 Spidroin Applications

1.2.1 Engineering and Industrial Applications

Spider silk has been used since pre-industrial times in a variety of applications such as a material for paintings, a component of musical instruments, crosshairs for rifles and clothing (Blench, 2009; Garner, 1955; Hock, 2008; Levene, 2012).

More recently synthetic spider silk has been utilised in clothing. Fabrics with spider silk are wear-resistant and tear-resistant. They are also very light (Spiber Inc., 2019).

The impressive tensile strength, elasticity, resistance to fibre fatigue and toughness of dragline silk in particular have been of interest in materials that need to withstand large and intermittent forces. An example is in components of car seats. Silk’s resilience and elasticity enable the seat to adapt to the driver’s acceleration during driving while maintaining its properties (Spiber Inc., 2019).

Dragline silk has also been studied as potential material for armour where it exceeds the strong synthetic fibre Kevlar in several properties, especially in flexibility (Hansel, 2019; Teule et al., 2012).

1.2.2 Medical Applications

Spidroins’ mechanical properties and biocompatibility make them particularly well suited for tissue engineering applications. In the case of musculoskeletal tissue engineering modifiable, resilient and adhesion-favourable scaffolds are needed to anchor different cell types. Spidroins can form matrices, films, sponges and hydrogels to enable the differentiation of mesenchymal stem cells into bone cartilage and muscle (Arndt et al., 2022; Chiasson et al., 2016; Kuhbier et al., 2010; Rawal et al., 2020; Yao et al., 2016). Silks are expected to undergo proteolytic cleavage gradually within an organism. Turnover rates of silk-based artificial scaffolds can be regulated by different methods of fibre assembly before implantation in the body (Dinjaski et al., 2018; Hennecke et al., 2013). Fibres from high molecular weight spidroins could be particularly valuable in materials that support engineered artificial tendons for implantation or for sutures in wounds (Cheng et al., 2022; Hennecke et al., 2013).

Electrospinning is a method where a charged spidroin solution is ejected under the influence of an electric field to produce fibres similar to natural silk (Zhong, 2016; Zhou et al., 2008). Lyophilisation (freeze-drying) is when water is removed from a frozen spidroin solution and this method can produce sponge-like structures (Chiasson et al., 2016; Schacht et al., 2016). Both of these methods can be adjusted to produce structures of variable dimensions (e.g. fibres between 200 nm - 500 nm).

Neural tissue-like organoids have been produced in the past by combining fibrous spidroin scaffolds with neural precursor stem cells. Furthermore, these have been successfully implanted in Rhesus macaques which further supports their biocompatibility (Baklaushev et al., 2019). Additionally, sponge structures from spidroins have been used as scaffolds for murine fibroblast cells, proving the potential for use in more organoid types (Moisenovich et al., 2011). While spidroins are favourable to cell adhesion, the addition of specific amino acid motifs such as the fibronectinmimicking RGD could further improve cell attachment (Wohlrab et al., 2012).

Hydrogels are structures with interconnected polymer chains which absorb and hold a proportionally large amount of water or aqueous solution. Such structures have been generated using recombinant spider silk. Since the aqueous solution within can contain a wide variety of drugs, various biologics have been proposed for adsorption (Kumari et al., 2018). For example, spidroin hydrogels containing growth factors can be used to stimulate cartilage differentiation and growth in mesenchymal stem cells (Kuhbier et al., 2010). Less complex applications where anti-cancer drug-containing hydrogels are topically administered to a tumour resection site have been explored (Chiu et al., 2014). Factors such as spidroin concentration, gelation temperature, ion concentration, and ultrasonic wavelengths can be altered to convert a hydrogel into liquid form. This way, spidroins can be injected into an organ and subsequently assume their hydrogel conformation inside of the body if needed (Bai et al., 2014). Combined spidroin and hyaluronic acid hydrogels have been also studied as potential scaffolds for neural tissue in spinal cord injury repair (Lin et al., 2021b).

Small structures like nanoparticles and microparticles have been produced from spider silk in the past. Since their turnover rates are predictable they have been proposed as potential intravenous drug delivery vehicles (Lammel et al., 2011). The nanoparticle generation process can again be moderated to produce variably sized structures (Lammel et al., 2008). However, research into this avenue has not progressed substantially over the past decade and the rest of the spidroin structures are favoured over micro- and nanoparticles for drug delivery (Florczak et al., 2021).

Fusion proteins between spidroins and a heparin-binding protein have been produced to create anticoagulant silk. This material takes advantage of the spidroins' natural antibacterial property and heparin's anticoagulant property. Such materials can be used to produce a variety of biocompatible coatings for medical devices (Mulinti et al., 2022).

1.3 Approaches in Spidroin Production

1.3.1 Collection from Live Spiders

In the past, spider silk has been an interesting material for human use but certain factors make naturally obtained spider silk infeasible for most uses. Devices where spiders are restrained and their silk is mechanically drawn from their glands have been available for three centuries (Figure 1.5A) (Bon, 1710). Attempting to collect spider webs from the wild has proven to be too laborious to be feasible on a large scale (Work and Emerson, 1982). Unlike silkworms which have been used in the manufacturing of textiles for centuries, spiders are very territorial and cannibalistic which renders any attempts of rearing them together impossible (Scheibel, 2004). However, extracted fibres from live spiders are still used to study the properties of natural silk (Figure 1.5B) (Lin et al., 2021a; Work and Emerson, 1982).

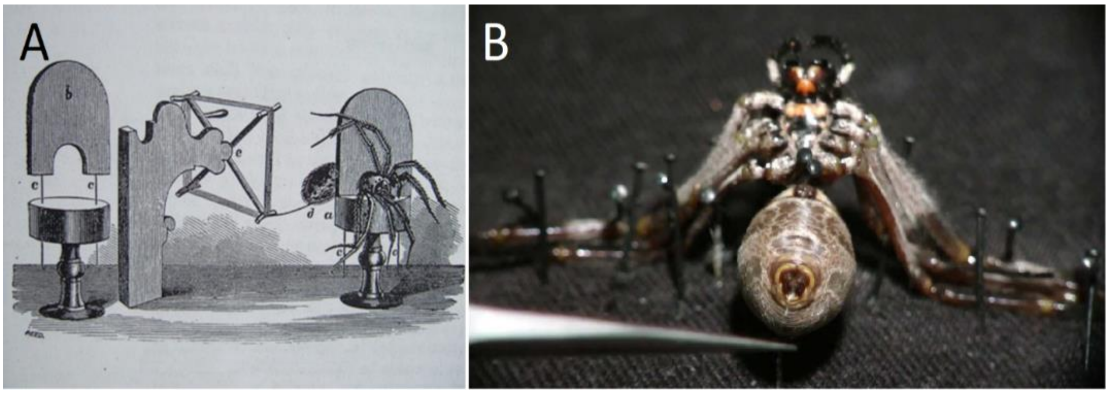


Figure 1.5 – **Live spider silk extraction methods**. (A) An early device for live spider silk extraction from immobilised spiders (Bon, 1710). (B) Contemporary silk extraction method (Oxford Silk Group, 2013).

Repurposed, natural, web-derived spidroins can be solubilised and then re-folded into fibres. This process allows some adjustment of the silk material that is to be made. However, such re-formed fibres maintain only a part of the mechanical strength and toughness of the original material (Shao et al., 2003).

1.3.2 Recombinant Approaches

A major obstacle in recombinant spidroin expression is sequence repetitiveness. Repetitive DNA sequences are especially prone to various unwanted recombination events, replication and transcriptional errors (Fahnestock and Bedzyk, 1997; Rising et al., 2007; Tang and Chilkoti, 2016). Indel mutations, transcription errors and translation pauses have also been observed particularly in *Escherichia coli* (Tokareva et al., 2013).

However, researchers have been able to express spidroins comparable in size and function to natural spidroins in *E. coli* at sizes of 284.9 kDa and 556kDa (Bowen et al., 2018; Xia et al., 2010). Especially in the latter case, the length of the recombinant product was associated with fibres high in strength (1.03 ± 0.11 GPa) and toughness (114 ± 51 MJ/m^3). Part of this approach involved metabolically engineering the *E. coli* to produce more glycyl- and alanyl-tRNA complexes to satisfy the requirements of spidroin expression. Attempts at spidroin expression without an optimised amino acyl-tRNA pool have been inefficient in *E. coli* (Cao et al., 2017; Fahnestock and Bedzyk, 1997). Overall, a major caveat of *E. coli* as a host is that many sequential purification steps are required in the extraction of the proteins (partly because the protein is expressed intracellularly). Additionally, their translation machinery is prone to premature translation termination (Bowen et al., 2018). This has been partly remedied with metabolic engineering but there still was a need to use split-intein-based protein polymerisation (proteins that covalently assemble at their termini to form larger concatemers) to achieve high protein sizes. This could render large-scale production too expensive to be cost-effective. Finally, spidroins from *E. coli* have shown issues with maintaining their solubility during expression and purification which further decreases their effective yields (Whittall et al., 2021).

Similarly, a short (33 kDa) recombinant spidroin was expressed in *E. coli*. While the overall yield was substantial at >20 g L^{-1}, the tensile strength of the resulting fibre was low (0.1 GPa) (Schmuck et al., 2021). Another bacterial expression system (*Corynebacterium glutamicum*) was used to similarly produce a high concentration (554.7 mg L^{-1}) but of only 16-repeat-long spidroins and therefore a much lower tensile strength (0.17 GPa) (Jin et al., 2022). *Salmonela typhimurium* has also been used as a host. This bacterium has the advantage of being able to secrete recombinant protein. However, the yields from this system were smaller than from metabolically engineered *E. coli* (Widmaier et al., 2009).

The yeasts *Pichia pastoris* and *Saccharomyces cerevisiae* can robustly and reliably express recombinant spidroins. However, their maximum spidroin sizes have been constrained to 65 kDa and 94 kDa respectively (Fahnestock and Bedzyk, 1997; Sidoruk et al., 2015).

Several plant host systems have been utilised to express recombinant spidroins including rice (*Oryza sativa*), tobacco (*Nicotiana tabacum*), potato (*Solanum tuberosum*), *Arabidopsis thaliana* and alfalfa (*Medicago sativa*) (Edlund et al., 2018; Hauptmann et al., 2013a; Park et al., 2019; Scheller et al., 2001; Yang et al., 2005).

In general, plants offer cheap and scalable production. However, spidroin yields have been relatively low and spidroin size has been limited to 127 kDa at most. This has been partly remedied by using the split intein protein polymerisation approach to produce flagelliform spidroins at 450 kDa (Weichert et al., 2016). There is also a unique disadvantage in plant recombinant systems – gene silencing. This is particularly prevalent in repetitive sequences and could severely impact transgenic spidroins (Stam et al., 1997). Furthermore, purification from plant tissues may be even more expensive than from *E. coli* (Hauptmann et al., 2013b).

The larvae of the silkmoth (*Bombyx mori*) known as silkworms have been used for centuries in the production of the silk that is used in textiles on a large scale nowadays. Silkworms are particularly suitable for large-scale production in part because they can be reared together in close proximity. Also, the silk fibroin genes that silkworms express naturally have elevated proportions of glycine (42.9%) and alanine (30%) and their amino acyl-tRNA pools have adapted accordingly (Asakura and Suzuki, 2014). Researchers have successfully attempted several strategies to produce spidroin proteins from these hosts. More recently, transgenic spidroins were produced in *B. mori* using CRISPR/Cas9 to introduce a spidroin-encoding transgene to the silkworm genome. This resulted in the expression of spider silk (alongside silkworm silk) which was comparable to native spider silk in mechanical properties (Zhang et al., 2019b).

In theory, mammalian cell lines should be capable of expressing large, repetitive genes more efficiently and reliably than yeast, plants or bacteria. Recombinant spidroins at sizes up to 140 kDa have been reported from bovine mammary epithelial alveolar cells (MAC) and baby hamster kidney cells (BHK). These systems took advantage of mammalian secretion pathways but ultimately they suffered from low spidroin yields (Heidebrecht and Scheibel, 2013).

Since certain cell types and organs within mammals readily secrete large quantities of protein, researchers have attempted to generate transgenic strains of mice and goats that would express spidroins in milk (Jones et al., 2015; Xu et al., 2007). However, their length did not exceed 66 kDa and their yields were smaller than in mammalian cell culture. So far expression of spidroins in milk has mostly been held back due to the expenses and time needed to produce transgenic mammalian strains. Such attempts will probably remain infeasible at least in the near future (Whittall et al., 2021).

Due to the difficulty in producing long recombinant spidroins alternative approaches have been developed that sacrifice molecular weight to obtain high yields and protein homogeneity. These short spidroins typically have 2-6 spidroin repeats and one of the termini in order to maintain solubility in aqueous solutions. Their small size allows efficient recombinant expression to occur from a bacterial host. Overall, these types of spidroins benefit from the biocompatibility of spidroins and have been developed for potential use in a biomedical context (Fredriksson et al., 2009; Li et al., 2021; Schmuck et al., 2021). Additionally, some of these proteins can be functionalised by the covalent binding of functional molecules to the protein such as antibiotics or growth factor ligands (Harvey et al., 2017; Zhao et al., 2014). Relative to larger spidroins, these shorter ones benefit from simple solubilisation and purification steps and high yields have been reported (up to 14 g/l). While these peptides may have some interesting mechanical properties, their corresponding fibres show tensile strengths that are much lower than that of longer proteins (Table 1.2).

Table 1.2 – **Summary of recombinant spidroin synthesis approaches** (adapted and improved from Whittall et al., 2021).

Expression Host	Spidroin Homolog	Recombinant Spidroin Size (kDa)	Maximum Yield	Tensile Strength	Reference
Escherichia coli	MaSp1/ MiSp1	33	14 g/l	0.1 GPa	(Schmuck et al., 2021)
	MaSp1	284.9	0.5 g/l	0.5 GPa	(Xia et al., 2010)
	MaSp1	556	1.24 g/l	1.03 GPa	(Bowen et al., 2018)
Corynebacterium glutamicum	MaSp1	168.9	0.5 g/l	0.17 GPa	(Jin et al., 2022)
Salmonella typhimurium	MaSp2	25-56	0.014 g/l	Not reported	(Widmaier et al., 2009)
Pichia pastoris	MaSp1	65	0.66 g/l	Not reported	(Fahnestock and Bedzyk, 1997)
Saccharomyces cerevisiae	MaSp1	94	0.4 g/l	Not reported	(Sidoruk et al., 2015)
Oryza sativa	MaSp1	22	Not reported	Not reported	(Park et al., 2019)
Nicotiana tabacum	FLAG	>460	190 mg/kg	Not reported	(Weichert et al., 2016)
Solanum tuberosum	MaSp1	99.8	0.5% of total soluble protein	Not reported	(Scheller et al., 2001)

Table 1.2 (continued) – **Summary of recombinant spidroin synthesis approaches** (adapted and improved from Whittall et al., 2021).

Expression Host	Spidroin Homolog	Recombinant Spidroin Size (kDa)	Maximum Yield	Tensile Strength	Reference
Arabidopsis thaliana	MaSp1	64-125	18% of total soluble protein	Not reported	(Yang et al., 2005)
Bombyx mori	MaSp1/ MiSp1	300	Not reported	1.2 GPa	(Zhang et al., 2019b)
Baby Hamster Kidney cells (*Mesocricetus auratus*)	MaSp1/ MaSp2	140	0.05 g/l	0.26 GPa	(Lazaris et al., 2002)
Mus musculus	MaSp1/ MaSp2	55	0.011 g/l	Not reported	(Xu et al., 2007)
***Capra hircus* (goat)**	MaSp1/ MaSp2	65	Not reported	Not reported	(Whittall et al., 2021)
***Nephila clavipes* (native form)**	MaSp1 MaSp2	>300	N/A	0.8–1.2 GPa	(Ko et al., 2001; Zemlin, 1968)

1.3.3 Recombinant Silk Processing

Provided that recombinant spidroins are solubilised into liquid form, they can be used to produce a wide variety of structures to be applied in medical or engineering contexts. While the processing of solubilised spider silk into solid form is a complex process in the absence of the natural gland, there have been attempts to mimic some or all of these conditions. If a high-quality, stable source of spider silk is available, processes like lyophilisation, shearing, wet-spinning and electrospinning can be used to produce various silk structures (Rising and Johansson, 2015; Thamm and Scheibel, 2017). A major factor in silk quality is spidroin length. Longer silk peptides are expected to have more sites to form the aforementioned non-covalent intramolecular interactions, leading to better cohesion and therefore stronger fibres (An et al., 2011; Ayoub et al., 2007; Lin et al., 2015b; Prince et al., 1995).

1.4 Rolling Circle Translation

Rolling circle translation (RCT) (Figure 1.6) is a method that could potentially resolve many limitations of recombinant spidroin production, particularly the caveats that originate from repetitive DNA sequences. This approach utilises a specialised circular RNA (circRNA) which is created by covalently linking the ends of a linear RNA molecule 5′ to 3′. The circRNA is designed to encode a short non-repetitive protein. However, through successive rounds of looped protein translation, such a molecule could help in expressing very large repetitive protein concatemers at sizes exceeding 300 kDa (Perriman and Ares, 1998; Zheng and Ling, 2019). This may be an invaluable approach in recombinant spidroin synthesis.

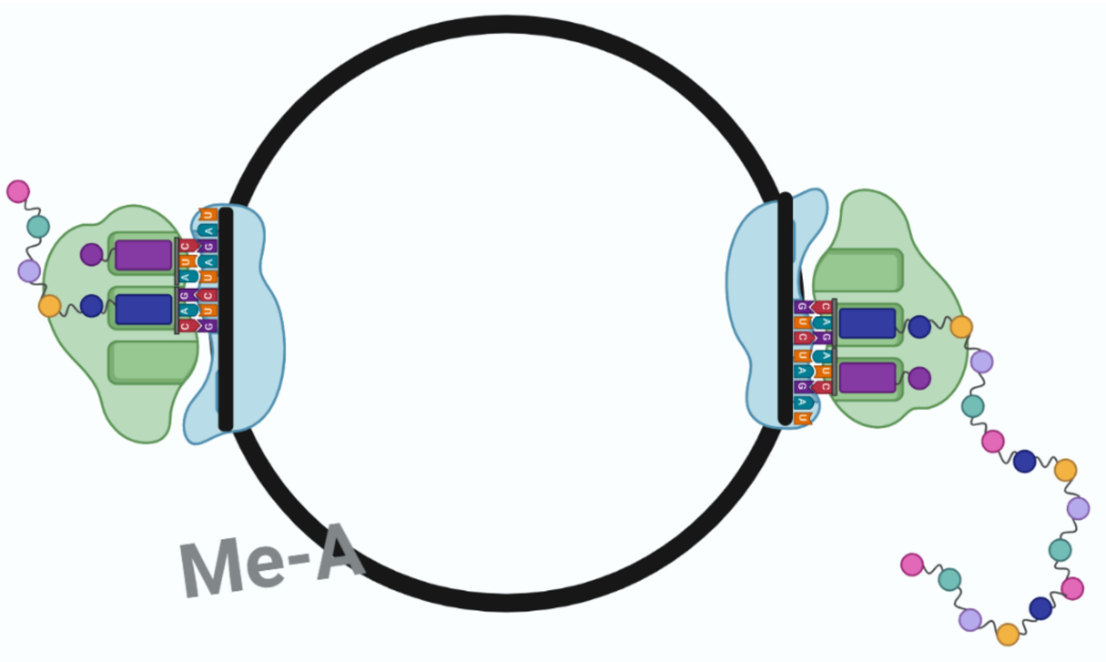


Figure 1.6 **– Schematic representation of a translating circRNA.** Translation initiation occurs at a ribosome recruitment site such as a methylated adenosine nucleotide (grey). The lack of stop codons ensures no translation termination will occur. A new iteration of the encoded peptide is added to the nascent peptide chain with each translation cycle.

An important requirement in the case of circRNA translation is a site for translation initiation. Essentially, the entirety of the circRNA is protein-coding and no part of it should contain in-frame stop codons. This is to ensure that translation is looped over and over in a cyclical manner thus producing its repetitive product. Additionally, the total length of the RNA must be a multiple of three to ensure that no change in frame follows the translation cycles. Translation termination most likely occurs stochastically after multiple rounds of translation. A circRNA where these conditions are met will be referred to here as an infinite open reading frame (IORF).

The main advantage of this approach is that it entirely avoids the use of repetitive nucleotide sequences which averts the aforementioned replication and transcription errors. Uniquely, this type of translation would result in a percentage of the overall peptide sequence deriving from RNA sequences that have no structural relevance and are instead required in translation initiation or extracellular export. Such non-structural peptide regions could result in novel properties for the spidroin peptides but are also very likely to disrupt some of the properties of silk compared to its natural counterpart.

1.5 Aims and Objectives

Despite the numerous recent advances in recombinant spidroin expression, there are several caveats to the established approaches. In general, current systems are capable of producing either high titres of low molecular weight spider silk proteins or high molecular weight peptides but at low levels. The unique aspects of RCT are expected to meet both of these goals.

This project aims to investigate whether a recombinant spidroin can be expressed through RCT (aim 1). Another aim is to characterise the recombinant spidroin and assess its average length and its expression efficiency (aim 2).

Additionally, several of the current approaches which use bacterial expression systems are hindered by the difficulty in protein purification after translation. Mammalian cultured cells were chosen as expression hosts. This is meant to investigate if mammalian RCT is a suitable strategy in recombinant spidroin production (aim 3).

The vast diversity of spiders and their silks has potentially left many undiscovered spidroin sequences. Knowing how intriguing and potentially useful known spidroins are, it may be particularly valuable to identify more members of this gene family. Therefore, the work described here aims to expand the collection of known spidroins by designing and implementing a pipeline that leverages spidroins' repetitive character in the identification of potentially previously overlooked sequences (aim 4). Furthermore, this pipeline inherently is suitable for the identification of other repetitive proteins. The work described here finally aims to test the designed pipeline in a context outside of the spidroins (aim 5). Another protein family – that of the squid suckerins (Chapter 5) shares similarities and differences with spidroins and was chosen for homolog identification.

1.6 Thesis Outline

In Chapter 3 there is a detailed description of the molecular cloning strategies employed in this work. The chapter begins with an introduction that explains and describes the different vital components that need to be cloned into the same plasmid in order to produce translating circRNA in mammalian cells. Next, the actual RNA and DNA sequence design is explained in detail with a focus on the DNA gene fragments that were to be utilised in the subsequent molecular cloning. Finally, the chapter shows the various cloning approaches that were implemented to generate a library of circRNA-generating DNA plasmids. These plasmids were then vital in the recombinant IORF expression of spidroins in the subsequent chapter.

Chapter 4 focuses on the transfection of mammalian host cells with the aforementioned plasmids and on the characterisation of their resulting circRNA and possible recombinant protein. In the introduction section, there is a description of the various processes that need to be considered in this recombinant expression: plasmid transfection, the intracellular transport of circRNA and dynamics of tRNA in IORF translation. A summary of previous recombinant methods in spidroin synthesis is also included with a focus on the hosts used. The chapter goes on to describe how the cloned plasmids were used to transfect mammalian cells and produce circRNA. A set of protein characterisation experiments are explained that attempted to detect recombinant spidroin. Finally, the dynamics of translation elongation and tRNA charging in the context of IORF translation were investigated.

Chapter 5 describes in detail the computational analysis undertaken to identify members of the spidroin superfamily. The chapter begins with an introduction about the known diversity of spidroins and an outline of the methods currently used in spidroin identification. Subsequently, the construction of a bioinformatics pipeline is described in detail. This is then implemented to search for new members of the spidroin family in a diverse set of spider transcriptomes. To further explore the capabilities of this pipeline a separate analysis of repetitive protein sequences is explained. The biology of the repetitive suckerin proteins is outlined in the introduction. Later the pipeline is implemented in the discovery of new suckerins. Their relationships were further characterised through phylogenetic analysis.

2 Materials and Methods

Unless specified otherwise, the reagents described here were obtained from Sigma or Thermo Fisher Scientific. The laboratory-grade chemicals were purchased in dry form and nuclease-free items were chosen when possible.

2.1 Mammalian Cell Culture Methods

2.1.1 Cell Lines and Maintenance

Four types of cell lines were used throughout this study - human breast cancer MCF-7 cells (Michigan Cancer Foundation - 7) (ATCC, Cat# HTB-22); *Cercopithecus aethiops* kidney fibroblasts COS-1 (ATCC, Cat# CRL-1650); mouse embryonic fibroblasts NIH/3T3 (ATCC, Cat# CRL-1658); human embryonic kidney cells HEK293T (ATCC, Cat# CRL-11268). DMSO-preserved cell lines were taken from liquid nitrogen storage and thawed rapidly in a 37°C water bath. Afterwards, they were transferred to T25 or T75 cell culture flasks (Sarstedt, Cat# 83.3911.002) and cultured in 5-10 ml Dulbecco's Modified Eagle media (DMEM), with 4.5 g/L - Glucose, with L-glutamine, without sodium pyruvate (Sigma Aldrich, Cat# D5671 or Lonza Bioscience, Cat# BE12-741F) supplemented with 10% foetal bovine serum (FBS – Fisher Scientific, Cat# 11550356).

Cell lines were maintained in incubators (Sanyo, MCO-17AIC Incusafe) at 37°C with 5% CO_2 and high humidity.

Upon reaching 70-90% confluence cell lines were passaged. Spent cell media was removed and cells were washed with phosphate-buffered saline (PBS - 137 mM NaCl, 2.7 mM KCl, 8 mM Na2HPO4, and 2 mM KH2PO4, pH~7.4). Next, 3.75 ml of room temperature 0.25% trypsin-EDTA (Fisher Scientific, Cat# 10779413) in PBS was added to the adherent cells. This was followed by pelleting of the cells via gentle centrifugation (10 minutes at 130 g) and resuspension in 10 ml DMEM-FBS to achieve a final confluence of 20-30%.

2.1.2 Transient Transfection

Host cells were cultured in 6-well plates (Sarstedt, Cat# 83.3901.300) with 3 ml DMEM-FBS 24h before transfection. Plasmid DNA (3 µg per well) was transfected into the aforementioned cell lines using 9 µg per well of non-liposomal reagents – either PEI MAX (Polysciences, Cat# 49553-93-7) or FuGENE 6 (Promega, Cat# E2693), according to the manufacturer’s instructions. DNA was diluted in reduced serum media Opti-MEM (Thermo Fisher Scientific, Cat# 31985062) at a concentration of 6.67 µg/ml. FuGENE 6 or PEI was diluted in Opti-MEM at a concentration of 20 µg/ml. Both of these dilutions were mixed at room temperature for 15-30 min to achieve a ratio of 3:1 transfection reagent to DNA. Subsequently, this mixture (200-300 µl) was introduced to the cultured cells and was incubated with them for at least 24h.

2.1.3 Fluorescence Microscopy

A Nikon Eclipse TS100 inverted microscope (Cat# NI-TS100) was used in conjunction with a CoolLED pE-100 illumination system (Cat# PE-100) to visualise GFP fluorescence in transfected mammalian cells. The number of fluorescent cells was counted in 2 fields of view per well and the average percentage of fluorescing cells was calculated by dividing the number of fluorescent cells by the total number of cells.

2.2 Bacterial Methods

To avoid contamination of reagents and instruments all bacterial media and glassware were autoclaved for 20 min at 121°C. Treatments or transfers of bacterial cultures between plates/tubes were carried out next to a Bunsen burner. Sterile disposable plastics were used throughout.

2.2.1 Strains Used and Preparation of Competent Cells

The bacterial strains used throughout this work were either XL 10-Gold ultracompetent *Escherichia coli* (Agilent, Cat# 200317) or DH5ɑ chemically competent *E. coli* (either from Thermo Fisher Scientific, Cat# 18265017 or prepared manually). Manual preparation of chemically competent DH5ɑ involved an initial growth of DH5ɑ cells in 5 ml lysogeny broth (LB – 1% Tryptone, 0.5% yeast extract, 1% NaCl) overnight at 37°C from a single starter colony. Next, the outgrown culture was used to inoculate 250 ml of fresh LB whose optical density (OD600) was measured with a spectrophotometer until it reached the range of 0.4-0.6. The cells were subsequently centrifuged at 4000 revolutions per minute (RPM) for 10 minutes at 4°C and their media were replaced with ice-cold transformation buffer 1 (TFB1 – 30 mM potassium acetate, 10 mM calcium chloride, 50 mM manganese chloride, 100 mM rubidium chloride, 15% glycerol, pH 5.8). They were incubated on ice for 5 minutes, centrifuged again and resuspended in transformation buffer 2 (TFB2 – 10 mM MOPS, 75 mM calcium chloride, 10 mM rubidium chloride, 15% glycerol, pH 6.5). After a second incubation on ice for 1 hour, the cells were aliquoted into microfuge tubes and snap-frozen by submerging them in liquid nitrogen for 1 minute.

2.2.2 Bacterial Transformation

Bacterial transformations with plasmid DNA and XL 10-Gold *E. coli* were performed as per the manufacturer's protocol. The ultracompetent cells were thawed on ice and aliquoted into 100 µl volumes. Each of these was supplemented with 4 µl β-mercaptoethanol and 50 ng of plasmid DNA was added to this. This was followed by 30 minutes of incubation on ice and the mixtures were heat-shocked at 42°C for 30 seconds. These were incubated on ice for 2 minutes and were supplemented with 900 µl of NZY broth (Agilent, Cat# 200317). Next, the cells were left to recover at 37°C with agitation for 1 hour. 200 µl of each transformation mixture were spread on ampicillin- (100 µg/ml) or kanamycin- (50 µg/ml) containing LB-agar plates and incubated overnight at 37°C.

Alternatively, DH5ɑ chemically competent *E. coli* were transformed by a different method. First, the competent cells were thawed on ice for 10 minutes. Next, they were incubated with plasmid DNA (50 ng DNA per 100 µl competent cells) on ice for 30 minutes. A heat shock was performed at 42°C for 30 seconds followed by 2 minutes of incubation on ice. Finally, the cells were supplemented with 900 µl LB and left to recover at 37°C for 1h with agitation. Again, 200 µl of each transformation mixture was spread on antibiotic-containing LB-agar plates and incubated overnight at 37°C.

2.2.3 Extraction of Plasmid DNA

For each plasmid of interest, small-scale DNA purification was performed using the NucleoSpin Plasmid QuickPure kit (Macherey-Nagel, Cat# 11902422). Single *E. coli* colonies were chosen and used to inoculate 3 ml of ampicillin- (100 µg/ml) or kanamycin- (50 µg/ml) containing LB overnight at 37°C with constant agitation. The cells were next pelleted by centrifugation at 4000 RPM for 10 minutes and suspended in 250 µl of the kit's resuspension buffer A1. Next, they were lysed with 300 µl of the lysis solution A2 (supplemented with alkaline protease) for 5 minutes at room temperature. The lysis solution was neutralised with 300 µl of buffer A3 and the entire resulting mixtures were centrifuged to remove debris. Next, the remaining aqueous supernatants were pipetted into the NucleoSpin Plasmid QuickPure kit's columns. After two washes with 450 µl of wash buffer AQ, the plasmid DNA was eluted in 20-50 µl of distilled water. Plasmid DNA concentrations were measured using the Nanodrop 1000 (Thermo Scientific).

2.3 Molecular Biology Methods

2.3.1 Agarose Gel Electrophoresis

Mixed solutions of nucleic acids were separated according to molecular weight through agarose gel electrophoresis. SYBR Safe and ethidium bromide 0.8-2% (w/v) agarose gels were prepared in 1x TAE (40 mM Tris, 40 mM acetic acid, 1 mM EDTA, pH 8.0) by briefly boiling the mixture and adding a fluorescent nucleic acid stain when the mixture reached 40-45°C. Either ethidium bromide (Sigma Aldrich, Cat# E1510) at a concentration of 0.5 µg/ml or 1x SYBR Safe (Thermo Fisher Scientific, Cat# S33102) were used as fluorescent stains. The gels were cast in plastic trays and left to cool to room temperature (typically 75-150 ml). Each sample of nucleic acid to be investigated by electrophoresis was supplemented with 6x of Purple loading dye (New England Biolabs, Cat# B7024S) or 6x TriTrack DNA Loading Dye (Thermo Fisher Scientific, Cat# R1161) to a final concentration of 1x. Next, the gels were submerged in an electrophoresis tank in 1x TAE and 12-50 µl of the samples were loaded into each well. Electrophoresis was carried out at 100-120V for 30-50 minutes followed by UV visualisation with Gel Doc XR+ (Bio-Rad, Cat# 1708195).

2.3.2 Polymerase Chain Reaction

To amplify plasmid DNA, cDNA or DNA fragments the Phusion High-Fidelity DNA polymerase was used (New England Biolabs, Cat# M0530) in polymerase chain reactions (PCRs). Generally, the manufacturer’s guidelines were followed. A total volume of 20 µl was prepared per reaction. Template DNA was added in quantities between 1 ng and 200 ng. Alternatively, colony screen PCRs were also performed by using a picked colony as the DNA template. Either the Phusion HF or GC buffers (New England Biolabs, Cat# M0530) were used at a final concentration of 1x, alongside 0.5 µM for each DNA primer, 3-5% DMSO, 200 µM of each deoxynucleotide and 0.4 units of Phusion DNA Polymerase. The reagents were mixed in a 0.2 ml thin-walled PCR tube and incubated in a Techne TC-512 gradient thermocycler (Cat# 11719372). Initial denaturation was for 30 seconds at 98°C followed by 22-35 cycles of 10-second denaturation at 98°C, 15-20 seconds of annealing at 55-72°C and 30-120 seconds of DNA extension at 72°C. A final extension step finished the reaction – 2 minutes at 72°C.

2.3.3 Extraction of Total RNA

For an eventual RNA circularisation assessment, total RNA was extracted from each well of the cultured cells using 1 ml TRIzol (Thermo Fisher Scientific, Cat# 15596026) following the manufacturer's protocol 24-72h after transfection. Cells were lysed in TRIzol and 0.1 ml of 1-bromo-3-chloropropane (Fisher Scientific, Cat# 11484190) was added to each sample. They were vortexed vigorously and allowed to stand for 15 minutes. Phase separation was carried out by centrifugation at 12000 g for 15 minutes at 2-8°C. The upper aqueous layer which contained RNA was supplemented with 0.5 ml isopropanol (Fisher Scientific, Cat# 184130250), mixed and allowed to stand for 10 minutes at room temperature. This was centrifuged at 12000 g for 15 minutes at 2-8°C to form an RNA pellet in each tube. These pellets were washed twice with 75% ethanol and resuspended in 50 µl distilled water. The resulting RNA concentrations and purity were measured with a Nanodrop 1000 spectrophotometer (Thermo Scientific, Cat# ND-1000).

2.3.4 cDNA Synthesis

The SuperScript IV Reverse transcriptase (Invitrogen, Cat# 18090010) was used to synthesise cDNA from RNA templates. For RNA circularisation assays 2 µM of gene-specific primers (both forward and reverse) were mixed with 2 mg of total RNA and 10 µmol of each dNTP in a microfuge tube (total volume 13 µl) and heated at 65°C to denature RNA and facilitate primer annealing. The primer sequences can be found in Table 4.1 (page 148). Next, 4 µl of 5x SuperScript IV reaction buffer, 1 µl of 0.1 M dithiothreitol (DTT), 8 units of RNase inhibitor (New England Biolabs, Cat# M0314 or Promega, Cat. # N2115) and 1 µl of SuperScript IV reverse transcriptase were added to the microfuge tube. This was incubated at 50-55°C to enable the reaction for 20 minutes and was followed by enzyme inactivation at 80°C for 10 minutes.

Alternatively, the template-specific DNA primers could be replaced with 1 µl of 50 µM random hexamer primers (Invitrogen, Cat# N8080127). This also requires an additional incubation at room temperature for 10 minutes before the incubation at 50-55°C.

Optionally, a ribonuclease R (RNase R) digestion step was included before the RNA denaturation step. 4 units of Lucigen's RNase R (Cat# RNR07250) were used to digest 4 mg of total RNA in a total reaction volume of 40 µl. This also contained the RNase R reaction buffer at a final concentration of 1x. Each RNA sample was digested at 37°C for 10 minutes. Afterwards, the enzyme was inactivated by heating at 65°C for 20 minutes.

2.3.5 Purification of Nucleic Acids

If purified samples of nucleic acids were needed such as for use in DNA ligation after restriction digestion, sample purification was undertaken either through ethanol precipitation or through column purification.

In the case of ethanol precipitation the DNA or RNA-containing sample was combined with 0.1 volumes of 3 M sodium acetate and 3 volumes of 4°C pure ethanol. This was left to precipitate overnight at -20°C. Next, this was centrifuged at 13000 RPM, at 2-8°C for 30 min to pellet the nucleic acids. Subsequently, the pellet was washed twice with 4°C 75% ethanol and after centrifugation at 16000 x g, at 2-8°C for 10 min the supernatant was removed by pipetting. The sample was left to air dry at room temperature for 15 min, followed by the addition of ultrapure water to re-dissolve the pellet.

Alternatively, a Monarch PCR & DNA cleanup kit was used (New England Biolabs, Cat# T1030) as per the manufacturer's guidelines. The DNA binding buffer from the kit was added to each sample at a ratio of 5:1 – buffer:sample. Next, the mixture was loaded on a centrifugation column from the kit and centrifuged for 1 minute at 16 000 g. This was washed twice with 200 µl of ethanol-containing wash buffer and eluted with distilled water in a microfuge tube (typically in 6-20 µl).

2.3.6 Gibson Assembly

A Gibson assembly reaction was attempted, using the Gibson Assembly master mix (New England Biolabs, Cat# E2611) to introduce a PCR amplified DNA fragment to a PCR-linearised plasmid vector. Initially, PCR primers were designed to add regions of homology to the ends of both an insert and a plasmid vector. These primers were designed using the NEBuilder v2.0.5 web tool (https://nebuilder.neb.com/). The four primers were 5'_fwd (5′ TGTGGTGGAATTCTGCAGATAAGCTTACAGTGTTGTGG 3′), 5'_rev (5′ TCCTGCACCTTGTCCATATCCACCTTGAC 3′), 3'_fwd (5′ GATATGGACAAGGTGCAGGAATTTCTGC 3′) and 3'_rev (5′ CGGCCGCCACTGTGCTGGATTCTAGAACAGTGTTGTGG 3′). The resulting DNA amplicons were separated by size through agarose electrophoresis (section 2.3.1) and they were purified using the Monarch Gel Extraction kit (section 2.3.7). The homologous insert and plasmid were combined at a 2:1 molar ratio and assembled with the Gibson Assembly master mix at 37°C for 1 hour. Finally, the reaction was terminated through heating at 50°C for 15 minutes.

2.3.7 Restriction Endonuclease Digestion

Various restriction endonucleases (from New England Biolabs) were used to produce DNA fragments with sticky ends. The reaction volume for each reaction was 50 μl and they were incubated for 20-60 minutes at 37°C. The specific buffers for each reaction were chosen using the NEBCloner web tool (version 1.12.0). The enzymes chosen are outlined in Chapter 3.

2.3.8 Agarose Gel DNA Extraction

DNA was extracted from agarose gels following restriction endonuclease digestion and prior to T4 DNA ligation. The Monarch DNA gel extraction kit was used (New England Biolabs, Cat# T1020) along with the corresponding manufacturer's protocol. Before casting the agarose gels, guanosine was added (Merck Life Sciences, Cat# G6264) to a final concentration of 1 mM as a UV protectant (Gründemann and Schömig, 2018). This agarose gel was used to resolve DNA bands and was visualised on a UV transilluminator at 70% light intensity. Individual bands of interest were excised with a scalpel and moved to an Eppendorf microfuge tube and dissolved in the kit's dissolving buffer for 10-15 minutes at 50-55°C. A ratio of 1:4 was used (gel:buffer) and typical total volumes ranged between 0.5 ml and 1.5 ml. Afterwards, the samples were loaded onto the kit's columns and centrifuged. Next, two washes were performed with 300 µl of the kit's wash buffer and the DNA samples were finally eluted in 13-20 µl distilled water. As before, plasmid DNA concentrations were measured using the Nanodrop 1000 (Thermo Scientific).

2.3.9 DNA Ligation

DNA ligations were performed with the T4 DNA ligase (New England Biolabs, Cat# M0202). Each ligation was performed in a 20 µl volume with a digested DNA vector and DNA insert fragment at a ratio of 1:3, 1:5 or 1:10. Also, the T4 DNA ligase buffer was added at a final concentration of 1X, alongside 400 units of T4 DNA ligase and 1 µl PEG8000 (New England Biolabs, Cat# M0202). The total amount of DNA per 20 µl did not exceed 100 ng. This mixture was then incubated in an ice bucket overnight. The gradual thawing of the ice ensures a range of temperatures for the reaction to occur successfully. Alternatively, the reactions can be transferred to 0.2 ml thin-wall PCR tubes and temperature cycling can be performed by a thermocycler. The Techne TC-512 gradient thermocycler was used to alternate between 2°C and 25°C with 1°C change per minute overnight.

2.3.10 *In Vivo* Assembly

In one experiment *in vivo* assembly (IVA) was used instead of restriction endonuclease digestion and T4 DNA ligation to insert a short DNA sequence into a plasmid vector (Bi and Liu, 1994; Bubeck et al., 1993; Garcia-Nafria et al., 2016). The primers IVA_FW (5′ CATCACCATCACCATCACCAAGGTGGGTACGGGCAAGGGAC 3′) and IVA_REV (5′ GTGATGGTGATGGTGATGTGCGAGTGCGGCAGCGACGAATG 3′) were used in a PCR with the pSP4GA plasmid as the template (22 ng per 20 µl reaction). The PCR proceeded as outlined in section 2.3.2. The number of cycles was 22. After the PCR a column DNA cleanup was performed as in section 2.3.4. A subsequent *DpnI* digestion was performed on the new DNA mixture (New England Biolabs, Cat# R0176). All of the amplified DNA, 5 µl of 10 CutSmart buffer (New England Biolabs, Cat# B6004) and 20 units of *DpnI* were incubated at 37°C for 30 minutes. The reaction was stopped by heating at 80°C for 10 minutes. *DpnI* digestion is necessary to destroy all of the circular plasmid within the sample. The linear PCR product plasmid is left intact since its *DpnI* restriction sites are not methylated. 5 µl of the digested mixture was then used to transform XL-10 Gold cells whose endogenous recA-independent recombination machinery is responsible for circularising the linear plasmid (its ends have short regions of homology due to the PCR step).

2.3.11 Extraction of Total Protein

Total protein was harvested from either the conditioned cell culture media or from the adherent cells on the 6-well plates. Adherent cells were lysed with 50 µl per well of 1x Laemmli buffer (0.0625 M Tris base, 2% SDS, 10% glycerol, 5% β-mercaptoethanol, 0.002 bromophenol blue) at room temperature with manual scraping. This was followed by DNA shearing using a Microlance needle 3 (Beckton Dickinson, 27 G ¾, Cat# 302200). The samples were centrifuged at 4^{o}C for 15 minutes at 20 000 g to remove the sheared DNA and the resulting supernatants were transferred to fresh tubes for future protein analysis.

Protein extraction form adherent cells was also performed with radioimmunoprecipitation assay buffer (RIPA buffer) (150 mM NaCl, 1% Nonidet P-40, 0.5% DOC, 0.1% SDS, 50 mM NaF, 50 mM Tris base, pH 7.4). 50 µl of RIPA buffer was added per 500 000 cells and these were scraped and pipetted with a pipette tip on ice until the cells were lysed. The lysate was incubated on ice for 15 minutes and subjected to DNA shearing using a Microlance needle 3. This was kept on ice for an additional 15 minutes and then centrifuged at 13,000 g for 5 minutes at 4°C. The resulting supernatant was transferred to a new tube and either stored at -20°C or immediately used in subsequent western blot analysis (supplemented with 4X Laemmli buffer).

Alternatively, TRIzol extraction was performed on adherent cells as per the manufacturer's protocol. Cells were lysed in the 6-well plates with scraping in 1 ml TRIzol at room temperature. After RNA has been extracted (according to the procedure in 2.1.3) the remaining phenol-ethanol fraction was further treated to obtain total protein. For every sample, 1.5 ml of 2-propanol was added to the phenol-ethanol fraction. These were allowed to stand for 10 minutes at room temperature before centrifugation at 12000 g for 10 minutes at 2-8°C. The subsequent protein pellet was washed three times with 0.3 M guanidine hydrochloride in 95% ethanol. Finally, the pellet was resuspended in 100-200 µl 4 M Urea or 1% sodium dodecyl sulphate (SDS). Protein quantification was performed with a Nanodrop 1000 (Thermo Scientific).

For several of the conditioned media samples, the above TRIzol method was used to extract protein. Otherwise, 1 ml of the conditioned media in each well was harvested daily 24, 48 or 72h post-transfection and condensed 10-20-fold using the Amicon Ultra-2 Centrifugal Filter Unit (Sigma Aldrich, Cat# UFC200324) with centrifugation at 2-8°C for 30-50 minutes.

2.3.12 SDS Polyacrylamide Electrophoresis

In order to separate a mixture of proteins within the same sample according to their size, sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) was performed. 5-15% gels were prepared by casting a liquid acrylamide solution in a Mini-PROTEAN (Biorad, 1658005EDU) glass plate where polymerisation occurred over 30-40 minutes. An initial running gel solution was cast (5-7 ml) which contained 5-15% acrylamide (from 30% stock, Protogel, Cat# A2-0072), 1% SDS, 0.4 M Tris-HCl (from 1.5 M, pH 8.8 stock), 1% ammonium persulphate and 0.001% tetramethylethylenediamine. A layer of isopropanol ensured the gel is level. After this had polymerised the isopropanol was aspirated and a second stacking gel solution was cast on top which contained 4% acrylamide, 1% SDS, 0.133 M Tris-HCl (from 0.5 M, pH 6.8 stock), 1% ammonium persulphate and 0.001% tetramethylethylenediamine. After the second polymerisation, the gel was placed in an electrophoresis tank with 1x running buffer (25 mM Tris-HCl; 192 mM glycine; 0.1% SDS). 5 µl of the BlueClassic prestained protein marker (Jena Bioscience, Cat# PS-107) were loaded onto the gel. Each protein sample was prepared by the addition of Laemmli loading buffer to a final concentration of 1X (0.0625 M Tris base, 2% SDS, 10% glycerol, 5% β-mercaptoethanol pH 6.8) then heated to 80°C for 10 minutes. The samples were loaded and the gel was run initially at 60 V for 30 minutes. Finally, the gel was run at 120 V until the dye front reached the end of the gel (1-3h).

2.3.13 Zinc Staining

After SDS-PAGE several gels were analysed through reversible zinc staining. The soluble imidazole and zinc ions interact to form crystals. Proteins in the gel also bind soluble zinc ions so the crystals are not formed in the presence of proteins. This results in a counterstain where all parts of the gel are stained except for where protein is present. The gel was removed from the glass plates and rinsed in distilled water. Next, it was shaken in 0.2 M imidazole for 5-10 minutes. The imidazole was then discarded and the gel was shaken in 0.3 M zinc chloride for 20-30 seconds. This solution was discarded again and the gel was rinsed again in distilled water. The gel was finally imaged on a Gel Doc XR+ (Bio-Rad, Cat# 1708195).

2.3.14 Silver Staining

Silver staining was also used to analyse proteins after several of the SDS-PAGE experiments. The gel was removed from the casting plates and the electrophoresis tank. Next, it was shaken in a fixing solution (50% acetone, 1.25% tricarboxylic acid, 0.015% formaldehyde) for 15 minutes then washed three times in distilled water for 5 minutes each. The gel was then washed in 50% acetone and pre-stained in a 1 mM thiosulphate solution for 1 minute. It was washed three times again for 5 minutes each in distilled water and then stained in darkness for 8 minutes with a staining solution (15 mM silver nitrate, 0.36% formaldehyde). This was washed twice in distilled water for 10 seconds and developed for 1-2 minutes in a developing solution (0.2 M sodium carbonate, 0.015% formaldehyde, 0.25 mM sodium thiosulphate). Finally, the developing solution was removed and the gel was shaken in 1% acetic acid for 1-2 minutes to stop the reaction. This was rinsed again in distilled water and imaged in a Gel Doc XR+ (Bio-Rad, Cat# 1708195).

2.3.15 Western Blotting

Western blotting was performed to visualise specific proteins after separation according to size on an SDS-PAGE. After the electrophoresis, the gel was removed from the casting plates and left to soak in 1X transfer buffer (10% ethanol, 0.2 M glycine, 0.25 M Tris base) for 10-15 minutes. At the same time, a nitrocellulose membrane (Bio-Rad, Cat# 1620094) was cut in the shape and size of the gel and also soaked in 1X transfer buffer alongside similarly-sized blotting paper (Bio-Rad, Cat# 1703965). Afterwards, these components were assembled in a transfer block in an electrophoresis tank with the gel oriented towards the cathode and the membrane toward the anode. This was used to transfer the protein from the gel onto the nitrocellulose membrane by electric current in 1X transfer buffer at 2-4°C for 2 hours at 60 V. Later the membrane was removed from the transfer block and rinsed with distilled water then reversibly stained briefly (10-20 seconds) with Ponceau S solution (0.5% Ponceau S, 1% acetic acid). If the transfer has been successful the Ponceau S staining would show protein bands on the membrane. Next, the Ponceau stain was removed by quick washes in Tris-buffered saline-tween (TBST - 150 mM NaCl, 50 mM Tris and 0.1% tween-20). The membrane was left to block in 5% non-fat milk (Marvel, Cat# M203) in TBST with constant agitation for 1 hour at room temperature or 2-8°C overnight. Next, a primary monoclonal antibody from mouse was diluted 5 000-50 000 times in TBST and 5% non-fat milk and used to probe the membrane for 1h at room temperature or overnight at 2-8°C. The primary antibodies used were all monoclonal mouse antibodies: anti-actin (Sigma, Cat# A5441), anti-tubulin (TUBB – VWR, Cat# BIRBORB95164-50), anti-histidine-tag (anti-HIS – Proteintech, Cat# 66005-1-Ig), anti-FLAG (Sigma, Cat# F1804) and anti-vinculin (Sigma, Cat# V4505). These are summarised in Table 2.1 below. Later the membrane was rinsed three times with TBST for 5 minutes each and then probed with an anti-mouse secondary antibody (Cell Signalling Technology, Cat# 7076). The latter was ordered conjugated with horseradish peroxidase (HRP). Next, it was washed again in TBST (four times for 5 minutes) and the membrane was left to develop in Pierce ECL Western Blotting substrate (Thermo Fisher Scientific, Cat# 32106). This reagent enables the secondary antibody’s HRP to produce a chemiluminescent signal. Finally, it was imaged in an ImageQuant LAS- 4000 luminescent image analyser (Fujifilm, Cat# 4347786).

Table 2.1 – **Antibodies used in the western blot analysis**.

Antibody Name	Host	Target	Type	Manufacturer	Cat. Number
Anti-TUBB	Mouse	TUBB (tubulin, beta class I)	Primary	VWR	BIRBORB95164-50
Anti-actin	Mouse	ACTB	Primary	Sigma	A5441
Anti-vinculin	Mouse	Vinculin	Primary	Sigma	V4505
Anti-HIS-tag	Mouse	His-tag	Primary	Protein Tech	66005-1-Ig
Anti-FLAG	Mouse	FLAG-tag	Primary	Sigma	F1804
Anti-mouse	Horse	IgG	Secondary HRP-conjugated	Cell Signalling Tech.	7076

2.3.16 tRNA Charging Assay

A tRNA charging assay was carried out to assess the effect of plasmid transfection on the tRNA-amino-acylation in HEK293 cells according to the protocol in Pavlova *et al*. (2020). Total RNA was extracted 72h after plasmid transfection using 1 ml TRIzol (Thermo Fisher Scientific, Cat# 15596026) and 2 µl of Glycoblue coprecipitant (Thermo Fisher Scientific, Cat# AM9515) per ~500 000 cells. The addition of 200 µl chloroform was used to aid the phase separation of RNA, DNA and protein. The aqueous RNA fractions were aspirated, moved to microfuge tubes and a volume of 2.5X ethanol was added per sample. These were then incubated at -20°C overnight.

Next, the precipitated RNA was pelleted by centrifugation at 18 600 g at 4^{o}C for 30 minutes. Each pellet was resuspended in 300 µl tRNA reprecipitation buffer (0.3 M sodium acetate, 10 mM EDTA, pH 4.5) and then 2.5X ice-cold ethanol was added. These samples were again precipitated at -20^{o}C overnight followed by pelleting at 18 600 g at 4^{o}C for 30 minutes.

Subsequently, the pellets were washed with 1 ml of 80% ethanol and then resuspended in 32 µl tRNA resuspension buffer (10 mM sodium acetate, 1 mM EDTA, pH 4.5). Then, 2 µg of each sample were used in oxidation reactions with 10 mM sodium periodate (20 minutes, room temperature, in darkness). For each sample, a corresponding mock oxidation reaction was carried out under the same conditions with 10 mM sodium chloride instead of periodate. These mock-oxidations serve to infer total tRNA levels later. These reactions were quenched with 2.2 µl of 2.5 M glucose (15 minutes, room temperature, in darkness). Each sample was then combined with 1.5 µl of GlycoBlue coprecipitant and 3X the volume of 100% ethanol. Precipitation at -20^{o}C overnight was followed by pelleting at 18 600 g at 4^{o}C for 30 minutes. Each pellet was then resuspended in 100 µl 50 mM Tris buffer (pH 9) to remove the aminoacylation from each tRNA. This reaction was quenched with 100 µl tRNA quench buffer (50 mM sodium acetate, 100 mM sodium chloride, pH 4.5). After the addition of 2.7 volumes of ethanol, precipitation was carried out at -20^{o}C overnight.

Next, the tRNA was pelleted at 18 600 g at 4°C for 30 minutes and resuspended in 10 µl ultrapure water. The samples were then adjusted to be at the same concentration. A 5′ adenylated and 3′ dideoxycytidinylated DNA oligo (Integrated DNA Technologies) was ligated to the 3′ end of every tRNA that had been protected from oxidation by a charged amino acid or that had not undergone oxidation. This ligation was carried out with 2.5 µl of the adjusted RNA samples, 0.5 µl of 100 µM tRNA adapter, 1x T4 RNA ligase buffer (New England Biolabs, Cat # M0373), 0.2 µl of 0.1 M DTT, 0.5 µl of 100% DMSO and 0.3 µl ultrapure water. The RNA in this mixture was denatured by heating to 90°C for 30 seconds followed by cooling on ice for 1 minute. Next, 0.2 µl of RNase inhibitor (Promega, Cat# N2115) and 0.3 µl of T4 RNA ligase 2, truncated KQ (New England Biolabs, Cat# M0373) were added to each sample mixture. These were incubated at room temperature overnight.

For cDNA synthesis, 1 µl of 5 µM CSQ_RT primer (Table 2.2) was added to each ligation. Similarly, 1 µl of 5 µM of the GAPDH_FW primer was added to a mixture of 2.5 µl unligated total RNA and 2.5 µl ultrapure water to serve as a housekeeping control. These mixtures were then incubated at 90°C for 30 seconds and 65°C for 5 minutes to anneal properly. For the subsequent cDNA synthesis, 4 µl of these mixtures were supplemented with 0.8 µl ultrapure water, 1.6 µl of 5X SuperScript RT IV buffer (Thermo Scientific, Cat# 18090050), 0.4 µl of 10 mM dNTPs, 0.4 µl of 0.1 M DTT, 0.4 µl of

SuperScript IV reverse transcriptase (Thermo Scientific, Cat# 18090050) and 0.4 µl of the RNase inhibitor. Each mixture was incubated at 55^{o}C for 10 minutes and then 80^{o}C for 10 minutes. Next, these were diluted 10 times to prepare them for quantitative PCR (qPCR) measurements. qPCRs were set up with 2.5 µl of each cDNA dilution, 0.2 µl of 10 µM of each qPCR primer (Table 2.2), 2.3 µl of ultrapure water and 5 µl of 2X Power SYBR Green mix (Applied Biosystems, Cat# 4368708). Quantitative real-time PCR was performed using a RotorGene Q (Qiagen, Cat# 9001863) machine with an annealing temperature of 63^{o}C.

After each run the tRNA charging ratios were calculated by subtracting the GAPDH Ct value and the average negative control Ct value from each reading (to generate a value known as ΔΔCt). This difference was then used in the formula $2^{-\Delta\Delta Ct}$ to obtain a relative abundance quantity. Finally, the $2^{-\Delta\Delta Ct}$ of each oxidised sample was divided by the $2^{-\Delta\Delta Ct}$ of the corresponding total tRNA sample (non-oxidised) to produce a ratio of tRNA charging.

Table 2.2 – **DNA oligo sequences used in the tRNA charging assay**

Name	DNA Sequence
tRNA adapter	5’-/rApp/TGGAATTCTCGGGTGCCAAGG/ddC/-3’
CSQ_RT	GCTGCCTTGGCACCCGAGAATTCCA
ArgACG_FW	GGGCCAGTGGCGCAATG
ArgACG_RV	GAGAATTCCATGGCGAGCCAGC
GlyGCC_FW	GCATTGGTGGTTCAGTGGTAGAATTC
GlyGCC_RV	GAGAATTCCATGGTGCATTGGCC
AlaAGC_FW	GAATTAGCTCAAGTGGTAGAGCGC
AlaAGC_RV	GAGAATTCCATGGTGGAGAATGNGG
GAPDH_FW	TCGGAGTCAACGGATTTGGT
GAPDH_RV	TTCCCGTTCTCAGCCTTGAC

2.4 Computational Methods

2.4.1 Figure Design

Most schematic figures in this thesis were created using BioRender’s web tool (available at biorender.com). The sucker ring tooth diagram (Section 5.1.3) was created with Adobe InDesign (Adobe Inc., 2022).

2.4.2 Plasmid Sequence Design

Several tools were used to facilitate the design of sequences for molecular cloning. The ExPASy translate web browser tool was used to translate nucleotide sequences into amino acid sequences in all six frames (Gasteiger et al., 2003). The Bioinformatics Sequence

Manipulation Suite 2 was used to perform reverse transcription *in silico* which was used to ensure accurate reverse complementarity of two intron sequences within the plasmid constructs (Stothard, 2000). The Human splice site finder (HSF 3.1) was used to predict splice sites within the designed sequences using default parameters (Desmet et al., 2009). The SignalP 5.0 tool was used to predict the efficiency of the planned trypsin signal peptide cleavage (Armenteros et al., 2019).

The Twist Bioscience web browser tool was used to visualise recombinant sequences that would make up the pre-circRNA. Codon optimisation was also performed with this tool for the specific reading frame of interest. The human and mouse amino acyl-tRNA pools were chosen for this optimisation (Twist Bioscience, 2022).

2.4.3 DNA Primer Sequence Design

For each planned PCR primer pair, a set of tools was used to assess the potential suitability for DNA amplification. The Bioinformatics Sequence Manipulation Suite 2’s PCR Primer Stats tool was used to screen primers for the presence of hairpin-forming sequences or self-annealing regions (Stothard, 2000). The NCBI Primer-blast tool was used to design several primer pairs from an input DNA sequence. The default settings were used and primer pair specificity was assessed against the *Homo sapiens* genome (Ye et al., 2012). The Thermo Fisher Multiple Primer Analyser was used to assess planned primer sequences for the formation of primer dimers (Thermo Fisher, 2022).

2.4.4 Consensus Sequence Identification of tRNA Isodecoders

To design primer pairs for qRT-PCR in the tRNA charging assay a consensus sequence of each tRNA isodecoder set was needed (a tRNA isodecoder set is the collection of different tRNA molecules that all share the same anticodon). The GeneCalc tool’s consensus sequence-finder was used. This produces a consensus nucleotide sequence from a selection of nucleotide sequences. A threshold value of 0.55 was set (Miks and Bińkowski, 2022). Sets of tRNA isodecoder sequences for the human AlaAGC and GlyGCC were chosen from the Genomic tRNA database (Chan and Lowe, 2016). An extra nucleotide triplet (CCA) was added to the end of each consensus of isodecoders. In nature, this is the result of a post-transcriptional tRNA modification ((Deutscher 1982). The consensus sequences were then used to design primer pairs as per Pavlova et al. (2020).

2.4.5 Statistical Analysis

A One-Way ANOVA ("analysis of variance") test was carried out to determine the statistical significance of the means of a set of tRNA charging assay data. The MS Excel Analysis ToolPak was used to perform a Single Factor ANOVA specifically. The chosen p-value cut-off point was 0.05.

2.4.6 Ribosome Frameshifting Prediction

The KnotInFrame tool was used to estimate the likelihood of -1 ribosomal frameshifting during IORF translation (Theis et al., 2008). Linear concatemer sequences of the circRNAs to be produced were inserted into KnotInFrame tool and default settings were used during the prediction.

2.4.7 Choice of Transcriptomes

Transcriptomic datasets were used to identify previously-unknown putative spidroin sequences. Publicly available paired-end read data were chosen from the Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra) for the following spider species (accession numbers are shown in parentheses): the giant white knee tarantula *Acanthoscurria geniculata* (SRR1024075), the red-tighed banana spider *Cupiennius coccineus* (SRR7028538), the black-and-white spiny spider *Gasteracantha kuhli* (DRR129307), the western black widow *Latrodectus hesperus* (SRR1219665), the black armoured trapdoor spider *Liphistius malayanus* (SRR1145736), the hermit spider *Nephilengys cruentata* (SRR3943479), the recently-described tarantula *Pamphobeteus verdolaga* (ERR2008012), the American house spider *Parasteatoda tepidariorum* (SRR1824487) and the wolf-spider *Pardosa pseudoannulata* (SRR6837902). High-quality paired-end read data from the daddy-longlegs spider *Pholcus phalangioides* were provided by Leah Ashley from the School of Life Sciences, University of Nottingham, UK.

PacBio generated single-molecule real-time (SMRT) long-read transcriptomic data of the tropical tent-web spider *Cyrtophora citricola* were provided by Ella Deutsch from the School of Life Sciences, University of Nottingham, UK. Similarly, PacBio SMRT-generated long-read transcriptomic data of the diving bell spider *Argyroneta aquatica* were provided by Charlotte Deall from the School of Life Sciences, University of Nottingham, UK.

Additionally, eleven assembled spider transcriptomes were obtained from the Transcriptome Shotgun Assembly (TSA) database (https://www.ncbi.nlm.nih.gov/genbank/tsa/). These assemblies are from - the giant white knee tarantula *Acanthoscurria geniculata* (GAZS01.1), the Darwin's bark spider *Caerostris darwini* (GGTX01.1), the six-spotted fishing spider *Dolomedes triton* (GGRN01.1), the western black widow *Latrodectus hesperus* (GBJN01.1), the golden silk orb-weaver *Nephila clavipes* (GFKT01.1), the hermit spider *Nephilengys cruentata* (GEWZ01.1), the tarantula *Pamphobeteus verdolaga* (HAHO01.1), the American house spider *Parastatoda tepidariorum* (IAAA01.1), the wolf-spider *Pardosa pseudoannulata* (GGRD01.1), the dark comb-footed spider *Steatoda grossa* (GBJQ01.1) and the castor bean tick *Ixodes ricinus* (GADI01.1).

For the discovery of previously non-annotated suckerins, pre-assembled publicly available squid transcriptomes were chosen from the TSA database based on their total size. Datasets from 2.7 to 239.6 megabases in size were considered to be more complete than smaller ones. This led to the selection of the following assemblies (accession numbers in parentheses): the Humboldt squid *Dosidicus gigas* (GHKL01.1), the southern bobtail squid *Euprymna tasmanica* (GEXE01.1), the southern pygmy squid *Idiosepius notoides* (GFNE01.1), the pelagic squid *Octopoteuthis deletron* (GGNB01.1), the common clubhook squid *Onychoteuthis banksia* (GHKK01.1), the golden cuttlefish *Sepia esculenta* (GGQU01.1), the pharaoh cuttlefish *Sepia pharaonis* (GEIE01.1), the spineless cuttlefish *Sepiella maindroni* (GFLT01.1), the striped pyjama squid

Sepioloidea lineolata (GEFX01.1), the purpleback flying squid *Sthenoteuthis oualaniensis* (GHKH01.1) and the firefly squid *Watasenia scintillans* (GEDZ01.1). Initially, the southern blue-ringed octopus *Hapalochlaena maculosa* transcriptome (GEXH01.1) was selected as a negative control for presumably absent suckerin peptides since it was considered a closely related outgroup. Upon the identification of a *Hapalochlaena* suckerin, the search was expanded to two additional octopus transcriptomes from the Mexican four-eyed octopus *Octopus maya* (GHBT01.1) and the vampire squid *Vampyroteuthis infernalis* (GGNA01.1) which is actually more closely related to octopuses. The bladder snail *Physella acuta* (GHAL01.1) and the giant lion's paw scallop *Nocipecten subnodosus* (GFNL01.1) were chosen as new outgroup negative controls. These transcriptomes had been *de novo* assembled with the Trinity tool before submission to TSA (Grabherr et al., 2011).

Two cladograms were made of the relationships between the arachnid species and the molluscs using the R Open Tree of Life (rotl – version 3.0.10) package (Hinchliff et al., 2015; Michonneau et al., 2016). This software assigns relationships between the different lineages based on their respective entries in the Open Tree of Life (https://tree.opentreeoflife.org/).

2.4.8 Transcriptome Assembly

The paired-end read data from SRA for *Cupiennius coccineus, Gasteracantha kuhli, Latrodectus hesperus, Liphistius malayanus* and *Pholcus phalangioides* were *de novo* assembled (without the use of a reference genome) into putative transcript contigs using the Trinity RNA-seq software (version 2.9.1) (Grabherr et al., 2011). The same was repeated for the *Argyroneta aquatic*a and *Cyrtophora citricola* long-read transcriptomes.

Trinity was limited to 25 GB maximum memory and six CPUs. For *Latrodectus* and *Parasteatoda* genome-guided assemblies were performed using publicly available genomes (NCBI IDs: 14107 and 13270). Quality trimming of SRA read data was performed using Trinity's built-in Trimmomatic tool with default options (version 0.39) (Bolger et al., 2014). This tool removes common adapter sequences from each read, low-quality reads or low-quality ends of reads. The Pholcus transcriptome was trimmed by Leah Ashley using the Trimmomatic tool using the same parameters (Joshi and Fass, 2011).

2.4.9 Transcriptome Assessment

Read data quality was assessed before and after each trimming using the FASTQC tool (version 0.11.9) (Andrews, 2010). The Bowtie2 tool (version 2.4.0) was used to align raw reads onto the assembled transcript contigs and quantify read support for the assembly (Langmead and Salzberg, 2012).

2.4.10 Open Reading Frame Estimation

The TransDecoder tool (version 5.5.0) was used under default settings to estimate which assembled transcripts corresponded to open reading frames and to translate these (Haas et al., 2013). A BLAST search is integrated into this process to also identify any pairwise homology to sequences known to be translated. Two separate databases of all annotated peptides from spiders and squids (not only spidroins and suckerins) were downloaded from UniProt for these BLAST searches (Bateman et al., 2021).

2.4.11 Profile HMM Library Construction

In order to identify sequences with homology to known spidroins or suckerins, a profile Hidden Markov Model (HMM) approach was used. Several spidroin profile HMMs were already available from the Protein Families database (Pfam) (Finn et al., 2016). Namely, the profile HMM for the spidroin N- (PF16763) and C-terminal domains (PF11260), CySp repeat domain (PF12042), Lamprin repeat domain (PF06403), CXCXC repeat domain (PF03128) and SSP160 repeat domain (PF06933) were downloaded.

Additionally, eight further known spidroin types were not represented in this database and their profile HMMs were constructed manually. The T-REKS tool (version HPC) was used first to create multiple sequence alignments (in Stockholm format) from peptide tandem repeats (Jorda and Kajava, 2009). Eight representative spidroin peptide sequences were chosen from the UniProtKB database – one per spidroin paralog: A0A4Y2R483 for aciniform spidroin, A0A4Y2BY01 for aggregate spidroin, A0A4Y2M3Y8 for FLAG spidroin, I6XQ31 for MaSp1, Q2VLH2 for MaSp2, A0A4Y2MH26 for MaSp3, A0A1V0D8S5 for MiSp and A0A1Y9T5P4 for pyriform spidroin. Using T-REKS individual repeats were clustered together based on sequence similarity to each other generating several alignments.

Using these alignments and the HMMER tool (version 3.3) (Eddy, 2011), profile HMMs were created for each of the eight representative spidroins (hmmbuild command). Then, a second set of multiple sequence alignments was constructed, using each profile HMM and a custom database of all annotated and available spidroin peptide sequences (obtained from Leah Ashley at the School of Life Science) through a relaxed (E-value threshold of 0.01) hmmsearch command. The new alignments were from multiple different orthologs per spidroin and were used to create a second set of HMMs which are more representative of the sequence diversity across the spider lineages. This set (together with the Pfam HMMs) was finally concatenated into a single profile HMM library.

To create a profile HMM for the only known suckerin domain an initial representative suckerin protein sequence (AGY36220.1, a 39 kDa suckerin protein from *Dosidicus gigas*) was chosen. Its repeat regions were aligned using T-REKS. An alignment from repeated regions that contain suckerin-specific repeats was chosen to manually generate a Stockholm format output. The latter was compiled into an initial profile HMM using the HMMER hmmbuild command. A database of all known suckerins was downloaded from NCBI Genbank. The hmmscan command was used to generate a new sequence alignment using the initial profile HMM as query and the database of previously annotated suckerins as the subject. From this later alignment a final profile HMM was constructed using the hmmbuild command again.

2.4.12 Domain Annotation

Searches for distant sequence homology were performed using the HMMER tool. The constructed spidroin HMM library or the single suckerin HMM was used to identify putative spidroins or suckerins in the transcriptome datasets with the hmmscan command.

Each putative suckerin or spidroin was independently screened for the presence of signal peptides with the Phobius (Kall et al., 2007) and SignalP 5.0 web tools (Armenteros et al., 2019). Additional analysis was performed with the web HMMER hmmsearch command for the presence of non-suckerin domains against the Pfam database (Potter et al., 2018).

2.4.13 Exclusion of Known Spidroins and Suckerins

Between transcriptome translation and domain annotation, screening was carried out to remove any known suckerin or spidroin sequences from the analysis. The Magic-BLAST tool was used for this exclusion with an E-value threshold of 0.0001 (Boratyn et al., 2019).

Each arachnid transcriptome was analysed for the presence of spidroins using Leah Ashley's aforementioned spidroin database. Each putative spidroin was checked for previous spidroin annotation using the web BLAST tools BLASTX, DELTA-BLAST, discontiguous MegaBLAST and tBLASTn (Boratyn et al., 2012; Camacho et al., 2009).

Similarly, a collection of previously annotated suckerin peptide sequences was obtained from the NCBI Proteins database through a search for the keyword "suckerin". These were concatenated into a multifasta database. Again, the Magic-BLAST tool was used to identify sequences with high pairwise homology (E-value threshold of 0.0001) to the previously annotated suckerin sequences from all translated mollusc transcriptomes using this database. These sequences were excluded from any downstream analysis to avoid redundant suckerin annotation. After domain annotation, each putative HMM-identified sequence was again inspected for any previous annotation using the web BLAST tools BLASTX, DELTA-BLAST, discontiguous MegaBLAST and tBLASTn.

2.4.14 Expression Quantification

Expression quantification was carried out in the transcriptomes that contained newly identified spidroins or suckerins. SRA transcriptomes had already been downloaded for several arachnid species. The remainder of newly-identified-sequence-containing transcriptomes were originally obtained pre-assembled from TSA (as described in section 2.4.7). Their corresponding SRA sequence reads were now downloaded as well. All trimmed reads were aligned to their respective assembled putative transcripts using Bowtie2 (Langmead and Salzberg, 2012). The resulting bam files were coordinate sorted using samtools (Li et al., 2009). Coordinate sorted bam files were used as input for the subsequent expression quantification. This was carried out using Trinity's built-in version of the RSEM tool (1.3.3) (Li and Dewey, 2011) to obtain expected counts and transcripts per million (TPM).

The RPL13 gene (which encodes a ribosomal protein) was used as a housekeeping control for arachnid transcriptomes. RPL13's spider homologs were identified by running hmmscan with an RPL13 Pfam profile HMM as the query (PF00572). Whenever multiple RPL13 candidates were identified, the most highly expressed one was considered. Read alignments were visualised using the Integrated Genomics Viewer (IGV - version 2.8.3) (Thorvaldsdottir et al., 2013).

General actin expression was used as a housekeeping control in the quantification of the mollusc transcriptomes. Multiple actin homologs were identified using hmmscan with the actin entry in Pfam (PF00022) as the query. Whenever multiple actin candidates were identified, the most highly expressed one was considered.

2.4.15 Phylogenetic Analysis

In order to investigate the phylogenetic relationships between the cephalopod suckerin sequences, a phylogenetic tree was constructed. The previously annotated suckerin peptide sequences were concatenated with the novel HMM and BLAST-discovered suckerin peptide sequences into a single multifasta file (89 total sequences). Read alignments were carried out using the web version of the Clustal Omega software (Lassmann et al., 2009). The MEGAX software (Kumar et al., 2018b) was used to create a maximum likelihood tree using the Whelan and Goldman substitution model (Whelan and Goldman, 2001) and 500 bootstrap replicates for node support as per Guerette et al. (2014) and Pattengale et al. (2009). No outgroup sequences were selected.

3 Producing a Self-splicing Spidroin RNA

3.1 Introduction

The process of rolling circle translation (RCT) can be leveraged to produce a long peptide chain of spidroin repeats. RCT was first described by Perriman and Ares (1998) who were able to translate a short circRNA into a continuous long chain of GFP within an *E. coli* host.

The overall process and name have been inspired by the phenomenon of rolling circle amplification (RCA) which had been discovered several years prior (Fire and Xu, 1995). In RCA a short, circular ssDNA molecule is continuously replicated by DNA polymerase to produce a concatemer of single-stranded DNA which is many times longer than the original template. While repetitive sequences are at risk of mutation during subsequent DNA replication within the host organism, in the case of short, circular sequences this potential issue is averted. The same is true in the case of RCT.

3.1.1 Circular RNA

Rolling circle translation would not be possible without circRNA. circRNA molecules result from covalently linking the 5′ and 3′ ends of linear RNA in a phosphodiester bridge to produce a circular nucleotide chain (Hsu and Cocaprados, 1979).

Eukaryotes are known to produce abundant amounts of circRNA, many of which are evolutionarily conserved. This suggests that there are natural functions for this class of RNA (Jeck et al., 2013; Memczak et al., 2013; Rybak-Wolf et al., 2015; Salzman et al., 2012). In mammals, circRNAs have been studied in more detail in recent years. While here their function is primarily unrelated to encoding proteins they still can affect gene expression as microRNA (miRNA) response elements. In short, circRNAs can anneal to some miRNAs and essentially negate their anti-translational activity similar to the activity of some long non-coding RNA (Hansen et al., 2013; Memczak et al., 2013; Niu et al., 2022; Yang et al., 2017b). More recently there has been increasing evidence which suggests that circRNA may encode certain functional peptides (Chen et al., 2021; Legnini et al., 2017; van Heesch et al., 2019; Zhang et al., 2018a; Zhang et al., 2018b). Additionally, circRNA molecules are primarily located in the cytoplasm and they have been observed to associate with ribosomes (Jeck et al., 2013; Li et al., 2018; Ragan et al., 2019; Salzman et al., 2012). Until recently all of the discovered translated circRNA molecules were thought to have a canonical translation termination i.e., no rolling circle translation had been observed in nature. However, more recently IORF translation was observed for the circRtn4 RNA which is derived from the reticulon 4 (RTN4) pre-mRNA. This is abundant in neuroendocrine cells where it may be involved in regulating neuroendocrine secretion (Mo et al., 2019). Another endogenous IORF has been observed where exons from the EGFR pre-mRNA circularise to produce circ-EGFR. RCT of this circRNA sustains EGFR’s signalling and contributes to glioblastoma tumorigenicity (Liu et al., 2021).

In the case of eukaryote-infecting viruses, circRNA translation has been observed much earlier. Examples include the rice yellow mottle virus (AbouHaidar et al., 2014) and hepatitis delta virus in humans (Wang et al., 1986). Potentially, the improved RNA stability due to exoribonuclease resistance (no exposed 5′ or 3′ ends are available on the circRNA) is beneficial to the infecting viruses (Qu et al., 2015).

There are numerous advantages of using circRNA beyond the synthesis of repetitive protein. circRNAs have increased stability over a longer time period relative to linear mRNA due to the aforementioned exoribonuclease degradation unsusceptibility (Shimizu et al., 2001). This is one of the reasons why translation efficiency can be increased greatly and extended in time. This may be particularly beneficial in cases where RNA is transfected into the cell and not produced endogenously (Wesselhoeft et al., 2018).

Furthermore, in the case of the infinite ORF, there is far less translation termination and subsequent re-initiation (a major rate-limiting step in translation) which additionally should increase the net translation efficiency (Costello et al., 2019; Lee et al., 2021). Additionally, pure exogenous circRNA shows promise in transfection because it appears to be largely non-immunogenic and non-cytotoxic even if completely unmodified. Finally, circRNAs are expressible *in vivo* where they show all of the aforementioned beneficial properties (Lucke et al., 2018; Wesselhoeft et al., 2019).

3.1.2 Translation Initiation

To enable translation from a circRNA certain conditions must be met. In eukaryotes, for the vast majority of mRNA, translation initiation takes advantage of an N^7-methylguanosine cap which has been added at their 5′ end during transcription (Figure 3.1) (Jackson et al., 2010). During canonical translation initiation, three eukaryotic translation initiation factors (eIFs), namely eIF4A, eIF4E and eIF4G, form a complex (eIF4F) which recruits the 40S ribosomal subunit and aids its association with the 5′ cap (Sonenberg et al., 1978). The resulting complex then begins scanning the RNA in search of a favourable start codon (AUG in a suitable context). Importantly, since circRNA has no 5′ end it has no 5′ cap structure. Therefore, canonical translation initiation cannot proceed.

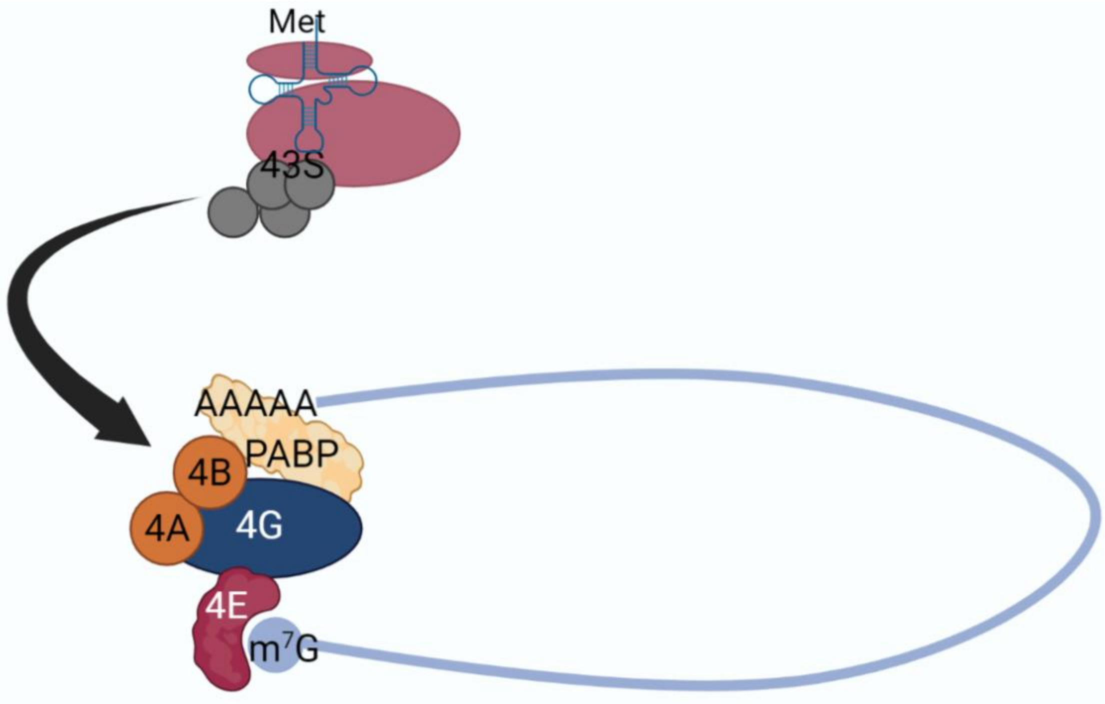


Figure 3.1 – **Schematic representation of N^7-methylguanosine cap-dependent translation initiation**. Most eukaryotic translation initiation is dependent on the presence of an N^7-methylguanosine cap (m^{7}G). This is bound by a complex of the eukaryotic initiation factors 4A, 4B, 4G and 4E. This complex then binds to and recruits a complex (43S) formed by the small ribosomal subunit, a methionyl-tRNA and several eIF proteins. Finally, the 40S subunit dissociates from the eIF proteins and begins scanning for the mRNA’s start codon.

However, there have been recorded many strategies where the ribosome is recruited without the need for a 5’ cap. A lot of these mechanisms are used by the eukaryotic cells for translation initiation in conditions of cell stress. In this case, global translation within a cell is suppressed by inhibiting cap-dependent translation specifically. In this context, some translation of stress-response genes and those critical for survival is still needed. Often elements within the 5’ untranslated region (UTR) of linear mRNA molecules exist that can recruit a ribosome independent of the cap structure specifically in such times of cell stress (Pelletier and Sonenberg, 1988; Walters and Thompson, 2016). Such elements have been co-opted to enable the translation of proteins from circRNA.

Cap-independent translation may be initiated by a single N^6-methyladenosine binding to the eukaryotic initiation factor 3 (eIF3) (Figure 3.2) (Coots et al., 2017; Yang et al., 2017a). This appears to be context-dependent but overall the consensus recognition sequence is just three nucleotides long (Meyer et al., 2015). This approach has been used successfully in the past to produce continuously translating circRNA (Costello et al., 2019).

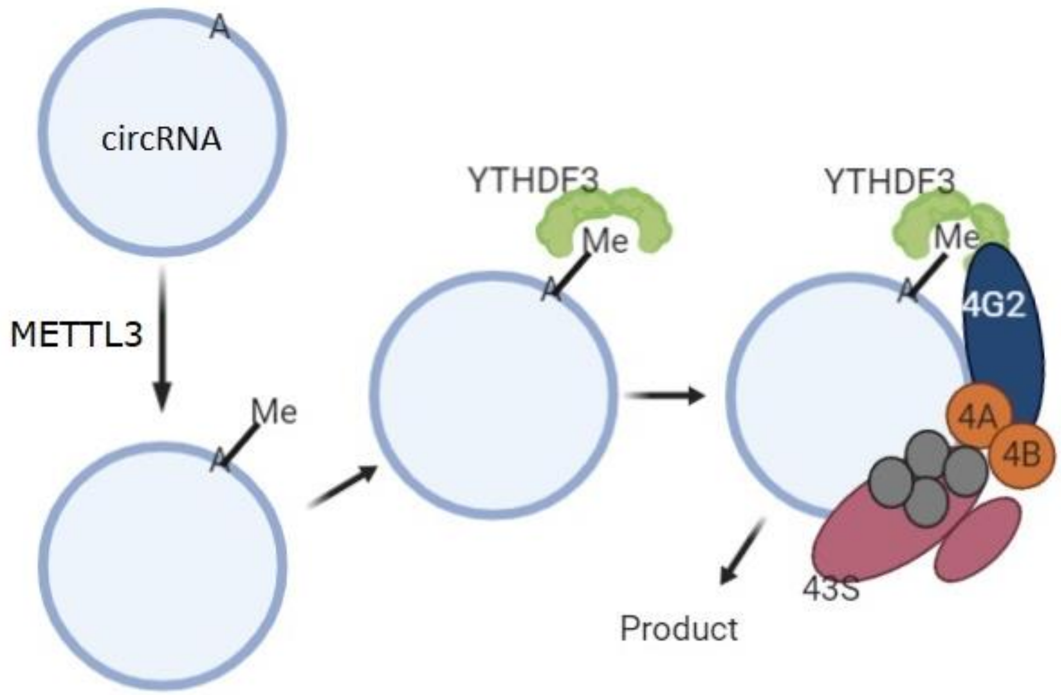


Figure 3.2 – **Schematic representation of N^6-methyladenosine-dependent translation initiation in circRNA**. A nascent RNA molecule is backspliced and methylated co-transcriptionally. In particular, METTL3 recognises an adenosine base-containing motif and methylates the adenosine. Subsequently, the m^{6}A structure is recognised by the reader YTHDF3 (YTH domain protein F3) which goes on to recruit the 43S ribosomal complex via the 4G2, 4A and 4B eukaryotic initiation factors.

Another example of cap-independent translation initiation is internal ribosome entry sites (IRESs) (Thakor and Holcik, 2012). IRESs have very diverse mechanisms and origins (Kanamori and Nakashima, 2001; Lozano et al., 2018) but all appear to be recognised by translation initiation factors due to their particular primary or secondary structures (Diaz-Toledano et al., 2017; Fricke et al., 2015). In the case of rolling circle translation, short IRES sequences would be preferable since they inevitably must be continuously translated**.** Reducing the relative number of IRES-derived amino acids within the peptide chain may improve the recombinant protein’s properties. Most importantly, IRESs that have no in-frame stop-codons must be selected to preserve the translatability of the IORF through multiple rounds of translation. These two requirements greatly limit the number of IRESs that can be used in RCT and no IRES can compete with the 5 bp m^6A motif in shortness.

3.1.3 Synthesis of circRNA

In order to translate any circRNA, it first needs to be transcribed and circularised. While the process of RNA circularisation does not have a direct impact on the process of RCT it is an important consideration nonetheless. There is a large variety of circularisation mechanisms that have been used *in vivo* and *in vitro*.

In eukaryotic cells, the majority of circRNA molecules are thought to arise from conventional pre-mRNA molecules through a process known as backsplicing.

This is thought to occur when a downstream 5′ splice donor site of an exon splices with the upstream 3′ splice acceptor site (Jeck et al., 2013; Memczak et al., 2013; Starke et al., 2015). Essentially this process results in both ends of an exon being joined instead of associating with another exon (Figure 3.3). Importantly, this reaction is carried out in eukaryotic cells' nuclei by large ribonucleoprotein complexes known as spliceosomes. Spliceosomes are responsible for the vast majority of splicing events in eukaryotes (Patel and Steitz, 2003).

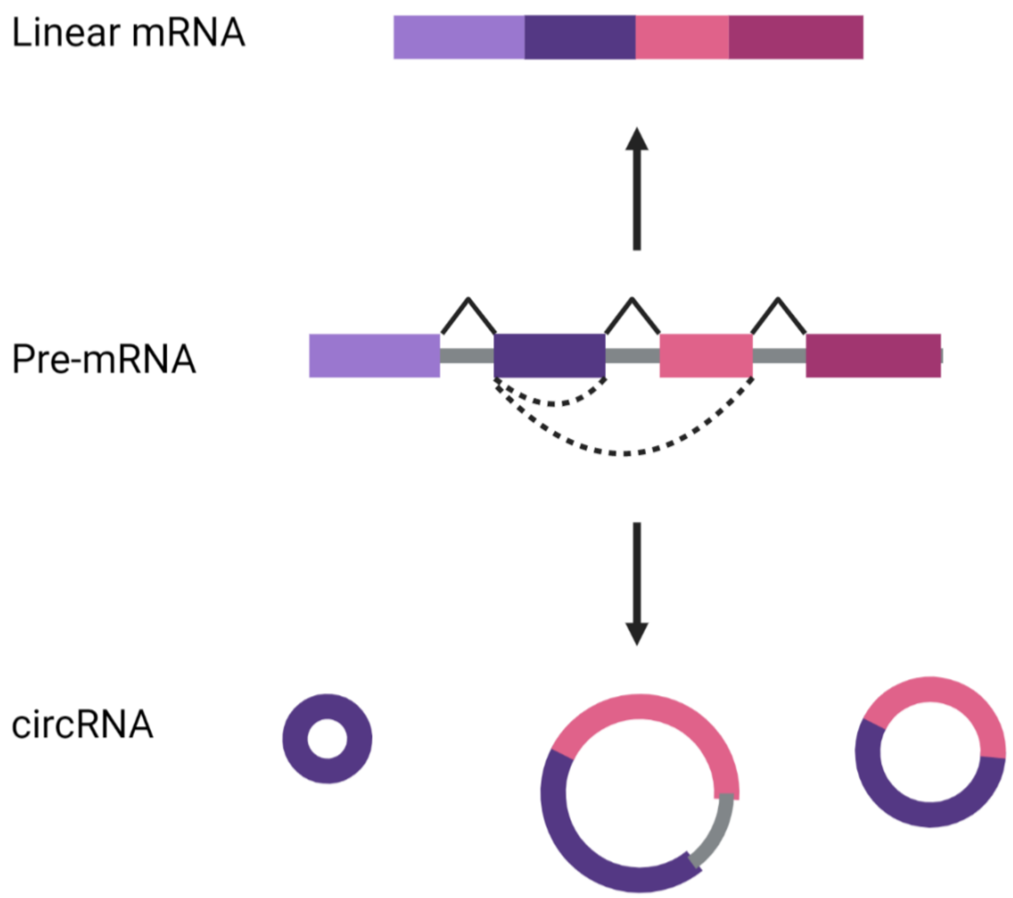


Figure 3.3 – **Schematic representation of canonical eukaryotic splicing and backsplicing**. A typical pre-mRNA molecule contains several intron and exon sequences. The introns may be spliced into a variety of alternative splice forms but typically this occurs in such a way that introns are spliced out of the mRNA (solid lines) and the 3′ ends of upstream exons are joined to the 5′ ends of downstream ones (top mRNA). However, in certain cases, the 5′ end of an exon may be joined to a downstream 3′ exon end (dashed lines). This produces a circular RNA molecule in a process known as backsplicing (bottom of image). Backsplicing can have several alternative outcomes. For example, the two ends of the same exon may be joined to form a single-exon circle (bottom left).

Alternatively, the 5′ end of an upstream exon can be joined to the 3′ end of a subsequent exon, producing a circle that contains two exons and retaining the intron between them (bottom centre). The intron that is retained in this way is still capable of canonical splicing and may be removed to produce a circular RNA from just two exons (bottom right).

In the event of large regions of complementarity between two intronic regions in the same pre-mRNA, a hairpin loop will be formed. Such complementarity aids backsplicing by favouring splicing between the splice donor and acceptor sites that have been brought close due to the hairpin loop formation (Graveley, 2005). This will be explained in more detail in subsection 3.2.

Indeed a relative enrichment of complementary regions within introns that flank known circularisable exons has been observed in animals (Ivanov et al., 2015; Li et al., 2015). Backsplicing typically produces circRNAs that contain a single exon or two exons with an unspliced intron. However, much longer circRNAs could also be produced (Jeck et al., 2013). Backsplicing has been utilised experimentally to circularise RNA in the context of RCT (Costello et al., 2019).

Self-splicing introns (also known as self-catalytic ribozymes) present an alternative for RNA circularisation. These have a complex structure that stimulates folding and autocatalysis at particular splice donor and acceptor sites (Wesselhoeft et al., 2018; Zhang et al., 2014). This type of intron generally requires helper proteins to undergo this splicing *in vivo* but under high salt conditions, self-splicing can occur *in vitro* as well. There are two distinct groups of these self-catalytic introns – group I and group II. They differ in their mode of action but overall they both result in the covalent joining of two exon ends (Jeck and Sharpless, 2014; Vanderveen et al., 1986). This enables RNA to be circularised *in vivo* or *in vitro* and then transfected into cells.

There are several ways to utilise these mechanisms and generate circRNA within the laboratory. The majority of the approaches rely on initially producing a pre-mRNA molecule that contains a single exon that is to be circularised as well as a pair of intronic sequences that flank the exon. Backsplicing can then proceed in a predictable way to produce a circRNA molecule.

However, for the purpose of producing circRNA for experiments, enzymatic circularisation *in vitro* is also available (Muller and Appel, 2017). This is a complex and time-consuming approach but enables circRNA to be produced without the need for cell culture.

This diversity of circularisation strategies allows circRNA translation to be tailored to the particular expression system *in vitro* or *in vivo*. Due to the availability of backsplicing strategies, it is feasible to create a transgene that would contain intronic sequences which flank a region composed of a spidroin coding sequence as well as a site for cap-independent translation. A cell that contains such a transgene will not require its circRNA to be transfected repeatedly potentially simplifying the synthesis of artificial spidroins.

3.1.4 Previous Applications of Rolling Circle Translation

Such expression systems have been used in the past to produce long concatemers of GFP (Abe et al., 2013; Perriman and Ares, 1998) and FLAG tags in *E. coli* and mammalian cell culture (Abe et al., 2015).

Costello and colleagues (2019) were able to leverage this expression system to produce a high titre of recombinant human erythropoietin protein (a blood-cell favourable growth factor). They referred to the process as continuous translation of circRNA (CTC). Importantly they used a protein cleavage amino acid motif to produce a homogeneous protein yield as opposed to a long sequence of linked proteins. However, the synthesis of long spidroin chains has not been attempted so far. No other large structural protein has been synthesised in this way so far either.

More recently a repetitive chain of GFP was synthesised in a bacterial host. This utilised a separate set of translation initiation strategies but relied on the same type of IORF. The result was that recombinant GFP quantities exceeded those of standard vectorial translation 1.5-fold (Lee et al., 2021).

The work described in the rest of this chapter sets out to produce the plasmid sequences necessary for further expression and characterisation of the circular translation of spidroins. First, the strategy for producing an infinitely-translated spidroin-encoding circRNA is outlined. The necessary considerations to enable an IORF are described as well as the strategies and sequences that will facilitate transcribed RNA to be circularised via backsplicing. Next, the silicon-based gene synthesis of the necessary fragments is explained. Finally, the stages of plasmid cloning are described. Several spidroin-encoding sequences are to be cloned into the pcDNA3.1 vector in three stages using a combination of cloning methods (*in vivo* assembly and restriction endonuclease cloning).

3.2 Design of IORF-Producing Sequences

To produce a spidroin-encoding IORF, three key sequences must be designed and cloned into a common construct: a spidroin nucleotide sequence, a sequence for ribosomal recruitment and intronic sequences for RNA circularisation.

For subsequent IORF construction and expression part of the MaSp1 gene (GenBank ID: AM490170.1) of the nursery web spider *Euprosthenops australis* was chosen. Due to the protein's impressive mechanical properties, it has been studied in depth in the past (Rising et al., 2007). In particular, dragline silk (partly composed of MaSp1 protein) from this spider is one of the strongest silks known (Madsen et al., 1999; Vollrath and Knight, 2001). Short, recombinant spidroins formed from *E. australis* sequences have been successfully used in the production of silk fibres (Hedhammar et al., 2010; Hedhammar et al., 2008; Stark et al., 2007). A part of the MaSp1 gene's central repeat domain was chosen, particularly its glycine-rich region (encoding amino acid motifs with a consensus of GQGGQGGYGGLGQGGYGQGAGSS) followed by a poly-alanine block (encoding 12-13 alanines). One combination of these two motifs will be referred to here as a GA repeat (from glycine-rich and poly-alanine block) (Figure 3.4).

A peptide chain of these GA repeats should produce a spidroin similar to the naturally-expressed ones but omitting the characteristic terminal domains. The N-terminus is the part of the natural spidroin that exits the ribosome first. Adding this sequence to the coding region of the circRNA would have produced a recombinant spidroin with a conventional N-terminus but also it would have been iterated during each cycle of translation. This would have introduced a large percentage of non-structural protein within the overall chain, potentially compromising its mechanical properties. Additionally, the C-terminus was not added because in RCT translation is expected to terminate stochastically and there is no way currently to regulate termination of RCT. Finally, either two or four GA repeats were chosen for cloning into circRNA-generating constructs.

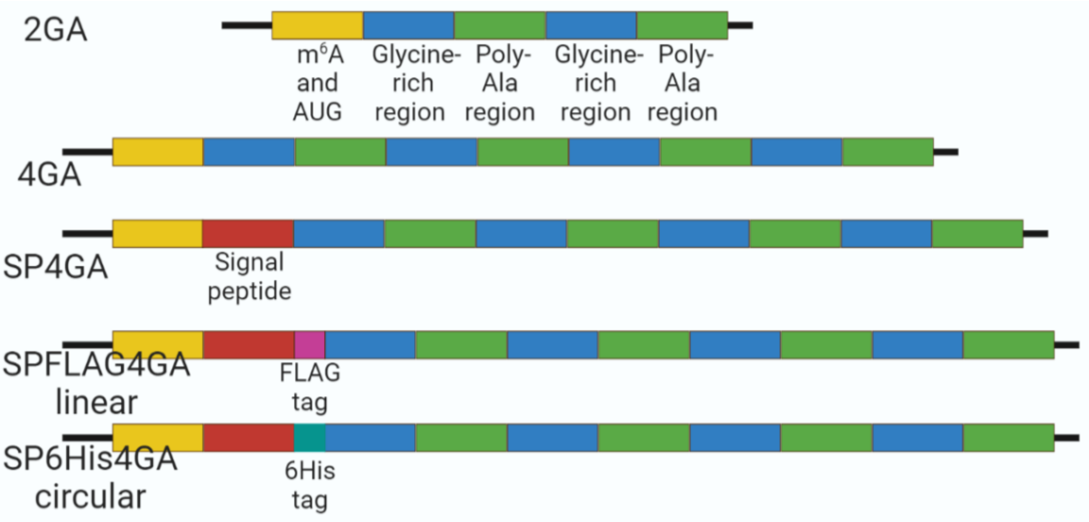


Figure 3.4 – **Library of cloned spidroin coding sequences described in this thesis**. A region with two m^6A methylation signals is shown in yellow. Glycine-rich regions are shown in blue. Poly-alanine tracts are shown in green. “GA” designates a single spidroin repeat that contains both a G-rich and poly-A region. The signal peptide (SP) sequence of human trypsin is shown in red. FLAG and 6His affinity tags are shown in pink and dark cyan respectively. All of these sequences aside from SPFLAG4GA are to be flanked by intronic sequences to enable backsplicing.

The other part of the coding sequence of an IORF is the region that enables cap-independent translation initiation. The particular sequence chosen in this work was a set of two RNA m^{6}A methylation motifs – "GGACU". The adenine nucleotide in this motif is to be methylated by the METTL3 methyltransferase prompting translation initiation to proceed (Yang et al., 2017a). Since this mechanism results in a ribonucleoprotein complex which begins scanning for a first AUG start codon, there is very little probability that translation initiation will begin at an incorrect codon (Zhou et al., 2018). If translation commences at an AUG codon in any other reading frame it would terminate quickly due to meeting an out-of-frame stop. Only one reading frame in this circRNA is infinite. Eventual translation termination would ensure that only continuous spidroins are expressed efficiently at high levels.

Additionally, the methylation motifs were flanked by spacer sequences that provide a sequence context favourable to both methylation at the GGACU motifs and initiation of translation at the downstream AUG codon (Yang et al., 2017a). The RNA methylation signal was chosen for ribosome recruitment due to its short nucleotide sequence. The overall length is 33 nucleotides (11 codons). Since this region needs to be translated in its entirety during every cycle of translation, such a short length would ensure minimal alteration of the spidroin primary structure. Additionally, this region's nucleotide length is a multiple of three (36 nt) and it is devoid of in-frame stop codons. The rest of the sequences to be part of the circRNA are specifically protein-coding and so are already at lengths of a multiple of three. Therefore the overall length of each circRNA to be produced is also a multiple of three and would enable continuous translation without a change in frame.

The spliceosome-dependent backsplicing approach has previously been reported to efficiently produce circRNA in a mammalian cell culture context (Costello et al., 2019). Additionally, this occurs *in vivo* after a vector has been transfected into the host cell. This ultimately ensures that a sufficient quantity of circRNA will be available to the host cell for the duration of transient transfection and without the need for laborious enzymatic *in vitro* circularisation.

The design of the intronic sequences was based on previous research by Costello *et al*. (2019). Specifically, the two intronic sequences that flank human actin-β's (ACTB) fourth exon were chosen. These have been predicted to contain splice signals with high splicing efficiency. In the spidroin-producing construct, these intronic sequences (235 nucleotides from the 5′ intron and 91 nucleotides from the 3′ intron) were designed to flank the m^{6}A-GA exon. Additionally, 80 nucleotides from the 5′ intron were added in reverse complementary sequence to the end of the 3′ intron. This ensures that a hairpin loop will be formed in the pre-mRNA which will bring the splice donor and acceptor sites together, thus increasing backsplicing efficiency exponentially (Figure 3.5). After the eventual backsplicing, a distinctive junction will be formed between the 3′ end of the spidroin encoding sequence and the 5′ end of the m^{6}A region.

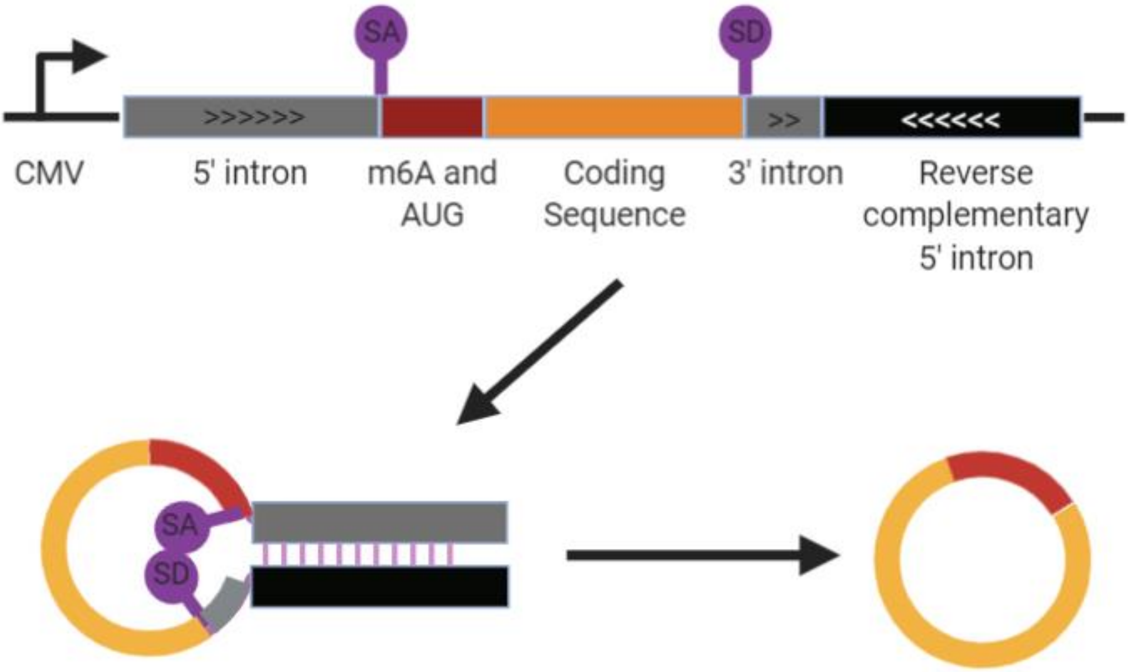
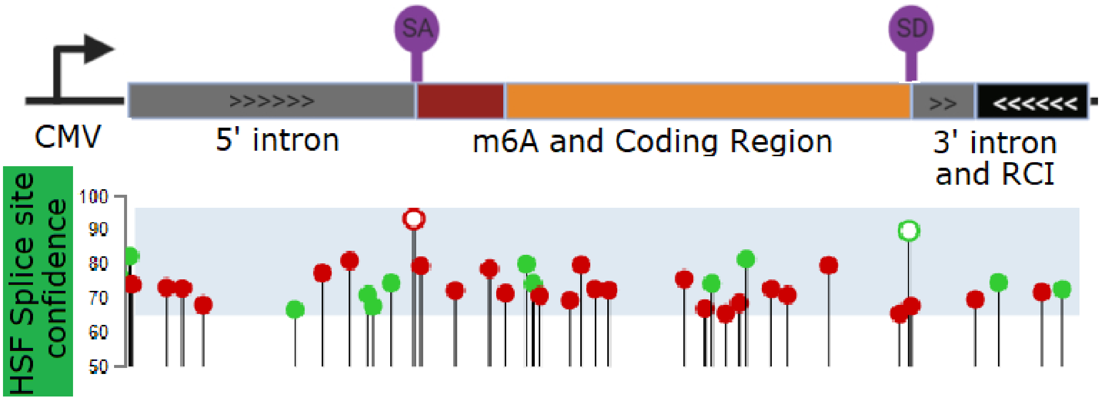


Figure 3.5 – **Backsplicing circularisation of 2GA.** The construct at the top shows a schematic representation of a pre-mRNA that is to be circularised through backsplicing. This is dependent on the annealing and hairpin formation of two complementary intronic sequences (black and grey). A splice donor site (SD) and a splice acceptor site (SA) are shown in purple within the hairpin, these are near each other and so they interact in backsplicing at high efficiency.

The splice sites' efficiency in this new context was assessed using human splice finder 3.1 (HSF) (Desmet et al., 2009). This showed that specifically the ACTB exon 4 splice signals have a substantially higher predicted splicing efficiency than the rest of the sequence (Figure 3.6).



	Motif	Splice confidence
Acceptor splice site	ctttctttgcagAA	93.34
Donor splice site	GCGgtgagt	89.87

Figure 3.6 – **Splice site analysis from Human Splice site Finder 3.1 (HSF) of the m^6A-GA ORF and the flanking intron sequence**. A representation of the pre-mRNA is shown on top including the reverse complementary intron (RCI). Any potential splice sites are depicted below this. Both diagrams are to scale. The two strongest splice signals are shown as white circles (or in purple above this). The remaining potential splice signals are shown in red or green. The table below summarises the confidence value that the HSF algorithm predicted for the two strongest splice signals.

Next, a library was designed which would have variations in GA coding sequence but not in the m^{6}A region or the intronic region. This would enable comparisons between different constructs in order to identify how particular peptide sequences impact the long repetitive protein when iterated after each cycle of translation. An initial construct with four GA repeats (will be referred to as 4GA) would serve as the basis for these comparisons. A construct with two repeats (2GA) would produce a repetitive spidroin with half the relative amount of spidroin compared to 4GA.

An addition of a 5' terminal signal peptide could enable the export of the nascent protein into the extracellular environment. This would prevent protein build-up within the cell which otherwise might be lethal or severely stressful (Dobson, 2003; Tyedmers et al., 2010). The particular sequence which was chosen was initially that of human trypsin preprotein's signal peptide (NCBI Reference Sequence: NP_002760.1). This was predicted to have a very high signal peptide activity and undergo signal protease cleavage in the context of the 4GA (immediately downstream). Signal peptide prediction was carried out using SignalP 5.0 (Armenteros et al., 2019). Additionally, at 15 codons in length, this particular sequence is at the low end of typical SPs (typically 15-30 codons/amino acids). This would ensure a minimal non-spidroin composition to the overall recombinant protein. The resulting sequence will be referred to as SP4GA.

Affinity tags were to be added to two constructs to facilitate western blot analysis. One of these would contain the 6His tag and would be added to SP4GA to produce a circularisable SP6HIS4GA RNA. The other would contain the FLAG tag and belong to a non-circularisable construct (contains neither intron). This would be named SPFLAG4GA.

After each complete sequence was designed they were planned for expression from the pcDNA3.1 (+) vector. This vector enables the constitutive transcription of recombinant sequences from the cytomegalovirus promoter (CMV) at a very high rate (Barrow et al., 2006).

3.3 Silicon-Based Gene Synthesis

To clone any of the aforementioned sequences into pcDNA3.1 (+) they first need to be synthesised. Since these sequences have diverse origins it would be much simpler to have them synthesised and delivered by a dedicated DNA synthesis service. Otherwise, cloning each part of the overall sequence would take numerous reactions. Silicon-based gene synthesis is a method of producing double-stranded DNA sequences between 300 and 1800 bps in size at a fraction of the cost of conventional DNA synthesis approaches. This can take up to two weeks which is substantially less than what it would have taken realistically if traditional cloning were involved. However, the caveats of this silicon-based synthesis approach are that it cannot produce sequences with high degrees of complementarity or repetitive sequences. Therefore, the DNA sequence of the entire circularisable RNA could not be produced in one fragment. The two introns would have to be produced in two separate fragments ("2GA short" and reverse complementary intron (RCI)) (Figure 3.5). Initially, short regions of overlap (~30-40 bps) between pcDNA3.1, "2GA short" and RCI were planned. These were to be taken advantage of during the subsequent Gibson Assembly (more detail in the next subsection). As a contingency, restriction endonuclease sites (*HindIII*, *XbaI*, *BstEII*) were planned in the event that Gibson Assembly would be unsuccessful. To produce the aforementioned construct library, additional fragments were ordered corresponding to each coding sequence. These were later to be cloned using *BamHI* and *BstEII* restriction endonucleases (Table S1).

To reduce GC content in the synthetic DNA fragments, codon optimisation was implemented only over the protein-coding region and specifically not over the ribosome binding region. Codon optimisation was carried out in the context of mammalian amino acyl-tRNA pools. Twist Bioscience's web app (Twist Bioscience, 2022) was used for codon optimisation.

3.4 Producing a Library of circRNA-generating Plasmid Constructs

3.4.1 Gibson Assembly of Circularisable Spidroin Exon into pcDNA3.1 (+)

Initially, the synthesised spidroin-encoding DNA fragments were designed to enable Gibson Assembly between a host vector (pcDNA3.1) and two insert sequences. Gibson Assembly has the advantage of inserting DNA sequences reliably, into the correct orientation into the vector with multiple inserts being added within the same reaction (Gibson et al., 2009).

After multiple attempts, no construct could be produced through this Gibson Assembly. Upon closer scrutiny, the complementarity between RCI and "2GA-short" most likely has been responsible for inefficient PCR reactions prior to the Gibson Assembly reaction.

3.4.2 Restriction Endonuclease Cloning of Circularisable Spidroin Exon into pcDNA3.1 (+)

Subsequently, Cloning was planned to proceed in three stages. First, the "2GA short" sequence was cloned into the pcDNA 3.1 (+) vector using the *HindIII* and *XbaI* restriction endonucleases. This was later confirmed with a PCR colony screen using a primer set (pcDNA3.1_fw and pcDNA3.1_rev) that flanks the insert (Figure 3.7) (Table S1). Amplicon band size was used to differentiate between the cloned plasmid and the empty pcDNA3.1. Sanger sequencing confirmed that cloning had been successful. Next, the *BsteII* and *XbaI* restriction endonucleases were used to add the RCI sequence to the pcDNA3.1-2GA-short plasmid. The complete result of this subcloning (in this order: 5'intron; m^{6}A methylation site; 2 GA repeats; 3' intron; reverse complementary intron) will be referred to here as the p2GA plasmid. Again, colony screen PCRs were performed this time with a primer pair where the forward primer is located within the RCI insert (RCI_screen_fw and pcDNA3.1_rev primers). Band size was used to differentiate between p2GA and its host plasmid (Figure 3.7). Similar to the previous cloning, Sanger sequencing confirmed the successful assembly of the new construct.

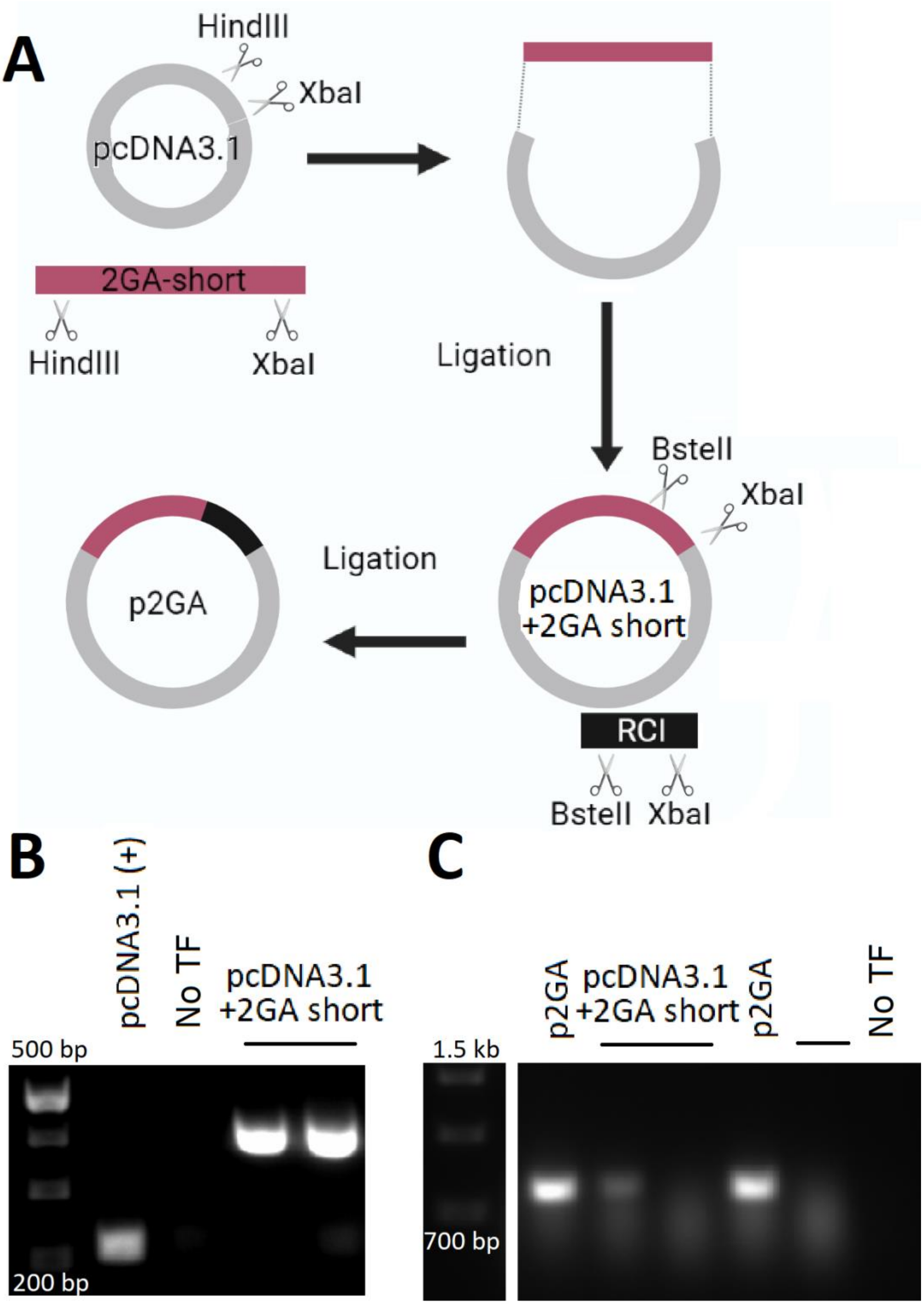


Figure 3.7 – **Molecular cloning of p2GA**. (A) Schematic representation of subcloning to produce p2GA. (B) Gel electrophoresis of colony screen PCR for insertion of "2GA short"

into pcDNA3.1. (C) Gel electrophoresis of colony screen PCR for p2GA. 1% agarose gels were prepared. 20 µl PCR reactions were loaded. ~50 ng template was used per well. The negative control contained no template DNA.

3.4.3 Subcloning of Protein-Coding Motifs into p2GA

At this stage, the p2GA vector provided intronic sequences within a mammalian expression vector which flanked two restriction enzyme sites (*BamHI* and *BsteII*) (Figure 3.7). Using this construct a wide variety of spidroin-encoding sequences could be cloned using these two sites. The aforementioned SP4GA and 4GA DNA sequences were ordered as double-stranded DNA fragments flanked by these restriction sites. After the cloning procedures were carried out, the resulting plasmids were named pSP4GA and p4GA. The result was assessed with a PCR and Sanger sequencing (Figure 3.8). The SP4GA_screen_fw, 4GA_screen_fw and pcDNA3.1_rev PCR primers were located on both sides of the *BsteII* restriction site (Table S1).

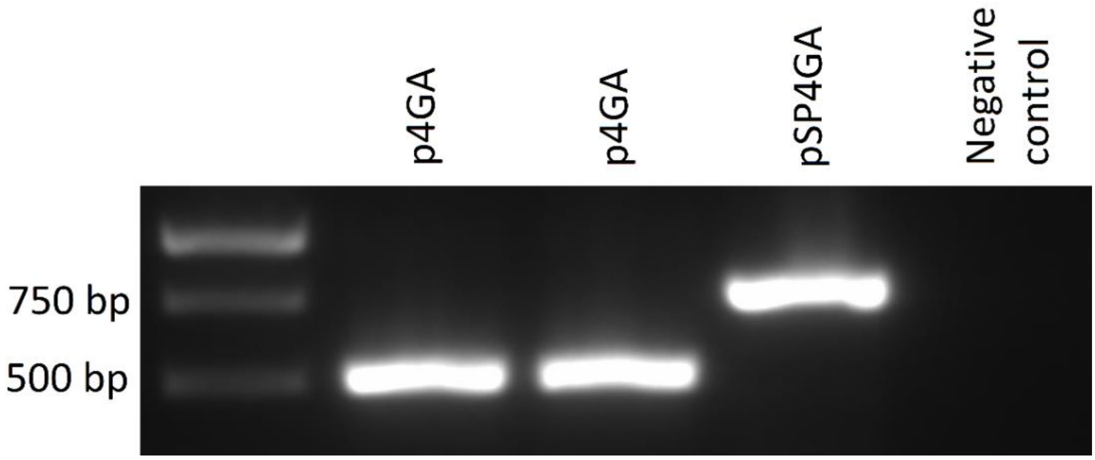


Figure 3.8 – **Colony screen PCR results of subcloning to produce pSP4GA and p4GA**. A 1% agarose gel was cast and 20 µl reactions were loaded per well. ~50 ng of template DNA were used per reaction while the negative control contained no template DNA.

3.4.4 Restriction Endonuclease Cloning of SPFLAG4GA into pcDNA3.1 (+)

The double-stranded SPFLAG4GA fragment was cloned into the pcDNA3.1 vector using the *BamHI* and *XbaI* restriction endonucleases (Figure 3.7). This fragment included an UAA stop codon followed by a guanosine base (UAAG) at the end of the spidroin coding sequence. The UAAG motif is known to produce a very consistent translation stop signal and would ensure that a short monomer spidroin is made (Cridge et al., 2018; McCaughan et al., 1995). This cloning resulted in the pSPFLAG4GA- plasmid. This construct should produce linear mRNA and not be spliced into a circle. The purpose of pSPFLAG4GA- is to express a ~14kDa protein that should be detectable through its FLAG tag. This would serve as a positive control in later spidroin expression. Cloning was assessed with Sanger sequencing and a colony screen PCR with a primer pair that surrounds the *XbaI* site - SP4GA_screen_fw and pcDNA3.1_rev (Figure 3.9, Table S1).

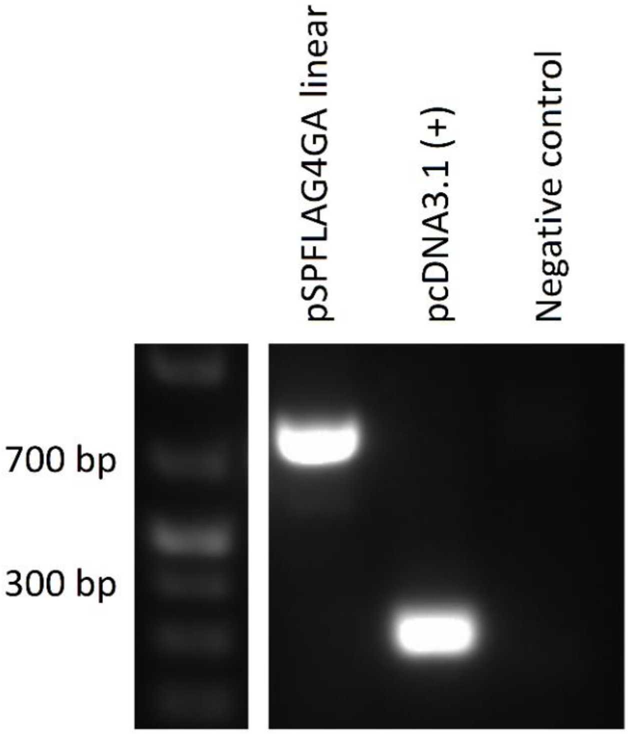


Figure 3.9 – **Restriction endonuclease cloning of SPFLAG4GA into pcDNA3.1**. Gels were cast with 1% agarose and 20 µl reactions were loaded per lane. Around 50 ng of pure plasmid DNA was used for the template in the PCR and the negative control contained no template DNA.

3.4.5 *In Vivo* Assembly of 6His Affinity Tag into pSP4GA

Finally, *in vivo* assembly (IVA) was implemented to insert the sequence encoding the 6His affinity tag into the pSP4GA vector. *In vivo* assembly takes advantage of a recombination mechanism which is common for most *E. coli* strains that are widely used in the lab for cloning (including DH5α). This only requires short homologous regions at the ends of a linear insert and a linearized plasmid to produce a single construct. Plasmid linearization and insert amplification can both take place during the same PCR. Subsequent digestion of this PCR reaction with the DNA methylation-dependent restriction endonuclease *DpnI* ensures that none of the circular empty vector plasmid is transformed into bacteria. These fragments are then transformed into competent *E. coli* whose recombination machinery assembles complete recombinant plasmids. Similarly to Gibson Assembly, IVA can produce a plasmid sequence out of several inserts all in the correct orientation. Additionally, IVA requires no additional procedures such as restriction enzyme digestion, gel extraction and *in vitro* ligation of DNA fragments. This significantly reduces the costs and the time to produce a successfully cloned plasmid (Bi and Liu, 1994; Bubeck et al., 1993; Garcia-Nafria et al., 2016).

To produce a complete pSP6HIS4GA construct, two primers were designed – IVA_fw and IVA_rev (Table S1). These included ~20 bp regions complementary to either pSP4GA’s SP sequence or the beginning of its spidroin-encoding DNA region (Figure 3.10 A). They also included two complementary sequences that together produce an 18 bp 6His coding sequence.

After the IVA procedure was carried out, successful cloning was confirmed through a PCR with a primer pair that partially covers the insert sequence – 6His_fw and pcDNA3.1_rev (Figure 3.10 B). Sanger sequencing further supported that the corresponding sequence has been cloned successfully.

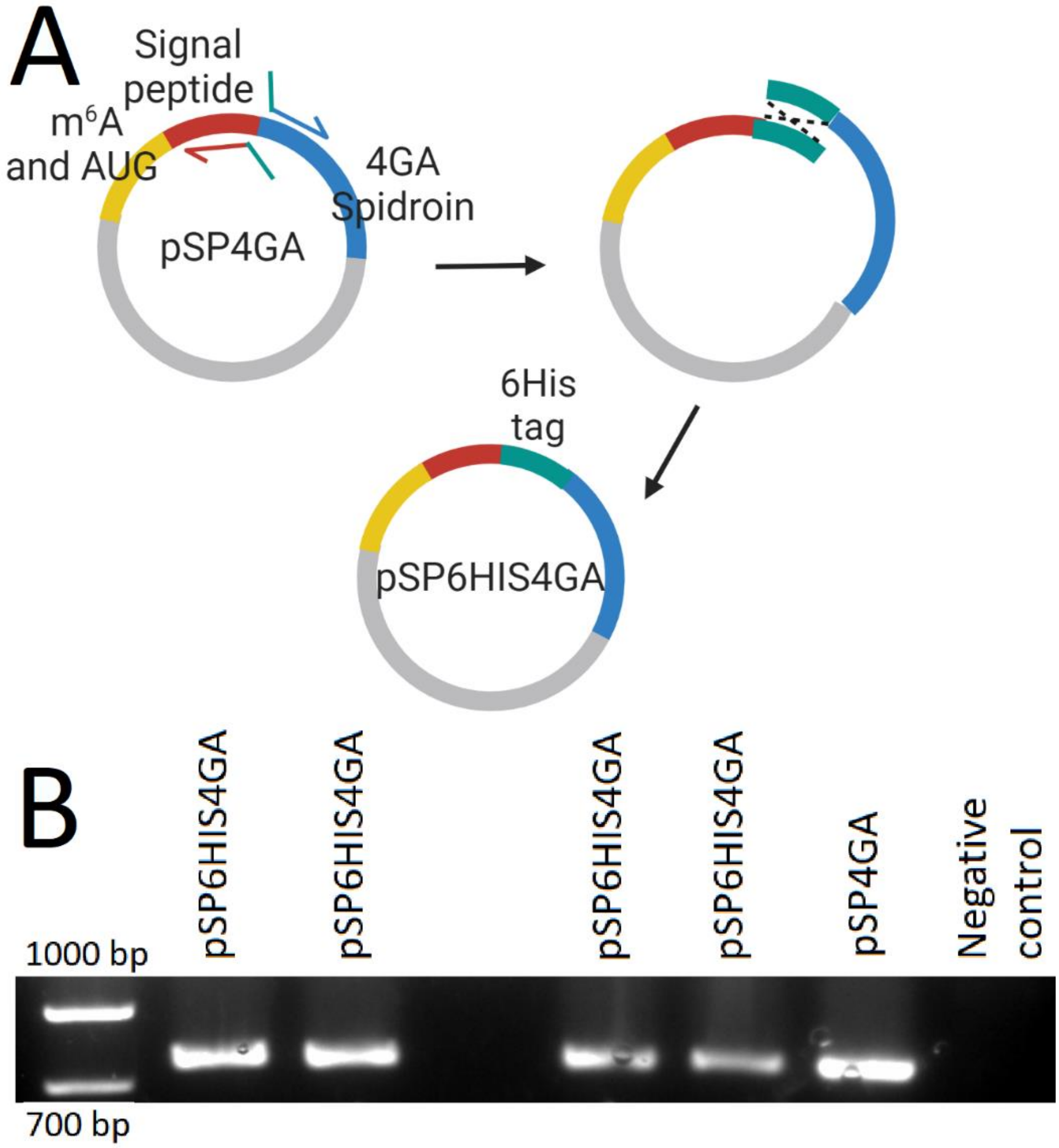


Figure 3.10 – (A) **Schematic representation of 6His IVA cloning into pSP4GA**. An initial PCR reaction amplifies the entire pSP4GA parent plasmid to produce a linear DNA which possesses identical 6HIS coding sequences on each end. These are recognised by homologous recombination machinery in *E. coli* after transformation and then produce a circular plasmid with one 6HIS coding sequence (dashed lines). (B) **Colony screen PCR of pSP6HIS4GA cloning**.

pSP4GA was used as a positive control (a different primer set was used for the latter).

In summary, a library of plasmid sequences has been successfully produced. These are to be used in downstream transfections into mammalian cell lines. The intron sequences within these plasmids will enable backsplicing that would produce circRNA molecules. These will contain an IORF that will express a type of repetitive spidroin protein. The spidroin coding sequence was further supplemented with signal peptide sequences for nuclear export or affinity tags to enable later western blot analysis. Molecular cloning proved to be laborious and required multiple subcloning approaches to finally produce circularisable spidroin-encoding sequences. In particular, the spidroin's repetitive sequence, its GC-rich content and the reverse complementarity between introns were the main complications that impacted molecular cloning.

3.5 Discussion

The results and methods presented here outline a robust and reliable approach toward the construction of circRNA-producing plasmids. Numerous considerations have been taken into account to produce a sequence that should generate circRNA in live cells.

The m^{6}A sites for ribosome recruitment and translation initiation chosen here were among the shortest possible that have been described in the literature. Upon further research, it may be found that this length of non-spidroin sequence within the overall peptide chain is detrimental to spidroin properties. In this case, it may be possible to further reduce the size of the ribosome binding region by removing one of the m^{6}A motifs described here. The original research that described these sites identified no large decrease in translation efficiency if only one of these sites is present (Yang et al., 2017a). Another benefit of these motifs is that m^{6}A methylation is known to improve translation efficiencies by resolving RNA secondary structures (Mao et al., 2019). It may be the case that an alternative translation initiation method would produce a larger amount of protein from the circRNA than the m^{6}A sites. However, no research so far has been conducted which compares translation efficiencies between m^{6}A sites, IRESs or any other cap-independent translation mechanism in the context of circRNA.

There are several other similar approaches that may enable ribosome recruitment to the circRNA. Abe et al. (2015) reported that IORF translation is possible even without a specialised ribosome recruitment motif. It is likely, however, that translation initiation in that context occurred due to the aforementioned m^{6}A-dependent pathway which was unknown at the time (Yang et al., 2017a). The majority of approaches use IRESs or similar sequences to reliably initiate translation (Thompson, 2012).

More recently it has been reported that GGGGCC repeat RNA sequences are capable of inducing translation at a considerable level comparable to several IRESs (Wang et al., 2021). The specific mechanism has not been studied in detail so far but it may even be possible to induce translation from the spidroin-coding sequences based on the fact that they are GC-rich.

circRNA produced enzymatically *in vitro* (as opposed to via splicing) can be transfected into cells. This is an alternative to backsplicing in the circularisation of RNA. However, since circRNA produced enzymatically has not been transcribed *in vivo*, it may lack certain modifications like RNA methylation so non-m^{6}A translation initiation methods will be needed in this case (Wesselhoeft et al., 2018). Additionally, byproducts of *in vitro* circularisation have been implicated in immunogenicity during the course of transfection into the host cell. Stringent purification of the circRNA is needed to avoid this (Pandey et al., 2019; Wesselhoeft et al., 2019). Furthermore, enzymatic *in vitro* RNA circularisation demands additional reagents and more manual work. These caveats are the reason why this approach was not chosen for the eventual expression of recombinant spidroins. The main advantage of *in vitro* circularisation is that it enables the use of various *in vitro* translation methods.

Hairpin-producing and GC-rich DNA sequences are sometimes difficult to amplify using conventional PCR approaches (Frey et al., 2008; Nelms and Labosky, 2011; Tsuneoka and Funato, 2020). Particularly in every annealing step of a PCR, the long complementary regions may partially anneal alongside the shorter primer sequences thus forming a steric obstacle in the way of active DNA polymerase, severely limiting amplification efficiency. This may have led to the unsuccessful Gibson cloning of the 2GA sequence into the pcDNA3.1 vector. Phusion and most other commercially available DNA polymerase enzymes are incapable of DNA helicase activity (Cao et al., 2018). This means that they are only capable of replicating single-stranded DNA and that the hairpin formation will interrupt DNA replication.

Additionally, PCR reactions aiming to amplify either the 2GA or 4GA sequences were often unsuccessful. This is largely consistent with previous observations of spidroin PCR amplification (Ayoub et al., 2013; Wang et al., 2019). The application of a high-fidelity DNA polymerase enzyme and the use of a reaction buffer optimised for GC-rich DNA sequences may have aided the process considerably but overall unsuccessful PCRs of the repeat regions led to numerous delays to the molecular cloning process. Even with low repeat numbers (2-4 GA repeats), it is still apparent that difficulties in DNA manipulation of repetitive sequences are a major obstacle.

Initially, during the course of Sanger sequencing of the spidroin library, very poor and inconclusive results were observed. This again was most likely caused by the presence of GC-rich and reverse complementary sequences. The use of dGTP chemistry largely improved sequence read quality (Kieleczawa, 2005). This suggests that future attempts at sequencing such constructs will greatly benefit from this approach.

Another important consideration in circRNA translation is the circle's overall size. A eukaryotic ribosome's footprint is slightly variable but in general is ~28-30 nt long (Guydosh and Green, 2014). This suggests that circRNA molecules that are too short would not be capable of being translated. Indeed it has been reported that circRNA at the size of 84 nt could not produce a repeated peptide product. The smallest circRNA to have undergone RCT is 126 nt in size (Abe et al., 2013). The circRNA sequences described here are all larger than 200 nt and so are expected to translate continuously.

Overall, a plasmid library was constructed which contains a variety of spidroin-encoding sequences. These include a variable number of spidroin (GA) repeats, signal peptide sequences and peptide affinity tags. These constructs were intended to be transfected into mammalian cell lines where they would be constitutively transcribed and the resulting RNA will be circularised through backsplicing. Introns with reverse complementarity have been designed that will enable efficient backsplicing. A sequence for m^{6}A RNA methylation was also included which will enable a ribosome to continuously translate a repetitive spidroin from the non-repetitive circRNA.

In the subsequent Chapter 4, the transfection of several of these plasmids will be described. Several experiments that aim to confirm or disprove successful RNA circularisation and protein expression were conducted.

4 Recombinant Expression of a continuously circRNA-translated Spidroin

4.1 Introduction

4.1.1 Host Cells

To express the recombinant spidroins from a continuously translating circRNA, a suitable host must be chosen. As outlined in Chapter 1, spidroin expression comes with a unique set of challenges, including a very specific metabolic demand for amino acids (particularly alanine, glycine, and glutamine) and a need for extracellular export. Additionally, continuous, m^{6}A-dependent circRNA translation is more easily accommodated but nonetheless it is important to ensure that host cells express specific RNA methylases and readers.

Certain cell lines have a long history of recombinant expression. HEK293 cells have the advantages of reliable growth and simple transfection protocols. Particularly, quick and efficient transient expression is possible with this cell line (Meissner et al., 2001; Smart and Thomas 2005). Their transfection efficiency after the use of some commercial transfection reagents can even exceed 90% (Geisse and Fux, 2009). This, along with a high recombinant protein titre makes them particularly suitable for transient expression of proteins (Dalton and Barton, 2014; Walsh, 2014).

On the other hand, NIH/3T3 fibroblasts are particularly adapted for the production of collagens. These proteins are naturally repetitive and are very glycine-rich (Traub and Piez, 1971). Because they have adapted to produce collagen, NIH/3T3 cells may provide a possible answer to part of the metabolic needs of recombinant spidroins. While they are not commonly used for recombinant protein expression, NIH/3T3 cells can be transfected with a variety of available reagents (Liu et al., 2019b).

Monkey kidney COS-1 cells have been used successfully in the past to express recombinant silk proteins (Grip et al., 2006). Also, they express the large T antigen. This is derived from the SV40 virus and is capable of replicating plasmids that have the SV40 origin of replication (Aruffo 2002; Weinander et al., 1995). This is exactly the case with the pcDNA3.1-derived plasmid constructs and may increase the availability of recombinant plasmid DNA within the cell.

Expression was also tested in breast cancer MCF-7 cells. These cells are primarily used to study breast cancer biology but they have certain advantages such as reliable culture growth and strong expression of METTL3 and YTHDF3 (Klinge et al., 2019; Pan et al., 2021; Sweeney et al., 2012).

The METTL3-dependent RNA methylation pathway is vital for the initiation of translation from the circRNA that has been designed. This includes the YTHDF2 and YTHDF3 proteins which are responsible for recognising the methylated RNA sites. All of these components are intact in HEK293, COS-1, NIH/3T3 and MCF-7 cells (Choe et al., 2018; Li et al., 2022; Linder et al., 2015; Mao et al., 2019; Zhuang et al., 2019).

4.1.2 Nuclear Export of circRNA

After they are spliced, circRNAs must somehow localise to the cytoplasm because the vast majority of them are found there (Jeck et al., 2013; Salzman et al., 2012). Transport over the nuclear membrane into the cytoplasm might occur through two evolutionarily conserved mechanisms. UAP56 and URH49 are two RNA helicases involved in the export of circular RNAs to the cytoplasm. UAP56 is responsible for exporting longer circRNA (>1300 nt) and URH49 is involved in short circRNA export (< 400 nt) (Huang et al., 2018; Li et al., 2019; Wan and Hopper, 2018). RNA binding proteins (RBPs) seem to bind specific nucleotide sequences of the nuclear circRNA and may somehow associate with the helicases (Zhang et al., 2019a). However, little more is known about this mechanism.

A separate m^{6}A-dependent nuclear export mechanism has been reported. The reader YTH domain-containing protein 1 (YTHDC1) specifically binds the m^{6}A modification. Knockdown of this component leads to the accumulation of circRNA in the nucleus (Chen et al., 2019). The subcellular localisation of circRNAs without this modification is not affected by this same knockdown (Di Timoteo et al., 2020). More details about this process are still to be discovered, however.

4.1.3 Ribosomal Frameshifting

Several mechanisms exist which may cause a shift in frame during translation elongation. This occurs when the ribosome skips one or two nucleotides (in either direction) as opposed to a full codon (Herr et al., 2000). Such an event would ultimately lead to an out-of-frame stop codon, terminating the translation.

While this event could occur stochastically, there are specific RNA sequences and secondary structures that can reliably induce ribosomal frameshifting (Leger et al., 2007; Namy et al., 2006)). This event is thought to occur through steric interactions between coding RNA structures, the translation elongation factors and the ribosome (Harger et al., 2002; Kim et al., 2001).

4.1.4 tRNA Charging in Recombinant Spidroin Expression

During active translation elongation, a nascent peptide increases in length because amino acids are being attached successively to the C-terminus (Grunberger et al., 1969). This process depends on amino-acylated tRNAs (charged tRNAs) that enter a ribosome's aminoacyl site (A-site) (Konevega et al., 2004). The actual charging of tRNA is carried out by aminoacyl-tRNA synthetases (ARS). These enzymes are specifically responsible for associating a specific anticodon-bearing tRNA molecule with a specific amino acid. ARSs also catalyse the formation of a covalent bond between the two (McClain, 1993; Schuber and Pinck, 1974). Active translation results in the expenditure of amino acids (added to a nascent peptide) and the release of tRNAs to be re-charged by ARSs (Kelly and Ibba, 2018; Swanson et al., 1988). While ARSs continuously charge tRNAs during translation, depletion of a certain amino acid could stall translation until the absence is corrected. Such depletion could trigger cellular stress responses (Pavlova et al., 2020). Aminoacyl-tRNA, amino acid and tRNA abundances have evolved for every cell type to meet metabolic demand under normal conditions (dos Reis et al., 2004; Frumkin et al., 2018; Gingold et al., 2012). However, in the event of recombinant protein translation, this amino acid depletion can overwhelm the cell's capacity for tRNA charging and amino acid synthesis and lead to translational stalling (Frumkin et al., 2018). This is a likely limiting factor in spidroin recombinant translation and has been partially addressed before by transfection with tRNA-gene-containing plasmids (Huh et al., 2021). However, a more direct experiment where tRNA charging levels are measured under spidroin expression conditions has not been reported before.

tRNAs that are to be aminoacylated by separate amino acids are termed isotypes (i.e., all tRNAs to be charged by arginine form one isotype). Different isotypes exist as diverse populations of molecules that differ in many ways even though they bind the same amino acid (Goodenbour and Pan, 2006; Grosjean et al., 2010). Several tRNAs within an isotype may share the same anticodon and these are termed isodecoders (i.e. all tRNAs that contain the ArgACG anticodon form one arginine tRNA isodecoder) (Geslain and Pan, 2010). Despite having the same anticodon, tRNAs may still differ in nucleotide sequence and post-transcriptional modifications. This variability complicates the analysis of tRNA charging, partly because isodecoder abundance can fluctuate between tissue types even in a single organism (Ehrlich et al., 2021).

tRNA amino-acylation can be monitored *in vivo* through a combination of methods. One northern blot-based technique requires radio-active amino acid labelling and is laborious but could reliably determine if spidroin circRNA-producing cells experience a deficiency in loaded tRNA^{Ala} and tRNA^{Gly} (Kohrer and RajBhandary, 2008; Niehues et al., 2015; Shitivelband and Hou, 2005).

However, for the experiments described here an *in vitro* approach was chosen to monitor the tRNA charging in mammalian cells after pSP6HIS4GA transfection. This technique ultimately depends on a qPCR measurement and compares the ratios of charged versus total tRNA in a sample (Pavlova et al., 2020).

The rest of this chapter focuses on the work towards expressing the spidroin-encoding circRNA within a mammalian cell culture context. First, the expression of a 2GA-repeat spidroin in breast cancer MCF-7 cells is outlined. Two protein staining methods in the detection of the recombinant spidroin are described. RT-PCRs to detect RNA circularisation are described also. Next, the attempted expression of a continuous SP6HIS4GA spidroin in HEK293s is described, again with protein and circRNA characterisation outlined. Subsequently, the attempted expression of the same spidroin in COS-1 and NIH/3T3s is described with details about protein and circRNA analysis. A computational analysis of the expected circRNA is described for the detection of ribosome frameshifting signals. Finally, a tRNA charging assay is described which assesses the aminoacylation in spidroin circRNA-producing cells.

4.2 Divergent Primer RT-PCR Strategy

For every attempted expression of any spidroin described here, RT-PCR was chosen to confirm RNA circularisation. The successful production of circRNA is an important prerequisite for continuous circular translation and had to be assessed. Several primer sets were designed to enable this. After backsplicing of pre-circRNA, the resulting circRNA should contain sequences that were previously located distally downstream of the 5' splice site but are now located immediately upstream of it (Figure 3.5). This allows a strategy where a primer pair with diverging orientations can amplify circRNA-derived templates but not those derived from linear RNA (Figure 4.1, Table 4.1). The particular annealing sites of the divergent primer pair are located within the GA spidroin repeats so as to be suitable for every type of circRNA designed here. Another (convergent) primer pair was also designed which amplifies any sequence that contains the spidroin construct – linear or circular. This serves as a positive control of sorts which shows if any spidroin RNA is present. Its annealing sites are located in the ribosome recruitment region and the spidroin repeats, again ensuring that all designed constructs can be amplified.

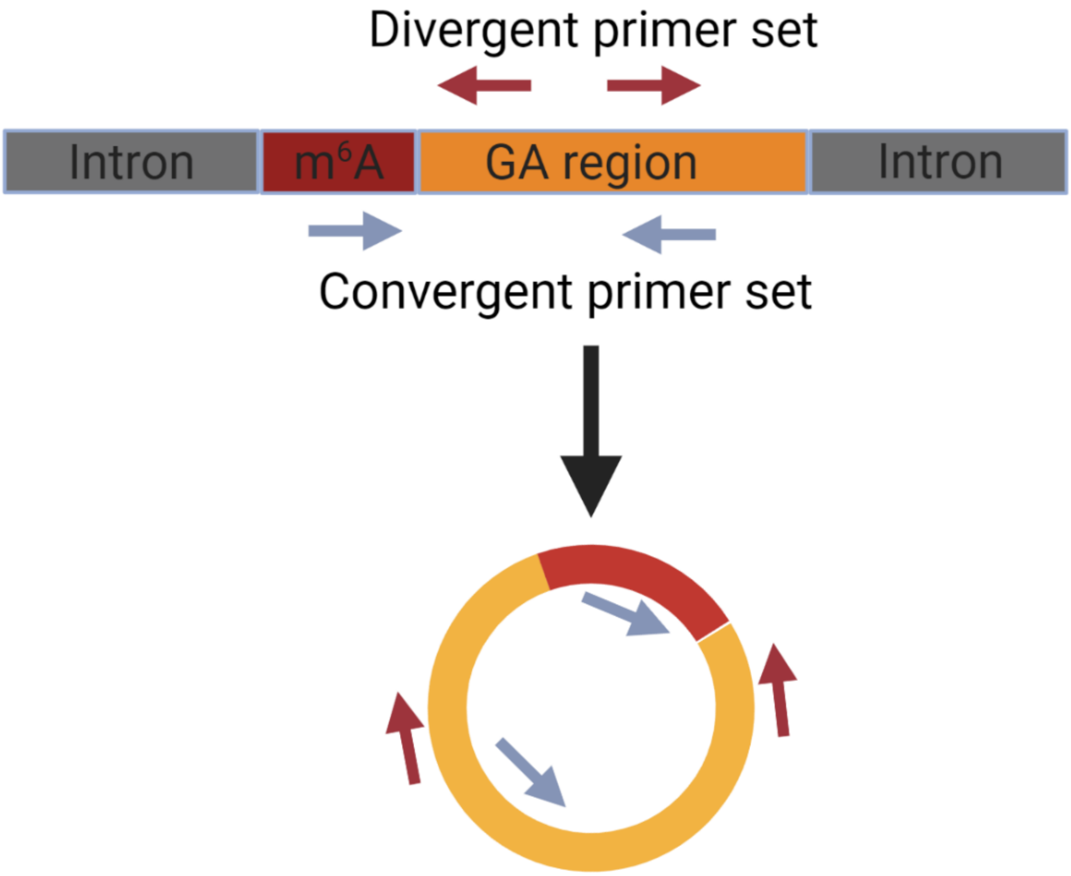


Figure 4.1 – **Schematic representation of convergent and divergent primer locations on the designed spidroin RNA molecules**. At the top is the linear form of the RNA before circularisation. The circular form can be seen at the bottom. The convergent primers are shown as blue arrows. The divergent primers are shown as red arrows.

Table 4.1 – **Primer sequences used in the circRNA RT-PCRs.**

Name	DNA Sequence
circRNA_divergent_FW	AAGGCGCAGGAATAAGTGCC
circRNA_divergent_RV	TACTAGAGCCCGTCCCTTGC
circRNA_converge_FW	GGATCCTGGACTAAAGCGGA
circRNA_converge_RV	TGTACTAGAGCCCGTCCCTTG
GAPDH_FW	GACAGTCAGCCGCATCTTCT
GAPDH_RV	TTAAAAGCAGCCCTGGTGAC
circHIPK3_FW	TATGTTGGTGGATCCTGTTCGGCA
circHIPK3_RV	TGGTGGGTAGACCAAGACTTGTGA

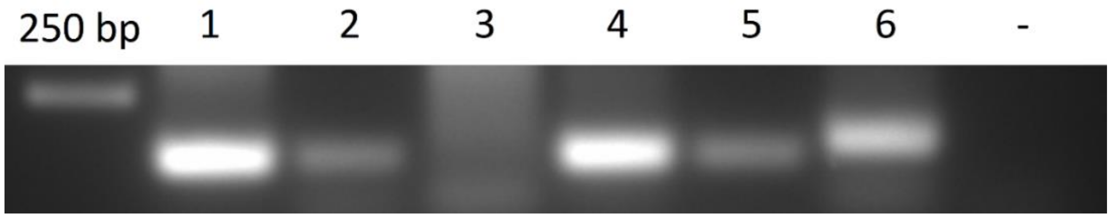
However, aside from backsplicing, canonical trans-splicing may also occur and produce a sequence that is capable of being detected via the divergent primer set but is not capable of circular translation. To avoid this confusion, an additional step was added to the circRNA extraction where linear RNA was degraded by the enzyme RNase R. This exoribonuclease can only degrade linear RNA, leaving circRNA intact. This ensures that divergent primers only amplify circRNA-derived templates.

A further pair of primers was used as a positive endogenous control for the presence of circRNA as described by Yaylak et al. (2019). circHIPK3 is a species of RNA that is abundantly expressed in the cell types described (Wen et al., 2021; Xu et al., 2021; Zheng et al., 2016). These primers were designed to amplify specifically from only the circular form of the HIPK3 RNA by again amplifying across the backsplice junction. An alternative set of positive endogenous control primers were designed to amplify the GAPDH linear mRNA.

4.3 Expression of p2GA in MCF-7 cells

After the p2GA construct had been produced (Sections 3.4.2 and 3.4.3) it was transfected into breast cancer MCF-7 cells using the PEI MAX reagent. At the same time, cells were also transfected with the precursor construct to 2GA which lacked the reverse complementary intron and so should be incapable of RNA circularisation. After, 72h RNA and protein from the conditioned media and cell lysates were purified with the phenol-chloroform approach. The total protein samples were resuspended in 4 M urea.

The resulting total RNA was used in subsequent RT-PCR experiments. The aforementioned primer sets were used to amplify cDNA from previously RNase R treated (linear RNA removed) or untreated (linear RNA present) samples. The results are shown in Figure 4.2. Lanes 1 and 4 demonstrate that regardless of the type of transfection, endogenous linear GAPDH mRNA was reverse transcribed and amplified. All amplicon lengths corresponded to what was expected. Upon the addition of RNase R, a loss of most of the GAPDH signal is observed (lanes 2 and 5). This demonstrates that part of or all of the linear RNA was degraded but that did not completely abolish the GAPDH signal. Only a partial degradation of the linear RNA or genomic DNA contamination could explain this.



1 – linear 2GA transfection, + reverse transcriptase (RT), GAPDH primers

2 – linear 2GA transfection, -RT, GAPDH primers

3 – linear 2GA transfection, +RT, divergent primers

4 – p2GA transfection, +RT, GAPDH primers

5 – p2GA transfection, -RT, GAPDH primers

6 – p2GA transfection, +RT, divergent primers

7 - no template, divergent primers

Figure 4.2 – **Agarose electrophoresis of RT-PCR assay for circRNA circularisation.** MCF-7 cells were transfected with PEI MAX and p2GA or p2GA linear plasmids. RNA was harvested from cell lysates by phenol-chloroform separation. RT-PCRs were prepared as described in the legend above. cDNA was synthesised and ~50 ng of this was used in 20 µl PCRs. 1% agarose gels were run. No template DNA was added in the negative control sample 7.

Lanes 3 and 6 show signals derived from divergent primers. These are meant to amplify only from 2GA RNA which has undergone splicing (backsplicing or trans-splicing). The RCI-less construct appears to not have produced such an RNA. On the other hand, the RT-PCR of the p2GA-derived RNA appears to have successfully produced a signal indicative of splicing. At this stage, it is unclear if this signal was from backsplicing or trans-splicing because RNase R efficiency was not independently confirmed. However, this experiment clearly shows that the RCI is very important in inducing the splicing.

Upon closer scrutiny, it was decided that GAPDH is an improper amplification target. It has several pseudogene homologous sequences in the human genome, many of which contain only exons and mimic the spliced form of the mRNA (Sun et al., 2012). The convergent spidroin primer set and the circHIPK3 primer set were chosen to control for subsequent RT-PCR experiments. The former pair would reverse transcribe from a linear RNA and show if RNase R degradation had been successful. The latter would serve as a positive control for intact circRNA and would not be affected by RNase R digestion.

Next, the total protein samples were subjected to SDS-PAGE to resolve them according to size. The spidroin RCT was expected to yield a mixed population of spidroin lengths. This was expected to appear as a smear from the size of the monomer (~14 kDa) to around the top of the gel (lengths exceeding 180 kDa). The smear was expected to be reflective of abundant protein which would be distinctive enough to be visible even in total protein samples.

Therefore, a zinc-imidazole negative stain was performed. Here, the imidazole and the zinc ions form a white precipitate which stains the polyacrylamide gels. However, proteins have a higher affinity for zinc so in their presence, this stain is not observed. Four different sample types were run on an SDS-PAGE gel – total protein from untransfected cells, from pmaxGFP transfected cells, from MCF-7 cells transfected with circRNA or linear RNA-producing spidroin plasmid. The results are shown in Figure 4.3. Equal amounts of protein were loaded but no difference was observed between any of the lanes. This suggests that either not enough recombinant spidroin had been produced to be visible or that the zinc-imidazole staining was not sensitive enough to differentiate between the different groups.

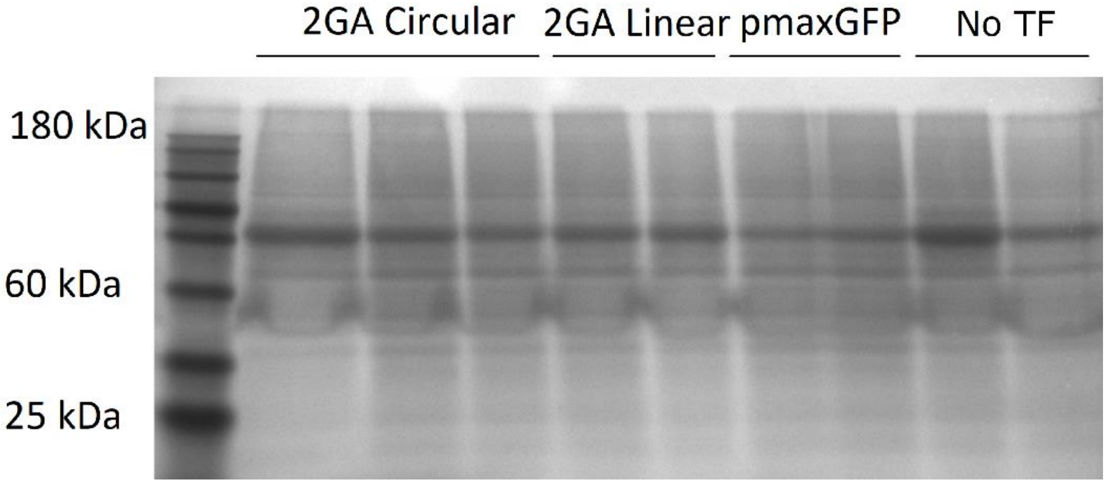


Figure 4.3 – **SDS-PAGE and zinc counterstain of total protein from cells transfected with p2GA, linear RNA-producing plasmid or pmaxGFP (using PEI reagent).** After transfection,

phenol-chloroform protein extraction and electrophoresis a $ZnCl_2$ – imidazole counterstain was performed. 10 µg total protein was loaded per well.

Therefore, another type of protein staining was attempted – silver staining. This method is much more laborious but its sensitivity is much higher than that of zinc-imidazole counterstaining. Again, the same samples were loaded on a gel and SDS-PAGE was performed. Silver staining was carried out and the results are presented in Figure 4.4. Again, no major difference was observed between any of the transfection groups. It is possible that not enough recombinant spidroin was expressed to be visualised in this stain. It is also possible that the amount of spidroin was small but that it is being overshadowed by the rest of the protein from the samples. Further investigation, therefore, needed to be carried out with a more specific method such as western blotting.

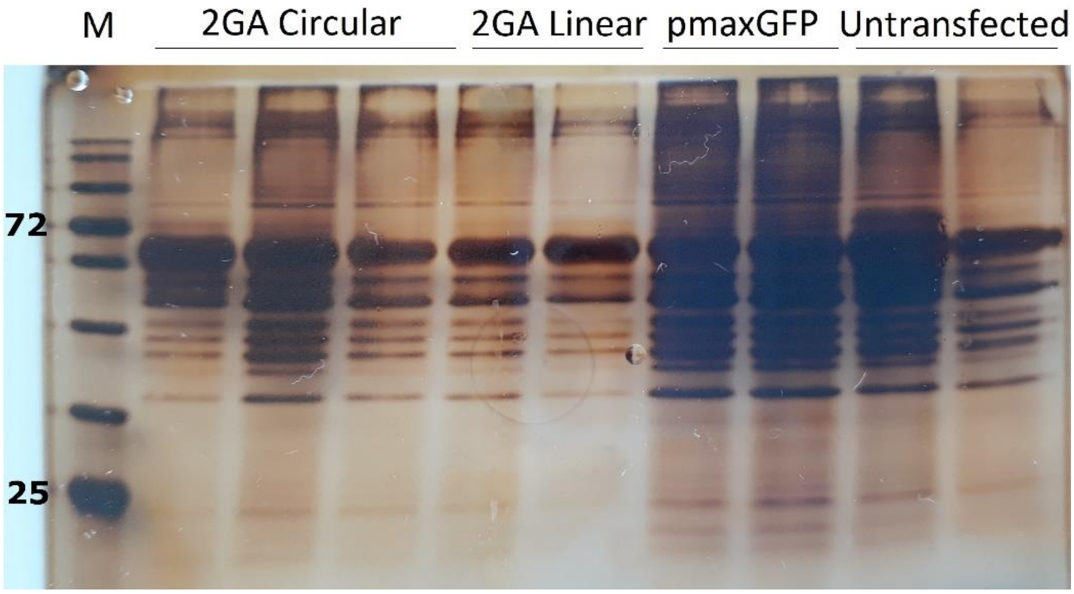


Figure 4.4 – **SDS-PAGE and silver stain of total protein from cells transfected with p2GA, 2GA linear RNA-producing plasmid or pmaxGFP**. After PEI transfection and phenol-chloroform protein extraction, 10 µg of protein was loaded per well. SDS-PAGE followed by a silver stain was carried out with the gel. The left-hand side labels refer to protein marker sizes in kDa.

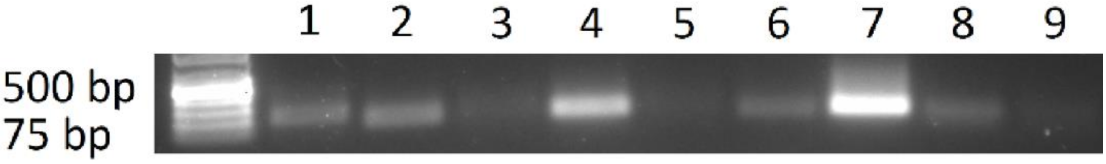
4.4 Initial Expression of pSP6HIS4GA in HEK293 cells

Next, HEK293 cells were transfected with spidroin plasmids. These cells have been commonly used in recombinant protein expression and so may be able to improve the recombinant spidroin yield. Due to the inability to detect expression from non-affinity-tagged spidroins with the gel staining approaches, only the pSP6HIS4GA and pSPFLAG4GA- plasmids were used for the remaining subsequent experiments.

Transfection was performed using the PEI MAX reagent again and RNA and protein were harvested after 72 hours. Phenol-chloroform extraction was used again and the total protein from both cell lysates and conditioned media were purified and resuspended in 4 M urea.

RNA circularisation was assessed again with the RT-PCR method. All RNA samples were treated with RNase R to degrade any linear RNA. The results are presented in Figure 4.5 below. These results are representative of several different transfections that will be outlined in this section. Lanes 1 and 5 show that a signal is observed from divergent primer amplification only after transfection with the circRNA-producing plasmid. Lanes 2 and 6 show that a spidroin-containing construct has been transfected into the cell. Linear RNA had already been degraded by RNase R but this does not affect the amplification with convergent primers from circRNA (lane 2) or from any contaminating DNA plasmid (lane 6). Lane 3 shows a negative control which demonstrates that reverse transcription is needed to be able to produce the signal observed in lane 2 and that it had not been amplified from genomic DNA. Lanes 4 and 7 demonstrate that an endogenous circRNA does produce a signal after reverse transcription but much less without it – lane 8 shows only a fraction of the intensity. The signal in lane 8 did not appear for every RT-

PCR and may have been caused by some previously undescribed backsplicing-generated pseudogene within genomic DNA. Nonetheless, Sanger sequencing of cDNA derived from pSP6HIS4GA-transfection RNA again confirmed that these samples contained the expected nucleotide sequences where the backsplice junction can be seen - the 3′ end of the spidroin encoding sequence is located immediately upstream of the 5′ end of the m^6A ribosome recruitment motif.



1 – pSP6HIS4GA transfection, +rev. transcriptase (RT), divergent primers

2 – pSP6HIS4GA transfection, +RT, convergent primers

3 – pSP6HIS4GA transfection, -RT, divergent primers

4 – pSP6HIS4GA transfection, +RT, circHIPK3 primers

5 – pSPFLAG4GA- transfection, +RT, divergent primers

6 – pSPFLAG4GA- transfection, +RT, convergent primers

7 – pSPFLAG4GA- transfection, +RT, circHIPK3 primers

8 – pSP6HIS4GA transfection, -RT, circHIPK3 primers

9 – No template, divergent primers

Figure 4.5 – **Agarose electrophoresis of RT-PCR for circRNA circularisation.** After transfection with pSP6HIS4GA or pSPFLAG4GA-, total RNA was harvested by phenol-chloroform separation. RT-PCR was performed and the details of each sample

are outlined in the legend above. 1% agarose gels were cast and 20 µl reactions were loaded per lane. cDNA volumes equivalent to 50 ng of total RNA were used as templates for these PCRs. The negative control sample contained no template DNA.

Western blot analysis was performed on the aforementioned protein samples. A composite western blot image is presented in Figure 4.6. 10 mg of total protein per lane were loaded. In every lane, clear bands can be seen at ~25, 30 and 40 kDa. Since they are also represented in the untransfected control lane, it can be assumed that they are not the result of spidroin translation. There is however a small cluster of bands at ~72 kDa. Strangely, this is only seen in the lysate lane and not in the transfected, conditioned media samples. It may be the result of a spidroin translation that was not successfully exported out into the surrounding media. However, it is also likely that this protein is not recombinant but intracellular and not exported. Such non-specific signal was not observed with the positive control anti-TUBB.

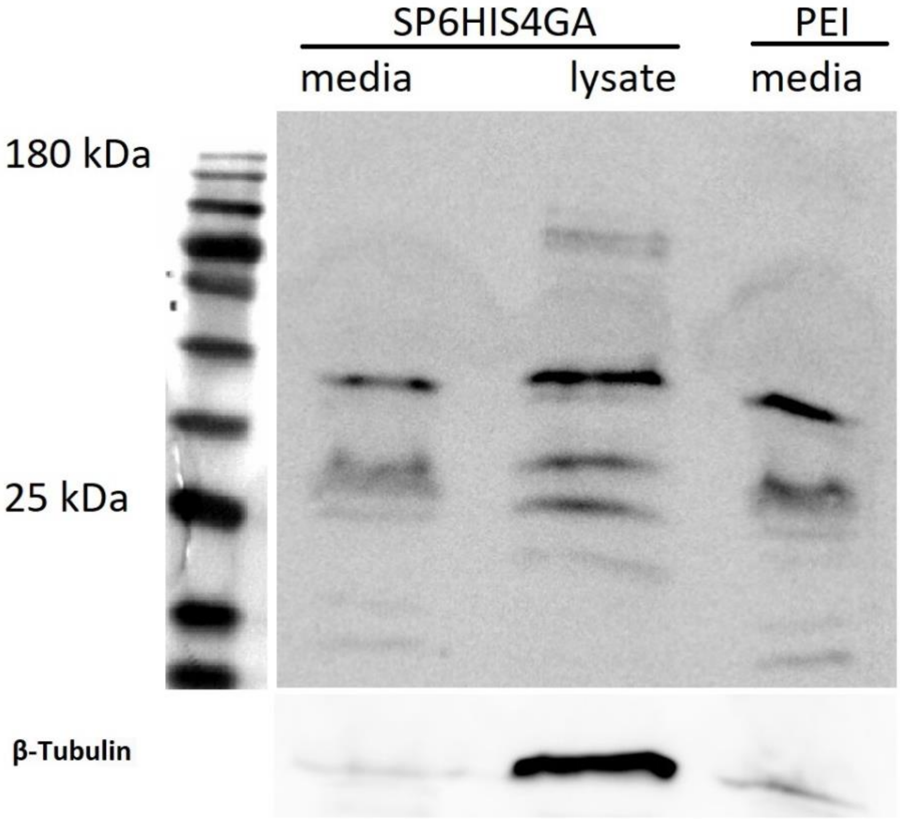


Figure 4.6 – **Western blot of protein from pSP6HIS4GA-transfected HEK293 cells.** After PEI transfection, phenol-chloroform purification and SDS-PAGE a western blot was performed. The incubation with anti-HIS-tag antibody was at 4^{o}C overnight. 10 mg of total protein was added per lane.

The observed bands may have been caused by endogenous histidine-rich proteins which are expressed in HEK293 cells. Such proteins could bind the anti-HIS-tag antibody and produce such signals (Mahmood and Xie, 2015). To determine this, another western blot was attempted with SP6HIS4GA. This time the incubation with the anti-HIS-tag antibody was reduced to 1.5h at room temperature based on a recommendation by the manufacturer. Additionally, conditioned media samples were collected at 24h intervals to determine if protein degradation after translation was responsible for the absent spidroin signal. The results are shown in Figure 4.7 below.

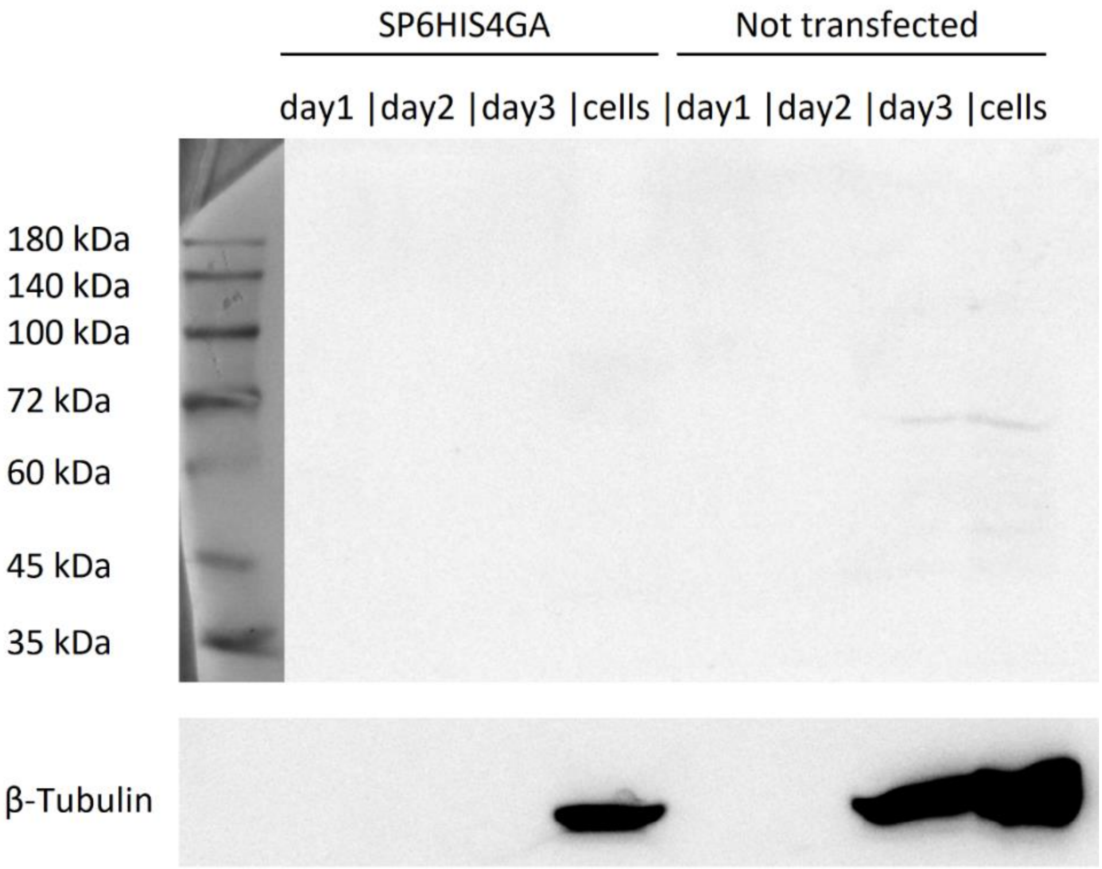


Figure 4.7 **– Second western blot of SP6HIS4GA transfection protein.** pSP6HIS4GA was transfected into HEK293 cells. Non-transfected control cells have only been treated with PEI. Conditioned media samples were collected at 24h intervals. All protein (from conditioned media or lysed cells) was harvested by phenol-chloroform separation. 10 mg of total protein was loaded

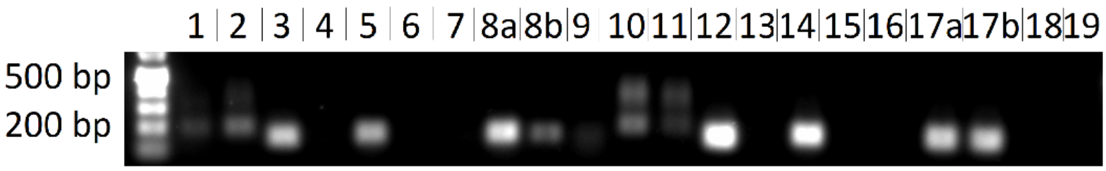
per well and SDS-PAGE was carried out. During western blot, protein was probed with anti-HIS-tag and anti-TUBB primary antibodies.

Some leftover signal can be seen in the last two lanes but overall it appears that the shorter duration of antibody incubation has removed most of the non-specific binding. Again, no spidroin expression can be observed. The anti-TUBB bands of five of the conditioned media samples are not visible. However, Ponceau S staining did confirm that protein was present in this sample. The amount of TUBB in the media may have been too diluted to produce an adequate signal, however. TUBB is not exported but rather most likely finds its way into the conditioned media during the death and lysis of some of the transfected cells. The results from the lanes with missing anti-TUBB bands are unreliable and the western blot needed to be repeated.

4.5 Expression of pSP6HIS4GA in COS-1 and NIH/3T3 cells

Based on the previous inability to detect a successfully expressed recombinant spidroin from circRNA, a new set of experiments was planned. To assess the possibility that some trait of HEK293 cell biology is interfering with the recombinant expression, new transfections were attempted in murine NIH/3T3 fibroblasts and monkey kidney COS-1 cells. Also, the previous transfection via PEI may have been inefficient and so a new transfection reagent was used – FuGENE 6. The pSP6HIS4GA plasmid was used alongside mock FuGENE 6-only transfection.

Again, the conditioned media was harvested and the cells were lysed to obtain RNA and protein. After, purification, all RNA samples were treated with RNase R to digest any linear RNA. RT-PCR was performed to assess RNA circularisation. The results can be seen in Figure 4.8. Lanes 1, 2, 10 and 11 show a band at ~200 bps. Its identity is most likely the amplicon from the divergent primers on the circRNA (expected size – 183 bps). An additional extra band can be seen in these samples at ~350 bps. This cannot be the result of trans-splicing since such a linear RNA would not be retained in the RNase R treated samples. Lanes 3 and 12 show a band at ~130 bps which is to be expected from the circRNA amplified with convergent primers (expected size – 123 bps). Lanes 4 and 13 show no bands which is to be expected if the circRNA had not been reverse transcribed. This shows that no contaminating gDNA is capable of producing an amplicon with the divergent primers. Lanes 5 and 14 show the amplification of the circHIPK3 circRNA (expected size ~130 bps). This is a positive control and shows that this RT-PCR is capable of detecting a circRNA molecule. Lanes 6 and 15 show that if the pSP6HIS4GA plasmid is missing there is no spidroin circRNA made. Lanes 7 and 16 similarly show that without pSP6HIS4GA no recombinant spidroin RNA is transcribed. Lanes 8 and 17 show that the circHIPK3 RNA is endogenous and expressed in these cell lines. Lanes 9 and 18 show that no reverse transcription results in no observable circHIPK3 signal so DNA contamination does not contribute to the signal from 8 and 17. Lane 19 shows that none of the bands observed are a result of reagent contamination. Taken together, these RT-PCRs show that the pSP6HIS4GA transfection is needed to generate spidroin circRNA in the COS-1 and NIH/3T3 cell lines. Sanger sequencing of the suspected spidroin circRNA again supported this conclusion as the 3′ end of the spidroin encoding sequence was found located upstream of the 5′ end of the m6A ribosome recruitment motif. This is to be expected in the case of successful RNA backsplicing.



1 – 3T3 lysate, pSP6HIS4GA transfection, +rev. transcriptase (RT), divergent primers

2 – 3T3 lysate, pSP6HIS4GA transfection, +RT, divergent primers

3 – 3T3 lysate, pSP6HIS4GA transfection, +RT, convergent primers

4 – 3T3 lysate, pSP6HIS4GA transfection, -RT, divergent primers

5 – 3T3 lysate, pSP6HIS4GA transfection, +RT, circHIPK3 primers

6 – 3T3 lysate, mock transfection, +RT, divergent primers

7 – 3T3 lysate, mock transfection, +RT, convergent primers

8a and b - 3T3 lysate, mock transfection, circHIPK3 primers

9 – 3T3 lysate, pSP6HIS4GA transfection, -RT, circHIPK3 primers

10 – COS-1 lysate, pSP6HIS4GA transfection, +RT, divergent primers

11 – COS-1 lysate, pSP6HIS4GA transfection, +RT, divergent primers

12 - COS-1 lysate, pSP6HIS4GA transfection, +RT, convergent primers

13 - COS-1 lysate, pSP6HIS4GA transfection, -RT, divergent primers

14 - COS-1 lysate, pSP6HIS4GA transfection, +RT, circHIPK3 primers

15 - COS-1 lysate, mock transfection, +RT, divergent primers

16 - COS-1 lysate, mock transfection, +RT, convergent primers

17a and b - COS-1 lysate, mock transfection, +RT, circHIPK3 primers

18 - COS-1 lysate, pSP6HIS4GA transfection, -RT, circHIPK3 primers

19 - no template, divergent primers

Figure 4.8 – **Agarose electrophoresis of RT-PCR assay for RNA circularisation.** Transfection of COS-1 and NIH/3T3 cells with pSP6HIS4GA or mock transfection was performed with FuGENE 6. Total RNA was harvested from cell lysates and cDNA was synthesised according to the legend outlined above. 20 µl PCRs were prepared and after amplification, these were loaded onto a 1%

agarose gel. cDNA from 50 ng of RNA was used as the template. No DNA template was added to the negative control in sample 19.

After, RNA circularisation had been confirmed the assessment of recombinant protein could begin. The methods for protein extraction were changed before the subsequent western blot. The conditioned media samples were condensed with Amicon centrifugal filters in order to obtain more concentrated protein samples. This would presumably increase the amount of TUBB per lane at the potential cost of overloading the polyacrylamide gel with protein. The preparation of total protein from cell lysates was performed with the RIPA buffer at ambient conditions to reduce any potential formation of hydrophobic spidroin bodies. In the previous experiments, protein purification may have caused an irreversible denaturation of the recombinant spidroin. Spidroins are known to be difficult to resolubilise after denaturation due to the formation of hydrophobic protein cores. RIPA buffer was subsequently chosen to dilute cell lysates. This may result in an increased amount of recombinant spidroin resolving during SDS-PAGE. The results are presented in Figure 4.9.

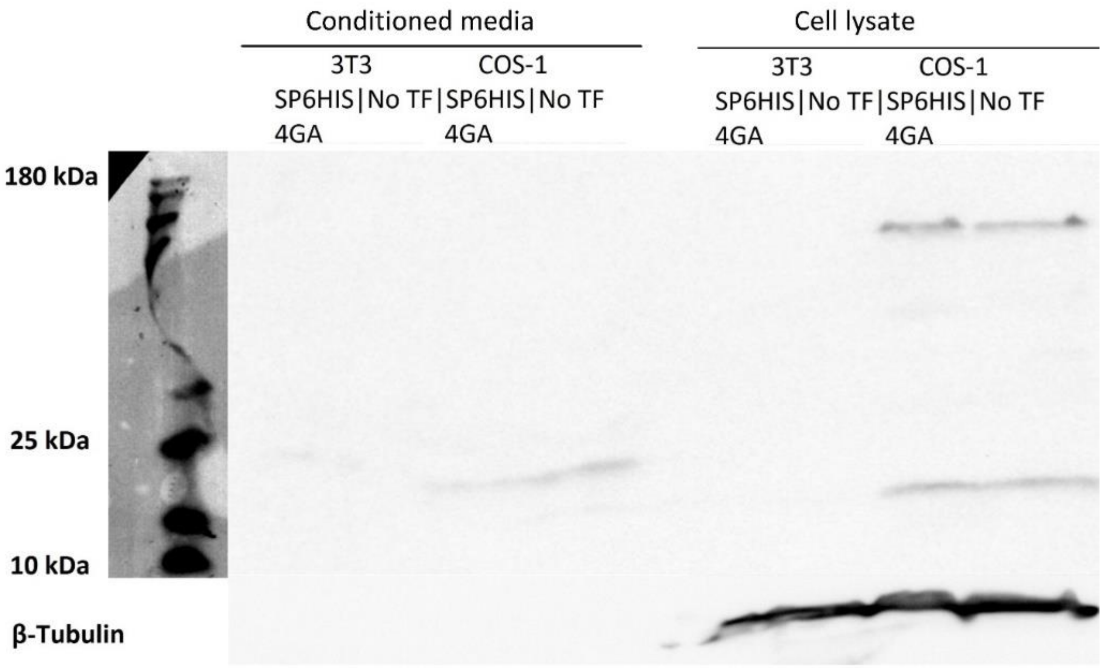


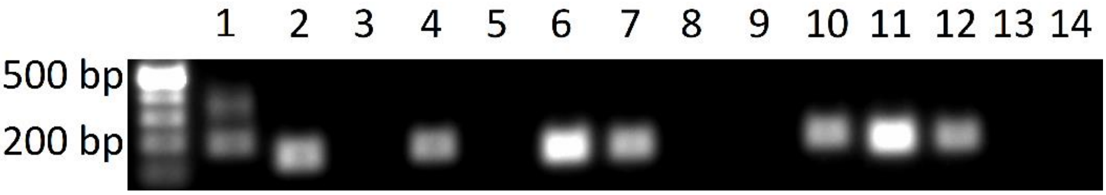
Figure 4.9 – **Western blot of protein from pSP6HIS4GA-transfected COS-1 and NIH/3T3 cells.** Transfection was performed with FuGENE 6. Conditioned media was condensed by centrifugal filtration. Cell lysates were diluted in RIPA buffer and all samples were run on an SDS-PAGE. 10 µl of each sample were loaded per well. Subsequent western blotting was performed with the anti-HIS-tag primary antibody at room temperature for 1 hour.

Again, no differences can be observed within the pairs of transfected and untransfected samples. Some bands did appear in the case of transfected and untransfected COS-1 cells and media. These can be ascribed to endogenous COS-1 proteins which non-specifically bind the anti-HIS-tag antibody. The conditioned media samples did not show any TUBB signal. This could have been caused again by low amounts of TUBB present in the lysates. These conditioned media results are thus unreliable. Large amounts of protein were loaded, however. This is what must have led to the displacement of several of the protein marker bands. This was also reflected in the Ponceau S staining. Overall, this transfection and western blot could not support the recombinant expression of circularly-translated spidroins.

4.6 Final Expression of pSP6HIS4GA and pSPFLAG4GA- in HEK293 cells

Since previous attempts at recombinant spidroin expression appeared unsuccessful a new experiment was attempted where HEK293 cells were used again. Since the COS-1 and NIH/3T3 cells were unable to express any noticeable amounts of spidroin a different approach was chosen for protein extraction and SDS-PAGE. This will be described later in this subsection. HEK293 cells were again chosen for transient recombinant expression. Transfection with FuGENE 6 and the pSP6HIS4GA and pSPFLAG4GA- or mock transfection was carried out.

After 48 h, total RNA was harvested via phenol-chloroform phase separation and RT-PCR was used to validate spidroin circRNA formation. RNase R digestion was performed for most samples except for one as will be described shortly. The results are available in Figure 4.10. Lanes 1, 8 and 13 demonstrate that the expected spidroin circRNA amplification signal can only be observed if the cells are transfected with the pSP6HIS4GA construct and not with the linear RNA-producing one or without transfection. Lanes 2 and 9 show that only the circRNA is capable of producing a signal with convergent primer amplification in these conditions. Most likely the recombinant spidroin linear RNA has been degraded by RNase R and thus is unavailable for reverse transcription. Lane 3 demonstrates again that reverse transcription is needed to produce the signal in lane 1 – i.e. that the signal is produced by RNA and not any contaminating DNA. Lanes 4, 5 and 10 show that the endogenous circRNA circHIPK3 also behaves in a similar way to the spidroin circRNA. More specifically, it too is observed in RNase R treated and reverse transcribed samples but not when RT is omitted. Lanes 6, 7, 11 and 12 demonstrate that the endogenous GAPDH DNA and possibly RNA are present within the sample. Lane 14 shows that none of the signals observed here has been caused by contamination of reagents. Overall, these results show that again, circRNA is produced from the RCI-containing construct and not from the cells themselves or alternative plasmid constructs.



1 – pSP6HIS4GA transfection, +rev. transcriptase (RT), divergent primers

2 – pSP6HIS4GA transfection, +RT, convergent primers

3 – pSP6HIS4GA transfection, -RT, divergent primers

4 – pSP6HIS4GA transfection, +RT, circHIPK3 primers

5 – pSP6HIS4GA transfection, -RT, circHIPK3 primers

6 – pSP6HIS4GA transfection, +RT, -RNase R, GAPDH primers

7 – pSP6HIS4GA transfection, +RT, GAPDH primers

8 – pSPFLAG4GA- transfection, +RT, divergent primers

9 – pSPFLAG4GA- transfection, +RT, convergent primers

10 – pSPFLAG4GA- transfection, +RT, circHIPK3 primers

11 – pSPFLAG4GA- transfection, +RT, - RNase R, GAPDH primers

12 – pSPFLAG4GA- transfection, +RT, GAPDH primers

13 – Mock transfection, +RT, divergent primers

14 – No template, divergent primers

Figure 4.10 **– Agarose gel of RT-PCR assessment of circRNA formation.** All samples were treated with RNase R unless otherwise stated (lane 6). HEK293 cell transfection was carried out with FuGENE 6 and pSP6HIS4GA, pSPFLAG4GA- or no plasmid (mock transfection). RNA was isolated form lysed cells via phenol-chloroform separation. RT-PCRs were performed as detailed in the legend above. The cDNA that had been produced from ~50 ng of total RNA was used in 20 µl PCR for gel electrophoresis (1% agarose)**.** The sample in lane 14 was prepared without any template DNA.

After the presence of circular spidroin RNA was confirmed, identification of the recombinant protein could be attempted. The secreted protein samples were collected as conditioned media, concentrated with the Amicon centrifugal column at 4^{o}C then suspended in Laemmli loading buffer directly. This would presumably stop any protease degradation of protein and potentially preserve the recombinant spidroin. Additionally, it is possible that in previous experiments the recombinant protein was too large to migrate through the polyacrylamide gel. This is unlikely since previous research on RCT has shown that part of the recombinant protein fraction is made up of shorter peptide chains that have been produced by a smaller number of translation cycles. These shorter peptides should have been able to migrate through the gel in previous experiments. Nonetheless, a lower percentage polyacrylamide gel (8%) was used before this final western blot in order to answer this question. After SDS-PAGE a western blot was performed. The results are available in Figure 4.11.

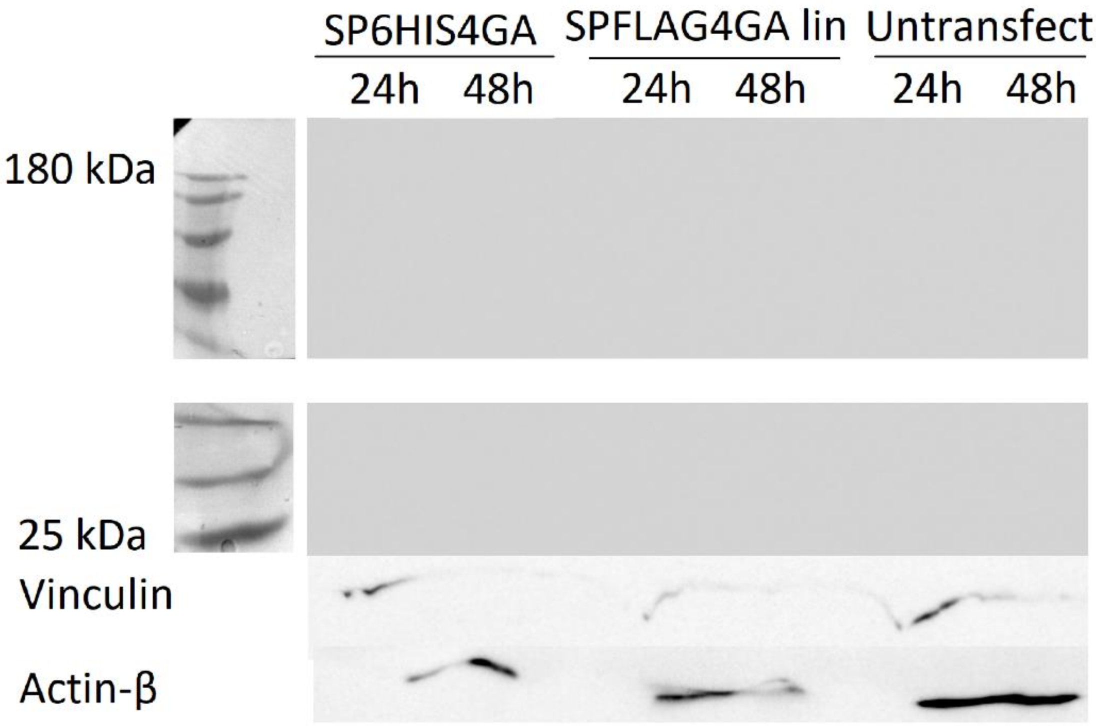


Figure 4.11 – **Western blot of secreted protein from HEK293 cells transfected with SP6HIS4GA and pSPFLAG4GA**. Conditioned media samples were obtained at 24 h intervals and condensed 6.66 times using the Amicon centrifugal filters. Immediately, Laemmli buffer was added to the concentrated protein. After an SDS-PAGE, a western blot was performed with anti-vinculin and anti-HIS-tag primary antibodies (top section) or anti-actin-β and anti-FLAG-tag (bottom section). The negative control was HEK293s, mock transfected with FuGENE 6 only.

The western blot results demonstrate no observable recombinant spidroin signal. Even though the protein was left to migrate through a less dense gel, no spidroin was able to be visualised. Most likely, difficulties in migration are not the cause of the observed result. It also appears that the change in protein extraction was not enough to produce a noticeable signal. It appears that the recombinant spidroin is not expressed in quantities sufficient enough to produce a western blot signal. Therefore, other explanations need to be analysed that could explain the observed lack of spidroin translation.

4.7 Detection of Ribosomal Frameshifting Signals in the circRNA Sequences

There may be specific translational events that might be causing the observed lack of translation or degradation of the recombinant spidroin. One possibility is that ribosome frameshifting events are causing translational termination. It is unknown how likely such an event is but is still worth investigating.

The software Knotinframe was specifically designed by Theis *et al*. (2008) to identify RNA features which cause -1 frameshifts. Specifically, -1 frameshifts are when the ribosome translates a codon and then moves back by one nucleotide to translate the first nucleotide of the same codon as the start of the subsequent codon for translation. For example in the AAATTTG heptamer, canonical translation may proceed from the AAA triplet to the TTT triplet.

However programmed frameshifting may cause the backward slippage to proceed from the AAA triplet to the AAT one.

Knotinframe was used to specifically search for potential primary and secondary sequence features that may be causing -1 frameshifting in the designed spidroin circRNAs. The software was used to identify such sequences in the entire recombinant spidroin construct library described in Chapter 3. A programmed -1 frameshift signal was found only in the SP6HIS4GA sequence. The results of this can be seen in Figure 4.12.

Sequence Name: SP6HIS4GA

Slippery sequence: C CCT TTT

Nucleotide Sequence (circular):
5′-AAGGATCCTGGACTAAAGCGGACTTGTCTCGAGATGAATCCCCTTT
TGA**TCCT**CACATTC**GTCGCTGC**CGCA**CTCG**C**ACATCACC**ATCACCAT
CACCAA**GGTGGGT**A**CGGG**CA**AGGGA****CGG**GCTCTAGTA**CTG**CA**GCA
GCGGC**CGCAGCTGCCGCTGCTGCTGCTGCCGGACAGGGTGGCCAG
GGTGGCTACGGTGGTCTCGGTCAGGGTGGATATGGACAAGGGGCCG
GTAGTAGTGCAGCCGCCGCTGCTGCCGCTGCAGCTGCTGCAGCAGC
AGGTCAAGGCGGTCAAGGTGGTTATGGTGGCCTGGGACAAGGTGGA
TATGGCCAGGGCGCTGGGTCAAGTGCTGCAGCTGCGGCTGCCGCGG
CCGCAGCTGCGGCAAGCGGACAAGGCGGTCAAGGTGGACAAGGCCA
AGGTGGATACGGACAAGGCGCAGGAATAAGTGCCGCCGCCGCCGCA
GCTGCAGCAGCCGCCGCTGCG-3′

Figure 4.12 – **The predicted -1 frameshift site in the SP6HIS4GA sequence**. The SP6HIS4GA sequence was analysed with the Knotinframe software. The circRNA sequence depicted

contains a ribosome slippage motif (C CCT TTT) and downstream RNA secondary structures (colour-coded complementary sequences in gold and shades of blue). After the slippage, a downstream stop codon (underlined) is now in frame.

A single, very strong frameshift signal was identified in the spidroin circRNA. Knotinframe specifically searches for a slippery site which is where the frameshift occurs and has the consensus sequence X XXY YYZ (here two distinct homotrimers are followed by a different nucleotide). The slippery site identified in the circular spidroin sequence was C CCT TTT. If -1 frameshifting were to occur here it would result in the CCT triplet being followed by a TTT (followed by the last T in the heptamer) triplet. The software also searches for a pseudoknot RNA secondary structure formed by two stem-loops (based on sequence information). Again this was identified in the region downstream of the slippery site. This RNA context allows a slippage that ultimately leads to a stop codon just three nucleotides later.

However, the programmed frameshifting signal that was identified could only act on a reading frame separate from the IORF. The planned translation initiation of the spidroin circRNA (Subsection 3.2) most likely has entirely excluded the opportunity for this programmed -1 frameshift. Knotinframe, therefore, did not show any signals that should affect the IORF of the circRNA. Other factors had to be investigated that could explain the lack of spidroin translation.

4.8 tRNA Charging in Spidroin-expressing Versus Non-expressing Cells

Due to the missing spidroin expression, an alternative potential mechanism of translation inhibition was investigated – depletion of amino acyl-tRNA levels. The tRNA charging assay was used to quantify the charged versus total tRNA in spidroin gene transfected cells and empty vector controls.

This assay employs a strategy whereby tRNA molecules are differentially oxidised depending on the presence of a charged amino acid. Amino-acylation protects the 3′ end of the tRNA which enables the enzymatic ligation of a DNA oligomer (after the removal of the amino acid). This DNA component can then be used to only amplify previously charged tRNA in an RT-PCR. The addition of a control sample (non-oxidised) allows total tRNA to be likewise ligated to DNA. A ratio can then be calculated of charged/total tRNA using a qPCR step.

For this final step, it is important to design a set of DNA primers that will accurately quantify the tRNA charging. Typically a spike-in phenylalanine control primer is used in conjunction with yeast-derived phenylalanine tRNA that is added to the total tRNA samples beforehand (Pavlova et al., 2020; Roberts et al., 2014). This serves as an internal control that allows the calibration of qPCR results so that experimental variability between samples does not affect final quantities. However, the yeast-derived phenylalanine tRNA reagent was not obtainable so human GAPDH expression was chosen as an internal control instead (measured with GAPDH primers in qPCR). The already available primer set for ArgACG tRNA was used as a control to monitor the charging of an amino acid not included in the recombinant spidroin (Table 2.2). Two additional primer sets were designed (AlaAGC and GlyGCC) with assistance from Natalya Pavlova from the Memorial Sloan Kettering Cancer Center, USA (Table 2.2). Each primer set was designed to bind the consensus sequence of a set of tRNAs (e. g. the AlaAGC primer set could anneal to all tRNA forms which contained the AGC anticodon) and to a ligated DNA adapter.

The GlyGCC and AlaAGC tRNA isodecoders were chosen to monitor the levels of alanine and glycine tRNA charging because these two are the most common amino acids in the expected recombinant spidroin. Also, another control primer set was used – ArgACG. Arginine is not present in the recombinant spidroin so arginyl-tRNA charging is expected not to be directly affected by the transfection.

The experiment was performed as outlined in section 2.3.14. After the qPCR step was carried out, tRNA charging ratios were calculated for each pair of samples. tRNA charging ratios above 1.00 were excluded from this analysis because the charged tRNA fraction cannot ever exceed the total tRNA levels. Ratios above 1.00 are theoretically impossible and potentially are the result of experimental error. The remaining data were compiled into a bar chart in Figure 4.13. The tRNA charging of AlaAGC appeared more pronounced in backbone-transfected samples relative to spidroin-expressing ones. Very little difference was observed for the other isodecoders which was expected for ArgACG but not GlyGCC. A statistical test (One-Way ANOVA) was carried out for each primer set to determine if the means of the different charging ratios are significantly different. None of the groups measured showed a difference whose p-value was below 0.05. Therefore these differences in tRNA charging do not appear to be statistically significant.

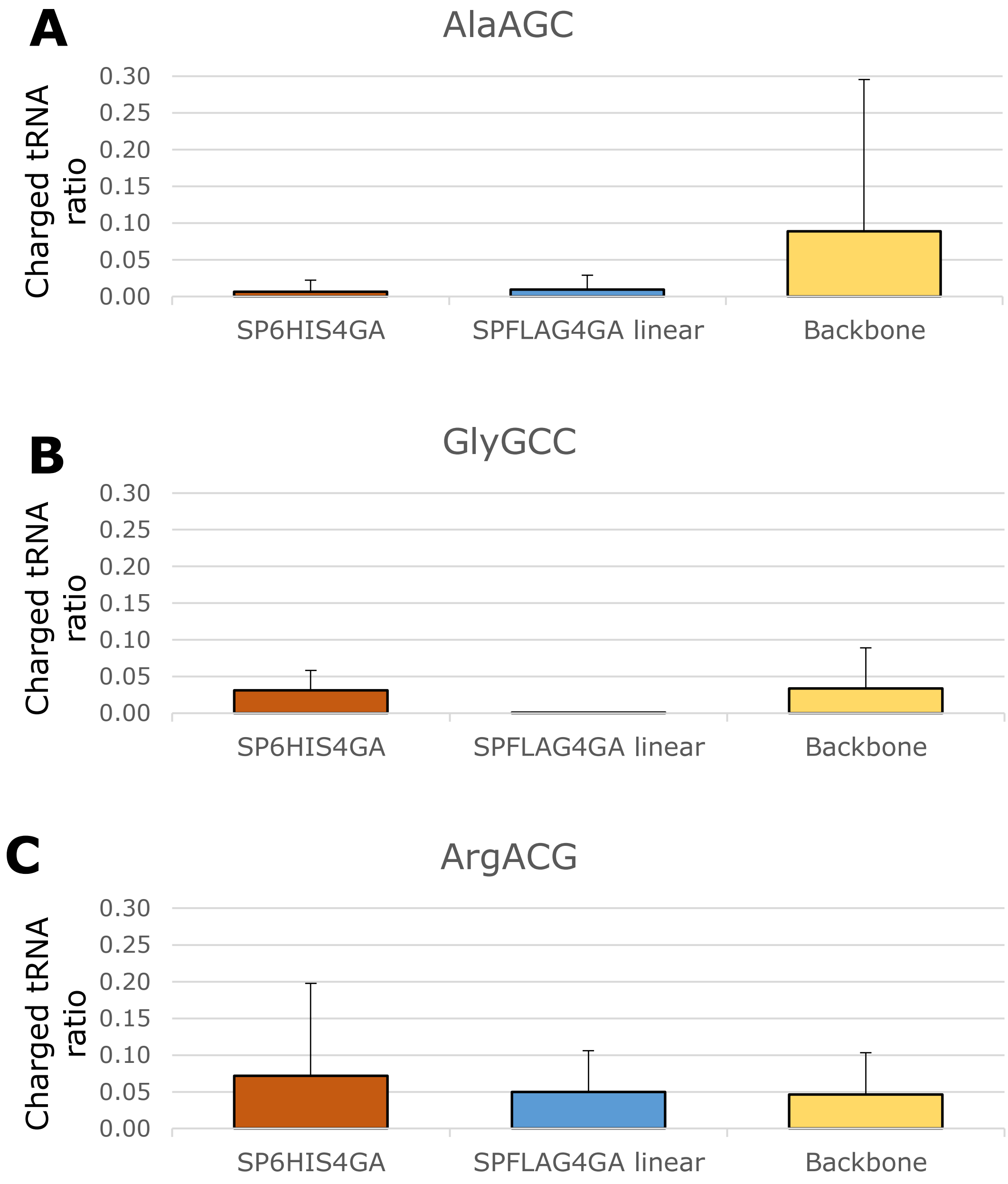


Figure 4.13 – **tRNA charging ratios in spidroin-expressing and empty vector-transfected HEK293 cells.** HEK293s were transfected with either SP6HIS4GA (red), SPFLAG4GA linear (blue) or the empty pcDNA3.1 vector (backbone, in yellow). After 48h the RNA of each sample was harvested with Trizol and Glycoblue coprecipitant. The assay was performed and tRNA charging of the indicated isodecoder groups was measured via qPCR. Primers

specific to AlaAGC (A), GlyGCC (B) and ArgACG (C) were used. The last three letters denote the anticodons targeted for each isodecoder. Ratios were calculated by dividing the relative amount of each qPCR-detected charged tRNA over the total amount of tRNA. Data are shown as mean ± SD of N = 3 biological replicates (samples that were cultured in plates, treated with plasmid and assayed independently).

4.9 Discussion

Overall, the results presented here could only confirm that circRNA has been produced. No recombinant spidroin was observed during any of the SDS-PAGE experiments. The reason for this lack of expression was further explored by the tRNA charging assay and identification of ribosome frameshifting sites. This analysis could not provide a conclusive answer for the lack of expressed spidroin but it did rule out specific tRNA aminoacyl depletion and certain ribosome frameshifting signals as the culprit.

Another factor in the lack of observed recombinant spidroins may have been low copy numbers of circRNA. If for some reason insufficient circRNA were produced it still would be detected through the sensitive RT-PCR approach. The presence of circRNA does not necessarily ensure adequate amounts of RNA. It is unknown what amount of circRNA is needed to produce detectable recombinant protein.

Several plasmids from the spidroin construct library were left unused and unexplored. These were ones that did not possess a spidroin affinity tag because they were specifically intended for identification through zinc staining and silver staining. In the future, if the spidroin RCT can occur, the tag-less constructs may be useful for their intended purpose – identifying how the coding sequence affects RCT and the recombinant spidroin. For example, the difference between 2GA and 4GA sequences was to investigate if more repeats per cycle would produce a spidroin with properties more similar to the natural ones.

When protein affinity tags are unavailable, spidroin-specific antibodies may be of use. However, these are typically designed to bind spidroin domains different from MaSp (Cadle, 2016; Huang et al., 2017). In the specific case of these plasmids, a MaSp-sensitive antibody would need to be available.

The method of zinc-imidazole gel staining was initially chosen because it is very quick and simple to implement while also being cheap and sensitive. It is sensitive enough to detect a protein band consisting of ~7 ng protein (Hardy et al., 1996). This is similar to the sensitivity of Coomassie staining (Gauci et al., 2013). Silver staining is much more sensitive – 0.25-0.5 ng bands can be typically detected (Rabilloud, 1992). However, this method again could not detect and resolve any recombinant spidroin. If spidroin yields are eventually optimised, the recombinant spidroin may be of sufficient quantities for zinc-staining or silver staining to be successful.

An alternative limiting factor in spidroin translation may have been a low plasmid transfection efficiency. It is unknown what efficiency the IORF translation will benefit most from but presumably, higher efficiency is preferable. However, a complete lack of transfection was not observed as circRNA was present within the host cells for at least 72h. Even at low levels, this circRNA was expected to be capable of translating at least some protein which would be detectable through the sensitive western blot technique (Canelle et al., 2005). Most likely there is an additional factor that has reduced or fully inhibited spidroin translation in this context post-transcriptionally. There may also be a compounding effect between different causes of reduced recombinant expression.

A lack of nuclear export for the spidroin circRNAs could explain the observed absence of protein and the presence of circRNA. However, the lack of understanding of the m^{6}A-related export mechanism precludes any predictions of export efficiency (Chen et al., 2019; Di Timoteo et al., 2020). Experiments that investigate circRNA localisation within a cell may shed light on this question.

The -1 ribosome frameshifting analysis was inherently fragmentary. Knotinframe is only able to detect part of the -1 events and these only represent a fraction of the total possible ribosomal frameshifts. A separate ribosomal frameshifting event may be responsible for the observed lack of spidroin translation but this may not be detectible by Knotinframe. Not many such prediction tools exist and even then such tools might overlook certain frameshifts. It would be more reliable to monitor frameshifting experimentally for example by dual-luciferase assays (Kelly et al., 2020a).

It is unknown what percentage of designed IORFs are affected by ribosomal frameshifting events so their processes need to be well understood. In one previous study, an observation was made that IORF translation terminates without a stop codon present after just one or two cycles (Mo et al., 2019). If something similar has occurred in the spidroin RCT, this would explain why no recombinant protein could be observed**.** More robust methods to detect this frameshifting will be invaluable in identifying potential issues in IORFs. These could include computational as well as biomolecular assays.

Other factors such as various translation stresses and stress responses could also explain the observed lack of recombinant spidroin. The mechanisms of circRNA degradation have not been studied in depth. It has been reported that an m^{6}A-dependent RNA degradation pathway exists but this also requires nucleotide sites for the binding of an endoribonuclease and a stabilising protein (Park et al., 2019). Such sites are absent from the spidroin-generating circRNA described here. Otherwise, an alternative mechanism degrades global circRNA after the cell recognises it has been invaded by a virus (Liu et al., 2019a). However, FUGENE 6 (which is non-viral) and plasmid DNA transfection is unlikely to trigger this response (Chong et al., 2021). Also, the RT-PCR RNA circularisation results strongly suggest that the expected circRNA is still present at least 72h after cell transfection. Therefore, it is unlikely that circRNA degradation has caused the lack of spidroin translation.

A molecular mechanism recognises stalled ribosomes and degrades the associated mRNA and the associated incomplete nascent peptide (termed no-go-decay - NGD). This occurs to prevent the propagation of potential deleterious peptides and to stop defective mRNAs from causing future ribosome stalling (Brandman et al., 2012; Choe et al., 2016; Ishimura et al., 2014). This process may have directly caused the degradation of the nascent recombinant spidroin and this may explain the observed absence of recombinant spidroin. There are several ways that ribosome stalling may occur.

Depletion of glycyl- and alanyl-tRNA complexes is known to be a major obstacle in recombinant spidroin production. This has been extensively reported in previous research and has been addressed successfully in *E. coli* (Cao et al., 2017). While several approaches have been implemented successfully before to engineer *E. coli* amino-acyl-tRNA levels, this approach is uncommon in mammalian cells. There is a high probability that amino-acyl-tRNA depletion is responsible for ribosomal stalling and premature translation termination which could have resulted in NGD and the low recombinant protein yields. The results of the tRNA charging assay were inconclusive since transfection did not appear to significantly affect alanyl and glycyl-tRNA charging. However, the cells that were to perform spidroin RCT were potentially very stressed from transfection, nutrient depletion in media and the presence of cell debris. In this context, a small and statistically non-significant reduction in alanyl-tRNA may have contributed to the observed lack of repetitive spidroin. This could act as another compounding factor that has reduced recombinant expression. Even if some recombinant protein is still translated it is feasible that the amount of protein that is made before amino-acyl-tRNA depletion occurs is insufficient to produce a signal detectable on a western blot.

Additionally, in the case of depleted glutamine and overabundance of free tRNA^{Gln} within the cell, a stress response is triggered whereby translation is repressed in order to maintain stable levels of glutamine. This occurs because of glutamine's importance for metabolism but serves as an obstacle for transgenic expression (Pavlova et al., 2020). In the case of the SP6HIS4GA peptide, for example, glutamine comprises ~9% of the overall amino acid residues. This may have led to the aforementioned glutamine depletion stress and may have contributed to the difficulty in recombinant spidroin synthesis. Therefore, it is vital to attempt the tRNA charging assay in the future with glycyl-tRNA-specific primers in order to assess glycine tRNA charging.

Co-transfectional overexpression of $tRNA^{Gln}$, $tRNA^{Ala}$ and $tRNA^{Gly}$ (Huh et al., 2021) alongside supplementation with the corresponding amino acids in media could ultimately remedy amino acyl-tRNA depletion. Amino acid complementation alone (without the cotemporaneous tRNA overexpression) would likely not be sufficient to improve protein yields (Cao et al., 2017; Klein et al., 2015). The amino acids that cultured cells have access to include large quantities of alanine and glycine (among others). Therefore it seems likely that metabolic stress to produce these particular amino acids is not the main factor at play.

tRNA depletion can also trigger a separate stress response where ribosome-ribosome collisions are recognised by a distinct cellular pathway. This could result in different cell fates such as reduced translation and apoptosis (Wu et al., 2020). This may be another factor that has contributed to the lack of detectable recombinant spidroin.

The existence of several more alanyl-tRNA, arginyl-tRNA and glycyl-tRNA isodecoders means that the tRNA analysis carried out here is inherently fragmentary. However, the remaining tRNA isodecoders account for less than 40% of tRNA charging due to the known existence of codon preference. The tRNA isotypes investigated here still serve as a proxy for the overall tRNA charging within a cell (Pavlova et al., 2020). The codon usage in the designed spidroin constructs was originally optimised to reflect tRNA abundances in mice and humans (Chapter 2.4.2).

Translation termination of IORFs has not been studied before. The mechanism responsible for this has not been uncovered so far. It is feasible that after many rounds of translation some form of frameshifting will occur stochastically which would induce termination at an out-of-frame stop codon. However, this has not been demonstrated experimentally and various currently unknown mechanisms may be involved.

What is known, however, is that proteins produced through RCT possess a heterogeneity in peptide lengths. In cases where consistent and reproducible protein lengths are needed the approach may need to be altered. Proteolytic cleavage or co-translational self-cleavage at each cycle of translation has been used to produce homogenous-sized peptides (Costello et al., 2019; Lee et al., 2021).

Systems other than mammalian cell culture have a proven record of recombinant spidroin expression. These could be adapted to the

RCT approach. Perhaps this combination would provide an increase in recombinant yields.

B. mori's advantage is that its native silk fibroin protein has a similar amino acid content to the spidroins. Therefore, the silkworm's silk gland already has an amino-acyl-tRNA pool that matches the spidroin's requirements (Zhang et al., 2019b).

COS-1 cells grown in cold-stress conditions (33°C) can be used to increase the level of transcription from a CMV promoter. This has led to the two-fold increase in recombinant protein in previous research and might be a possible consideration for any CMV-dependent RCT in the future (Lin et al., 2015a).

The addition of a strong signal peptide sequence as described in Chapter 3 is usually sufficient to ensure the export of the nascent protein out of the mammalian cell. However, it has been reported in the past that different signal peptides have different export efficiencies and that ones from the same species may be outperformed by SPs from very divergent animals. It is advisable to compare several types of SP in the future using a simple reporter such as a luciferase protein (Stern et al., 2007). While the trypsin SP used here was predicted to be very efficient in export, it is still possible that it did not perform as expected. This could have led to an accumulation of the spidroin protein within the host cells. This could have triggered a stress response such as the unfolded protein response which may degrade the unfolded spidroin, limit translation or even trigger cell death (Gardner et al., 2013). Any of these responses may explain the observed lack of expressed spidroin in these experiments.

In the tRNA charging assay primers were not designed for an AlaGGC tRNA because one does not exist. GCC is a wobble codon and does not possess its own tRNA. An enzymatic post-transcriptional modification changes the adenosine in Ala**A**GC to an inosine nucleotide. This naturally reduces the specificity of the AlaAGC tRNA because inosine can bind to all four bases in the 3' ends of mRNA codons (GCX) (Gerber et al., 1998; Zhou et al., 2013). Therefore, tRNA charging assay primers for alanine-tRNA were designed for the AlaAGC tRNA.

The AlaAGC and GlyGCC primer sets had not been previously optimised. The corresponding qPCRs yielded quite variable readings. This could have contributed to the error within the qPCR signal. Future primer optimisation may rectify this.

In previous analyses of tRNA charging or RT-qPCR more generally, a spike-in control RNA has been used to ensure robust qPCR results (Pavlova et al., 2020; Roberts et al., 2014). A phenylalanine tRNA from yeast was to be added to the tRNA samples ahead of reverse transcription. However, due to supply shortages, this control had to be replaced with the aforementioned GAPDH. Compared to an endogenous loading control, the abundance of spike-in RNA does not vary regardless of the original host cell's state. For every mg of sample RNA, an invariable amount of spike-in RNA can be added that provides a robust loading control for subsequent estimation of charging ratios. In the future, this phenylalanine tRNA form yeast can be used to obtain more reliable tRNA charging ratios.

Overall, a series of gene expression experiments was carried out. These aimed to ascertain if spidroin recombinant expression had occurred. However, no such protein was detected under any of the conditions described. Two subsequent experiments were carried out to identify the reasons for deficient RCT. The results of the tRNA charging assay were inconclusive due to a lack of statistical significance. Additionally, a computational analysis of the recombinant spidroin circRNA sequence could not conclusively show a direct cause for ribosome frameshifting. These experiments do not exhaust every possible factor that may be limiting protein synthesis in this context, however. There are many possible avenues left to explore in future research.

In the following Chapter 5, a different aspect of spidroin biology will be investigated. To increase the amount of knowledge we have about spidroin sequences a computational pipeline was constructed and implemented. Its design, use and results will be described shortly. The pipeline uses publicly available transcriptome and spidroin sequence data to identify previously unknown spidroin homologs. The same pipeline was used on a separate family of repetitive proteins (the suckerins) to further demonstrate its utility and test its capabilities.

5 Transcriptomic Identification of Spidroin and Suckerin Sequences from Publicly Available Data

5.1 Introduction

5.1.1 Spidroin Diversity

As mentioned previously the spidroin gene family is very diverse. Potentially, thousands of spidroin homologs exist. Most spiders use their silk for multiple applications throughout their lifetimes. For example, the aquatic spiders *Argyroneta aquatica* (diving bell spider) produce a sac structure from silk which they use to store air when under water, to store prey that is being digested or to raise their offspring (Schutz et al., 2007). The spider *Cyrtophora citricola* (tent-web spider) builds a tent-like web and uses it for prey capture and raising its offspring (Rypstra, 1979; Yip et al., 2019). Both of these species are thought to possess several spidroin genes that would enable the building of these complex structures. The same is expected for many other spider species. The orb-weaving spiders produce up to seven different types of silk which in turn are composed of even more spidroins (Hormiga and Griswold, 2014; Kono et al., 2019; Rising and Johansson, 2015). The hundreds of millions of years of spider evolution have given us more than 50 000 described spider species today (World Spider Catalog, 2014).

Many more spiders are left to be discovered so the potential number of total spidroin genes is difficult to estimate. Considering the advantages that known spidroins possess (Sections 1.1 and 1.2), identifying more spidroins could increase the repertoire of advantageous sequences for biotechnology to exploit.

5.1.2 Current State of Spidroin Discovery

Spidroin research has been focusing on sequence data from its beginning. The majority of research in this sphere is dependent on some knowledge of nucleotide and peptide sequences (Kono et al., 2019). As mentioned, a large number of spidroins have been discovered already (Chen et al., 2012; Kono et al., 2021; Wang et al., 2019). Typically, the discovery of such sequences relies on a sequenced genome or transcriptome. A wide variety of DNA or RNA sequencing technologies have been used. However, new spidroin discovery is complicated by the low number of available complete spider genomes. The large divergence times between some lineages further limit the possibility of genome-guided transcriptome assemblies (Schwager et al., 2017). Many more spider transcriptomes than genomes are available publicly.

There are several transcriptome sequencing strategies but this chapter will briefly describe two of them. Short-read RNA sequencing is the older of the two and relies on fragmenting an RNA sample into short fragments (50-300 bases) which are sequenced to produce short sequence reads. This is usually the cheaper option that is still most used nowadays (Grabherr et al., 2011).

Alternatively, a full transcript or a large part of it may be sequenced in one reaction. This is referred to as long-read sequencing and can produce reads at large as 10 000 – 100 000 bases. One of the most commonly used families of long-read sequencing technologies is PacBio's SMRT. Here a long nucleic acid fragment is ligated to hairpin-producing adapters (one adapter per end). This generates a circular sequence when annealing is abolished. A DNA polymerase is used to replicate this circle in a manner similar to the aforementioned RCA (Section 3.1). This allows the circle to be replicated many times after many cycles. The addition of each nucleotide is recorded, producing repeats of nucleotide sequence interspersed with the adapters. Each time the sequence is read it helps produce a reliable consensus sequence. This allows us to confidently estimate the actual nucleotide sequence of the original sample, even if it was very long (Rhoads and Au, 2015).

After a transcriptome has been sequenced, typically in some fragmented form of individual reads (especially after Illumina sequencing), it needs to be assembled into full transcripts that represent the biological RNA molecule sequences. *De novo* transcriptome assembly tools have been preferred in spider transcriptomics. Here, instead of aligning transcript fragments to a reference genome, the transcript fragments are assembled based on sequence overlap entirely. This is not without its caveats but there is no alternative when genomes are unavailable. Multiple spider RNA sequencing datasets are available publicly (Conesa et al., 2016). A large amount of research also focuses on spider venom but the sequenced transcriptomes from this are deposited into public databases, thus increasing the overall number of available datasets. Another advantage of RNA sequencing in spidroin discovery is that it almost completely excludes pseudogenes from the analysis (Troskie et al., 2021). Otherwise, pseudogenes from genomic datasets may have some non-biologically relevant differences from transcribed spidroins that may be misconstrued as novel functional spidroins (Kono et al., 2019).

After it is assembled a transcriptome needs to be annotated. This process is when the transcripts are assigned some identity and/or function based on their sequence (Raghavan et al., 2022). In the annotation of spidroins overwhelmingly some form of sequence homology to prior spidroins has been used. The majority of this has been through the Basic Local Alignment Search Tool (BLAST) (Camacho et al., 2009). This tool uses pairwise alignment which means that the identity of each base (or amino acid) in a sequence is aligned to another sequence. Penalties are assigned based on the number of mismatches and in the end, the most similar sequences give the most significant similarity value. A major caveat of this process is that it does not always identify homologous sequences that have diverged long ago. Over millions of years of evolution, sequences with a common ancestor experience many mutations including base substitutions. This could result in a homologous sequence that has too many different bases/amino acids to be recognised by BLAST. However, an alternative method that identifies sequence homology between peptides exists – profile Hidden Markov Models (HMMs). A profile HMM essentially is a statistical representation of the alignment of multiple homologous sequences. The profile HMM includes a consensus sequence for the alignment and estimations for the likelihood that a particular amino acid (or a gap) will be present at a particular position. This information is combined with estimations for how likely each amino acid is to mutate into a different one. This enables certain software to use profile HMMs and to easily detect distantly related homologous sequences. Other advantages are that high-quality profile HMMs are relatively easy to construct by non-experts and that many profile HMMs are readily available in databases such as Pfam (Eddy, 1998, 2011; Finn et al., 2016; Johnson et al., 2010). This approach will be beneficial for the subsequent spidroin discovery that will be explained in this chapter.

5.1.3 Suckerins

Despite their unique set of properties, spidroins share some similarities with other protein families. A good example is the suckerin superfamily. The suckerins are structural proteins which make up the sucker ring teeth (SRT) found in squids. These structures are found on the suckers of the animals' tentacles (Figure 5.1) and possess several similarities to spidroins**.** Both suckerins and (some) spidroins are very repetitive proteins that possess alternating alanine-rich regions and glycine-rich regions (Hiew and Miserez, 2017; Miserez et al., 2009). This limited similarity is part of the reason why suckerins were chosen for new sequence discovery as will be described later in this chapter.

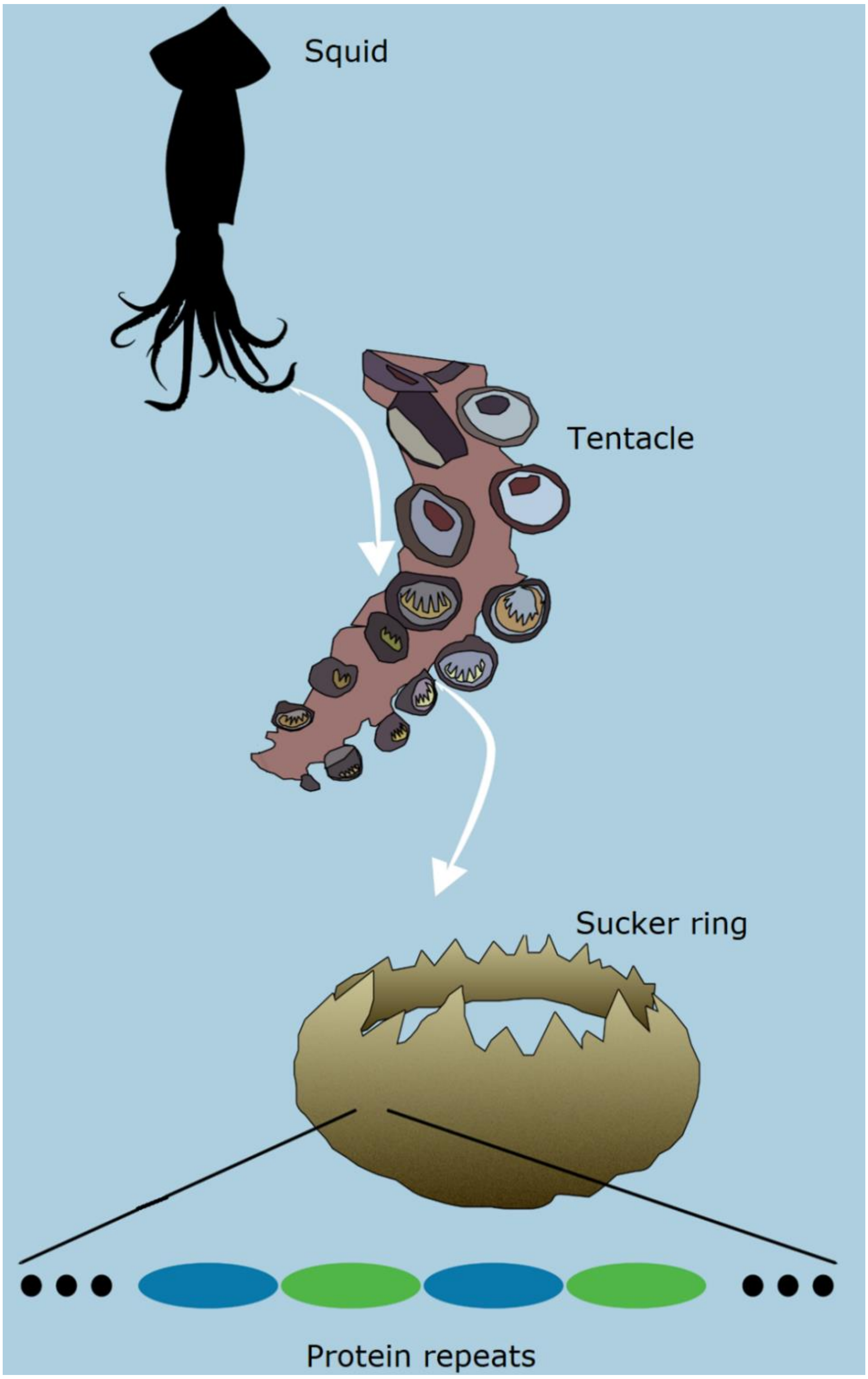


Figure 5.1 – **The sucker ring teeth are located within the suckers of the squid's tentacles**. They have sharp, jagged surfaces that enable the capture of fish. The proteins are composed of alternating alanine- and glycine-rich regions (green and blue).

Decapodiformes (squids and cuttlefish) are primarily active, open-water predators and their sucker ring teeth are predominantly used for piercing into their prey (Nachtigall, 1974). While the outer edge of this structure is sharp and jagged, major adhesive properties have also been recorded. These properties presumably enable the teeth to remain in the flesh of fast and slippery prey such as fish (Miserez et al., 2009).

SRT are entirely formed by proteins, more specifically, members of the suckerin superfamily (Guerette et al., 2014; Miserez et al., 2009). Suckerins are predominantly composed of several amino acid repeats which are generally 30-70 amino acids long and iterated 2-13 times (Guerette et al., 2014; Guerette et al., 2013). The repeats consist of alanine, valine, serine, threonine and histidine-rich regions, which assemble into β-sheet structures similar to how alanine-rich sequences are organised in members of the spidroin proteins (Kumar et al., 2016). The motifs alternate with longer GGY-rich regions that form an amorphous mesh structure again similar to glycine-rich regions within some spidroins (Hiew and Miserez, 2017; Hiew et al., 2017; Miserez et al., 2009). Additionally, there often is a single proline residue that separates the two types of suckerin repeats, potentially confining β-sheets to just the alanine-rich regions (Monsellier and Chiti, 2007).

Sucker ring teeth naturally need to endure somewhat large compression and shearing forces during prey capture so they possess relatively high stiffness and resistance to stress for an unmineralised animal-derived material (Rieu et al., 2016) as well as high resistance to wearing and cracking (Gebbie et al., 2017; Hiew and Miserez, 2017). Furthermore, suckerins are highly hydrophilic and relatively soluble (Miserez et al., 2009). This is caused in part by the presence of weak intramolecular and intermolecular interactions and the absence of covalent intermolecular bonds. Conversely, stacking interactions between the aromatic rings in suckerin amorphous regions contribute to the easy self-assembly of suckerins (Kumar et al., 2018a). Therefore, suckerins can readily form and reform their structure in aqueous solution after heat denaturation (Miserez et al., 2009; Rieu et al., 2016; Latza et al., 2015). In other rigid tissues primarily formed out of protein, such as spider fangs and marine worm jaws, it is far more difficult to dissolve and then re-form their former structures (Lichtenegger et al., 2002; Politi et al., 2012). These properties enable native suckerins to be used as a material in 3D printing. Successive suckerin denaturation and refolding do not compromise its mechanical properties (Latza et al., 2015).

Suckerins have also been studied for their potential biocompatibility, processability and especially for maintaining adhesion in wet environments (Bhagat and Becker, 2017; Lee et al., 2011). The latter property especially has been sought after in tissue engineering where only a few such materials are known (Latza et al., 2015). Non-biological alternatives such as cyanoacrylates and formaldehyde have been investigated for their adhesive properties but their cytotoxicity limits their usefulness (Papatheofanis, 1989). Suckerins' adhesive properties can be further improved by the incorporation of non-canonical amino acids such as DOPA (Deepankumar et al., 2020). The abundance of advantageous properties of suckerins has inspired many hypothetical scaffolds in tissue engineering and drug delivery (Sanchez -Ferrer et al., 2018; Buck et al., 2019; Hiew et al., 2019; Deepankumar et al., 2020).

While repetitive, suckerin sequences are quite short which greatly facilitates recombinant expression. His-tagged suckerins have been successfully expressed and purified from *E. coli* (Ding et al., 2014).

One of the major prerequisites in biomimetic synthesis, including that of suckerins, is the determination of their peptide sequences (Guerette et al., 2013). Suckerins have not been researched as well as other protein families such as the spidroins. Few genomes are available for squid lineages and cephalopods as a whole. Few studies have focused on suckerin discovery and only between 30 and 40 members of this family are known (Guerette et al., 2014). The suckerins that have been discovered belong to the Decapodiform lineage which diverged from their closest relatives, the Octopodiformes (octopuses), approximately 336 million years ago (Hedges et al., 2015). However, phylogenetic analysis of the protein family members has indicated that these peptides may have emerged no later than 354 million years ago (Guerette et al., 2014). Furthermore, the SRT structures can be observed in many if not most members of the Decapodiformes (Arkhipkin and Laptikhovsky, 2008; Burgess, 1982; Clarke and Maul, 1962; Roper, 1968; Sin et al., 2009). Therefore, it is reasonable to expect that a large number of suckerins remain to be discovered in the Decapodiform lineage. These additional suckerins may possess novel properties that could be of use in biotechnology.

The remainder of this chapter focuses on the computational work toward the discovery of spidroin and suckerin sequences from publicly available transcriptome data. For this purpose, a pipeline was designed which incorporates a variety of publicly available sequence analysis tools. The pipeline accepts transcriptome datasets and annotates novel spidroin sequences as the output. In particular, this pipeline accepts tandem repeat alignments as a basis for the creation of profile HMMs and therefore the identification of novel spidroins. The spidroins' and suckerins' repetitive structures make them especially amenable to analysis with this pipeline. First, the design of the pipeline will be described. Next, its implementation on the spidroin family and then in the suckerin family will be explained. Finally, the phylogenetic analysis of the newly-described suckerins will be shown.

5.2 Construction of Pipeline for the Discovery of Repetitive Proteins from Available Transcriptomes

The first step in the transcriptome analysis was to design a suitable pipeline (Figure 5.2) which could identify sequences similar to a representative repetitive protein. It is important that a representative protein is chosen which possesses some repeat units at least longer than six amino acids and is iterated at least twice.

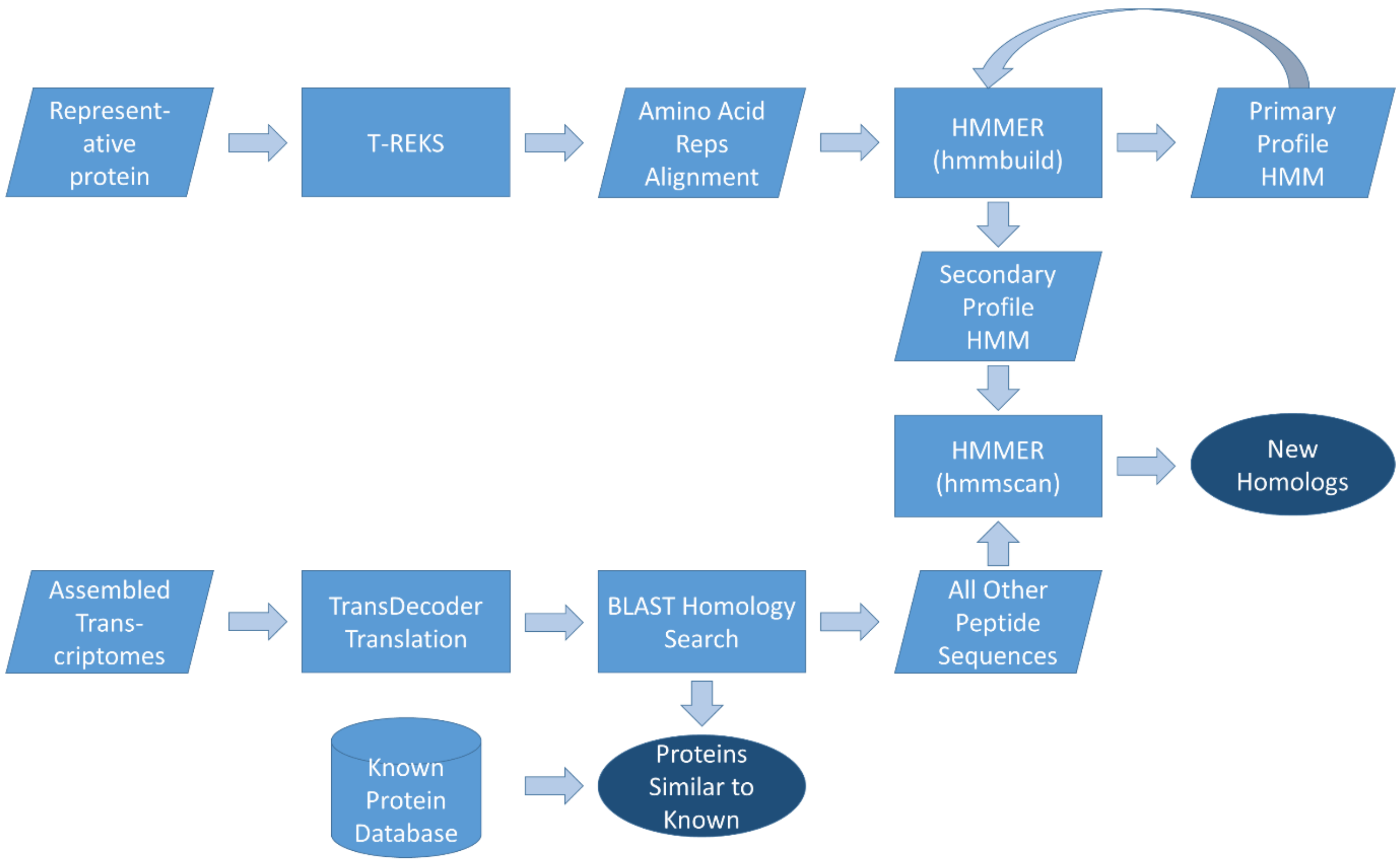


Figure 5.2 – **Flowchart of the sequence discovery pipeline**.

The pipeline was designed to take advantage of the repetitive regions within the proteins studied. After a representative protein was chosen its repeats were aligned onto each other using the tool T-REKS. This employs a K-means clustering algorithm to produce sequence alignments from the different repeats. In short, T-REKS starts by searching for a random sequence of two residues. When several such sequences are identified, T-REKS will record the distance between them. It assumes that if repeat sequences exist, the distance between the iterations will be equal. It then goes through cycles of expanding the presumed tandem repeat sequences until the total length of tandem repeats is reached (Jorda and Kajava, 2009).

T-REKS has several advantages over other similar software. It is very efficient at identifying repeat sequences without finding false positives and can do so without a reference sequence. Most importantly, T-REKS can account for mismatches between the repeats and it has been designed to specifically detect repeats at the size range similar to the MaSp repeats (Jorda and Kajava, 2009).

The output of T-REKS was manually converted to a Stockholm sequence alignment format. This was used in the next step – the construction of a profile HMM. Using the tool HMMER an initial profile HMM was constructed that represents the consensus of the repeat sequence alignment and also contains information about their sequence diversity (Eddy, 2011). However, since this is only based on a single peptide it has limited information about the sequence diversity within the whole protein family. Therefore, a second step is implemented where the initial profile HMM is compared to a database of currently known proteins homologous to the initial representative one. This outputs a new sequence alignment which essentially contains all of the currently available information on the possible variability between individual repeat units. The new sequence alignment is used to construct a new profile HMM (using HMMER). This is the final model that contains as much information as possible about the repeats of the protein of interest.

Next, the transcriptomes that are to be investigated need to be prepared. They must be assembled into full transcripts prior to analysis. After assembly, these datasets are translated into protein sequences with the tool TransDecoder. This tool has been specifically designed to process *de novo* transcriptome assembly outputs and is optimal even without the use of a guiding genome. TransDecoder particularly looks for the longest possible open reading frame in each transcript. A BLAST search is integrated into this process to identify transcripts that translate into previously annotated peptide sequences. Significant homology to previously known peptides from spiders or squids adds a small number of additional peptide sequences to the output of TransDecoder that may otherwise be omitted by the tool (Haas et al., 2013).

After the completion of transcriptome translation, all resulting peptide sequences are analysed for sequence homology to known spidroins or suckerins. Sequences with high levels of homology will have previously been identified using traditional methods, so will not be of particular interest in this search for novel sequences. This separates the translated transcriptomes into two parts – one with sequences that are pairwise homologous to known spidroins (these are excluded from most downstream analysis) and all other sequences fall into the second part (these are further analysed for the presence of novel suckerins and spidroins).

Finally, the pipeline uses the constructed profile HMMs to screen the translated transcriptomes. Each sequence in a transcriptome is analysed with HMMER to determine if it is similar to the profile HMM.

5.3 Discovery of Spidroins

5.3.1 Choice of Transcriptomes

Several sort-read spider transcriptomes were collected from the SRA and TSA databases for downstream analysis. Samples from basal lineages like liphistiids and tarantulas were chosen as well as more derived lineages like several orb-weavers and cobweb spiders (Figure 5.3). This selection will potentially identify a diverse collection of spidroin genes. Additionally, not focusing on a specific clade of spiders allows a larger number of transcriptomes to be investigated.

Both assembled and unassembled transcriptomes were chosen for this analysis. Additionally, the short-read transcriptome of the castor bean tick (*Ixodes ricinus*) was analysed with the same pipeline. Since this animal does not produce spidroins or any such fibrous proteins it was chosen as an outgroup for negative control. The tick is a distant relative of the spiders and is also an arachnid.

Two long-read spider transcriptomes, those of *Argyroneta aquatica* and *Cyrtophora citricola* were investigated as well. These spiders contribute to the overall diversity of potential spidroin sequences and their unique web structures as mentioned before could provide some interesting spidroins for discovery. The relationships between all of the chosen arachnid species are shown in Figure 5.3 below.

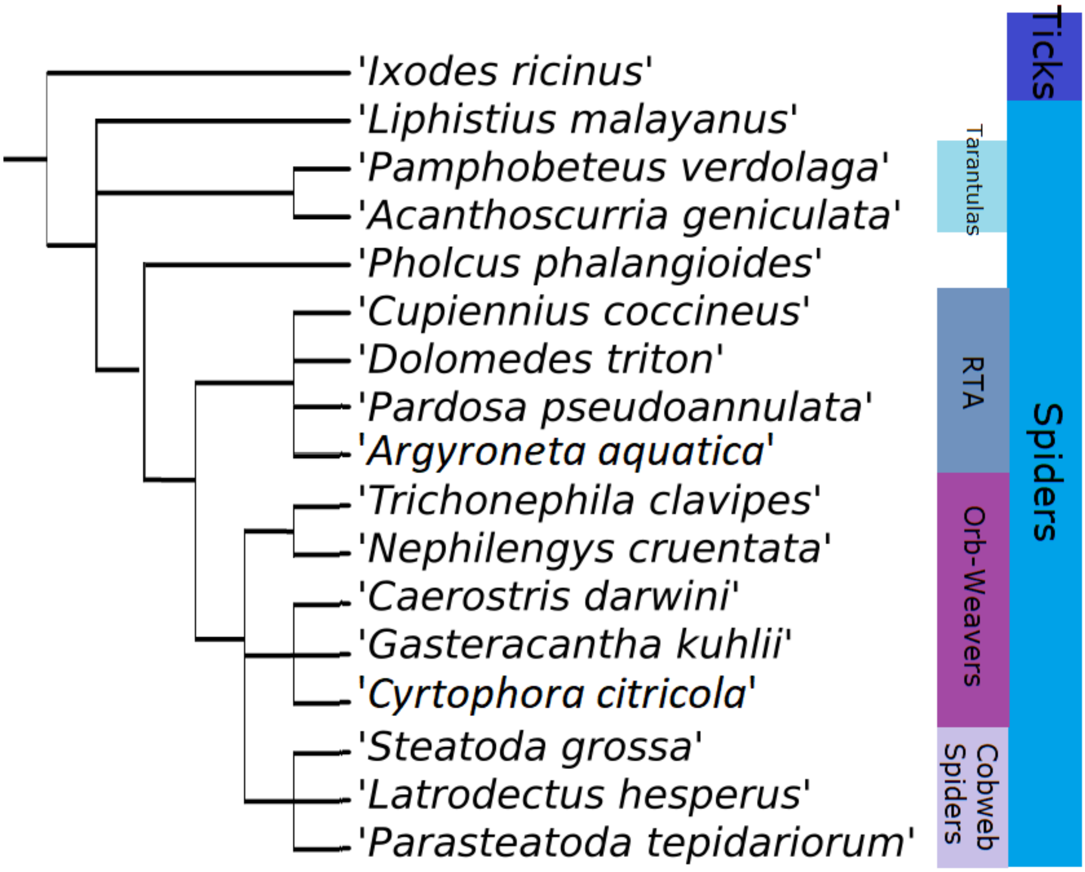


Figure 5.3 – **Cladogram of the taxonomic relationships between the species used in this study**. The RTA spiders are named for an anatomical feature – the retrolateral tibial apophysis (Coddington and Levi, 1991).

5.3.2 Transcriptome Assembly

Transcriptome assembly was performed for seven spider transcriptomes. *De novo* assembly was carried out for six of them and genome-guided assembly for two of them using the Trinity tool. Trinity works in three different stages to assemble full-length transcripts from individual reads. In the first stage (named Inchworm) a random read is chosen and extended in the 5′ and 3′ directions based on overlap with other reads to produce contigs. The next stage (named Chrysalis) assembles the contigs into de Bruijn graphs. These are representations of alternative splicing which should correspond to the full transcriptional complexity for a given gene. Lastly, the third stage (named Butterfly) resolves these graphs into full-length transcripts labelled as alternatively spliced isoforms (Grabherr et al., 2011).

The *Latrodectus* transcriptome was assembled through both of these approaches (Table 5.1). Overall, the number of estimated transcripts was on the order of tens of thousands. The N50 value is an estimation of the length of the shortest contig out of the longest contigs that together constitute at least 50% of the total assembly. This metric is similar but not identical to the median contig length. In general, a large N50 is indicative of a high-quality assembly. The N50 of these assemblies ranged from ~300 to ~1600. This suggests a large degree of fragmentation of the *Cupiennius* and *Latrodectus* (genome-guided) assemblies. The read alignment rate was estimated by aligning the assembled transcriptome to the corresponding trimmed reads and is indicative of assembly completeness. Overall, a high alignment rate was observed. In the case of genome-guided *Latrodectus* assembly, these metrics were considered insufficient and it was not considered for downstream analysis.

Table 5.1 – **Summary of transcriptome assembly metrics.**

Spider (assembly method)	Total "genes" (and total transcripts)	Tissue of origin	Contig N50	Read alignment rate
***Cupiennius* (*de novo*)**	94930 (259924)	Whole body	313	82.97%
***Gasteracantha* (*de novo*)**	124731 (194905)	Whole body	632	88.97%
***Latrodectus* (*de novo*)**	18913 (35380)	Whole Body	803	98.10%
***Latrodectus* (genome guided)**	5407 (8254)	Whole body	325	5.85%
***Liphistius* (*de novo*)**	53505 (69057)	Whole body	1120	95.52%
***Pamphobeteus* (*de novo*)**	72281 (199797)	Venom gland	968	96.99%
***Parasteatoda* (genome guided)**	132574 (151295)	Whole body	1596	94.18%
***Pholcus* (*de novo*)**	52189 (63582)	Silk glands	1640	97.65%
***Cyrtophora* (long-read *de novo*)**	65659 (89303)	Silk glands	1788	98.71%
***Argyroneta* (long-read *de novo*)**	60850 (83655)	Silk glands	2249	98.34%

The remainder of the short-read transcriptomes in the spider analysis were downloaded pre-assembled from the TSA database. Also, in the case of the long-read transcriptomes from *Argyroneta* and *Cyrtophora*, even though their reads may fit the length of a complete transcript, *de novo* Trinity assembly was again implemented.

The assembled long-read transcriptomes achieved a comparable number of total “genes” and transcripts to short-read transcriptomes. However, the long-read assemblies surpassed the rest in the metrics of contig N50 which represents overall contig length and read alignment rate which represents assembly completeness. This indicates that long-read sequencing possesses some advantages over short-read sequencing. The difference between *Argyroneta, Cyrtophora* and *Latrodectus (de novo)* was very small, however. The observed differences in assembly may also reflect the biology of the spiders in question.

5.3.3 Transcriptome Annotation

After all transcriptomes were translated with TransDecoder, the HMMER tool was used to identify the assembled contigs containing spidroin protein domains. This allows the identification of distant homology between sequences that may be otherwise omitted by pairwise homology search algorithms such as BLAST. A profile HMM is made from a collection of homologous sequence data and is aligned to subject sequences with HMMER again (Eddy, 2011).

To search for as many spidroin domains as possible at the same time an HMM database was constructed using both publicly available profile HMMs and personally constructed ones. The Protein Families database (Pfam) contained entries for the unique spidroin N- and C-termini, as well as CySp repeats (Berezikov et al., 1998; Bochicchio et al., 2001; Case et al., 1997). The rest were constructed manually, based on the presence of tandem repeats, as described previously (Section 2.4.11, Section 5.2), from eight representative spidroin sequences (AcSp, AgSp, FLAG spidroin, MaSp1, MaSp2, MaSp3, MiSp, PySp).

This concatenated database was used to scan each assembled transcriptome and output domain annotations. Whenever both terminal domains were found at the ends of the same contig, an assumption was made that a full spidroin peptide is found since this domain organisation is conserved in spidroins (Heiby et al., 2017).

Overall, 9 unique peptide sequences were identified in six species (accession numbers in parentheses): one sequence in *Pholcus* (DN1767), two in *Latrodectus* (GBJN01059188.1 and GBJN01052811.1), two in *Nephilengys* (GEWZ01013945.1 and GEWZ01016889.1), one in *Pardosa* (GGRD01219347.1), two in *Acanthoscurria* (GAZS01040257.1 and GAZS01047594.1) and one in *Argyroneta* (54229_201110_222647/13697856). All spider transcriptomes contained many additional spidroin fragments that contained spidroin domains according to hmmscan. All of the identified spidroins are summarised in Table 5.2 below. These generally did not exceed 800 nt in length and were not shown to contain both termini. The *Cyrtophora*, *Caerostris* and *Gasteracantha* transcriptomes in particular contained a large number of fragments, none of these contained both termini in the same transcript. These spidroin fragments have not been analysed further at the time of writing. No putative spidroins or fragments were found in the tick (*Ixodes*) transcriptome, which acted as a negative control. The *Argyroneta* spidroin was identified through BLAST as a known TuSp homolog. These findings are summarised in Table 5.2 below.

Table 5.2 – **Summary of HMMER3 spidroin annotation**.

Spider	Spidroin Domains Identified	Full-Length Spidroins	Unique Spidroins Identified
Acanthoscurria	101	2	2
Argyroneta	350	1	0
Caerostris	1,035	0	0
Cupiennius	47	0	0
Cyrtophora	1,568	0	0
Dolomedes	250	0	0
Gasteracantha	1,574	0	0
Ixodes	0	0	0
***Latrodectus* (genome guided)**	14	0	0
***Latrodectus* (*de novo*)**	446	1	0
***Latrodectus* (TSA)**	172	1	1
Liphistius (de novo)	50	0	0
Liphistius (TSA)	0	0	0
Nephila	318	0	0
Nephilengys	127	2	1
Pamphobeteus	18	0	0
***Parasteatoda* (genome guided)**	344	0	0
***Parasteatoda* (TSA)**	74	0	0
Pardosa	221	1	1
Pholcus	27	1	1
Steatoda	155	0	0
Total	6,891	9	6

BLASTP and TBLASTN searches were performed with each putative spidroin's peptide sequence. One *Argyroneta*, One *Latrodectus* (GBJN01059188.1) and one *Nephilengys* (GEWZ01013945.1) sequence were found to be previously annotated as spidroins. The rest were presumed to be undescribed sequences (Table 5.2 above).

5.3.4 Characterisation of Novel Spidroins

The unique putative spidroins were investigated for the presence of signal peptides using the Phobius and SignalIP web tools (Armenteros et al., 2019; Kall et al., 2004). Signal peptides are estimated at the N-terminal end of GAZS01047594.1, GGRD01219347.1 and DN1767.

T-REKS under relaxed conditions (E-value threshold 0.7) was used to detect the tandem repeat motifs present in the six peptide sequences. No large repeat regions were found for any of the putative spidroins. Nonetheless, a variety of small repeat motifs characteristic of spidroins was found in these sequences (Figure 5.4).

Single iterations of the characteristic long CySp repeats were found in GAZS01040257.1 and GAZS01047594.1 in *Acanthoscurria*. Three such repeats were found within DN1767_c0_g2_i1 in *Pholcus*. They were accompanied by the TTX, XQQ and SSX motifs. No large-scale repeats were found in the three other spidroin central domains. GBJN01052811.1 in *Latrodectus* and GGRD01219347.1 in *Pardosa* contained short spidroin motifs but the majority of their sequences could not be identified as typical for a specific spidroin. GEWZ01016889.1 in *Nephilengys* contained a large number of MaSp2 repeat motifs (Figure 5.4).

GAZS01040257.1 *Acanthoscurria geniculata*

IPEYQSGVLEAGQFSNFATAASAASQESSSITSITTTSSASAAA
AQAAT**AAGAAGTAGSSAASSIFSQTMMSYLLQSTAIASAFSQAKSSS**
SAYAIAFAMTQAAANQMGLSNYARALSL
AAAGAVSEVRLGGSLYDYLYALVRALSQFLFGYGLITSA
NENSFGTTMATALANAASSSEASAAALASSSAVADVAV
RGAAGAVSGGGAASRSVGLIGSGFPGSVSSTPFLLNSSLGGAGGSGQVP
VLPLAGGLLPPSS

GAZS01047594.1 *Acanthoscurria geniculata*

DNEIDQNLPQEEYPGFITQVGSADESEIFGSVTAGNTAQQVS
TEATQSSSTTQT**YSGQTASSSPITFEAITPSASEFFIQHLT**
SLLMENLKFVSQYNSIASSGQVSAYASTSAENVAYSIDQGSIASLM
ASAVSQATADISALEGPSSFVHAFASA
LAQILSTSGVLNWSNVIELASAFGNSLFTAISTASSSAT
SEWSTAGSQISTLQTSSAAVTGAFASSEAADANYAGLEN
PFLPAFGAGDGLNNFDFLTPMPGISGVPSSSELSPYSSLISGV

GBJN01052811.1 *Latrodectus hesperus*

PSPSQPNNPQSPQVSIALSLGNNNPQPYPGNQFADSSANSSPSQYPFNNN
NQNSATGGSFFSNPSAFAQVNVPGNSENPTIRSD
SVRNYATGGTDNYPAKDETYQNGPPQSAIKINVQSNPMQEN
ANPYQNEQSQMRFIRPIPSYYYYDYPGGYGPQVNSKNDNEPS
AAAAASVQINPVAPEQTPLSSSPDSSAS

GEWZ01016889.1 *Nephilengys cruentata*

KENSYTDNSEVFSRNFGGPSGLGGNSAAAAAGGDGGSRQGGYIGNIGGQ
GGLSGDSVAAAASAGGNGGSRQGGYGGSFG
GPGGESAAAAAGGNGGSGQGGYGGTLGRPGSNSAARGDG
GSGQRGYGGNFGGPSGLGGNSAAAATVGDGGFGQGGYSG
NIGGQGGLSGDSVAAAASAGGNGGSRQGGYGGNFGGPGD
ESAAAATGENGGSGQEGYGGNFRGPGGLGGISIAVTSSAGG
DGGSGQGGYSGNIGGQG

GGRD01219347.1 *Pardosa pseudoannulata*

FAEEAASSDTYNELDIGLDYGFGQSIGYGNAGTFSSSGSGTAGSSSTATST
STLTSTSTSTGAAAGSAAGAAGAAGAAFGAG
FGGFSGAGGFTNNLLPNLGGISLPTDFTSLLSSPVGLNSQQ

DN1767 *Pholcus phalangioides*

WNFVFCIY**NELFSSSAFTNVFTSNLPYERALTYIRRISAS**
LAGSFFFGRMTEYDISSAFSKALTFLNPVPTTEEFAYMF
GQELSKMISNQGMFERSNPCRLASSVAKEITKNLNNPP
NQNTGQNIPLSKSSDVEEKPPLPNSLVMAKPPLSNSLGVMDK
PPLPNSLRVMDKPPLPYLVTVPSLNDDFKLRRKLD**FERKLRT**
RLLSSEIFYTAFPNEVKYNKLTDYAFAVGSFLPATPEYR
SLNISRLSTALTDSLLSVKPPAVANQYATAVVTVISSFM
EDKGLLNDTNVNKQASRVGTIILNALCTQITTNFQICARK
QRLRTRIVPKFPNIPEPTPNVVSPSTNVKNDEGPTELSFPNNIYKSSTVFDN
MAGDNKEP

Figure 5.4 – **Excerpts from central domains within the newly identified spidroins**. The terminal domains are not shown. Short amino acid repeats characteristic of spidroins are shown in red, yellow, blue, green orange and purple. Higher complexity CySp repeats are highlighted in bold.

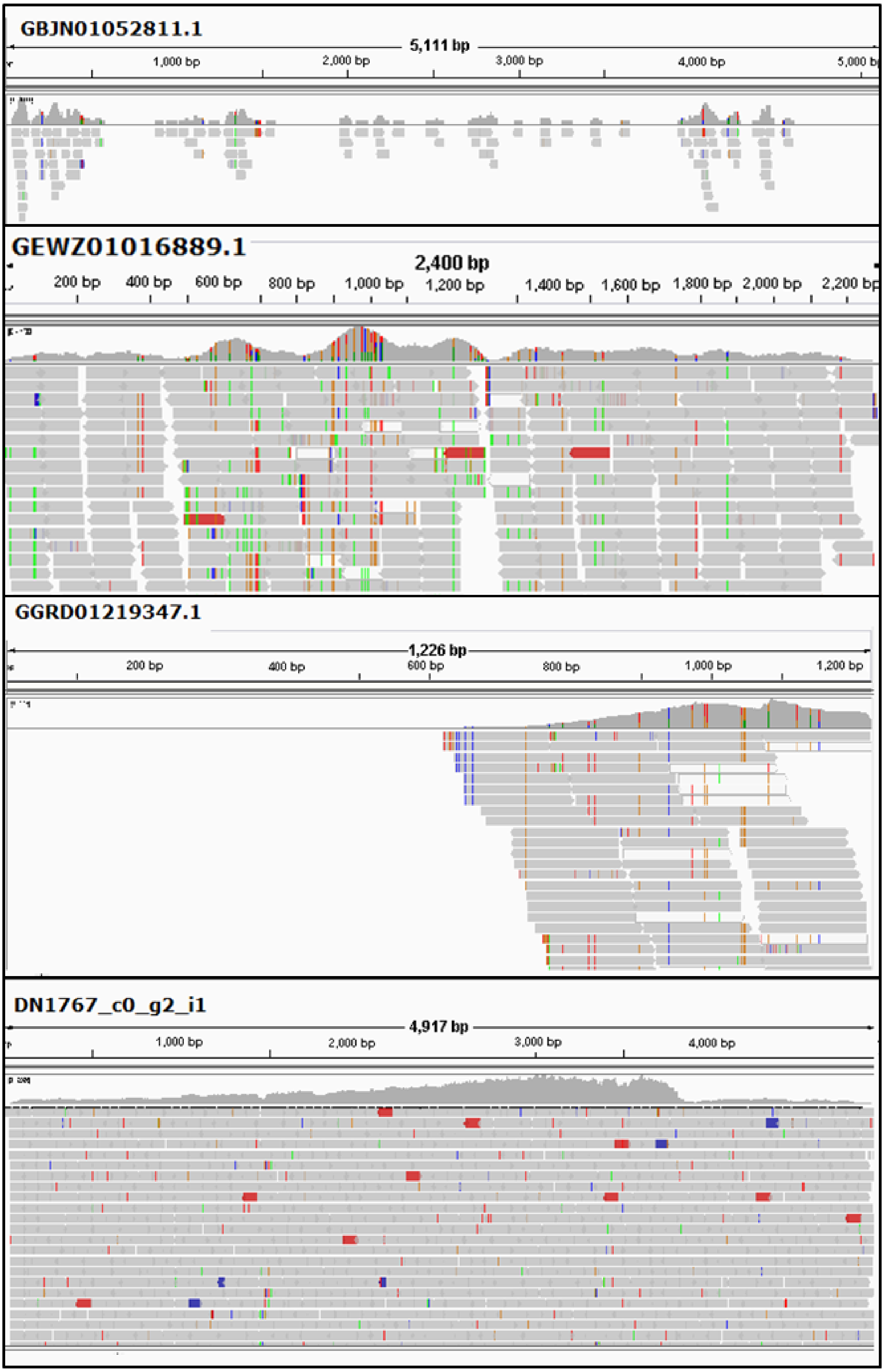
Expression quantification was performed for putative spidroin-containing transcriptomes (*Acanthoscurria, Latrodectus* (*de novo* assembled), *Nephilengys, Pardosa and Pholcus*). Spidroin expression in each sample was compared to that of a housekeeping gene – the ribosomal protein RPL13 (Scharlaken et al., 2008) (Table 5.3).

The *Acanthoscurria* putative spidroin transcripts with zero expression were considered to not be biologically relevant. The *Latrodectus* and *Pardosa* putative spidroin expression levels were much lower than that of their corresponding RPL13 transcripts. The *Nephylengys* putative spidroin had a comparable expression to that of RPL13 but the *Pholcus* DN1767 was much more highly expressed than RPL13.

Table 5.3 – **Expression quantification of putative spidroins in transcripts per million (TPM).**

Organism	Transcript	Length (bps)	TPM
Acanthoscurria	RPL13	558	116.63
Acanthoscurria	GAZS01040257.1	1940	0.00
Acanthoscurria	GAZS01047594.1	1940	0.00
Latrodectus	RPL13	815	46.28
Latrodectus	GBJN01052811.1	5146	1.35
Nephilengys	RPL13	712	25.58
Nephilengys	GEWZ01016889.1	2420	20.21
Pardosa	RPL13	765	360.79
Pardosa	GGRD01219347.1	1234	4.70
Pholcus	RPL13	760	5.95
Pholcus	DN1767_c0_g2_i1	4950	175.60

Next, read coverage was visualised for the putative spidroins that showed non-zero expression using IGV (Figure 5.5). Overall it was observed that the assembled spidroins do not possess a uniform expression throughout the length of each transcript. In the case of GEWZ01016889.1 and DN1767, several peaks of read coverage were observed at regions which include the characteristic spidroin repeat motifs. This hints at repeat motifs that may have been erroneously recognised as identical transcripts by Trinity's algorithm. For GGRD01219347.1 read coverage was observed only over the 3’ end of the transcript which may indicate an incorrect assembly.


GBJN01052811.1
5,111 bp
1,000 bp
2,000 bp
3,000 bp
4,000 bp
5,000 bp
GEWZ01016889.1
2,400 bp
200 bp
400 bp
600 bp
800 bp
1,000 bp
1,200 bp
1,400 bp
1,600 bp
1,800 bp
2,000 bp
2,200 bp
GGRD01219347.1
1,226 bp
200 bp
400 bp
600 bp
800 bp
1,000 bp
1,200 bp
DN1767_c0_g2_i1
4,917 bp
1,000 bp
2,000 bp
3,000 bp
4,000 bp

↑Figure 5.5 – **Read coverage visualisation of putative spidroin transcripts**. Bowtie2 was used to align raw reads to each assembled transcriptome. The Integrated Genomics Viewer was used to visualise whether read coverage is uniform throughout four putative spidroin transcripts (Thorvaldsdottir et al., 2013).

5.4 Discovery of Suckerins

5.4.1 Choice of Transcriptomes

A diverse selection of sixteen mollusc transcriptomes was chosen for this analysis (Figure 5.6). The majority of transcriptomes were of decapodiform animals (squids and allies) since this is the clade that is known to possess the SRT structures. Initially, an octopus (*Hapalochlaena*) transcriptome was chosen as an outgroup control but the discovery of putative suckerins in this sample, as will be described shortly, warranted further attention for this group. Two other large octopus transcriptomes were available in the TSA – that of the more closely related *Octopus maya* and the more distantly related vampire squid (not a true squid) *Vampyroteuthis*. The distantly related molluscs *Physella* (a terrestrial snail) and *Nodipecten* (a bivalve) were later chosen as outgroups based on their relatively complete transcriptomes. Overall, the largest and most complete transcriptomes were chosen whose size ranges between 2.7 and 239.6 Mbases. Wherever possible, transcriptomes based on tentacle-extracted RNA were selected. For this part of the research, only pre-assembled transcriptomes were used as the SRA database did not possess any additional species that could expand the repertoire of transcriptomes.

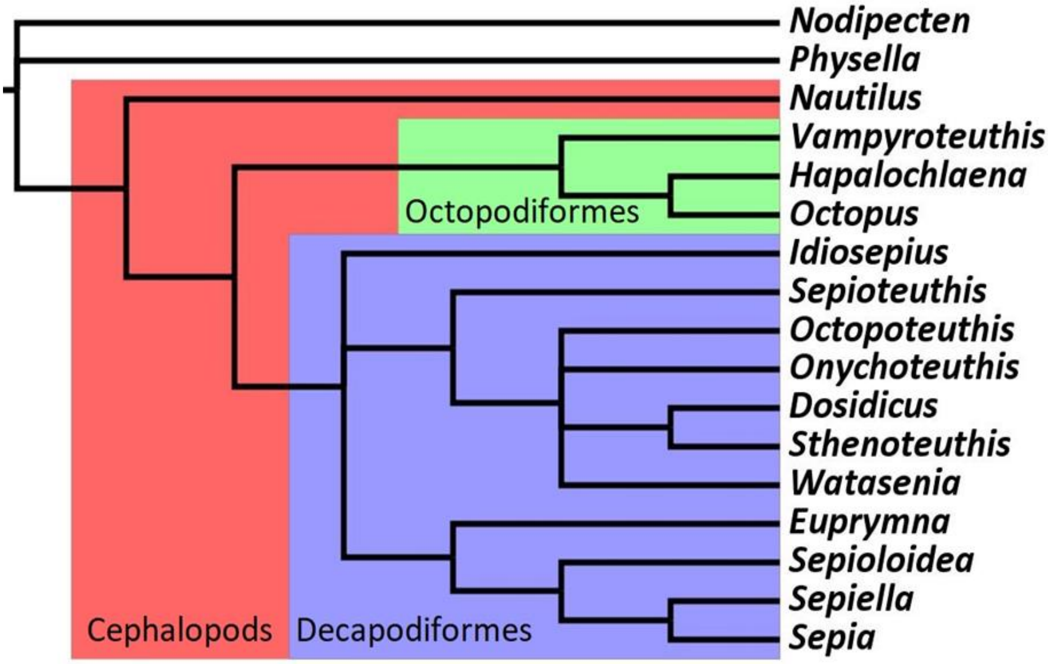


Figure 5.6 – **Cladogram of mollusc genera whose transcriptomes were used in this analysis.** *Nautilus* represents the most basal extant cephalopod lineage (Bonnaud et al., 1997) and is only included for clarity. This tree was constructed using the NCBI Taxonomy Browser and FigTree v1.4.4 (Rambaut, 2006; Schoch et al., 2020).

5.4.2 Transcriptome Annotation

Initially, a profile HMM was constructed as described previously in sections 2.4.11 and 5.3.3. To begin this construction a representative suckerin peptide was chosen (AGY36220.1, a 39 kDa suckerin protein from *Dosidicus gigas*). Its size is similar to the average for suckerins and it lacks non-suckerin peptide sequences. Again, T-REKS was used to produce an initial sequence alignment of repeats and HMMER was used to construct an initial profile HMM from this. Afterwards, a database of 39 previously known suckerins was downloaded from Genbank. This database was used to align the initial profile HMM to all the various known suckerins. This ensures that a multiple sequence alignment will be produced which accounts for the sequence diversity within the suckerin family. Again, this multiple sequence alignment was used to generate a second, more comprehensive profile HMM.

Later, this profile HMM was compared to the filtered squid transcriptomes. In total, 34 sequences with large pairwise homology (e-value below 0.00001) to known suckerins over more than 80% of their length were identified in six of the cephalopod transcriptomes. These findings have been summarised in Table 5.4 below. While most of these have not been described previously, their structure and function will likely be equivalent to that of known suckerins. Additionally, the single BLAST-identified suckerin from

Sepia esculenta was virtually identical to a previously known *S. esculenta* suckerin. These were omitted from most downstream analysis to focus on the unique profile HMM-identified homologs.

Table 5.4 – **Summary of the identified suckerin homologs**.

Clade	Genus	Suckerin sequences identified with BLAST	Suckerin sequences identified with HMMER
Squids	*Dosidicus*	0	0
	Euprymna	0	0
	Idiosepius	1	7
	Octopoteuthis	13	16
	Onychoteuthis	0	0
	Sepia esculenta	1	0
	Sepia pharaonis	9	3
	Sepiella	9	3
	Sepioloidea	1	2
	Sthenoteuthis	0	0
	Watasenia	0	0
Octopuses	*Hapalochlaena*	0	1
	Octopus	0	0
	Vampyroteuthis	0	0
Bivalves	*Nodipecten*	0	0
Snails	*Physella*	0	0
Total	16	34	32

After this exclusion, the remainder of each transcriptome was translated with the TransDecoder tool. HMMER3 was used to identify suckerin domains using a constructed profile HMM which contained both a GGY-rich region and an Ala-rich region. This yielded 32 putative suckerin sequences from six species of cephalopods (Table 5.4). The various iterations of the BLAST tool were used to identify previously known suckerins within those found by HMMER. Regions within GFNE01035555.1 and GFNE01035556.1 from *Idiosepius* were identified as having pairwise homology to a known suckerin (suckerin-6 from *Sepioteuthis lessoniana*) over a small part of their sequence (less than 40%) and these sequences were retained for downstream analysis.

5.4.3 Characterisation of the New Suckerins

The HMMER identified peptides ranged in size between 111 and 482 residues which is consistent with the variability within known suckerins. They contained between 2 and 9 iterations of repeats. Sixteen were found to contain N-terminal signal peptide sequences which is consistent with secreted proteins such as suckerins (Figure 5.7). Overall, the repeats identified in the novel suckerins were very similar in sequence to the previously known suckerins. Signal peptide sequences, GGY-rich regions and Ala-rich regions were annotated in the novel suckerins. A web-based HMMER search against the ~18 000 domain profiles in Pfam did not identify any other protein domains within the suckerins, apart from the occasional short disordered region.

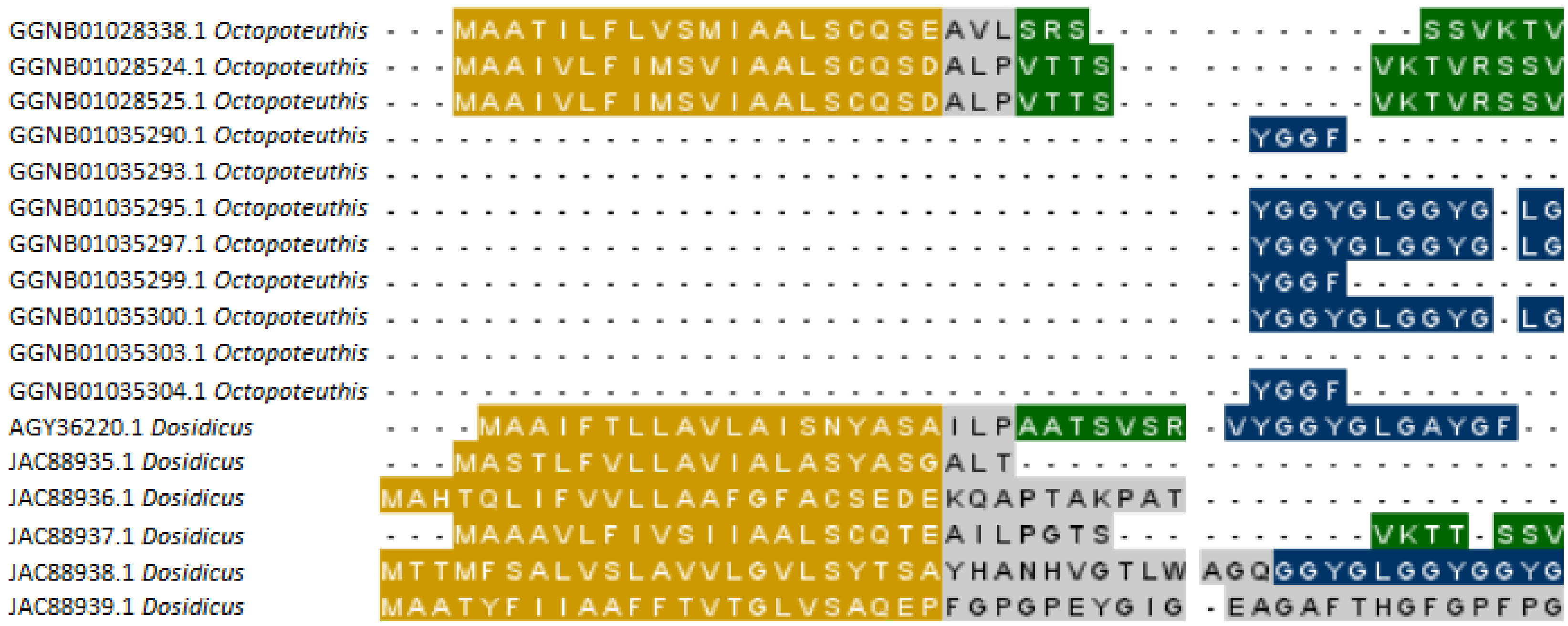

GGNB01028338.1 *Octopoteuthis* ---MAATILFLVSMIAALSCQSEAVLSRS---- ---------SSVKTV
GGNB01028524.1 *Octopoteuthis* ---MAAIVLFIMSVIAALSCQSDALPVTTS--- -------VKTVRSSV
GGNB01028525.1 *Octopoteuthis* ---MAAIVLFIMSVIAALSCQSDALPVTTS--- -------VKTVRSSV
GGNB01035290.1 *Octopoteuthis* --------------------------------- --YGGF---------
GGNB01035293.1 *Octopoteuthis* --------------------------------- ---------------
GGNB01035295.1 *Octopoteuthis* --------------------------------- --YGGYGLGGYG-LG
GGNB01035297.1 *Octopoteuthis* --------------------------------- --YGGYGLGGYG-LG
GGNB01035299.1 *Octopoteuthis* --------------------------------- --YGGF---------
GGNB01035300.1 *Octopoteuthis* --------------------------------- --YGGYGLGGYG-LG
GGNB01035303.1 *Octopoteuthis* --------------------------------- ---------------
GGNB01035304.1 *Octopoteuthis* --------------------------------- --YGGF---------
AGY36220.1 *Dosidicus* ----MAAIFTLLAVLAISNYASAILPAATSVSR -VYGGYGLGAYGF--
JAC88935.1 *Dosidicus* ---MASTLFVLLAVIALASYASGALT------- ---------------
JAC88936.1 *Dosidicus* MAHTQLIFVVLLAAFGFACSEDEKQAPTAKPAT ---------------
JAC88937.1 *Dosidicus* ---MAAAVLFIVSIIAALSCQTEAILPGTS--- -------VKTT-SSV
JAC88938.1 *Dosidicus* MTTMFSALVSLAVVLGVLSYTSAYHANHVGTLW AGQGGYGLGGYGGYG
JAC88939.1 *Dosidicus* MAATYFIIAAFFTVTGLVSAQEPFGPGPEYGIG -EAGAFTHGFGPFPG

Figure 5.7 – **Signal peptide and repeat architecture in several suckerin protein sequences**. Eleven putative suckerins from *Octopoteuthis deletron* are shown alongside six previously known suckerins from *Dosidicus gigas*. Signal peptide sequences (orange) were estimated with the Phobious and and SignalIP 5.0 web tools (Armenteros et al., 2019; Kall et al., 2007). GGY-rich and A, T, S, V-rich regions are shown in blue and green respectively. Non annotated regions are shown in grey.

Additionally, the suckerin identified in the *Hapalochlaena* transcriptome (GEXH01164213.1) did not exhibit regions similar to the characteristic suckerin repeats. Instead, it showed very strong sequence homology to undetermined regions in the previously identified suckerin-6 from the bigfin reef squid (*Sepioteuthis lessoniana*) as well as GFNE01035555.1 and GFNE01035556.1 from *Idiosepius* (Figure 5.8).

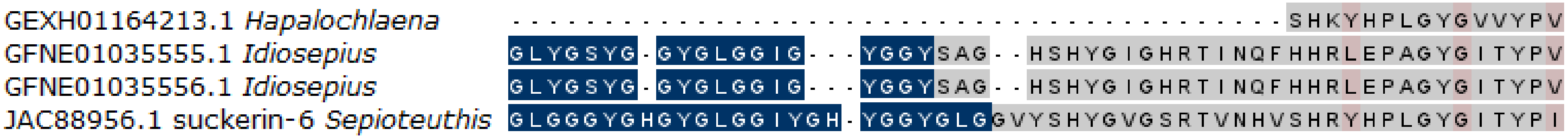

GEXH01164213.1 *Hapalochlaena* --SHKYHPLGYGVVYPV
GFNE01035555.1 *Idiosepius* GLYGSYG-GYGLGGIG---YGGYSAG--HSHYGIGHRTINQFHHRLEPAGYGITYPV
GFNE01035556.1 *Idiosepius* GLYGSYG-GYGLGGIG---YGGYSAG--HSHYGIGHRTINQFHHRLEPAGYGITYPV
JAC88956.1 suckerin-6 *Sepioteuthis* GLGGGYGHGYGLGGIYGH-YGGYGLGGVYSHYGVGSRTVNHVSHRYHPLGYGITYPI

Figure 5.8 - **Partial multiple sequence alignment including several suckerin peptide sequences**. The majority of the putative suckerin GEXH01164213.1 was uncharacterised and nearly identical to the uncharacterised regions in suckerin-6 from the bigfin reef squid (*Sepioteuthis lessoniana*), GFNE01035555.1 and GFNE01035556.1 from *Idiosepius*. Part of this is shown on top in grey. The bottom three suckerins exhibit suckerin-characteristic glycine-rich repeats (blue).

5.4.4 Expression Quantification

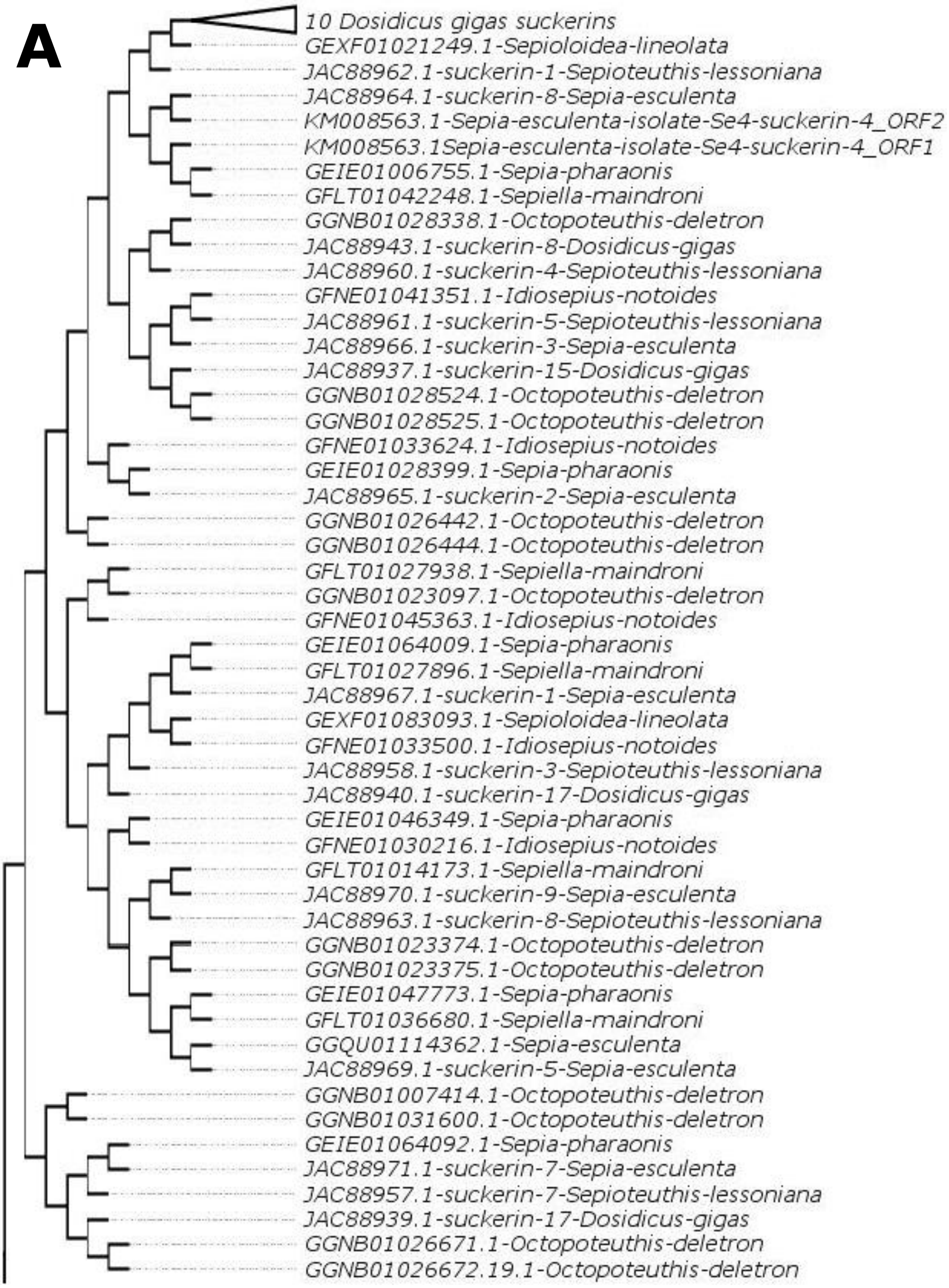
For the six transcriptomes thought to contain novel non-pairwise homologous suckerins (*Idiosepius, Octopoteuthis, Sepia pharaonis, Sepiella, Sepioloidea* and *Hapalochlaena*) expression quantification was carried out using RSEM. The expression levels of the cephalopods' actin homologs were recorded as housekeeping controls (Table 5.5). Very few putative suckerins showed zero expression levels (measured in transcripts per million). Several transcripts showed expression below 1 TPM which does not appear biologically relevant, especially considering that their corresponding actin expression was thousands of times higher. However, several putative suckerin transcripts appear to show much higher expression, especially considering that corresponding assembled actin transcripts are much longer than suckerin transcripts.

Table 5.5 – **Expression quantification of putative suckerins in transcripts per million (TPM).**

Organism	Source of RNA	Transcript	Length (bps)	TPM
Hapalo-chlaena	Heart	Actin	3873	5018
		GEXH01164213.1	580	0.40
Idiosepius	Adhesive gland	Actin	2173.00	9454
		GFNE01030216.1	801	0.60
		GFNE01033624.1	1305	0.17
		GFNE01035555.1	819	0.46
		GFNE01035556.1	1255	0.60
		GFNE01036157.1	811	0.10
		GFNE01041351.1	971	0.47
		GFNE01045363.1	571	0.00
Octopo-teuthis	Tentacles	Actin	1874	3186
		GGNB01007414.1	742	116
		GGNB01023097.1	661	5.60
		GGNB01023374.1	1664	0.78
		GGNB01023375.1	1633	4.59
		GGNB01025375.1	605	95
		GGNB01028338.1	815	34
		GGNB01028524.1	1012	13
		GGNB01028525.1	985	2.48
		GGNB01035290.1	658	87
		GGNB01035293.1	472	241
		GGNB01035295.1	792	0.00
		GGNB01035297.1	855	56
		GGNB01035299.1	717	42
		GGNB01035300.1	796	2.76
		GGNB01035303.1	476	0.00
		GGNB01035304.1	654	1.07
Sepia pharaonis	Whole-body	Actin	1903	25525
		GEIE01046349.1	731	285
		GEIE01047773.1	532	143
		GEIE01051091.1	509	52.70
Sepiella	Eyestalk, peduncle, tentacle, gill, muscle and ovary	Actin	2160	12351.35
		GFLT01027896.1	1150	1497.90
		GFLT01027938.1	1774	1008.35
		GFLT01036680.1	632	564.77
Sepiolo-idea	Mantle	Actin	1355	6081.60
		GEXF01021249.1	341	0.17
		GEXF01025591.1	363	0.30

5.4.5 Phylogenetic Analysis of Suckerin Sequences

Phylogenetic analysis was carried out to identify the relationships between both the novel and previously known members of the suckerin family. The 32 novel HMM-identified suckerins, the 39 previously known suckerins and 18 additional BLAST-identified novel suckerins were aligned based on their amino acid sequence. Using the MEGAX software a maximum likelihood tree was constructed (with the Whelan and Goldman substitution model) (Figure 5.9). This approach was chosen because it accounts for numerous homoplasy events (which may bring about apparent sequence similarity without shared descent) that are expected to have arisen over the many millions of years of evolution. All of the novel suckerin candidates have some degree of similarity to the previously known suckerins according to the phylogenetic analysis. For this 500 bootstrapping replicates were used. This number ensures that 500 different possible tree topologies are taken into consideration and that the most representative tree is constructed finally. Based on the relationship within this tree, several broad clades can be seen. These clades may represent clusters of orthologous suckerins whose functions may be the similar across species.


A
10 Dosidicus gigas suckerins
GEXF01021249.1-Sepioloidea-lineolata
JAC88962.1-suckerin-1-Sepioteuthis-lessoniana
JAC88964.1-suckerin-8-Sepia-esculenta
KM008563.1-Sepia-esculenta-isolate-Se4-suckerin-4_ORF2
KM008563.1Sepia-esculenta-isolate-Se4-suckerin-4_ORF1
GEIE01006755.1-Sepia-pharaonis
GFLT01042248.1-Sepiella-maindroni
GGNB01028338.1-Octopoteuthis-deletron
JAC88943.1-suckerin-8-Dosidicus-gigas
JAC88960.1-suckerin-4-Sepioteuthis-lessoniana
GFNE01041351.1-Idiosepius-notoides
JAC88961.1-suckerin-5-Sepioteuthis-lessoniana
JAC88966.1-suckerin-3-Sepia-esculenta
JAC88937.1-suckerin-15-Dosidicus-gigas
GGNB01028524.1-Octopoteuthis-deletron
GGNB01028525.1-Octopoteuthis-deletron
GFNE01033624.1-Idiosepius-notoides
GEIE01028399.1-Sepia-pharaonis
JAC88965.1-suckerin-2-Sepia-esculenta
GGNB01026442.1-Octopoteuthis-deletron
GGNB01026444.1-Octopoteuthis-deletron
GFLT01027938.1-Sepiella-maindroni
GGNB01023097.1-Octopoteuthis-deletron
GFNE01045363.1-Idiosepius-notoides
GEIE01064009.1-Sepia-pharaonis
GFLT01027896.1-Sepiella-maindroni
JAC88967.1-suckerin-1-Sepia-esculenta
GEXF01083093.1-Sepioloidea-lineolata
GFNE01033500.1-Idiosepius-notoides
JAC88958.1-suckerin-3-Sepioteuthis-lessoniana
JAC88940.1-suckerin-17-Dosidicus-gigas
GEIE01046349.1-Sepia-pharaonis
GFNE01030216.1-Idiosepius-notoides
GFLT01014173.1-Sepiella-maindroni
JAC88970.1-suckerin-9-Sepia-esculenta
JAC88963.1-suckerin-8-Sepioteuthis-lessoniana
GGNB01023374.1-Octopoteuthis-deletron
GGNB01023375.1-Octopoteuthis-deletron
GEIE01047773.1-Sepia-pharaonis
GFLT01036680.1-Sepiella-maindroni
GGQU01114362.1-Sepia-esculenta
JAC88969.1-suckerin-5-Sepia-esculenta
GGNB01007414.1-Octopoteuthis-deletron
GGNB01031600.1-Octopoteuthis-deletron
GEIE01064092.1-Sepia-pharaonis
JAC88971.1-suckerin-7-Sepia-esculenta
JAC88957.1-suckerin-7-Sepioteuthis-lessoniana
JAC88939.1-suckerin-17-Dosidicus-gigas
GGNB01026671.1-Octopoteuthis-deletron
GGNB01026672.19.1-Octopoteuthis-deletron

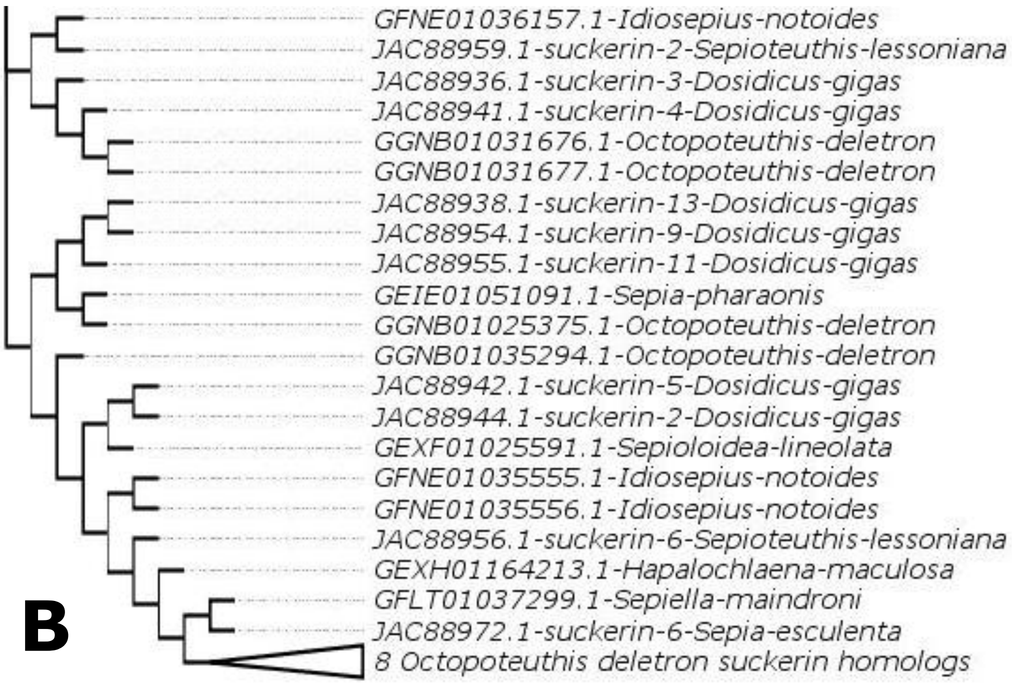

GFNE01036157.1-Idiosepius-notoides
JAC88959.1-suckerin-2-Sepioteuthis-lessoniana
JAC88936.1-suckerin-3-Dosidicus-gigas
JAC88941.1-suckerin-4-Dosidicus-gigas
GGNB01031676.1-Octopoteuthis-deletron
GGNB01031677.1-Octopoteuthis-deletron
JAC88938.1-suckerin-13-Dosidicus-gigas
JAC88954.1-suckerin-9-Dosidicus-gigas
JAC88955.1-suckerin-11-Dosidicus-gigas
GEIE01051091.1-Sepia-pharaonis
GGNB01025375.1-Octopoteuthis-deletron
GGNB01035294.1-Octopoteuthis-deletron
JAC88942.1-suckerin-5-Dosidicus-gigas
JAC88944.1-suckerin-2-Dosidicus-gigas
GEXF01025591.1-Sepioloidea-lineolata
GFNE01035555.1-Idiosepius-notoides
GFNE01035556.1-Idiosepius-notoides
JAC88956.1-suckerin-6-Sepioteuthis-lessoniana
GEXH01164213.1-Hapalochlaena-maculosa
GFLT01037299.1-Sepiella-maindroni
JAC88972.1-suckerin-6-Sepia-esculenta
8 Octopoteuthis deletron suckerin homologs
B

Figure 5.9 – **Phylogenetic tree of newly discovered and previously known suckerin amino acid sequences**. The tree was constructed using the Maximum Likelihood approach and the Whelan and Goldman substitution model (Whelan and Goldman, 2001). The tree with the highest log likelihood (-31823.77) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the JTT model, and then selecting the topology with superior log likelihood value. This analysis involved 89 amino acid sequences. There were a total of 1019 positions in the final dataset. Evolutionary analyses were conducted in MEGA X (Kumar et al., 2018b). The newly discovered suckerin accession

numbers start with “G” and the previously described start with the letters “J” or “A”. The values at each node represent the bootstrap support percentages. The tree is split into two subtrees (A and B) to account for space.

5.5 Discussion

Overall the pipeline described here successfully discovered and analysed several putative spidroin and suckerin sequences that up to now have been mostly outside the scope of pairwise-similarity assisted annotation. This approach is independent of genome availability and is applicable for transcriptomes from a variety of sources. Most importantly this method enables the identification of domains that are tandemly repeated in proteins even if they have lost much of their apparent sequence homology.

Genome-guided transcriptome assemblies are greatly affected by the completeness of the genome. More complete genomes are expected to yield larger, better-assembled contigs. For the *Parasteatoda* genome, 1,445,396,121 basepairs were represented by 16,533 individual scaffolds (Schwager et al., 2017) whereas for *Latrodectus* 1,233,806,489 basepairs were represented by 161,595 scaffolds (GenBank assembly accession: GCA_000697925.2). This large difference in genome completeness may have been the major reason for the difference in quality of the two genome-guided transcriptome assemblies. The use of more complete genomes could potentially increase the number of complete novel spidroins to be discovered.

While the profile HMM approach can find very distant sequence homologs it is unable to map together two spidroin fragments that belong to the same transcript (Jorda and Kajava, 2009). The sequences identifiable as full spidroins by HMMER represent only a fraction of the total possible spidroins that are annotated (Table 5.4). These differences are part of the reason why pairwise sequence alignments are still irreplaceable in domain annotation.

Interestingly, the sequence of GEWZ01016889.1 from *Nephilengys* was characterised as a spidroin by its termini but its central domain could not be identified as that of a spidroin by either HMMER or BLAST. Manual identification of repeat patterns was needed. GEWZ01016889.1 most resembles MaSp2 but is substantially different from other MaSp2 sequences. This may suggest a new function for this gene but it is difficult to speculate at this stage.

The presence of signal peptides in GAZS01047594.1, GGRD01219347.1 and TRINITY_DN1767 further supports their annotation as spidroins. Spidroins are typically secreted into the lumen of the silk gland, a process which is ultimately regulated by the presence of a signal peptide (Rising and Johansson, 2015).

It is expected for some transcripts to have low or zero expression values due to incomplete assembly or because they do not recruit both ends of paired-end reads. This type of low expression is artificial and not due to natural expression variation. However, it is estimated that biologically relevant transcripts will be well assembled so GAZS01040257.1 and GAZS01047594.1 may not be biologically relevant (Haas et al., 2013).

Regions with increased coverage relative to the rest of the transcript were found in GEWZ01016889.1 and DN1767. It is possible that these differences are not caused by natural expression variance and are actually due to erroneous Trinity assembly. This has been reported in previous attempts at spidroin discovery (Huang et al., 2017). In spidroin transcripts, the presence of repeats in these regions is expected (Kono et al., 2019). Furthermore, Trinity is suboptimal in repeat region determination since it may resolve identical reads as belonging to identical (not adjacent) parts of the mRNA or as sequential repeats. This ambiguity is a major caveat in the use of Trinity for repeat sequence assembly (Kono et al., 2019; Lima et al., 2017). Whenever a genome is available, more suitable assembly algorithms may be used.

The levels of the housekeeping protein RPL13 may be artificially high when measuring expression from whole abdomens or whole bodies relative to samples from pooled silk glands only. RPL13 is expected to be expressed in every tissue in the body but spidroins are not (Scharlaken et al., 2008).

Previously described spidroins expressed in the same tissues that are investigated are often used to compare the expression of newly identified ones (Kono et al., 2019). Additionally, for this analysis, a selection of transcriptomes from various tissue samples was selected. Silk-gland-only specimens may be omitting rare events where spidroins are expressed in other tissues (Rind et al., 2011). The analysis of expression was meant only as a rough estimate of which spidroins are strongly transcribed and is not meant to be conclusive.

Technologies such as the PacBio SMRT enable the direct (but error-prone) sequencing of very long single RNA molecules. This has been used before for the direct sequencing of *Argiope trifasciata* AgSp RNA. Long reads could also serve as scaffolds to use in the assembly of short Illumina reads (Stellwagen and Renberg, 2019). Such transcriptomes were expected to be particularly well suited for analysis using the pipeline described here. Using the *Cyrtophora* and *Argyroneta* long-read transcriptomes, however, did not yield any non-pairwise homologous full spidroin transcripts (with both termini). It appears that the BLAST screening step has successfully collected all full spidroins and none were left for profile HMM identification. The fragments that were identified with HMMER were still incomplete spidroins. It appears that long-read sequencing has been unable to sequence full spidroins and has produced numerous fragments (ones still longer than those from short-read sequencing).

Long-read sequencing is expected to provide sequences for all spidroins in an organism. However, if all transcribed spidroins have been identified previously, the pipeline presented here would not be able to detect any new spidroins. This could be because the spidroins of *Argyroneta* and *Cyrtophora* are not distinct enough from previously described ones so they are detectable entirely by pairwise homology algorithms (e.g. BLAST family). A similar observation has been made in the *A. ventricosus* transcriptome (Zhou et al., 2021).

Short spidroins with a small repeat domain have been described before (Starrett et al., 2012). If these sequences are encoded in spiders, they may represent more ancestral spidroins or ones which have underwent large-scale deletions in the central domain.

Overall, the pipeline identified almost 6 900 spidroin transcripts (complete or fragments) and out of these only nine were pairwise homologous to known spidroins. This demonstrates that the chosen parameters were selective enough to find pairwise-homologous spidroins but not so much that spidroins would be considered pairwise homologous when they were not.

As for the suckerins, the uncharacterised regions within them may actually represent a previously unknown functional domain. This seems to be evolutionarily conserved since it is present in the distant octopus lineage. It would be intriguing to identify its function. This will additionally help to explain why it is conserved and why it may be expressed in octopus heart tissue. This could also shed further light on the evolutionary history of the suckerins and any prior functions before sucker ring teeth.

The wide range of lengths in the newly discovered suckerins is consistent with what has been observed previously (between 5-57 kDa). Based on this variability in length and repeat number it is difficult to judge whether the novel suckerins are complete or not. Unlike spidroin proteins where terminal domains are well-known, for suckerins there is much less certainty. The presence of signal peptides is the only indicator that a sequence is possibly at the 5′ end of the coding region (Guerette et al., 2014). It is feasible that several of the identified sequences are actually fragments of the same gene. However, the absence of available cephalopod genomes, especially for most of the species analysed here, precludes the possibility of mapping the sequences to a genome. While long repeated sequences are expected to adversely affect *de novo* assembly reliability (Lima et al., 2017), shorter repetitive genes such as the suckerins may be incorrectly assembled less often simply because there is less opportunity for error by algorithms such as that of Trinity. Therefore, this pipeline is particularly suitable for the suckerin gene family.

The data presented here is entirely based on RNA-seq. This only represents the suckerin sequences that are expressed in the tissues sampled. There may be a large number of suckerin pseudogenes left to be uncovered from genomic data sets. Also, the low expression of most suckerins relative to actin could be explained by the fact that actin is expressed in all cells of the body whereas suckerins are expected to be confined to just the arm suckers.

The *Dosidicus* transcriptome did not show a single suckerin sequence even though 21 such sequences are expected based on previous research (Guerette et al., 2014). However, these results are less surprising when considering that the assembled transcriptome entry came from internal organ RNA samples. Conversely, the large number of the *Octopoteuthis* suckerins identified is reminiscent of the large number of such sequences in the *Dosidicus* genome (Guerette et al., 2014). The numbers of newly identified suckerins for the rest of the species described here are more reminiscent of the suckerin repertoires in *Sepia esculenta* and *Sepioteuthis lessoniana*. This apparent completeness may be attributable to the fact that the *Octopoteuthis* transcriptome has been collected specifically from tentacle tissue.

Especially interesting was the discovery of a potential suckerin-like sequence within the transcriptome of an octopus – *Hapalochlaena*. However, the *Hapalochlaena* transcriptome used in this analysis came from the heart tissue which may suggest an unrelated function of suckerins in their evolutionary history. This particular sequence did not exhibit the same repeat organisation as in other suckerin peptides and may also hint at a separate function. This suckerin homolog showed a large degree of peptide sequence similarity with the uncharacterised regions in the previously described suckerin-6 from *Sepioteuthis lessoniana* as well as GFNE01035555.1 and GFNE01035556.1 from *Idiosepius* (Figure 5.8). The *Hapalochlaena* suckerin was not expressed very highly when compared to the corresponding actin expression. The presence of a suckerin in non-tentacle tissues additionally suggests that tentacles should not be the only part of the squid anatomy that is analysed for the presence of suckerins.

Additionally, based on the grouping of this sequence with other suckerins, it is very probable that it is part of the suckerin family. Even if it is not expressed highly, the presence of such a gene in the octopus lineage is plausible. The origin of suckerins, 354 million years ago (Guerette et al., 2014), predates the squid-octopus divergence from 336 million years ago (Hedges et al., 2015). Additionally, a small number of ancestral suckerins is expected based on previous research which identified six large clades of suckerins (Guerette et al., 2014). The similarity of the Hapalochlaena suckerin to the squid suckerins from one of these clades additionally suggests that they were present before the squid-octopus divergence. Only three octopus species were analysed for the presence of suckerins. Potentially many more suckerins remain to be discovered in this group but very few octopus transcriptomes are publicly available.

The peptide sequence from this suckerin has been conserved very strongly for at least 354 million years. The lack of characteristic suckerin repeats in the *Hapalochlaena* suckerin may suggest an unrelated function to SRT composition. Conversely, the absence of sucker ring teeth in octopuses does not exclude the possibility of an octopus suckerin. Octopuses are known to have a chitinous cuticle on their suckers which serves to protect them (Girod and 1884; Naef, 1923; Nixon et al., 1977; Packard and 1988). Additionally, they have a large number of small chitinous pegs on their suckers called "denticles" which serve to improve attachment (Kier et al., 1990; Nixon et al., 1977). It is feasible that the chitinous denticles in octopuses are homologous to the sucker ring teeth.

Except for one sequence from *Octopoteuthis* (GGNB01007414.1), the putative suckerins appear to fall into one of six suckerin clades. In many instances, a novel suckerin is grouped more closely with a previously known one. It is difficult to speculate further on the exact mechanical property of each putative suckerin, however, as most of the known suckerins have not been studied in great detail.

The phylogenetic relationships between these six clades were observed to be somewhat different compared to what has been previously reported in Guerette et al. (2014). Likely, the addition of new sequence information in the form of novel peptide sequences has changed the topology of the phylogenetic relationships somewhat. The tree reported here was made using the maximum likelihood approach which is proposed to be more reliable than the neighbour-joining method that has been illustrated in the previous phylogenetic analysis (Tateno et al., 1994; Yoshida and Nei, 2016).

When the timescale of sequence divergence is hundreds of millions of years, multiple homoplasy events can occur. This effectively creates sequence similarity in biological sequences that are not closely related. The maximum likelihood tree-building method uses substitution models (such as the Whelan and Goldman model used here) which estimate how likely such events are to occur and corrects tree topology accordingly. The neighbour-joining tree-building method only relies on sequence similarity for its estimations (Brandley et al., 2009; Hasegawa and Fujiwara, 1993).

As noted before, Trinity deals well with assembling divergent repeat sequences but struggles with repeat regions. In some assembled transcripts identical repeats may have been perceived to derive from different transcript molecules while they may in fact come from adjacent repeats within the same mRNA. Isolating the proteins and performing mass spectrometry on them would provide the exact size of each putative suckerin and resolve such ambiguity.

A recent study identified two sets of suckerin genes from the genomes of two squid species that are not described here (Albertin et al., 2022). Their analysis of an octopus genome could not identify an octopus suckerin. Possibly this is due to their reliance on BLAST tools and the Pfam database which are unable on their own to identify more distantly-related suckerins.

In short, the work described in this chapter aimed to identify repetitive protein homologs without relying on sequence pairwise similarity. With the repetitive structure of MaSp proteins in consideration, a pipeline was assembled from publicly available software primarily in the Linux environment. The pipeline accepts assembled transcriptome datasets and some prior information on homologous protein families. It constructs robust profile HMMs and yields homologous peptide sequences (translated *in silico*) as the output. Using this, several spidroin and suckerin homologs were found in spiders and cephalopods respectively. The new sequences identified range widely in their primary structure but nonetheless share much with their gene families. These new sequences greatly expanded the number of known suckerins while also helping to identify homologous sequences which lack the actual repeats further expanding what may be considered a suckerin.

6 Discussion and Future Work

6.1 Spidroin Rolling Circle Translation

The difficulty in expressing recombinant spidroins in mammalian cells has been known in advance (Heidebrecht and Scheibel, 2013; Jones et al., 2015; Whittall et al., 2021; Xu et al., 2007). The work described here was intended as a proof of concept for the potential to generate the repetitive spidroin from the RCT method. After analysing the results of this work it has become apparent that mammalian cells are not the most suitable hosts for recombinant spidroin expression, especially in the RCT context. They are not easily metabolically engineered to meet the amino-acyl-tRNA demands of the nascent spidroin. Mammalian cells are much less cost-effective and much more difficult to culture in large batches relative to *E. coli*. However, a major benefit of mammalian recombinant expression is its ability to add variable post-translational modifications (dos Santos-Pinto et al., 2016; dos Santos-Pinto et al., 2014; Kannicht et al., 2013; Satapathy et al., 2020). All of these factors taken together suggest that alternative expression hosts will need to be used in spidroin RCT. The possible steps that can be taken to rescue some spidroin RCT from mammalian cells do not justify the associated difficulties when systems such as *E. coli* and *B. mori* are available. The backsplicing approach would be equally suitable for *B. mori* RNA circularisation. The splicing in insects does not differ significantly from that of mammals (Rogozin et al., 2003). While the reverse complementary intron can remain unchanged, some consideration of splice site sequences is in order. Similar tools to HSF exist which can predict the splice site efficiency in a wide variety of eukaryotes including insects (Reese et al., 1997; Scalzitti et al., 2021).

Recently, research has been published that also implemented the RCT approach in the expression of spidroins. Lee *et al.* (2021) were able to replicate the circular RNA translation approach in *E. coli* through group I intron *in vivo* RNA circularisation. Their splicing and RNA circularisation could not and did not involve backsplicing. Their translation initiation occurred via an optimised ribosomal binding site. They were able to produce a multimeric sequence of GFP that was then separated into monomers by protease cleavage. Their attempt at spidroin translation was less successful, however. Their largest spidroin did not exceed 90 kDa. This was thought to occur due to the buildup of insoluble protein within the cell which decreases the availability of spidroins after extractions and acts as a stress on the cells themselves. Very similar work by Liu *et al*. (Liu et al., 2022) was also carried out recently. They again used *E. coli* as the expression host. Through a group I intron they were able to produce recombinant spidroin circRNA *in vivo*. Very similarly they achieved spidroin lengths at 90-110 kDa.

The advantages of *E. coli* in scalable cell culture, high recombinant protein yields and robustness are not to be underestimated. However, if this host is going to be used in the future important solutions need to be implemented to facilitate the secretion of the nascent spidroin and to meet the metabolic needs of amino-acyl-tRNA complexes. Certain strategies for the former do exist but similar to the signal peptide in mammals they leave an added peptide sequence in subsequent translation cycles (Kleiner-Grote et al., 2018).

For the purpose of fibre formation, the native N-terminus appears very important as terminus-terminus interactions are directly responsible for bringing two spidroins together (Ries et al., 2014). It may be necessary to include this domain within the IORF. Presumably, this would lead to long artificial spidroins but interspersed with N-termini. It is unclear how the N-terminus would interfere with the fibre structure and how the repetitive region would interfere with terminus-terminus interactions.

There are approaches in repetitive protein production which rely on enzymatic or spit intein-based methods. In the case of recombinant spidroins, large peptide sequences have been produced before through a split-intein synthesis with the spidroin motifs repeated up to 192 times (Bowen et al., 2018). Using the RCT approach, this protein length could be replicated after 48 cycles of translation with the 4GA IORF. It is currently unknown if such protein lengths are attainable with IORF spidroin translation.

So far there are no cell lines available from either spiders' spidroin glands or *B. mori's* (Lee et al., 2012; Whittall et al., 2021). However, transgenic silkworms may be a suitable host for *in vivo* RCT recombinant spidroin expression. Furthermore, these animals have a history of successful genetic manipulation in the lab so are already suited to transgenic experiments and such techniques are well established (Chen et al., 2018; Zhang et al., 2019b).

The circRNA-producing plasmids described here may be adapted to a *B. mori* host and transfected *in vivo*. This may produce a recombinant spidroin and eliminate the need for transfecting long DNA fragments into the animal genome. However, even successful RCT spidroins in *B. mori* are unlikely to match the properties of the previously described, termini-containing recombinant spidroins (Zhang et al., 2019b). For expression to be successful, the CMV promoter would need to be replaced with an endogenous or exogenous insect-specific promoter (e.g. the *Autographa californica* multiple nuclear polyhedrosis virus (AcMNPV)) (Wu et al., 2000; Xu et al., 2018). A cell line derived from spider silk glands would largely eliminate the need for spidroin transfection since the spidroin genes are endogenously present.

Alternatively, intein sequences can be introduced into the IORF. An intein sequence surrounding the m^{6}A and SP region could enable the rest of the peptide sequence (in this case the GA repeats) to become ligated together, essentially excising any non-spidroin peptide sequences (Sharma et al., 2006; Wood and Camarero, 2014). This could potentially negate many of the disadvantages of RCT by removing the majority of peptide sequences that may otherwise compromise recombinant protein structure or function. Furthermore, this process can be regulated in the context of recombinant expression.

Some natural spider silk also possesses a small amount of a non-spidroin protein termed SpiCE (spider silk constituting element). This has been shown to have a role in the silk fibres' remarkable mechanical properties (Kono et al., 2021). Perhaps the addition of such proteins to a solution of recombinant spidroins would enable the production of even more durable fibres.

6.2 Transcriptomic Identification of Spidroins and Suckerins

The computational pipeline described here was designed to identify repetitive protein homologs for established protein families. However, it may also be useful in cases where only one unique example of a repetitive protein is found, when it is unknown if a whole gene family exists (e.g. when a repeat protein is found for the first time). In such situations, the pipeline could omit the secondary profile HMM step and immediately proceed after the construction of the initial profile HMM. This would provide a very flexible method of detecting somewhat derived sequence homologs. The utility of BLAST cannot be overshadowed, however. Its speed and access to millions of datasets are still unparalleled.

This research has focused on spidroin sequences which contain both terminal domains. However, it is not impossible that among the thousands of identified spidroin sequences, there is a new, biologically relevant protein that does not contain termini and only shares homology with known spidroins over the repeat domain. Such a spidroin may have quite a different function to what is currently known but more investigation would be needed to

It is not currently known what kinds of effects the non-spidroin circRNA regions could have on their respective recombinant protein. It would be interesting to investigate this, particularly in the case of the spidroin whose regions have important structural roles and may or may not become disrupted. It may also be useful to add the spidroin terminal domains to the IORF. At least in the case of the N-terminus, this would ensure that it is capable of dimerising with other spidroin monomers, enabling more silk-like properties (Askarieh et al., 2010; Sarr et al., 2022). However, it is impossible to ensure that a C-terminus is found at the end of the IORF-translated spidroin. Too little is known about its translation termination. Additionally, it is expected that iterating te terminal domains inside the spidroin every cycle of translation would lead to a spidroin of inferior quality.

Hybrid peptides have been produced with a suckerin central domain and spidroin termini (Ramos et al., 2021). This could potentially result in a very interesting peptide with a combination of suckerins' and spidroins' properties. However, this has only been used for the production of drug delivery nanocapsules. More difficult-to-construct structures like fibres or 3D printed scaffolds have not been demonstrated.

A much larger and more repetitive suckerin molecule could be expressed. Since native suckerins are typically very short, such a high molecular weight suckerin might exhibit some elastic properties similar to spidroin proteins in spiders (Hiew and Miserez, 2017; Hiew et al., 2017; Miserez et al., 2009). It would be interesting to implement the RCT approach in the expression of such long suckerins. As with the spidroins, suckerins may be naturally amenable for expression within silkworms. However,

The Eastern Oyster (*Crassostrea virginica*) has been reported to contain 22 genes which similarly possess poly-alanine tracts and tandem repeat peptide sequences. These proteins have been implicated in shell formation but have not been studied in great detail (Zeng and Guo, 2022). The computational pipeline presented here may be of particular use in identifying more members of this potential protein family and expanding the knowledge about these proteins. Many other repetitive protein families exist and these may further benefit from homolog discovery and recombinant expression through the approaches outlined in this work.

6.3 Conclusion

The overarching aim of this project was to facilitate the recombinant expression of spidroins. The approach chosen was continuous translation of circRNA. A construct was designed which allows a short spidroin-encoding sequence to be transcribed, circularised by backsplicing, recognised by a ribosome and translated continuously due to the absence of in-frame stop codons. Subsequent experiments were performed that sought to confirm the circularisation and translation of the recombinant spidroin RNA. While RNA circularisation was supported by the results, protein translation could not be observed in any form. Several possibilities for this were discussed and the potential events of amino-acyl-tRNA depletion and programmed -1 ribosomal frameshifts were examined. These approaches could not produce a conclusive answer for the lack of spidroin RCT and many other possible explanations for this observation still exist.

In line with the aim of improving recombinant spidroin expression, a pipeline was designed to identify new spidroin homologous sequences that may be of use in biotechnology. This pipeline relies on the repeat regions present in the spidroin central domain and leverages them for the construction of profile HMMs. The pipeline was implemented on several publicly available transcriptomes and yielded thousands of sequences, six of which were considered both unknown and complete. The pipeline was additionally tested in a separate context – the cephalopod suckerin proteins. As a result, a large number of suckerins were discovered. This includes even a previously unexpected octopus suckerin. This unique sequence provides valuable information about the history of the suckerin protein family. Some of the new peptides discovered may be suitable for biotechnological applications and further investigation for recombinant expression.

7 Appendix

7.1 Supplemantary Tables

Table S1 – **Primer and insert sequences used in the cloning of the spidroins plasmid library.**

Name	DNA Sequence
pcDNA3.1 fw	CGCAAATGGGCGGTAGGCGTG
pcDNA3.1 rev	TAGAAGGCACAGTCGAGG
RCI_screen fw	CTCGATATCGGAAGCCACTG
SP4GA_screen fw	TTGATCCTCACATTCGTCGC
4GA_screen_fw	CTGGGACAAGGCGGTTATG
IVA_fw	CATCACCATCACCATCACCAAGGTGGGTACGGGCAAGGGAC
IVA_rev	GTGATGGTGATGGTGATGTGCGAGAGTGCGGCAGCGACGAATG
6His_screen_fw	GCACATCACCATCACCATCA
RCI	GAAGTGCCATTCCGCCTGACCTAGCGCTTACCCGGTTTACGACCTCTGGTTACCCCTCGGGGCTGTGCTGTGGAAGCTAAGTCCTGCCCTCGATATCGGAAGCCACTGGGGACAGCCAGGCCAGACGGGGGACATGCAGAAAGTGCAAAGAACACGGCTAAGTGTGCTGGGGTCTTGGGATGGGGAGTCTGTTCAGACCTACTGTGCACCTACTTAATACACACTCCAAGGCCGCTTTACACCAGCCTCATGGCCTTGTCACACGAGCCAGTGTTAGTACCTACACCCACAACACTGTTCTAGACCGGTGAGCTGTCCAGTCGCGTTCAAGGCTAGGTGGAGGCTCAGTG
2GA short	GAAGTGCCATTCCGCCTGACCTAAGCTTACAGTGTTGTGGGTGTAGGTACTAACACTGGCTCGTGTGACAAGGCCATGAGGCTGGTGTAAAGCGGCCTTGGAGTGTGTATTAAGTAGGTGCACAGTAGGTCTGAACAGACTCCCCATCCCAAGACCCCAGCACACTTAGCCGTGTTCTTTGCACTTTCTGCATGTCCCCCGTCTGGCCTGGCTGTCCCCAGTGGCTTCCCCAGTGTGACATGGTGTATCCTTTCTTTGCAGAAGGATCCTGGACTAAAGCGGACTTGTCTCGAGATGCAAGGTGGTTACGGGCAAGGGACGGGCTCTAGTACAGCGGCAGCCGCAGCCGCGGCGGCCGCTGCTGCCGCTTCTGGACAAGGCGGTCAAGGCGGCCAAGGTCAAGGCGGATACGGCCAGGGAGCAGGAATATCCGCTGCTGCAGCCGCCGCTGCCGCCGCCGCCGCTGCGGTGAGTGGAGACTGTCTCCCGGCTCTGCCTGACATGAGGGTTACCCCTCGGGGCTGTGCTGTGGAAGCTAAGTCCTGCCCTCGATATCGGAAGCCACTGGGGACAGCCAGGCCAGACGGGGGACATGCAGAAAGTGCAAAGAACACGGCTAAGTGTGCTGGGGTCTTGTCTAGAAGGCTAGGTGGAGGCTCAGTG

Table S1 (continued) – **Primer and insert sequences used in the cloning of the spidroins plasmid library**.

Name	DNA Sequence
SP4GA	GAAGTGCCATTCCGCCTGACCTTTGGCGAATCCCTGGATGAGTAATAAACAGCCATAGGAGGGTCTCGTGACTTGCAGAAGGATCCTGGACTAAAGCGGACTTGTCTCGAGATGAATCCCCTTTTGATCCTCACATTCGTCGCTGCCGCACTCGCACAAGGTGGGTACGGGCAAGGGACGGGCTCTAGTACTGCAGCAGCGGCCGCAGCTGCCGCTGCTGCTGCTGCCGGACAGGGTGGCCAGGGTGGCTACGGTGGTCTCGGTCAGGGTGGATATGGACAAGGGGCCGGTAGTAGTGCAGCCGCCGCTGCTGCCGCTGCAGCTGCTGCAGCAGCAGGTCAAGGCGGTCAAGGTGGTTATGGTGGCCTGGGACAAGGTGGATATGGCCAGGGCGCTGGGTCAAGTGCTGCAGCTGCGGCTGCCGCGGCCGCAGCTGCGGCAAGCGGACAAGGCGGTCAAGGTGGACAAGGCCAAGGTGGATACGGACAAGGCGCAGGAATAAGTGCCGCCGCCGCCGCAGCTGCAGCAGCCGCCGCTGCGGTGAGTGGAGACTGTCTCCCGGCTCTGCCTGACATGAGGGTTACCCCTCGGGAGGCTAGGTGGAGGCTCAGTG
4GA	GAAGTGCCATTCCGCCTGACCTTTTCTTTGCAGAAGGATCCTGGACTAAAGCGGACTTGTCTCGAGATGCAAGGCGGATACGGGCAAGGGACGGGCTCTAGTACAGCCGCAGCAGCCGCGGCAGCAGCAGCTGCTGCGGCAGGACAGGGTGGGCAGGGCGGGTATGGCGGACTGGGCCAAGGTGGGTATGGTCAAGGAGCAGGTAGTTCTGCAGCAGCAGCTGCCGCTGCTGCCGCTGCCGCAGCTGCTGGCCAGGGCGGTCAGGGTGGATATGGTGGTCTGGGACAAGGCGGTTATGGACAAGGCGCTGGAAGCTCTGCTGCTGCTGCAGCCGCTGCAGCCGCTGCCGCTGCTAGTGGACAAGGCGGACAAGGTGGCCAAGGTCAAGGCGGTTACGGTCAAGGTGCCGGAATTTCCGCAGCTGCCGCAGCAGCAGCAGCGGCAGCAGCCGCGGTGAGTGGAGACTGTCTCCCGGCTCTGCCTGACATGAGGGTTACCCCTCGAGGCTAGGTGGAGGCTCAGTG
SPFLAG4GA linear	GAAGTGCCATTCCGCCTGACCTCCCAAGCTGGCTAGCGTTTAAACTTAAGCTTAAGGATCCTGGACTAAAGCGGACTTGTCTCGAGATGAATCCCCTTTTGATCCTCACATTCGTCGCTGCCGCACTCGCAGACTATAAAGATGATGATGACAAGCAAGGTGGGTACGGGCAAGGAACAGGTTCCAGCACTGCAGCAGCGGCCGCAGCTGCCGCTGCTGCTGCTGCCGGACAGGGTGGCCAGGGTGGCTACGGTGGTCTCGGTCAGGGTGGATATGGACAAGGGGCCGGTAGTAGTGCAGCCGCCGCTGCTGCCGCTGCAGCTGCTGCAGCAGCAGGTCAAGGCGGTCAAGGTGGTTATGGTGGCCTGGGACAAGGTGGATATGGCCAGGGCGCTGGGTCAAGTGCTGCAGCTGCGGCTGCCGCGGCCGCAGCTGCGGCAAGCGGACAAGGCGGTCAAGGTGGACAAGGCCAAGGTGGATACGGACAAGGCGCAGGAATAAGTGCCGCCGCCGCCGCAGCTGCAGCAGCCGCCGCTTAAGCTTGTCTAGAGGGCCCGTTTAAACCCGCTGATCAGCCTCGACTGTGCCTTCTAGTTGTTATTTCCGGTTCGTCCATTCGTCCAGTACATAAAAGAGAGGCTAGGTGGAGGCTCAGTG

7.2 PIP Statement

Note to examiners:

This statement is included as an appendix to the thesis in order that the thesis accurately captures the PhD training experienced by the candidate as a BBSRC Doctoral Training Partnership student.

The Professional Internship for PhD Students is a compulsory 3-month placement which must be undertaken by DTP students. It is usually centred on a specific project and must not be related to the PhD project. This reflective statement is designed to capture the skills development which has taken place during the student's placement and the impact on their career plans it has had.

PIPS Reflective Statement

My 3-month placement took part in the Sheffield-based company IN-PART. It lasted from August to October 2021. My work was entirely web-based. I was engaged in connecting industry with universities in order to facilitate cooperation in research. More specifically I worked in two of the divisions within IN-PART.

One side of my work involved reading about patented technologies that IN-PART's university clients have developed. I was then responsible for identifying specific industry professionals (specifically those involved in research and development or technology licencing) and notifying them about available technologies for licencing. Work in IN-PART’s other division involved searching for and collecting large amounts of information about university-developed technologies and academic research. I was also responsible for compiling large reports with this information. Most of the topics that I dealt with fall into the broad areas of biology, medicine and engineering.

During my PIP I was able to identify hundreds of potential technologies and research opportunities that could be of interest to IN-PART’s industry clients. I was able to compile many such opportunities into large reports that are then sent to the clients. I also facilitated several direct interactions between industry and universities. This may have contributed to the potential eventual commercialisation of university-owned patents and technologies. In either case, the overall results of these activities could be financing for academic research, further development of a technology or technology licencing. The long-term outcomes of my work would be difficult to pinpoint but several products have been developed in the past as a consequence of IN-PART’s assistance.

The main skills that I developed in IN-PART are team work, analytical skills and decision-making. Most of the work at IN-PART is done in specific stages and usually different people are responsible for each stage. This means that my contribution is both contingent on and foundational for my colleagues’ work. Also, I had to familiarise myself with many different technologies and research opportunities quickly and accurately.

I experienced working in a business environment with carefully planned deadlines and allocated workloads. I had to cope with large amounts of information and to meticulously compile relevant information. Working in a customer-oriented environment also taught me to identify what is a priority and what needs to be done more urgently due to numerous shifting deadlines. Report-writing for IN-PART also taught me new skills in document design.

This placement has given me much insight into the current trends within the spheres of industrial R&D, particularly within biology. This will help me choose a career within more promising avenues of research. I also learned much about companies that are involved in various research fields that are of interest to me, giving me the opportunity to contact potential employers.

I have decided that purely desk-based work similar to what I did at IN-PART is not suitable for me at this stage. I will most likely be pursuing a career in a wet lab. Work in the private sector appeals to me relative to academia because it feels much more dynamic. Also, having more direction and feedback on my work will be beneficial for me and I believe I can find the right amount in industry. My PIP at IN-PART helped me understand my preferences.

References

Abe, N., Hiroshima, M., Maruyama, H., Nakashima, Y., Nakano, Y., Matsuda, A., Sako, Y., Ito, Y., and Abe, H. (2013). Rolling Circle Amplification in a Prokaryotic Translation System Using Small Circular RNA. Angew Chem-Int Edit *52*, 7004-7008.

Abe, N., Matsumoto, K., Nishihara, M., Nakano, Y., Shibata, A., Maruyama, H., Shuto, S., Matsuda, A., Yoshida, M., Ito, Y.*, et al.* (2015). Rolling Circle Translation of Circular RNA in Living Human Cells. Scientific Reports *5*, 9.

AbouHaidar, M.G., Venkataraman, S., Golshani, A., Liu, B.L., and Ahmad, T. (2014). Novel coding, translation, and gene expression of a replicating covalently closed circular RNA of 220 nt. Proceedings of the National Academy of Sciences of the United States of America *111*, 14542-14547.

Albertin, C.B., Medina-Ruiz, S., Mitros, T., Schmidbaur, H., Sanchez, G., Wang, Z.Y., Grimwood, J., Rosenthal, J.J.C., Ragsdale, C.W., Simakov, O.*, et al.* (2022). Genome and transcriptome mechanisms driving cephalopod evolution. Nat Commun *13*, 2427.

An, B., Hinman, M.B., Holland, G.P., Yarger, J.L., and Lewis, R.V. (2011). Inducing beta-Sheets Formation in Synthetic Spider Silk Fibers by Aqueous Post-Spin Stretching. Biomacromolecules *12*, 2375-2381.

Andersson, M., Chen, G.F., Otikovs, M., Landreh, M., Nordling, K., Kronqvist, N., Westermark, P., Jornvall, H., Knight, S., Ridderstrale, Y.*, et al.* (2014). Carbonic Anhydrase Generates CO2 and H+ That Drive Spider Silk Formation Via Opposite Effects on the Terminal Domains. PLoS Biol *12*.

Andrews, S. (2010). FastQC A Quality Control Tool for High Throughput Sequence Data (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

Arkhipkin, A.I., and Laptikhovsky, V.V. (2008). Discovery of the fourth species of the enigmatic chiroteuthid squid Asperoteuthis (Cephalopoda : Oegopsida) and extension of the range of the genus to the South Atlantic. J Molluscan Stud *74*, 203-207.

Armenteros, J.J.A., Tsirigos, K.D., Sonderby, C.K., Petersen, T.N., Winther, O., Brunak, S., von Heijne, G., and Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. Nature Biotechnology *37*, 420-+.

Arndt, T., Jaudzems, K., Shilkova, O., Francis, J., Johansson, M., Laity, P.R., Sahin, C., Chatterjee, U., Kronqvist, N., Barajas-Ledesma, E.*, et al.* (2022). Spidroin N-terminal domain forms amyloid-like fibril based hydrogels and provides a protein immobilization platform. Nat Commun *13*.

Aruffo , A. (2002). Transient Expression of Proteins Using COS Cells. Current Protocols in Molecular Biology.

Asakura, T., and Suzuki, Y. (2014). Encyclopedia of Polymeric Nanomaterials: Silk Fibroin (Berlin, Heidelberg: Springer).

Askarieh, G., Hedhammar, M., Nordling, K., Saenz, A., Casals, C., Rising, A., Johansson, J., and Knight, S.D. (2010). Self-assembly of spider silk proteins is controlled by a pH-sensitive relay. Nature *465*, 236-U125.

Ayoub, N.A., Garb, J.E., Kuelbs, A., and Hayashi, C.Y. (2013). Ancient Properties of Spider Silks Revealed by the Complete Gene Sequence of the Prey-Wrapping Silk Protein (AcSp1). Mol Biol Evol *30*, 589-601.

Ayoub, N.A., Garb, J.E., Tinghitella, R.M., Collin, M.A., and Hayashi, C.Y. (2007). Blueprint for a High-Performance Biomaterial: Full-Length Spider Dragline Silk Genes. Plos One *2*, 13.

Bai, S.M., Zhang, X.L., Lu, Q., Sheng, W.Q., Liu, L.J., Dong, B.J., Kaplan, D.L., and Zhu, H.S. (2014). Reversible Hydrogel-Solution System of Silk with High Beta-Sheet Content. Biomacromolecules *15*, 3044-3051.

Baklaushev, V.P., Bogush, V.G., Kalsin, V.A., Sovetnikov, N.N., Samoilova, E.M., Revkova, V.A., Sidoruk, K.V., Konoplyannikov, M.A., Timashev, P.S., Kotova, S.L.*, et al.* (2019). Tissue Engineered Neural Constructs Composed of Neural Precursor Cells, Recombinant Spidroin and PRP for Neural Tissue Regeneration. Scientific Reports *9*, 18.

Barrow, K.M., Perez-Campo, F.M., and Ward, C.M. (2006). Use of the Cytomegalovirus Promoter for Transient and Stable Transgene Expression in Mouse Embryonic Stem Cells. In Embryonic Stem Cell Protocols: Volume 1: Isolation and Characterization, K. Turksen, ed. (Totowa, NJ: Humana Press), pp. 283-294.

Bateman, A., Martin, M.J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E.H., Britto, R., Bursteinas, B.*, et al.* (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res *49*, D480-D489.

Becker, N., Oroudjev, E., Mutz, S., Cleveland, J.P., Hansma, P.K., Hayashi, C.Y., Makarov, D.E., and Hansma, H.G. (2003). Molecular nanosprings in spider capture-silk threads. Nature Materials *2*, 278-283.

Berezikov, E., Blinov, A.G., Scherbik, S., Cox, C.K., and Case, S.T. (1998). Structure and polymorphism of the Chironomus thummi gene encoding special lobe-specific silk protein, ssp160. Gene *223*, 347-354.

Bhagat, V., and Becker, M.L. (2017). Degradable Adhesives for Surgery and Tissue Engineering. Biomacromolecules *18*, 3009-3039.

Bi, X., and Liu, L.F. (1994). RecA-Independent and RecA-Dependent Intramolecular Plasmid Recombination - Differential Homology Requirement and Distance Effect. J Mol Biol *235*, 414-423.

Blench, R. (2009). A guide to the musical instruments of Cameroun: classification, distribution, history and vernacular names (Kay Williamson Educational Foundation).

Bochicchio, B., Pepe, A., and Tamburro, A.M. (2001). On (GGLGY) synthetic repeating sequences of lamprin and analogous sequences. Matrix Biol *20*, 243-250.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114-2120.

Bon, F.X. (1710). I. A discourse upon the usefulness of the silk of spiders. by Monsieur Bon, President of the Court of Accounts, Aydes and Finances, and President of the Royal Society of Science at Montpellier. Communicated by the Author. Philosophical Transactions of the Royal Society *27*.

Bonnaud, L., BoucherRodoni, R., and Monnerot, M. (1997). Phylogeny of cephalopods inferred from mitochondrial DNA sequences. Mol Phylogenet Evol *7*, 44-54.

Boratyn, G.M., Schaffer, A.A., Agarwala, R., Altschul, S.F., Lipman, D.J., and Madden, T.L. (2012). Domain enhanced lookup time accelerated BLAST. Biol Direct *7*, 15.

Boratyn, G.M., Thierry-Mieg, J., Thierry-Mieg, D., Busby, B., and Madden, T.L. (2019). Magic-BLAST, an accurate RNA-seq aligner for long and short reads. BMC Bioinformatics *20*, 19.

Bosia, F., Buehler, M.J., and Pugno, N.M. (2010). Hierarchical simulations for the design of supertough nanofibers inspired by spider silk. Phys Rev E *82*, 7.

Boutry, C., and Blackledge, T.A. (2008). The Common House Spider Alters the Material and Mechanical Properties of Cobweb Silk in Response to Different Prey. Journal of Experimental Zoology Part a-Ecological and Integrative Physiology *309A*, 542-552.

Bowen, C.H., Dai, B., Sargent, C.J., Bai, W.Q., Ladiwala, P., Feng, H.B., Huang, W.W., Kaplan, D.L., Galazka, J.M., and Zhang, F.Z. (2018). Recombinant Spidroins Fully Replicate Primary Mechanical

Properties of Natural Spider Silk. Biomacromolecules *19*, 3853-3860.

Brandley, M.C., Warren, D.L., Leache, A.D., and McGuire, J.A. (2009). Homoplasy and Clade Support. Syst Biol *58*, 184-198.

Brandman, O., Stewart-Ornstein, J., Wong, D., Larson, A., Williams, C.C., Li, G.W., Zhou, S., King, D., Shen, P.S., Weibezahn, J.*, et al.* (2012). A Ribosome-Bound Quality Control Complex Triggers Degradation of Nascent Peptides and Signals Translation Stress. Cell *151*, 1042-1054.

Bratzel, G., and Buehler, M.J. (2012). Sequence-structure correlations in silk: Poly-Ala repeat of N. clavipes MaSp1 is naturally optimized at a critical length scale. J Mech Behav Biomed Mater *7*, 30-40.

Bubeck, P., Winkler, M., and Bautsch, W. (1993). Rapid Cloning By Homologous Recombination In-Vivo. Nucleic Acids Res *21*, 3601-3602.

Buchli , H.R. (2015). Hunting Behavior in the Ctenizidae. *American Zoologist 9*, 175–193.

Buck, C.C., Dennis, P.B., Gupta, M.K., Grant, M.T., Crosby, M.G., Slocik, J.M., Mirau, P.A., Becknell, K.A., Comfort, K.K., and Naik, R.R. (2019). Anion-Mediated Effects on the Size and Mechanical Properties of Enzymatically Crosslinked Suckerin Hydrogels. Macromol Biosci *19*, 8.

Burgess, L.A. (1982). Four new species of squid (Oegopsida: Enoploteuthis) from the central Pacific and a description of adult Enoploteuthis reticulata. Fishery Bulletin *80*, 703-734.

Cadle, K.A. (2016). The Role the N-terminal Domain Plays in Spidroin Assembly. In Genetics and Biochemistry (South Carolina: *Clemson University*).

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST plus : architecture and applications. BMC Bioinformatics *10*, 9.

Canelle, L., Bousquet, J., Pionneau, C., Deneux, L., Imam-Sghiouar, N., Caron, M., and Joubert-Caron, R. (2005). An efficient proteomics-based approach for the screening of autoantibodies. Journal of Immunological Methods *299*, 77-89.

Cao, H., Parveen, S., Ding, D., Xu, H.J., Tan, T.W., and Liu, L. (2017). Metabolic engineering for recombinant major ampullate spidroin 2 (MaSp2) synthesis in Escherichia coli. Scientific Reports *7*, 8.

Cao, Y., Kim, H.-J., Li, Y., Kong, H., and Lemieux, B. (2018). Helicase-Dependent Amplification of Nucleic Acids. Current Protocols in Molecular Biology *104*, 15.11.11 – 15.11.12.

Case, S.T., Cox, C., Bell, W.C., Hoffman, R.T., Martin, J., and Hamilton, R. (1997). Extraordinary conservation of cysteines among homologous Chironomus silk proteins sp185 and sp220. J Mol Evol *44*, 452-462.

Catalog, W.S. (2014). World Spider Catalog (wsc.nmbe.ch: Natural History Museum Bern).

Chan, P.P., and Lowe, T.M. (2016). GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Res *44*, D184-D189.

Chaw, R.C., Saski, C.A., and Hayashi, C.Y. (2017). Complete gene sequence of spider attachment silk protein (PySp1) reveals novel linker regions and extreme repeat homogenization. Insect Biochem Mol Biol *81*, 80-90.

Chen, C.K., Cheng, R., Demeter, J., Chen, J., Weingarten-Gabbay, S., Jiang, L.H., Snyder, M.P., Weissman, J.S., Segal, E., Jackson, P.K.*, et al.* (2021). Structured elements drive extensive circular RNA translation. Mol Cell *81*, 4300-+.

Chen, G.F., Liu, X.Q., Zhang, Y.L., Lin, S.Z., Yang, Z.J., Johansson, J., Rising, A., and Meng, Q. (2012). Full-Length Minor Ampullate Spidroin Gene Sequence. Plos One *7*, 11.

Chen, R.X., Chen, X., Xia, L.P., Zhang, J.X., Pan, Z.Z., Ma, X.D., Han, K., Chen, J.W., Judde, J.G., Deas, O.*, et al.* (2019). N-6-methyladenosine modification of circNSUN2 facilitates cytoplasmic export and stabilizes HMGA2 to promote colorectal liver metastasis. Nat Commun *10*.

Chen, W.J., Wang, F., Tian, C., Wang, Y.C., Xu, S., Wang, R.Y., Hou, K., Zhao, P., Yu, L., Lu, Z.S.*, et al.* (2018). Transgenic Silkworm-Based Silk Gland Bioreactor for Large Scale Production of Bioactive Human Platelet-Derived Growth Factor (PDGF-BB) in Silk Cocoons. International Journal of Molecular Sciences *19*.

Cheng, C., Qiu, Y.M., Tang, S.M., Lin, B.Y., Guo, M.Z., Gao, B.B., and He, B.F. (2022). Artificial Spider Silk Based Programmable Woven Textile for Efficient Wound Management. Adv Funct Mater *32*.

Chiasson, R., Hasan, M., Al Nazer, Q., Farokhzad, O.C., and Kamaly, N. (2016). The Use of Silk in Nanomedicine Applications. In Nanomedicine, K.A. Howard, T. VorupJensen, and D. Peer, eds. (New York: Springer), pp. 245-278.

Chiu, B., Coburn, J., Pilichowska, M., Holcroft, C., Seib, F.P., Charest, A., and Kaplan, D.L. (2014). Surgery combined with controlled-release doxorubicin silk films as a treatment strategy in an orthotopic neuroblastoma mouse model. Br J Cancer *111*, 708-715.

Choe, J., Lin, S.B., Zhang, W.C., Liu, Q., Wang, L.F., Ramirez-Moya, J., Du, P., Kim, W., Tang, S.J., Sliz, P.*, et al.* (2018). mRNA circularization by METTL3-eIF3h enhances translation and promotes oncogenesis. Nature *561*, 556-+.

Choe, Y.J., Park, S.H., Hassemer, T., Korner, R., Vincenz-Donnelly, L., Hayer-Hartl, M., and Hartl, F.U. (2016). Failure of RQC machinery causes protein aggregation and proteotoxic stress. Nature *531*, 191-+.

Chong, Z.X., Yeap, S.K., and Ho, W.Y. (2021). Transfection types, methods and strategies: a technical review. Peerj *9*.

Clarke, M.R., and Maul, G.E. (1962). A Description of the "Scaled" Squid *Lepidoteuthis Grimaldi*. Proceedings of the Zoological Society of London *139*, 97-118.

Coddington, J.A., and Levi, H.W. (1991). SYSTEMATICS AND EVOLUTION OF SPIDERS (ARANEAE). Annu Rev Ecol Syst *22*, 565-592.

Coekin, T. (2021). Influencing the degradation rate of recombinant spider silk in the presence of matrix metalloproteinases. In Faculty of Science (University of Nottingham).

Colgin, M.A., and Lewis, R.V. (1998). Spider minor ampullate silk proteins contain new repetitive sequences and highly conserved non-silk-like "spacer regions". Protein Sci *7*, 667-672.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szczesniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X.G.*, et al.* (2016). A survey of best practices for RNA-seq data analysis. Genome Biol *17*, 19.

Coots, R.A., Liu, X.M., Mao, Y.H., Dong, L.M., Zhou, J., Wan, J., Zhang, X.Q., and Qian, S.B. (2017). m(6)A Facilitates eIF4F-Independent mRNA Translation. Mol Cell *68*, 504-+.

Correa-Garhwal, S.M., Clarke, T.H., Janssen, M., Crevecoeur, L., McQuillan, B.N., Simpson, A.H., Vink, C.J., and Hayashi, C.Y. (2019). Spidroins and Silk Fibers of Aquatic Spiders. Scientific Reports *9*, 12.

Costello, A., Lao, N.T., Barron, N., and Clynes, M. (2019). Continuous translation of circularized mRNA improves recombinant protein titer. Metab Eng *52*, 284-292.

Cridge, A.G., Crowe-McAuliffe, C., Mathew, S.F., and Tate, W.P. (2018). Eukaryotic translational termination efficiency is influenced by the 3 ' nucleotides within the ribosomal mRNA channel. Nucleic Acids Res *46*, 1927-1944.

Dalton, A.C., and Barton, W.A. (2014). Over-expression of secreted proteins from mammalian cell lines. Protein Sci *23*, 517-525.

Deepankumar, K., Lim, C., Polte, I., Zappone, B., Labate, C., De Santo, M.P., Mohanram, H., Palaniappan, A., Hwang, D.S., and Miserez, A. (2020). Supramolecular beta-Sheet Suckerin-Based Underwater Adhesives. Adv Funct Mater *30*, 11.

Desmet, F.O., Hamroun, D., Lalande, M., Collod-Beroud, G., Claustres, M., and Beroud, C. (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res *37*.

Deutscher , M.P. (1982). The Enzymes, Vol 15 (Paris, San Diego, San Francisco, St Paul, Sydney, Tokyo, Toronto: Academic Press Inc.).

Di Timoteo, G., Dattilo, D., Centron-Broco, A., Colantoni, A., Guarnacci, M., Rossi, F., Incarnato, D., Oliviero, S., Fatica, A., Morlando, M.*, et al.* (2020). Modulation of circRNA Metabolism by m(6)A Modification. Cell Reports *31*.

Diaz-Toledano, R., Lozano, G., and Martinez-Salas, E. (2017). In-cell SHAPE uncovers dynamic interactions between the untranslated regions of the foot-and-mouth disease virus RNA. Nucleic Acids Res *45*, 1416-1432.

Ding, D.W., Guerette, P.A., Hoon, S., Kone, K.W., Cornvik, T., Nilsson, M., Kumar, A., Lescar, J., and Miserez, A. (2014). Biomimetic Production of Silk-Like Recombinant Squid Sucker Ring Teeth Proteins. Biomacromolecules *15*, 3278-3289.

Dinjaski, N., Ebrahimi, D., Qin, Z., Giordano, J.E.M., Ling, S.J., Buehler, M.J., and Kaplan, D.L. (2018). Predicting rates of in vivo degradation of recombinant spider silk proteins. J Tissue Eng Regen Med *12*, E97-E105.

Dobson, C.M. (2003). Protein folding and misfolding. Nature *426*, 884-890.

dos Reis, M., Savva, R., and Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res *32*, 5036-5044.

dos Santos-Pinto, J.R.A., Arcuri, H.A., Lubec, G., and Palma, M.S. (2016). Structural characterization of the major ampullate silk spidroin-2 protein produced by the spider Nephila clavipes. Biochimica Et Biophysica Acta-Proteins and Proteomics *1864*, 1444-1454.

dos Santos-Pinto, J.R.A., Lamprecht, G., Chen, W.Q., Heo, S., Hardy, J.G., Priewalder, H., Scheibel, T.R., Palma, M.S., and Lubec, G. (2014). Structure and post-translational modifications of the web silk protein spidroin-1 from Nephila spiders. Journal of Proteomics *105*, 174-185.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics *14*, 755-763.

Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Comput Biol *7*, 16.

Edlund, A.M., Jones, J., Lewis, R., and Quinn, J.C. (2018). Economic feasibility and environmental impact of synthetic spider silk production from escherichia coli. New Biotech *42*, 12-18.

Ehrlich, R., Davyt, M., Lopez, I., Chalar, C., and Marin, M. (2021). On the Track of the Missing tRNA Genes: A Source of Non-Canonical Functions? Frontiers in Molecular Biosciences *8*.

Eisoldt, L., Smith, A., and Scheibel, T. (2011). Decoding the secrets of spider silk. Mater Today *14*, 80-86.

Eisoldt, L., Thamm, C., and Scheibel, T. (2012). The role of terminal domains during storage and assembly of spider silk proteins. Biopolymers *97*, 355-361.

Elettro, H., Neukirch, S., Vollrath, F., and Antkowiak, A. (2016). In-drop capillary spooling of spider capture thread inspires hybrid fibers with mixed solid-liquid mechanical properties. Proceedings of the National Academy of Sciences of the United States of America *113*, 6143-6147.

Fahnestock, S.R., and Bedzyk, L.A. (1997). Production of synthetic spider dragline silk protein in Pichia pastoris. Appl Microbiol Biotechnol *47*, 33-39.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A.*, et al.* (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res *44*, D279-D285.

Fire, A., and Xu, S.Q. (1995). ROLLING REPLICATION OF SHORT DNA CIRCLES. Proceedings of the National Academy of Sciences of the United States of America *92*, 4641-4645.

Thermo Fisher. (2022). Multiple Primer Analyser (https://www.thermofisher.com/uk/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/multiple-primer-analyzer.html: Thermo Fisher Scientific).

Florczak, A., Deptuch, T., Kucharczyk, K., and Dams-Kozlowska, H. (2021). Systemic and Local Silk-Based Drug Delivery Systems for Cancer Therapy. Cancers *13*.

Fredriksson, C., Hedhammar, M., Feinstein, R., Nordling, K., Kratz, G., Johansson, J., Huss, F., and Rising, A. (2009). Tissue Response to Subcutaneously Implanted Recombinant Spider Silk: An in Vivo Study. Materials *2*, 1908-1922.

Frey, U.H., Bachmann, H.S., Peters, J., and Siffert, W. (2008). PCR-amplification of GC-rich regions: 'slowdown PCR'. Nat Protoc *3*, 1312-1317.

Fricke, M., Dunnes, N., Zayas, M., Bartenschlager, R., Niepmann, M., and Marz, M. (2015). Conserved RNA secondary structures and long-range interactions in hepatitis C viruses. Rna *21*, 1219-1232.

Frumkin, I., Lajoie, M.J., Gregg, C.J., Hornung, G., Church, G.M., and Pilpel, Y. (2018). Codon usage of highly expressed genes affects proteome-wide translation efficiency. Proceedings of the National Academy of Sciences of the United States of America *115*, E4940-E4949.

Garb, J.E., Ayoub, N.A., and Hayashi, C.Y. (2010). Untangling spider silk evolution with spidroin terminal domains. BMC Evol Biol *10*.

Garb, J.E., and Hayashi, C.Y. (2005). Modular evolution of egg case silk genes across orb-weaving spider superfamilies. Proceedings of the National Academy of Sciences of the United States of America *102*, 11379-11384.

Garcia-Nafria, J., Watson, J.F., and Greger, I.H. (2016). IVA cloning: A single-tube universal cloning system exploiting bacterial In Vivo Assembly. Scientific Reports *6*.

Gardner, B.M., Pincus, D., Gotthardt, K., Gallagher, C.M., and Walter, P. (2013). Endoplasmic Reticulum Stress Sensing in the Unfolded Protein Response. Cold Spring Harbor Perspectives in Biology *5*.

Garner, L.E. (1955). The Tommy Dot: Example of using spider silk for telescopic rifle sights In Popular Science (New York, USA: Popular Science Publishing Co.), pp. 215-218.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D., and Bairoch, A. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res *31*, 3784-3788.

Gatesy, J., Hayashi, C., Motriuk, D., Woods, J., and Lewis, R. (2001). Extreme diversity, conservation, and convergence of spider silk fibroin sequences. Science *291*, 2603-2605.

Gauci, V.J., Padula, M.P., and Coorssen, J.R. (2013). Coomassie blue staining for high sensitivity gel-based proteomics. Journal of Proteomics *90*, 96-106.

Gebbie, M.A., Wei, W., Schrader, A.M., Cristiani, T.R., Dobbs, H.A., Idso, M., Chmelka, B.F., Waite, J.H., and Israelachvili, J.N. (2017). Tuning underwater adhesion with cation-pi interactions. Nat Chem *9*, 473-479.

Geisse, S., and Fux, C. (2009). Recombinant Protein Production By Transient Gene Transfer Into Mammalian Cells. Guide to Protein Purification, Second Edition *463*, 223-238.

Gerber, A., Grosjean, H., Melcher, T., and Keller, W. (1998). Tad1p, a yeast tRNA-specific adenosine deaminase, is related to the mammalian pre-mRNA editing enzymes ADAR1 and ADAR2. Embo J *17*, 4780-4789.

Geslain, R., and Pan, T. (2010). Functional Analysis of Human tRNA Isodecoders. J Mol Biol *396*, 821-831.

Gibson, D.G., Young, L., Chuang, R.Y., Venter, J.C., Hutchison, C.A., and Smith, H.O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods *6*, 343-U341.

Gingold, H., Dahan, O., and Pilpel, Y. (2012). Dynamic changes in translational efficiency are deduced from codon usage of the transcriptome. Nucleic Acids Res *40*, 10053-10063.

Girod, and , P. (1884). Recherches sur la peau des céphalopodes (Research on the Skin of Cephalopods). Arch Zool Exp Gen, 2379 - 2401.

Goodenbour, J.M., and Pan, T. (2006). Diversity of tRNA genes in eukaryotes. Nucleic Acids Res *34*, 6137-6146.

Gosline, J.M., Guerette, P.A., Ortlepp, C.S., and Savage, K.N. (1999). The mechanical design of spider silks: From fibroin sequence to mechanical function. J Exp Biol *202*, 3295-3303.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q.D., *et al.* (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology *29*, 644-U130.

Graveley, B.R. (2005). Mutuallv exclusive splicing of the insect Dscam Pre-mRNA directed by competing intronic RNA secondary structures. Cell *123*, 65-73.

Grip , S., Rising , A., Nimmervoll , H., Storckenfeldt , E., McQueen-Mason , S.J., Pouchkina-Stantcheva , N., Vollrath , F., Engström , W., and Fernandez-Arias , A. (2006). Transient Expression of a

Major Ampullate Spidroin 1 Gene Fragment from *Euprosthenops* sp. in Mammalian Cells. Cancer Genomics Proteomics *3*, 83-87.

Grosjean, H., de Crecy-Lagard, V., and Marck, C. (2010). Deciphering synonymous codons in the three domains of life: Co-evolution with specific tRNA modification enzymes. Febs Letters *584*, 252-264.

Group, O.S. (2013). Spider Silk Harvesting (Youtube: Oxford Silk Group).

Grunberger, D., Weinstein, I.B., and Jacobson, K.B. (1969). Codon Recognition By Enzymatically Mischarged Valine Transfer Ribonucleic Acid. Science *166*, 1635-+.

Gründemann, D., and Schömig, E. (2018). Protection of DNA During Preparative Agarose Gel Electrophoresis Against Damage Induced by Ultraviolet Light. Biotechniques *21*.

Guerette, P.A., Ginzinger, D.G., Weber, B.H.F., and Gosline, J.M. (1996). Silk properties determined by gland-specific expression of a spider fibroin gene family. Science *272*, 112-115.

Guerette, P.A., Hoon, S., Ding, D.W., Amini, S., Masic, A., Ravi, V., Venkatesh, B., Weaver, J.C., and Miserez, A. (2014). Nanoconfined beta-Sheets Mechanically Reinforce the Supra-Biomolecular Network of Robust Squid Sucker Ring Teeth. ACS Nano *8*, 7170-7179.

Guerette, P.A., Hoon, S., Seow, Y., Raida, M., Masic, A., Wong, F.T., Ho, V.H.B., Kong, K.W., Demirel, M.C., Pena-Francesch, A.*, et al.* (2013). Accelerating the design of biomimetic materials by integrating RNA-seq with proteomics and materials science. Nature Biotechnology *31*, 908-+.

Guinea, G.V., Elices, M., Plaza, G.R., Perea, G.B., Daza, R., Riekel, C., Agullo-Rueda, F., Hayashi, C., Zhao, Y., and Perez-Rigueiro, J. (2012). Minor Ampullate Silks from Nephila and Argiope Spiders: Tensile Properties and Microstructural Characterization. Biomacromolecules *13*, 2087-2098.

Guydosh, N.R., and Green, R. (2014). Dom34 Rescues Ribosomes in 3 ' Untranslated Regions. Cell *156*, 950-962.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M.*, et al.* (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc *8*, 1494-1512.

Hansel, B. (2019). Kraig Biocraft Laboratories selects new high performance recombinant spider silk fordelivery to US ARMY (Kraiglabs.com: Kraig Biocraft Laboratories).

Hansen, T.B., Jensen, T.I., Clausen, B.H., Bramsen, J.B., Finsen, B., Damgaard, C.K., and Kjems, J. (2013). Natural RNA circles function as efficient microRNA sponges. Nature *495*, 384-388.

Hardy, E., Pupo, E., Casalvilla, R., Sosa, A.E., Trujillo, L.E., Lopez, E., and CastellanosSerra, L. (1996). Negative staining with zinc-imidazole of gel electrophoresis separated nucleic acids. Electrophoresis *17*, 1537-1541.

Harger, J.W., Meskauskas, A., and Dinman, J.D. (2002). An 'integrated model' of programmed ribosomal frameshifting. Trends BiochemSci *27*, 448-454.

Harvey, D., Bardelang, P., Goodacre, S.L., Cockayne, A., and Thomas, N.R. (2017). Antibiotic Spider Silk: Site-Specific Functionalization of Recombinant Spider Silk Using "Click" Chemistry. Adv Mater *29*.

Hasegawa, M., and Fujiwara, M. (1993). Relative Efficiencies Of The Maximum-Likelihood, Maximum Parsimony, And Neighbor-Joining Methods For Estimating Protein Phylogeny. Mol Phylogenet Evol *2*, 1-5.

Hauptmann, V., Weichert, N., Menzel, M., Knoch, D., Paege, N., Scheller, J., Spohn, U., Conrad, U., and Gils, M. (2013a). Native-sized spider silk proteins synthesized in planta via intein-based multimerization. Transgenic Res *22*, 369-377.

Hauptmann, V., Weichert, N., Rakhimova, M., and Conrad, U. (2013b). Spider silks from plants - a challenge to create native-sized spidroins. Biotechnol J *8*, 1183-1192.

Hawthorn, A.C., and Opell, B.D. (2002). Evolution of adhesive mechanisms in cribellar spider prey capture thread: evidence for van der Waals and hygroscopic forces. Biol J Linnean Soc *77*, 1-8.

Hayashi, C.Y., Blackledge, T.A., and Lewis, R.V. (2004). Molecular and mechanical characterization of aciniform silk: Uniformity of iterated sequence modules in a novel member of the spider silk fibroin gene family. Mol Biol Evol *21*, 1950-1959.

Hayashi, C.Y., and Lewis, R.V. (1998). Evidence from flagelliform silk cDNA for the structural basis of elasticity and modular nature of spider silks. J Mol Biol *275*, 773-784.

Hayashi, C.Y., and Lewis, R.V. (2001). Spider flagelliform silk: lessons in protein design, gene structure, and molecular evolution. Bioessays *23*, 750-756.

Hayashi, C.Y., Shipley, N.H., and Lewis, R.V. (1999). Hypotheses that correlate the sequence, structure, and mechanical properties of spider silk proteins. Int J Biol Macromol *24*, 271-275.

Hedges, S.B., Marin, J., Suleski, M., Paymer, M., and Kumar, S. (2015). Tree of Life Reveals Clock-Like Speciation and Diversification. Mol Biol Evol *32*, 835-845.

Hedhammar, M., Bramfeldt, H., Baris, T., Widhe, M., Askarieh, G., Nordling, K., von Aulock, S., and Johansson, J. (2010). Sterilized Recombinant Spider Silk Fibers of Low Pyrogenicity. Biomacromolecules *11*, 953-959.

Hedhammar, M., Rising, A., Grip, S., Martinez, A.S., Nordling, K., Casals, C., Stark, M., and Johansson, J. (2008). Structural properties of recombinant nonrepetitive and repetitive parts of major ampullate spidroin 1 from Euprosthenops australis: Implications for fiber formation. Biochemistry *47*, 3407-3417.

Heiby, J.C., Rajab, S., Rat, C., Johnson, C.M., and Neuweiler, H. (2017). Conservation of folding and association within a family of spidroin N-terminal domains. Scientific Reports *7*, 11.

Heidebrecht, A., Eisoldt, L., Diehl, J., Schmidt, A., Geffers, M., Lang, G., and Scheibel, T. (2015). Biomimetic Fibers Made of Recombinant Spidroins with the Same Toughness as Natural Spider Silk. Adv Mater *27*, 2189-+.

Heidebrecht, A., and Scheibel, T. (2013). Recombinant Production of Spider Silk Proteins. Advances in Applied Microbiology, Vol 82 *82*, 115-153.

Hennecke, K., Redeker, J., Kuhbier, J.W., Strauss, S., Allmeling, C., Kasper, C., Reimers, K., and Vogt, P.M. (2013). Bundles of Spider Silk, Braided into Sutures, Resist Basic Cyclic Tests: Potential Use for Flexor Tendon Repair. Plos One *8*, 10.

Herr, A.J., Atkins, J.F., and Gesteland, R.F. (2000). Coupling of open reading frames by translational bypassing. Annu Rev Biochem *69*, 343-372.

Hiew, S.H., and Miserez, A. (2017). Squid Sucker Ring Teeth: Multiscale Structure-Property Relationships, Sequencing, and Protein Engineering of a Thermoplastic Biopolymer. ACS Biomater Sci Eng *3*, 680-693.

Hiew, S.H., Mohanram, H., Ning, L.L., Guo, J.J., Sanchez-Ferrer, A., Shi, X.Y., Pervushin, K., Mu, Y.G., Mezzenga, R., and Miserez, A. (2019). A Short Peptide Hydrogel with High Stiffness Induced by 3(10)-Helices to beta-Sheet Transition in Water. Adv Sci *6*, 11.

Hiew, S.H., Sanchez-Ferrer, A., Amini, S., Zhou, F., Adamcik, J., Guerette, P., Su, H.B., Mezzenga, R., and Miserez, A. (2017). Squid Suckerin Biomimetic Peptides Form Amyloid-like Crystals with Robust Mechanical Properties. Biomacromolecules *18*, 4240-4248.

Hinchliff, C.E., Smith, S.A., Allman, J.F., Burleigh, J.G., Chaudhary, R., Coghill, L.M., Crandall, K.A., Deng, J., Drew, B.T., Gazis, R.*, et* *al.* (2015). Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proceedings of the National Academy of Sciences of the United States of America *112*, 12764-12769.

Hinman, M.B., and Lewis, R.V. (1992). Isolation Of A Clone Encoding A 2nd Dragline Silk Fibroin - Nephila-Clavipes Dragline Silk Is A 2-Protein Fiber. J Biol Chem *267*, 19320-19324.

Hock, L. (2008). Cobweb Art a Triumph of Whimsy Over Practicality. In Northwestern Now (Evanston, Ill.: Northwestern).

Hormiga, G., and Griswold, C.E. (2014). Systematics, Phylogeny, and Evolution of Orb-Weaving Spiders. In Annual Review of Entomology, Vol 59, 2014, M.R. Berenbaum, ed. (Palo Alto: Annual Reviews), pp. 487-512.

Hsu, M.T., and Cocaprados, M. (1979). Electron-Microscopic Evidence For The Circular Form Of Rna In The Cytoplasm Of Eukaryotic Cells. Nature *280*, 339-340.

Huang, C., Liang, D.M., Tatomer, D.C., and Wilusz, J.E. (2018). A length-dependent evolutionarily conserved pathway controls nuclear export of circular RNAs. Genes Dev *32*, 639-644.

Huang, W.D., Zhang, Y., Chen, Y.F., Wang, Y., Yuan, W.S., Zhang, N., Lam, T.J., Gong, Z.Y., Yang, D.W., and Lin, Z. (2017). From EST to novel spider silk gene identification for production of spidroin-based biomaterials. Scientific Reports *7*.

Huh, D., Passarelli, M.C., Gao, J., Dusmatova, S.N., Goin, C., Fish, L., Pinzaru, A.M., Molina, H., Ren, Z.J., McMillan, E.A.*, et al.* (2021). A stress-induced tyrosine-tRNA depletion response mediates codon-based translational repression and growth suppression. Embo J *40*.

Humenik, M., Scheibel, T., and Smith, A. (2011). Spider Silk: Understanding the Structure-Function Relationship of a Natural Fiber. In Molecular Assembly in Natural and Engineered Systems, Vol 103, S. Howorka, ed. (San Diego: Elsevier Academic Press Inc), pp. 131-185.

Abobe Inc. (2022). Adobe InDesign (https://adobe.com/products/indesign: Adobe).

Spiber Inc. (2019). Endeavor Spiber Inc. (Spiber.jp: Spiber Inc.).

Ishimura, R., Nagy, G., Dotu, I., Zhou, H.H., Yang, X.L., Schimmel, P., Senju, S., Nishimura, Y., Chuang, J.H., and Ackerman, S.L. (2014). Ribosome stalling induced by mutation of a CNS-specific tRNA causes neurodegeneration. Science *345*, 455-459.

Ivanov, A., Memczak, S., Wyler, E., Torti, F., Porath, H.T., Orejuela, M.R., Piechotta, M., Levanon, E.Y., Landthaler, M., Dieterich, C.*, et al.* (2015). Analysis of Intron Sequences Reveals Hallmarks of Circular RNA Biogenesis in Animals. Cell Reports *10*, 170-177.

Jackson, R.J., Hellen, C.U.T., and Pestova, T.V. (2010). The mechanism of eukaryotic translation initiation and principles of its regulation. Nat Rev Mol Cell Biol *11*, 113-127.

Jeck, W.R., and Sharpless, N.E. (2014). Detecting and characterizing circular RNAs. Nature Biotechnology *32*, 453-461.

Jeck, W.R., Sorrentino, J.A., Wang, K., Slevin, M.K., Burd, C.E., Liu, J.Z., Marzluff, W.F., and Sharpless, N.E. (2013). Circular RNAs are abundant, conserved, and associated with ALU repeats. Rna *19*, 141-157.

Jenkins, J.E., Creager, M.S., Butler, E.B., Lewis, R.V., Yarger, J.L., and Holland, G.P. (2010). Solid-state NMR evidence for elastin-like beta-turn structure in spider dragline silk. Chemical Communications *46*, 6714-6716.

Jin, Q., Pan, F., Hu, C.F., Lee, S.Y., Xia, X.X., and Qian, Z.G. (2022). Secretory production of spider silk proteins in metabolically engineered Corynebacterium glutamicum for spinning into tough fibers. Metab Eng *70*, 102-114.

Johnson, L.S., Eddy, S.R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics *11*.

Jones, J.A., Harris, T.I., Tucker, C.L., Berg, K.R., Christy, S.Y., Day, B.A., Gaztambide, D.A., Needham, N.J.C., Ruben, A.L., Oliveira, P.F.*, et al.* (2015). More Than Just Fibers: An Aqueous Method for the Production of Innovative Recombinant Spider Silk Protein Materials. Biomacromolecules *16*, 1418-1425.

Jorda, J., and Kajava, A.V. (2009). T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. Bioinformatics *25*, 2632-2638.

Jorge, I., Ruiz, V., Lavado-Garcia, J., Vazquez, J., Hayashi, C., Rojo, F.J., Atienza, J.M., Elices, M., Guinea, G.V., and Perez-Rigueiro, J. (2022). Expression of spidroin proteins in the silk glands of golden orb-weaver spiders. Journal of Experimental Zoology Part B-Molecular and Developmental Evolution *338*, 241-253.

Joshi, N.A., and Fass, J.N. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software] (https://github.com/najoshi/sickle.).

Kall, L., Krogh, A., and Sonnhammer, E.L.L. (2004). A combined transmembrane topology and signal peptide prediction method. J Mol Biol *338*, 1027-1036.

Kall, L., Krogh, A., and Sonnhammer, E.L.L. (2007). Advantages of combined transmembrane topology and signal peptide prediction - the Phobius web server. Nucleic Acids Res *35*, W429-W432.

Kanamori, Y., and Nakashima, N. (2001). A tertiary structure model of the internal ribosome entry site (IRES) for methionine-independent initiation of translation. Rna *7*, 266-274.

Kannicht, C., Ramstrom, M., Kohla, G., Tiemeyer, M., Casademunt, E., Walter, O., and Sandberg, H. (2013). Characterisation of the post-translational modifications of a novel, human cell line-derived recombinant human factor VIII. Thrombosis Research *131*, 78-88.

Kelly, J.A., Olson, A.N., Neupane, K., Munshi, S., San Emeterio, J., Pollack, L., Woodside, M.T., and Dinman, J.D. (2020a). Structural and functional conservation of the programmed-1 ribosomal frameshift signal of SARS coronavirus 2 (SARS-CoV-2). J Biol Chem *295*, 10741-10748.

Kelly, P., and Ibba, M. (2018). Aminoacyl-tRNA Quality Control Provides a Speedy Solution to Discriminate Right from Wrong. J Mol Biol *430*, 17-19.

Kelly, S.P., Huang, K.P., Liao, C.P., Khasanah, R.A.N., Shih-Sen Chien, F., Hu, J.S., Wu, C.L., and Tso, I.M. (2020b). Mechanical and structural properties of major ampullate silk from spiders fed carbon nanomaterials. Plos One *15*, 13.

Kieleczawa, J. (2005). Simple Modifications of the Standard DNA Sequencing Protocol Allow for Sequencing Through siRNA Hairpins and Other Repeats. Journal of Biomolecular Techniques *16*.

Kier , W.M., Smith , and M., A. (1990). The Morphology and Mechanics of Octopus Suckers. The Biological Bulletin *178*, 126--136.

Kim, Y.G., Maas, S., and Rich, A. (2001). Comparative mutational analysis of cis-acting RNA signals for translational frameshifting in HIV-1 and HTLV-2. Nucleic Acids Res *29*, 1125-1131.

Klein, T., Niklas, J., and Heinzle, E. (2015). Engineering the supply chain for protein production/secretion in yeasts and mammalian cells. Journal of Industrial Microbiology & Biotechnology *42*, 453-464.

Kleiner-Grote, G.R.M., Risse, J.M., and Friehs, K. (2018). Secretion of recombinant proteins from E. coli. Engineering in Life Sciences *18*, 532-550.

Klinge, C.M., Piell, K.M., Tooley, C.S., and Rouchka, E.C. (2019). HNRNPA2/B1 is upregulated in endocrine-resistant LCC9 breast cancer cells and alters the miRNA transcriptome when overexpressed in MCF-7 cells. Scientific Reports *9*.

Ko, K.K., Kawabata, S., Inoue, M., Niwa, M., Fossey, S., and Song, J.W. (2001). Engineering properties of spider silk. Paper presented at: Symposium on Advanced Fibers, Plastics, Laminates and

Composites held at the 2001 MRS Fall Meeting (Boston, Ma: Materials Research Society).

Kohrer, C., and RajBhandary, U.L. (2008). The many applications of acid urea polyacrylamide gel electrophoresis to studies of tRNAs and aminoacyl-tRNA synthetases. Methods *44*, 129-138.

Konevega, A.L., Soboleva, N.G., Makhno, V.I., Semenkov, Y.P., Wintermeyer, W., Rodina, M.V., and Katunin, V.I. (2004). Purine bases at position 37 of tRNA stabilize codon-anticodon interaction in the ribosomal A site by stacking and Mg2+-dependent interactions. Rna *10*, 90-101.

Kono, N., Nakamura, H., Mori, M., Yoshida, Y., Ohtoshi, R., Malay, A.D., Moran, D.A.P., Tomita, M., Numata, K., and Arakawa, K. (2021). Multicomponent nature underlies the extraordinary mechanical properties of spider dragline silk. Proceedings of the National Academy of Sciences of the United States of America *118*.

Kono, N., Nakamura, H., Ohtoshi, R., Moran, D., Shinohara, A., Yoshida, Y., Fujiwara, M., Mori, M., Tomita, M., and Arakawa, K. (2019). Orb-weaving spider Araneus ventricosus genome elucidates the spidroin gene catalogue. Scientific Reports *9*, 13.

Kuhbier, J.W., Allmeling, C., Reimers, K., Hillmer, A., Kasper, C., Menger, B., Brandes, G., Guggenheim, M., and Vogt, P.M. (2010). Interactions between Spider Silk and Cells - NIH/3T3 Fibroblasts Seeded on Miniature Weaving Frames. Plos One *5*, 9.

Kumar, A., Kannan, S., Lescar, J., Verma, C., and Miserez, A. (2016). Squid's Suckerin Proteins in Bits & Bytes. Biophys J *110*, 341A-341A.

Kumar, A., Mohanram, H., Kong, K.W., Goh, R., Hoon, S., Lescar, J., and Miserez, A. (2018a). Supramolecular propensity of suckerin proteins is driven by -sheets and aromatic interactions as revealed by solution NMR. Biomater Sci *6*, 2440-2447.

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018b). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol *35*, 1547-1549.

Kumari, S., Bargel, H., Anby, M.U., Lafargue, D., and Scheibel, T. (2018). Recombinant Spider Silk Hydrogels for Sustained Release of Biologicals. ACS Biomater Sci Eng *4*, 1750-1759.

Kummerlen, J., van Beek, J.D., Vollrath, F., and Meier, B.H. (1996). Local structure in spider dragline silk investigated by two-dimensional spin-diffusion nuclear magnetic resonance. Macromolecules *29*, 2920-2928.

La Mattina, C., Reza, R., Hu, X., Falick, A.M., Vasanthavada, K., McNary, S., Yee, R., and Vierra, C.A. (2008). Spider minor ampullate silk proteins are constituents of prey wrapping silk in the cob weaver Latrodectus hesperus. Biochemistry *47*, 4692-4700.

Lammel, A., Schwab, M., Hofer, M., Winter, G., and Scheibel, T. (2011). Recombinant spider silk particles as drug delivery vehicles. Biomaterials *32*, 2233-2240.

Lammel, A., Schwab, M., Slotta, U., Winter, G., and Scheibel, T. (2008). Processing conditions for the formation of spider silk microspheres. ChemSusChem *1*, 413-416.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-U354.

Lassmann, T., Frings, O., and Sonnhammer, E.L.L. (2009). Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. Nucleic Acids Res *37*, 858-865.

Latza, V., Guerette, P.A., Ding, D.W., Amini, S., Kumar, A., Schmidt, I., Keating, S., Oxman, N., Weaver, J.C., Fratzl, P.*, et al.* (2015). Multi-scale thermal stability of a hard thermoplastic protein-based material. Nat Commun *6*, 8.

Lawrence, B.A., Vierra, C.A., and Mooref, A.M.F. (2004). Molecular and mechanical properties of major ampullate silk of the black widow spider, Latrodectus hesperus. Biomacromolecules *5*, 689-695.

Lazaris, A., Arcidiacono, S., Huang, Y., Zhou, J.F., Duguay, F., Chretien, N., Welsh, E.A., Soares, J.W., and Karatzas, C.N. (2002). Spider silk fibers spun from soluble recombinant silk produced in mammalian cells. Science *295*, 472-476.

Lee, B.P., Messersmith, P.B., Israelachvili, J.N., and Waite, J.H. (2011). Mussel-Inspired Adhesives and Coatings. In Annual Review of Materials Research, Vol 41, D.R. Clarke, and P. Fratzl, eds. (Palo Alto: Annual Reviews), pp. 99-132.

Lee, J.M., Kawakami, N., Mon, H., Mitsunobu, H., Iiyama, K., Ninaki, S., Maenaka, K., Park, E.Y., and Kusakabe, T. (2012). Establishment of a Bombyx mori nucleopolyhedrovirus (BmNPV) hyper-sensitive cell line from the silkworm e21 strain. Biotechnology Letters *34*, 1773-1779.

Lee, S.O., Xie, Q., and Fried, S.D. (2021). Optimized Loopable Translation as a Platform for the Synthesis of Repetitive Proteins. Acs Central Science *7*, 1736-1750.

Leger, M., Dulude, D., Steinberg, S.V., and Brakier-Gingras, L. (2007). The three transfer RNAs occupying the A, P and E sites on the ribosome are involved in viral programmed-1 ribosomal frameshift. Nucleic Acids Res *35*, 5581-5592.

Legnini, I., Di Timoteo, G., Rossi, F., Morlando, M., Briganti, F., Sthandier, O., Fatica, A., Santini, T., Andronache, A., Wade, M.*, et al.* (2017). Circ-ZNF609 Is a Circular RNA that Can Be Translated and Functions in Myogenesis. Mol Cell *66*, 22-+.

Levene, D. (2012). Golden cape made with silk from a million spiders - in pictures (The Guardian: The Guardian).

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 16.

Li, E.J., Xia, M.Y., Du, Y., Long, K.L., Ji, F., Pan, F.Y., He, L.F., Hu, Z.G., and Guo, Z.G. (2022). METTL3 promotes homologous recombination repair and modulates chemotherapeutic response in breast cancer by regulating the EGF/RAD51 axis. Elife *11*.

Li, F., Bian, C., Li, D.Q., and Shi, Q. (2021). Spider Silks: An Overview of Their Component Proteins for Hydrophobicity and Biomedical Applications. Protein and Peptide Letters *28*, 255-269.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data, P. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Li, X., Yang, L., and Chen, L.L. (2018). The Biogenesis, Functions, and Challenges of Circular RNAs. Mol Cell *71*, 428-442.

Li, Z.G., Kearse, M.G., and Huang, C. (2019). The nuclear export of circular RNAs is primarily defined by their length. RNA Biol *16*, 1-4.

Li, Z.Y., Huang, C., Bao, C., Chen, L., Lin, M., Wang, X.L., Zhong, G.L., Yu, B., Hu, W.C., Dai, L.M.*, et al.* (2015). Exon-intron circular RNAs regulate transcription in the nucleus. Nature Structural & Molecular Biology *22*, 256-264.

Lichtenegger, H.C., Schoberl, T., Bartl, M.H., Waite, H., and Stucky, G.D. (2002). High abrasion resistance with sparse mineralization: Copper biomineral in worm jaws. Science *298*, 389-392.

Lima, L., Sinaimeri, B., Sacomoto, G., Lopez-Maestre, H., Marchet, C., Miele, V., Sagot, M.F., and Lacroix, V. (2017). Playing hide and seek with repeats in local and global de novo transcriptome assembly of short RNA-seq reads. Algorithms Mol Biol *12*, 19.

Lin, C.B., Lin, Y.H., Chen, W.Y., and Liu, C.Y. (2021a). Photonic Nanojet Modulation Achieved by a Spider-Silk-Based Metal-Dielectric Dome Microlens. Photonics *8*.

Lin, C.H., Ekblad-Nordberg, A., Michaelsson, J., Gotherstrom, C., Hsu, C.C., Ye, H., Johansson, J., Rising, A., Sundstrom, E., and Akesson, E. (2021b). In Vitro Study of Human Immune Responses to Hyaluronic Acid Hydrogels, Recombinant Spidroins and Human

Neural Progenitor Cells of Relevance to Spinal Cord Injury Repair. Cells *10*.

Lin, C.Y., Huang, Z., Wen, W., Wu, A., Wang, C.Z., and Niu, L. (2015a). Enhancing Protein Expression in HEK-293 Cells by Lowering Culture Temperature. Plos One *10*.

Lin, S.C., Ryu, S., Tokareva, O., Gronau, G., Jacobsen, M.M., Huang, W.W., Rizzo, D.J., Li, D., Staii, C., Pugno, N.M.*, et al.* (2015b). Predictive modelling-based design and experiments for synthesis and spinning of bioinspired silk fibres. Nat Commun *6*, 12.

Lin, Z., Huang, W.D., Zhang, J.F., Fan, J.S., and Yang, D.W. (2009). Solution structure of eggcase silk protein and its implications for silk fiber formation. Proceedings of the National Academy of Sciences of the United States of America *106*, 8906-8911.

Linder, B., Grozhik, A.V., Olarerin-George, A.O., Meydan, C., Mason, C.E., and Jaffrey, S.R. (2015). Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. Nat Methods *12*, 767-U114.

Liu, C.X., Li, X., Nan, F., Jiang, S., Gao, X., Guo, S.K., Xue, W., Cui, Y.G., Dong, K.G., Ding, H.H.*, et al.* (2019a). Structure and Degradation of Circular RNAs Regulate PKR Activation in Innate Immunity. Cell *177*, 865-+.

Liu, H., Yang, Z., Xun, Z., Gao, Z.H., Sun, Y.P., Yu, J.K., Yang, T.Z., Zhao, X.Y., Cai, C.F., and Ding, P.T. (2019b). Nuclear delivery of plasmid DNA determines the efficiency of gene expression. Cell Biology International *43*, 789-798.

Liu, L., Wang, P.J., Zhao, D.D., Zhu, L., Tang, J.L., Leng, W.C., Su, J.C., Liu, Y., Bi, C.H., and Zhang, X.L. (2022). Engineering Circularized mRNAs for the Production of Spider Silk Proteins. Appl Environ Microbiol *88*.

Liu, Y., Li, Z.J., Zhang, M.L., Zhou, H.K., Wu, X.J., Zhong, J., Xiao, F.Z., Huang, N.N., Yang, X.S., Zeng, R.*, et al.* (2021). Rolling-translated EGFR variants sustain EGFR signaling and promote glioblastoma tumorigenicity. Neuro-Oncology *23*, 743-756.

Lozano, G., Francisco-Velilla, R., and Martinez-Salas, E. (2018). Deconstructing internal ribosome entry site elements: an update of structural motifs and functional divergences. Open Biol *8*, 11.

Lucke, M., Mottas, I., Herbst, T., Hotz, C., Romer, L., Schierling, M., Herold, H.M., Slotta, U., Spinetti, T., Scheibel, T.*, et al.* (2018). Engineered hybrid spider silk particles as delivery system for peptide vaccines. Biomaterials *172*, 105-115.

Madsen, B., Shao, Z.Z., and Vollrath, F. (1999). Variability in the mechanical properties of spider silks on three levels: interspecific, intraspecific and intraindividual. Int J Biol Macromol *24*, 301-306.

Mahmood, N., and Xie, J.Y. (2015). An endogenous 'non-specific' protein detected by a His-tag antibody is human transcription regulator YY1. Data in Brief *2*, 52-55.

Malay, A.D., Arakawa, K., and Numata, K. (2017). Analysis of repetitive amino acid motifs reveals the essential features of spider dragline silk proteins. Plos One *12*, 16.

Mao, Y.H., Dong, L.M., Liu, X.M., Guo, J.Y., Ma, H.H., Shen, B., and Qian, S.B. (2019). m(6)A in mRNA coding regions promotes translation via the RNA helicase-containing YTHDC2. Nat Commun *10*.

McCaughan, K.K., Brown, C.M., Dalphin, M.E., Berry, M.J., and Tate, W.P. (1995). Translational Termination Efficiency In Mammals Is Influenced By The Base Following The Stop Codon. Proceedings of the National Academy of Sciences of the United States of America *92*, 5431-5435.

McClain, W.H. (1993). Rules That Govern Transfer-Rna Identity In Protein-Synthesis. J Mol Biol *234*, 257-280.

Meissner, P., Pick, H., Kulangara, A., Chatellard, P., Friedrich, K., and Wurm, F.M. (2001). Transient gene expression: Recombinant protein production with suspension-adapted HEK293-EBNA cells. Biotechnol Bioeng *75*, 197-203.

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., Maier, L., Mackowiak, S.D., Gregersen, L.H., Munschauer, M.*, et al.* (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. Nature *495*, 333-338.

Meyer, K.D., Patil, D.P., Zhou, J., Zinoviev, A., Skabkin, M.A., Elemento, O., Pestova, T.V., Qian, S.B., and Jaffrey, S.R. (2015). 5' UTR m(6)A Promotes Cap-Independent Translation. Cell *163*, 999-1010.

Michalik, P., Piorkowski, D., Blackledge, T.A., and Ramirez, M.J. (2019). Functional trade-offs in cribellate silk mediated by spinning behavior. Scientific Reports *9*.

Michonneau, F., Brown, J.W., and Winter, D.J. (2016). rot1: an R package to interact with the Open Tree of Life data. Methods Ecol Evol *7*, 1476-1481.

Miks, S., and Bińkowski, J. (2022). Tool to Obtain Consensus Sequence (https://gene-calc.pl/sequences-analysis-tools/consensus-sequence: Gene Calc).

Miserez, A., Weaver, J.C., Pedersen, P.B., Schneeberk, T., Hanlon, R.T., Kisailus, D., and Birkedal, H. (2009). Microstructural and Biochemical Characterization of the Nanoporous Sucker Rings from Dosidicus gigas. Adv Mater *21*, 401-+.

Mo, D.D., Li, X.P., Raabe, C.A., Cui, D., Vollmar, J.F., Rozhdestyensky, T.S., Skryabin, B.V., and Brosius, J. (2019). A universal approach to investigate circRNA protein coding function. Scientific Reports *9*.

Moisenovich, M.M., Pustovalova, O.L., Arhipova, A.Y., Vasiljeva, T.V., Sokolova, O.S., Bogush, V.G., Debabov, V.G., Sevastianov, V.I., Kirpichnikov, M.P., and Agapov, II (2011). In vitro and in vivo biocompatibility studies of a recombinant analogue of spidroin 1 scaffolds. Journal of Biomedical Materials Research Part A *96A*, 125-131.

Monsellier, E., and Chiti, F. (2007). Prevention of amyloid-like aggregation as a driving force of protein evolution. EMBO Rep *8*, 737-742.

Moon, M.J. (2018). Fine structure of the aggregate silk nodules in the orb-web spider Nephila clavata. Animal Cells and Systems *22*, 421-428.

Mulinti, P., Diekjurgen, D., Kurtzeborn, K., Balasubramanian, N., Stafslien, S.J., Grainger, D.W., and Brooks, A.E. (2022). Anti-Coagulant and Antimicrobial Recombinant Heparin-Binding Major Ampullate Spidroin 2 (MaSp2) Silk Protein. Bioengineering-Basel *9*.

Muller, S., and Appel, B. (2017). In vitro circularization of RNA. RNA Biol *14*, 1018-1027.

Nachtigall, W. (1974). Biological Mechanisms of Attachment, 1 edn (Berlin, Heidelberg: Springer).

Naef, A. (1923). Die Cephalopoden (The Cephalopods). R Freidläner and Sohn,   Berlin

Namy, O., Moran, S.J., Stuart, D.I., Gilbert, R.J.C., and Brierley, I. (2006). A mechanical explanation of RNA pseudoknot function in programmed ribosomal frameshifting. Nature *441*, 244-247.

Nelms, B.L., and Labosky, P.A. (2011). A predicted hairpin cluster correlates with barriers to PCR, sequencing and possibly BAC recombineering. Scientific Reports *1*.

Niehues, S., Bussmann, J., Steffes, G., Erdmann, I., Kohrer, C., Sun, L.T., Wagner, M., Schafer, K., Wang, G.X., Koerdt, S.N.*, et al.* (2015). Impaired protein translation in Drosophila models for Charcot-Marie-Tooth neuropathy caused by mutant tRNA synthetases. Nat Commun *6*.

Niu, M.T., Ju, Y., Lin, C., and Zou, Q. (2022). Characterizing viral circRNAs and their application in identifying circRNAs in viruses. Brief Bioinform *23*.

Nixon, , M., and Dilly, P.N. (1977). Sucker surfaces and prey capture (*Symp. Zool. Soc. Lond.*), pp. 38447 -38511.

Packard, and , A. (1988). The skin of cephalopods (Coleoids): General and special adaptations (Academic Press, Inc., San Diego).

Pan, X.P., Hong, X.L., Li, S.M., Meng, P., and Xiao, F. (2021). METTL3 promotes adriamycin resistance in MCF-7 breast cancer cells by accelerating pri-microRNA-221-3p maturation in a m6A-dependent manner. Experimental and Molecular Medicine *53*, 91-102.

Pandey, P.R., Rout, P.K., Das, A., Gorospe, M., and Panda, A.C. (2019). RPAD (RNase R treatment, polyadenylation, and poly(A) plus RNA depletion) method to isolate highly pure circular RNA. Methods *155*, 41-48.

Papatheofanis, F.J. (1989). Cyto-Toxicity Of Alkyl-2-Cyanoacrylate Adhesives. J Biomed Mater Res *23*, 661-668.

Park , J.-E., Jeong , Y.J., Park , J.B., Kim , H.Y., Yoo , Y.H., Lee , K.S., Yang , W.T., Kim , D.H., and Kim, J.-M. (2019). Dietary Exposure to Transgenic Rice Expressing the Spider Silk Protein Fibroin Reduces Blood Glucose Levels in Diabetic Mice: The Potential Role of Insulin Receptor Substrate-1 Phosphorylation in Adipocytes. Development and Reproduction *23*, 223 - 229.

Park, O.H., Ha, H., Lee, Y., Boo, S.H., Kwon, D.H., Song, H.K., and Kim, Y.K. (2019). Endoribonucleolytic Cleavage of m(6)A-Containing RNAs by RNase P/MRP Complex. Mol Cell *74*, 494-+.

Patel, A.A., and Steitz, J.A. (2003). Splicing double: Insights from the second spliceosome. Nat Rev Mol Cell Biol *4*, 960-970.

Pattengale, N.D., Alipour, M., Bininda-Emonds, O.R.P., Moret, B.M.E., and Stamatakis, A. (2009). How Many Bootstrap Replicates Are Necessary? In Annual International Conference on Research in Computational Molecular Biology, pp. 184-200.

Pavlova, N.N., King, B., Josselsohn, R.H., Violante, S., Macera, V.L., Vardhana, S.A., Cross, J.R., and Thompson, C.B. (2020). Translation in amino-acid-poor environments is limited by tRNA(Gl)(n) charging. Elife *9*.

Alfaro, R.E., Griswold, C.E. and Miller, K.B. (2018). Comparative spigot ontogeny across the spider tree of life. *PeerJ*, 6, pp. 4233.

Pelletier, J., and Sonenberg, N. (1988). Internal Initiation Of Translation Of Eukaryotic Messenger-RNA Directed By A Sequence Derived From Poliovirus RNA. Nature *334*, 320-325.

Perriman, R., and Ares, M. (1998). Circular mRNA can direct translation of extremely long repeating-sequence proteins in vivo. Rna *4*, 1047-1054.

Politi, Y., Priewasser, M., Pippel, E., Zaslansky, P., Hartmann, J., Siegel, S., Li, C.H., Barth, F.G., and Fratzl, P. (2012). A Spider's Fang: How to Design an Injection Needle Using Chitin-Based Composite Material. Adv Funct Mater *22*, 2519-2528.

Potter, S.C., Luciani, A., Eddy, S.R., Park, Y., Lopez, R., and Finn, R.D. (2018). HMMER web server: 2018 update. Nucleic Acids Res *46*, W200-W204.

Prince, J.T., McGrath, K.P., Digirolamo, C.M., and Kaplan, D.L. (1995). Construction, Cloning, And Expression Of Synthetic Genes Encoding Spider Dragline Silk. Biochemistry *34*, 10879-10885.

Qu, S.B., Yang, X.S., Li, X.L., Wang, J.L., Gao, Y., Shang, R.Z., Sun, W., Dou, K.F., and Li, H.M. (2015). Circular RNA: A new star of noncoding RNAs. Cancer Lett *365*, 141-148.

Rabilloud, T. (1992). A Comparison Between Low Background Silver Diammine And Silver-Nitrate Protein Stains. Electrophoresis *13*, 429-439.

Ragan, C., Goodall, G.J., Shirokikh, N.E., and Preiss, T. (2019). Insights into the biogenesis and potential functions of exonic circular RNA. Scientific Reports *9*.

Raghavan, V., Kraft, L., Mesny, F., and Rigerte, L. (2022). A simple guide to de novo transcriptome assembly and annotation. Brief Bioinform *23*.

Rambaut, A. (2006). FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Ramos, R., Koh, K., Gabryelczyk, B., Chai, L.X., Kanagavel, D., Yan, X.B., Ganachaud, F., Miserez, A., and Bernard, J. (2021). Nanocapsules Produced by Nanoprecipitation of Designed Suckerin-Silk Fusion Proteins. Acs Macro Letters *10*, 628-634.

Rawal, A., Rhinehardt, K.L., and Mohan, R.V. (2020). Molecular Dynamics Investigation of Self-Association of Synthetic Collagen and Spider Silk Composite System for Biomaterial Applications. Mrs Advances *5*, 797-804.

Reese, M.G., Eeckman, F.H., Kulp, D., and Haussler, D. (1997). Improved splice site detection in Genie. Journal of Computational Biology *4*, 311-323.

Rhoads, A., and Au, K.F. (2015). PacBio Sequencing and Its Applications. Genomics Proteomics & Bioinformatics *13*, 278-289.

Ries, J., Schwarze, S., Johnson, C.M., and Neuweiler, H. (2014). Microsecond Folding and Domain Motions of a Spider Silk Protein Structural Switch. J Am Chem Soc *136*, 17136-17144.

Rieu, C., Bertinetti, L., Schuetz, R., Salinas-Zavala, C.C.A., Weaver, J.C., Fratzl, P., Miserez, A., and Masic, A. (2016). The role of water on the structure and mechanical properties of a thermoplastic natural block co-polymer from squid sucker ring teeth. Bioinspir Biomim *11*, 10.

Rind, F.C., Birkett, C.L., Duncan, B.J.A., and Ranken, A.J. (2011). Tarantulas cling to smooth vertical surfaces by secreting silk from their feet. J Exp Biol *214*, 1874-1879.

Rising, A., and Johansson, J. (2015). Toward spinning artificial spider silk. Nature Chemical Biology *11*, 309-315.

Rising, A., Johansson, J., Larson, G., Bongcam-Rudloff, E., Engstroem, W., and Hjalmt, G. (2007). Major ampullate spidroins from Euprosthenops australis: multiplicity at protein, mRNA and gene levels. Insect Molecular Biology *16*, 551-561.

Roberts, A.D., Finnigan, W., Kelly, P.P., Faulkner, M., Breitling, R., Takano, E., Scrutton, N.S., Blaker, J.J., and Hay, S. (2020). Non-covalent protein-based adhesives for transparent substrates-bovine serum albumin vs. recombinant spider silk. Mater Today Bio *7*, 7.

Roberts, T.C., Coenen-Stass, A.M.L., and Wood, M.J.A. (2014). Assessment of RT-qPCR Normalization Strategies for Accurate Quantification of Extracellular microRNAs in Murine Serum. Plos One *9*.

Rogozin, I.B., Wolf, Y.I., Sorokin, A.V., Mirkin, B.G., and Koonin, E.V. (2003). Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Current Biology *13*, 1512-1517.

Roper, C.F.E. (1968). Preliminary descriptions of two new species of the bathypelagic squid Bathyteuthis (Cephalopoda: Oegopsida) . *Proceedings of the Biological Society of Washington*, 161–172.

Rybak-Wolf, A., Stottmeister, C., Glazar, P., Jens, M., Pino, N., Giusti, S., Hanan, M., Behm, M., Bartok, O., Ashwal-Fluss, R.*, et al.* (2015). Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. Mol Cell *58*, 870-885.

Rypstra, A.L. (1979). Foraging Flocks of Spiders: A Study of Aggregate Behavior in Cyrtophora citricola Forskål (Araneae; Araneidae) in West Africa. Behavioral Ecology and Sociobiology *5*.

Sahni, V., Blackledge, T.A., and Dhinojwala, A. (2010). Viscoelastic solids explain spider web stickiness. Nat Commun *1*, 4.

Salzman, J., Gawad, C., Wang, P.L., Lacayo, N., and Brown, P.O. (2012). Circular RNAs Are the Predominant Transcript Isoform from Hundreds of Human Genes in Diverse Cell Types. Plos One *7*.

Sanchez-Ferrer, A., Adamcik, J., Handschin, S., Hiew, S.H., Miserez, A., and Mezzenga, R. (2018). Controlling Supramolecular Chiral Nanostructures by Self-Assembly of a Biomimetic beta-Sheet-Rich Amyloidogenic Peptide. ACS Nano *12*, 9152-9161.

Sarr, M., Kitoka, K., Walsh-White, K.A., Kaldmaee, M., Metlans, R., Tars, K., Mantese, A., Shan, D.P., Landreh, M., Rising, A.*, et al.* (2022). The dimerization mechanism of the N-terminal domain of spider silk proteins is conserved despite extensive sequence divergence. J Biol Chem *298*.

Satapathy, S., Dabbs, R.A., and Wilson, M.R. (2020). Rapid high-yield expression and purification of fully post-translationally modified recombinant clusterin and mutants. Scientific Reports *10*.

Scalzitti, N., Kress, A., Orhand, R., Weber, T., Moulinier, L., Jeannin-Girardon, A., Collet, P., Poch, O., and Thompson, J.D. (2021). Spliceator: multi-species splice site prediction using convolutional neural networks. BMC Bioinformatics *22*.

Schacht, K., Vogt, J., and Scheibel, T. (2016). Foams Made of Engineered Recombinant Spider Silk Proteins as 3D Scaffolds for Cell Growth. ACS Biomater Sci Eng *2*, 517-525.

Scharlaken, B., de Graaf, D.C., Goossens, K., Brunain, M., Peelman, L.J., and Jacobs, F.J. (2008). Reference gene selection for insect expression studies using quantitative real-time PCR: The head of the honeybee, Apis mellifera, after a bacterial challenge. J Insect Sci *8*, 10.

Scheibel, T. (2004). Spider silks: recombinant synthesis, assembly, spinning, and engineering of synthetic proteins. Microb Cell Fact *3*, 10.

Scheller, J., Guhrs, K.H., Grosse, F., and Conrad, U. (2001). Production of spider silk proteins in tobacco and potato. Nature Biotechnology *19*, 573-577.

Schmuck, B., Greco, G., Barth, A., Pugno, N.M., Johansson, J., and Rising, A. (2021). High-yield production of a super-soluble miniature spidroin for biomimetic high-performance materials. Mater Today *50*, 16-23.

Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., Leipe, D., McVeigh, R., O'Neill, K., Robbertse, B.*, et al.* (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database, 21.

Schuber, F., and Pinck, M. (1974). Chemical Reactivity Of Aminoacyl-Transfer Rna Ester Bond .1. Influence Of Ph And Nature Of Acyl Group On Rate Of Hydrolysis. Biochimie *56*, 383-390.

Schutz, D., Taborsky, M., and Drapela, T. (2007). Air bells of water spiders are an extended phenotype modified in response to gas composition. Journal of Experimental Zoology Part a-Ecological and Integrative Physiology *307A*, 549-555.

Schwager, E.E., Sharma, P.P., Clarke, T., Leite, D.J., Wierschin, T., Pechmann, M., Akiyama-Oda, Y., Esposito, L., Bechsgaard, J., Bilde, T.*, et al.* (2017). The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. BMC Biol *15*, 27.

Selden, P.A., Shear, W.A., and Sutton, M.D. (2008). Fossil evidence for the origin of spider spinnerets, and a proposed arachnid order. Proceedings of the National Academy of Sciences of the United States of America *105*, 20781-20785.

Shao, Z.Z., Vollrath, F., Yang, Y., and Thogersen, H.C. (2003). Structure and behavior of regenerated spider silk. Macromolecules *36*, 1157-1161.

Sharma, S.S., Chong, S.R., and Harcum, S.W. (2006). Intein-mediated protein purification of fusion proteins expressed under high-cell density conditions in E-coli. Journal of Biotechnology *125*, 48-56.

Shimizu, Y., Inoue, A., Tomari, Y., Suzuki, T., Yokogawa, T., Nishikawa, K., and Ueda, T. (2001). Cell-free translation reconstituted with purified components. Nature Biotechnology *19*, 751-755.

Shitivelband, S., and Hou, Y.M. (2005). Breaking the stereo barrier of amino acid attachment to tRNA by a single nucleotide. J Mol Biol *348*, 513-521.

Sidoruk, K.V., Davydova, L.I., Kozlov, D.G., Gubaidullin, D.G., Glazunov, A.V., Bogush, V.G., and Debabov, V.G. (2015). Fermentation optimization of a Saccharomyces cerevisiae strain producing 1F9 recombinant spidroin. Applied Biochemistry and Microbiology *51*, 766-773.

Simmons, J.R., Xu, L.L., and Rainey, J.K. (2019). Recombinant Pyriform Silk Fiber Mechanics Are Modulated by Wet Spinning Conditions. ACS Biomater Sci Eng *5*, 4985-4993.

Sin, Y.W., Yau, C., and Chu, K.H. (2009). Morphological and genetic differentiation of two loliginid squids, Uroteuthis (Photololigo) chinensis and Uroteuthis (Photololigo) edulis (Cephalopoda: Loliginidae), in Asia. J Exp Mar Biol Ecol *369*, 22-30.

Smart, T.G., and Thomas , P. (2005). HEK293 cell line: A vehicle for the expression of recombinant proteins. Journal of Pharmacological and Toxicological Methods *51*, 187-200.

Sonenberg, N., Morgan, M.A., Merrick, W.C., and Shatkin, A.J. (1978). Polypeptide In Eukaryotic Initiation-Factors That Crosslinks Specifically To 5'-Terminal Cap In Messenger-RNA. Proceedings of the National Academy of Sciences of the United States of America *75*, 4843-4847.

Spiess, K., Lammel, A., and Scheibel, T. (2010). Recombinant Spider Silk Proteins for Applications in Biomaterials. Macromol Biosci *10*, 998-1007.

Sponner, A., Unger, E., Grosse, F., and Weisshart, K. (2005). Differential polymerization of the two main protein components of dragline silk during fibre spinning. Nature Materials *4*, 772-775.

Stam, M., Mol, J.N.M., and Kooter, J.M. (1997). The silence of genes in transgenic plants. Annals of Botany *79*, 3-12.

Stark, M., Grip, S., Rising, A., Hedhammar, M., Engstrom, W., Hjalm, G., and Johansson, J. (2007). Macroscopic fibers self-assembled from recombinant miniature spider silk proteins. Biomacromolecules *8*, 1695-1701.

Starke, S., Jost, I., Rossbach, O., Schneider, T., Schreiner, S., Hung, L.H., and Bindereif, A. (2015). Exon Circularization Requires Canonical Splice Signals. Cell Reports *10*, 103-111.

Starrett, J., Garb, J.E., Kuelbs, A., Azubuike, U.O., and Hayashi, C.Y. (2012). Early Events in the Evolution of Spider Silk Genes. Plos One *7*, 14.

Stellwagen, S.D., and Renberg, R.L. (2019). Toward Spider Glue: Long Read Scaffolding for Extreme Length and Repetitious Silk Family Genes AgSp1 and AgSp2 with Insights into Functional Adaptation. G3-Genes Genomes Genet *9*, 1909-1919.

Stern, B., Olsen, L.C., Trösse, C., Ravneberg, H., and Pryme, I. (2007). Improving mammalian cell factories: The selection of signal peptide has a major impact on recombinant protein synthesis and secretion in mammalian cells. Trends in Cell and Molecular Biology.

Stothard, P. (2000). The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques *28*, 1102-+.

Sun, Y., Li, Y., Luo, D.Z., and Liao, D.J. (2012). Pseudogenes as Weaknesses of ACTB (Actb) and GAPDH (Gapdh) Used as Reference Genes in Reverse Transcription and Polymerase Chain Reactions. Plos One *7*.

Swanson, R., Hoben, P., Sumnersmith, M., Uemura, H., Watson, L., and Soll, D. (1988). Accuracy Of Invivo Aminoacylation Requires Proper Balance Of Transfer-Rna And Aminoacyl-Transfer RNA-Synthetase. Science *242*, 1548-1551.

Sweeney , E.E., McDaniel , R.E., Maximov , P.Y., Fan, P.i., and Jordan , V.C. (2012). Models and mechanisms of acquired antihormone resistance in breast cancer: significant clinical progress despite limitations. Hormone Molecular Biology and Clinical Investigation *9*.

Tang, N.C., and Chilkoti, A. (2016). Combinatorial codon scrambling enables scalable gene synthesis and amplification of repetitive proteins. Nature Materials *15*, 419-+.

Tateno , Y., Takezaki , N., and Nei, M. (1994). Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. Mol Biol Evol *11*.

Teule, F., Miao, Y.G., Sohn, B.H., Kim, Y.S., Hull, J.J., Fraser, M.J., Lewis, R.V., and Jarvis, D.L. (2012). Silkworms transformed with chimeric silkworm/spider silk genes spin composite silk fibers with improved mechanical properties. Proceedings of the National Academy of Sciences of the United States of America *109*, 923-928.

Thakor, N., and Holcik, M. (2012). IRES-mediated translation of cellular messenger RNA operates in eIF2 alpha-independent manner during stress. Nucleic Acids Res *40*, 541-552.

Thamm, C., and Scheibel, T. (2017). Recombinant Production, Characterization, and Fiber Spinning of an Engineered Short Major Ampullate Spidroin (MaSp1s). Biomacromolecules *18*, 1365-1372.

Theis, C., Reeder, J., and Giegerich, R. (2008). KnotInFrame: prediction of-1 ribosomal frameshift events. Nucleic Acids Res *36*, 6013-6020.

Thompson, S.R. (2012). So you want to know if your message has an IRES? Wiley Interdiscip Rev-RNA *3*, 697-705.

Thorvaldsdottir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform *14*, 178-192.

Tokareva, O., Michalczechen-Lacerda, V.A., Rech, E.L., and Kaplan, D.L. (2013). Recombinant DNA production of spider silk proteins. Microbial Biotechnology *6*, 651-663.

Traub, W., and Piez, K.A. (1971). Advances in Protein Chemistry: The Chemistry and Structure of Collagen, Vol 25 (New York, London: Academic Press, Inc.).

Tremblay, M.L., Xu, L., Lefevre, T., Sarker, M., Orrell, K.E., Leclerc, J., Meng, Q., Pezolet, M., Auger, M., Liu, X.Q.*, et al.* (2015). Spider wrapping silk fibre architecture arising from its modular soluble protein precursor. Scientific Reports *5*.

Troskie, R.L., Jafrani, Y., Mercer, T.R., Ewing, A.D., Faulkner, G.J., and Cheetham, S.W. (2021). Long-read cDNA sequencing identifies functional pseudogenes in the human transcriptome. Genome Biol *22*.

Tsuneoka, Y., and Funato, H. (2020). Modified in situ Hybridization Chain Reaction Using Short Hairpin DNAs. Frontiers in Molecular Neuroscience *13*.

Twist Bioscience (2022). What does the Twist Codon Optimization tool do? (https://www.twistbioscience.com/faq/using-your-twist-account/what-does-twist-codon-optimization-tool-do: Twist Bioscience).

Tyedmers, J., Mogk, A., and Bukau, B. (2010). Cellular strategies for controlling protein aggregation. Nat Rev Mol Cell Biol *11*, 777-788.

van Heesch, S., Witte, F., Schneider-Lunitz, V., Schulz, J.F., Adami, E., Faber, A.B., Kirchner, M., Maatz, H., Blachut, S., Sandmann, C.L.*, et al.* (2019). The Translational Landscape of the Human Heart. Cell *178*, 242-+.

Vanderveen, R., Arnberg, A.C., Vanderhorst, G., Bonen, L., Tabak, H.F., and Grivell, L.A. (1986). Excised Group-II Introns In Yeast Mitochondria Are Lariats And Can Be Formed By Self-Splicing Invitro. Cell *44*, 225-234.

Vienneau-Hathaway, J.M., Brassfield, E.R., Lane, A.K., Collin, M.A., Correa-Garhwal, S.M., Clarke, T.H., Schwager, E.E., Garb, J.E., Hayashi, C.Y., and Ayoub, N.A. (2017). Duplication and concerted evolution of MiSp-encoding genes underlie the material properties of minor ampullate silks of cobweb weaving spiders. BMC Evol Biol *17*, 18.

Vollrath, F., and Knight, D.P. (1999). Structure and function of the silk production pathway in the Spider nephila edulis. Int J Biol Macromol *24*, 243-249.

Vollrath, F., and Knight, D.P. (2001). Liquid crystalline spinning of spider silk. Nature *410*, 541-548.

Vollrath, F., Porter, D., and Holland, C. (2013). The science of silks. MRS Bull *38*, 73-80.

Walsh, G. (2014). Biopharmaceutical benchmarks 2014. Nature Biotechnology *32*, 992-1000.

Walters, B., and Thompson, S.R. (2016). Cap-independent Translational Control of Carcinogenesis. Frontiers in Oncology *6*.

Wan, Y., and Hopper, A.K. (2018). Size matters: conserved proteins function in length-dependent nuclear export of circular RNAs. Genes Dev *32*, 600-601.

Wang, K.K., Wen, R., Jia, Q.P., Liu, X.Q., Xiao, J.H., and Meng, Q. (2019). Analysis of the Full-Length Pyriform Spidroin Gene Sequence. Genes *10*, 12.

Wang, K.S., Choo, Q.L., Weiner, A.J., Ou, J.H., Najarian, R.C., Thayer, R.M., Mullenbach, G.T., Denniston, K.J., Gerin, J.L., and Houghton, M. (1986). Structure, Sequence And Expression Of The Hepatitis Delta (Delta) Viral Genome. Nature *323*, 508-514.

Wang, S.P., Latallo, M.J., Zhang, Z., Huang, B., Bobrovnikov, D.G., Dong, D.Y., Livingston, N.M., Tjoeng, W., Hayes, L.R., Rothstein, J.D.*, et al.* (2021). Nuclear export and translation of circular repeat-containing intronic RNA in C9ORF72-ALS/FTD. Nat Commun *12*.

Weichert, N., Hauptmann, V., Helmold, C., and Conrad, U. (2016). Seed-Specific Expression of Spider Silk Protein Multimers Causes Long-Term Stability. Frontiers in Plant Science *7*.

Weinander, R., Mosialou, E., Dejong, J., Tu, C.P.D., Dypbukt, J., Bergman, T., Barnes, H.J., Hoog, J.O., and Morgenstern, R. (1995). Heterologous Expression Of Rat-Liver Microsomal Glutathione Transferase In Simian Cos Cells And Escherichia-Coli. Biochem J *311*, 861-866.

Wen, R., Liu, X.Q., and Meng, Q. (2017). Characterization of full-length tubuliform spidroin gene from Araneus ventricosus. Int J Biol Macromol *105*, 702-710.

Wen, Y., Li, B., He, M., Teng, S.L., Sun, Y.X., and Wang, G.B. (2021). circHIPK3 promotes proliferation and migration and invasion via regulation of miR-637/HDAC4 signaling in osteosarcoma cells. Oncology Reports *45*, 169-179.

Wesselhoeft, R.A., Kowalski, P.S., and Anderson, D.G. (2018). Engineering circular RNA for potent and stable translation in eukaryotic cells. Nat Commun *9*, 10.

Wesselhoeft, R.A., Kowalski, P.S., Parker-Hale, F.C., Huang, Y.X., Bisaria, N., and Anderson, D.G. (2019). RNA Circularization Diminishes Immunogenicity and Can Extend Translation Duration In Vivo. Mol Cell *74*, 508-+.

Whelan, S., and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol *18*, 691-699.

Whittall, D.R., Baker, K.V., Breitling, R., and Takano, E. (2021). Host Systems for the Production of Recombinant Spider Silk. Trends Biotechnol *39*, 560-573.

Widmaier, D.M., Tullman-Ercek, D., Mirsky, E.A., Hill, R., Govindarajan, S., Minshull, J., and Voigt, C.A. (2009). Engineering the Salmonella type III secretion system to export spider silk monomers. Mol Syst Biol *5*, 9.

Wohlrab, S., Muller, S., Schmidt, A., Neubauer, S., Kessler, H., Leal-Egana, A., and Scheibel, T. (2012). Cell adhesion and proliferation on RGD-modified recombinant spider silk proteins. Biomaterials *33*, 6650-6659.

Wolff, J.O., Grawe, I., Wirth, M., Karstedt, A., and Gorb, S.N. (2015). Spider's super-glue: thread anchors are composite adhesives with synergistic hierarchical organization. Soft Matter *11*, 2394-2403.

Wood, D.W., and Camarero, J.A. (2014). Intein Applications: From Protein Purification and Labeling to Metabolic Control Methods. J Biol Chem *289*, 14512-14519.

Work, R.W., and Emerson, P.D. (1982). An Apparatus And Technique For The Forcible Silking Of Spiders. J Arachnol *10*, 1-10.

Wu, C.C.C., Peterson, A., Zinshteyn, B., Regot, S., and Green, R. (2020). Ribosome Collisions Trigger General Stress Responses to Regulate Cell Fate. Cell *182*, 404-+.

Wu, T.Y., Liono, L., Chen, S.L., Chen, C.Y., and Chao, Y.C. (2000). Expression of highly controllable genes in insect cells using a modified tetracycline-regulated gene expression system. Journal of Biotechnology *80*, 75-83.

Xia, X.X., Qian, Z.G., Ki, C.S., Park, Y.H., Kaplan, D.L., and Lee, S.Y. (2010). Native-sized recombinant spider silk protein produced in metabolically engineered Escherichia coli results in a strong fiber. Proceedings of the National Academy of Sciences of the United States of America *107*, 14059-14063.

Xu, H.T., Fan, B.L., Yu, S.Y., Huang, Y.H., Zhao, Z.H., Lian, Z.X., Dai, Y.P., Wang, L.L., Liu, Z.L., Fei, J.*, et al.* (2007). Construct synthetic gene encoding artificial spider dragline silk protein and its expression in milk of transgenic mice. Anim Biotechnol *18*, 1-12.

Xu, J., Dong, Q.L., Yu, Y., Niu, B.L., Ji, D.F., Li, M.W., Huang, Y.P., Chen, X., and Tan, A.J. (2018). Mass spider silk production through targeted gene replacement in Bombyx mori. Proceedings of the National Academy of Sciences of the United States of America *115*, 8757-8762.

Xu, M., and Lewis, R.V. (1990). Structure Of A Protein Superfiber - Spider Dragline Silk. Proceedings of the National Academy of Sciences of the United States of America *87*, 7120-7124.

Xu, Q., Cheng, D.M., Li, G.R., Liu, Y., Li, P., Sun, W.Q., Ma, D.Y., and Ni, C.H. (2021). CircHIPK3 regulates pulmonary fibrosis by facilitating glycolysis in miR-30a-3p/FOXK2-dependent manner. International Journal of Biological Sciences *17*, 2294-2307.

Yamaura, K., Okumura, Y., Ozaki, A., and Matsuzawa, S. (1985). Flow-Induced Crystallization Of Bombyx-Mori L Silk Fibroin From Regenerated Aqueous-Solution And Spinnability Of Its Solution. Applied Polymer Symposia, 205-220.

Yang, J.J., Barr, L.A., Fahnestock, S.R., and Liu, Z.B. (2005). High yield recombinant silk-like protein production in transgenic plants through protein targeting. Transgenic Res *14*, 313-324.

Yang, Y., Fan, X.J., Mao, M.W., Song, X.W., Wu, P., Zhang, Y., Jin, Y.F., Chen, L.L., Wang, Y., Wong, C.C.L.*, et al.* (2017a). Extensive translation of circular RNAs driven by N-6-methyladenosine. Cell Research *27*, 626-641.

Yang, Z.G., Awan, F.M., Du, W.W., Zeng, Y., Lyu, J.J., Wu, D., Gupta, S., Yang, W.N., and Yang, B.B. (2017b). The Circular RNA Interacts with STAT3, Increasing Its Nuclear Translocation and Wound Repair by Modulating Dnmt3a and miR-17 Function. Molecular Therapy *25*, 2062-2074.

Yao, D.Y., Liu, H.F., and Fan, Y.B. (2016). Silk scaffolds for musculoskeletal tissue engineering. Exp Biol Med *241*, 238-245.

Yaylak, B., Erdogan, I., and Akgul, B. (2019). Transcriptomics Analysis of Circular RNAs Differentially Expressed in Apoptotic HeLa Cells. Frontiers in Genetics *10*.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T.L. (2012). Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics *13*.

Yip, E.C., Rao, D., Smith, D.R., and Lubin, Y. (2019). Interacting maternal and spatial cues influence natal - dispersal out of social groups. Oikos *128*, 1793-1804.

Yoshida, R., and Nei, M. (2016). Efficiencies of the NJp, Maximum Likelihood, and Bayesian Methods of Phylogenetic Construction for Compositional and Noncompositional Genes. Mol Biol Evol *33*, 1618-1624.

Zemlin, J.C. (1968). A Study of the Mechanical Behavior of Spider Silks (Defence Technical Information Center: Collaboative Research INC Waltham, MA).

Zeng, D., and Guo, X.M. (2022). Mantle Transcriptome Provides Insights into Biomineralization and Growth Regulation in the Eastern Oyster (Crassostrea virginica). Marine Biotechnology *24*, 82-96.

Zhang, J., Zhang, X.L., Li, C.D., Yue, L.Y., Ding, N., Riordan, T., Yang, L., Li, Y., Jen, C., Lin, S.*, et al.* (2019a). Circular RNA profiling provides insights into their subcellular distribution and molecular characteristics in HepG2 cells. RNA Biol *16*, 220-232.

Zhang, M.L., Huang, N.N., Yang, X.S., Luo, J.Y., Yan, S., Xiao, F.Z., Chen, W.P., Gao, X.Y., Zhao, K., Zhou, H.K.*, et al.* (2018a). A novel protein encoded by the circular form of the SHPRH gene suppresses glioma tumorigenesis. Oncogene *37*, 1805-1814.

Zhang, M.L., Zhao, K., Xu, X.P., Yang, Y.B., Yan, S., Wei, P., Liu, H., Xu, J.B., Xiao, F.Z., Zhou, H.K.*, et al.* (2018b). A peptide encoded by circular form of LINC-PINT suppresses oncogenic transcriptional elongation in glioblastoma. Nat Commun *9*, 17.

Zhang, X.L., Xia, L.J., Day, B.A., Harris, T.I., Oliveira, P., Knittel, C., Licon, A.L., Gong, C.L., Dion, G., Lewis, R.V.*, et al.* (2019b). CRISPR/Cas9 Initiated Transgenic Silkworms as a Natural Spinner of Spider Silk. Biomacromolecules *20*, 2252-2264.

Zhang, X.O., Wang, H.B., Zhang, Y., Lu, X.H., Chen, L.L., and Yang, L. (2014). Complementary Sequence-Mediated Exon Circularization. Cell *159*, 134-147.

Zhao, A.C., Zhao, T.F., Nakagaki, K., Zhang, Y.S., SiMa, Y.H., Miao, Y.G., Shiomi, K., Kajiura, Z., Nagata, Y., Takadera, M.*, et al.* (2006). Novel molecular and mechanical properties of egg case silk from wasp spider, Argiope bruennichi. Biochemistry *45*, 3348-3356.

Zhao, H.S., Heusler, E., Jones, G., Li, L.H., Werner, V., Germershaus, O., Ritzer, J., Luehmann, T., and Meinel, L. (2014). Decoration of silk fibroin by click chemistry for biomedical application. Journal of Structural Biology *186*, 420-430.

Zheng, K., and Ling, S.J. (2019). De Novo Design of Recombinant Spider Silk Proteins for Material Applications. Biotechnol J *14*, 11.

Zheng, Q.P., Bao, C.Y., Guo, W.J., Li, S.Y., Chen, J., Chen, B., Luo, Y.T., Lyu, D.B., Li, Y., Shi, G.H.*, et al.* (2016). Circular RNA profiling reveals an abundant circHIPK3 that regulates cell growth by sponging multiple miRNAs. Nat Commun *7*.

Zhong, W. (2016). Advances in Smart Medical Textiles, *Langenhove, L. van* edn (Woodhead Publishing).

Zhou, J., Wan, J., Shu, X.E., Mao, Y.H., Liu, X.M., Yuan, X., Zhang, X.Q., Hess, M.E., Bruning, J.C., and Qian, S.B. (2018). N-6-Methyladenosine Guides mRNA Alternative Translation during Integrated Stress Response. Mol Cell *69*, 636-+.

Zhou, S.B., Peng, H.S., Yu, X.J., Zheng, X.T., Cui, W.G., Zhang, Z.R., Li, X.H., Wang, J.X., Weng, J., Jia, W.X.*, et al.* (2008). Preparation and characterization of a novel electrospun spider silk fibroin/poly(D,L-lactide) composite fiber. J Phys Chem B *112*, 11209-11216.

Zhou, S.Y., Dong, Q.L., Zhu, K.S., Gao, L., Chen, X., and Xiang, H. (2021). Long-read transcriptomic analysis of orb-weaving spider Araneus ventricosus indicates transcriptional diversity of spidroins. Int J Biol Macromol *168*, 395-402.

Zhou, W.B., Karcher, D., and Bock, R. (2013). Importance of adenosine-to-inosine editing adjacent to the anticodon in an Arabidopsis alanine tRNA under environmental stress. Nucleic Acids Res *41*, 3362-3372.

Zhuang, M.R., Li, X.B., Zhu, J.D., Zhang, J., Niu, F.G., Liang, F.H., Chen, M.X., Li, D., Han, P., and Ji, S.J. (2019). The m(6)A reader YTHDF1 regulates axon guidance through translational control of Robo3.1 expression. Nucleic Acids Res *47*, 4765-4777.