

Spider Silk:

From super strong webs to customisable synthetic protein

Charlotte Rose Smith

(née Deall)

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

September 2021

Supervisors:

Professor Sara Goodacre

Professor Neil Thomas


Charlotte's
Webs

Acknowledgements

First, I have to thank the BBSRC and the University of Nottingham for the funding that has made this project possible. To my amazing supervisor Sara, thank you for all those long chats, tea at the lake and your patience and humour as we navigated 'Life's Rich Tapestry'. You are an inspiration for all women in science, with an infectious optimism and unfailing ability to work with a range of people – always with a smile. I hope that more people can learn from you. To my second supervisor Neil, who provided invaluable advice, some hard chemistry questions and opinions that I could trust, thank you.

I want to thank Ella, for being the best friend (and bridesmaid!) I could ask for, helping both in the lab and out. I will always treasure fond memories of dealing with errant tarantulas at outreach events, too-warm fieldwork trips and navigating new cities together. Also, to Ryan, without whom I wouldn't have got most of my protein characterisation data, and would probably still be stuck in Oxford! Thank you to Hannah, for always being ready to do something to take our minds off lab work. I am grateful to all those Spiderlab members past and present, from Life Sciences to CBS, for great conversations, pub trips and shared spider experiences. To the undergraduate students with whom I had my first supervisory experiences, thank you for bearing with me! I also want to thank Professor John Brookfield, for his patience and help with my first experiences of genetic analysis. A lot of this wouldn't have been possible without the arachnologists that helped me to really love this field and willingly shared their time, experience and material; special thanks to Milan Rezác and Marco Isaia.

Every day is made a happier place by Percy and Ivor, who don't understand how much they do. I am grateful indeed to my parents, who have supported me from the beginning, for being role models I can only hope to aspire to be. To Christopher, who is both the best and most annoying little brother I could wish for, thanks for being on my team.

And finally, my sincere thanks go wholeheartedly to my husband Matt, for being my rock throughout it all. This section isn't long enough to list all the ways that you have helped, but I wouldn't have made it here without you. For your unfailing enthusiasm, optimism, love and support I will be eternally grateful.

COVID19 impact statement

In 2020, people's lives across the world were fundamentally changed; both professionally and personally due to the COVID-19 pandemic. From March 2020 this project experienced a number of disruptions and roadblocks to completing the planned activities in the funded period.

The major impact was the lack of accessibility of laboratories and specialist equipment, which meant reduced experimental evidence for some of the investigations in Chapter 3, especially in regard to correlation of genetic sequence with physical structures. I was also unable to complete physical diving bell characterisation and further *Argyroneta aquatica* silk observations, as well as PCR amplification of various transcripts of questionable fidelity from the long-read dataset. It was also not possible to travel to and access the Research Complex at Harwell to continue protein characterisation and additional construct expression as set out in Chapter 4.

The project path was adapted quite significantly in response to these challenges, as work was shifted to have a smaller focus on practical lab work. This included the acquisition of long read data in early 2021 and more in-depth sequencing analysis which made up the bulk of Chapters 2 and 3, as well as the omission of gas exchange data and planned venom/silk physical characterisation experiments. Much of this, and other future work, is discussed in Chapter 5.

Declaration

I declare that this thesis is the result of my own work, unless otherwise stated, and that the work contained within has not been presented as part of any other degree.



Charlotte Rose Smith

Abstract

Spider silks display a remarkable array of properties, making them an area of great interest in biotechnological research. Silk has many of the desired properties needed for a biomaterial; it is flexible, biocompatible and can have remarkable tensile strength, which makes it as strong and tough as some man-made materials such as Kevlar and steel. This combination of physical properties and biocompatibility makes it an especially interesting candidate for both bioengineering and medical applications. Due to this, and the relative difficulty of farming spiders due to their cannibalistic nature, much interest has focussed on the recombinant production of synthetic spider silk.

Unusual silks also have the potential to be invaluable in the pursuit of unique and tailored biomaterials. For example, the water spider *Argyroneta aquatica* uses its silk to create a diving bell, from where it spends most of its life submerged. The silk used to create this structure is gas permeable, which makes it an interesting case for further investigation. By learning from nature, we can hope to recapitulate these complex materials, adapting and applying them to biotechnological, medical, or industrial applications.

In this thesis, three bodies of work have been undertaken to further the knowledge of natural spider silk and the production of silk mimics in the laboratory. Each research chapter explores the diverse theme of spider silk in extreme environments. This research has created a pipeline for the identification, characterisation and translation of unusual natural silks into recombinant proteins.

Using both short and long read RNA sequencing technology, Chapters 2 and 3 investigate the amino acid makeup of the silk of spider species which live in extreme abiotic environments. Chapter 2 presents the first sequenced silk genes from linyphiid and nesticid spiders and the first genetic evidence that these spiders use at least four of the most well-characterised silk types, as well as uncovers a potentially novel silk type. It is shown here, using transcriptomic assemblies, that the silks used

by spiders in acidic environments assemble in the expected way, while harbouring some diverged repetitive region motifs which may be key to their function.

Chapter 3 uses SMRT long read sequencing to present the first full length silk gene sequences from the water spider, *Argyroneta aquatica,* and the colonial tent-weaver, *Cyrtophora citricola*. In addition, short read RNA sequencing is used to create several robust transcriptomes from a range of spiders and tissue types. These sequences reveal environment and species specific silk adaptations, and proposes that specific acidic and basic residues could be important environmental adaptations. This long read data set also highlights the presence of frameshifts in silk sequences, which could be a possible novel way of naturally increasing spidroin diversity at the genetic level.

Recent technological advancements have enabled the recombinant production of artificial miniaturised silk proteins in *Escherichia coli*, and Chapter 4 uses the knowledge gained from natural silk transcripts to design and express a specialised *A. aquatica* spidroin. Several techniques are used, including Analytical Ultra Centrifugation and Mass Spectrometry, to characterise the dimerisation properties of the C- terminal domain, highlighting the importance of the repetitive region in fibre formation. This research demonstrates the utility of a modular approach to expressing recombinant silk proteins, and several novel constructs are designed to represent other silk types.

These studies highlight several approaches for the investigation into the silk genes of spiders that live in usual environments, and shows that the principle of recombinant production of miniaturised spidroins can be applied to more diverse silk types, to create different proteinaceous materials. With further refining, this pipeline could be used to efficiently design and create a range of functionalised recombinant spidroins, using natural spidroins as direct inspiration, with properties tailored for use as silk-based biomaterials.

Graphical Abstract

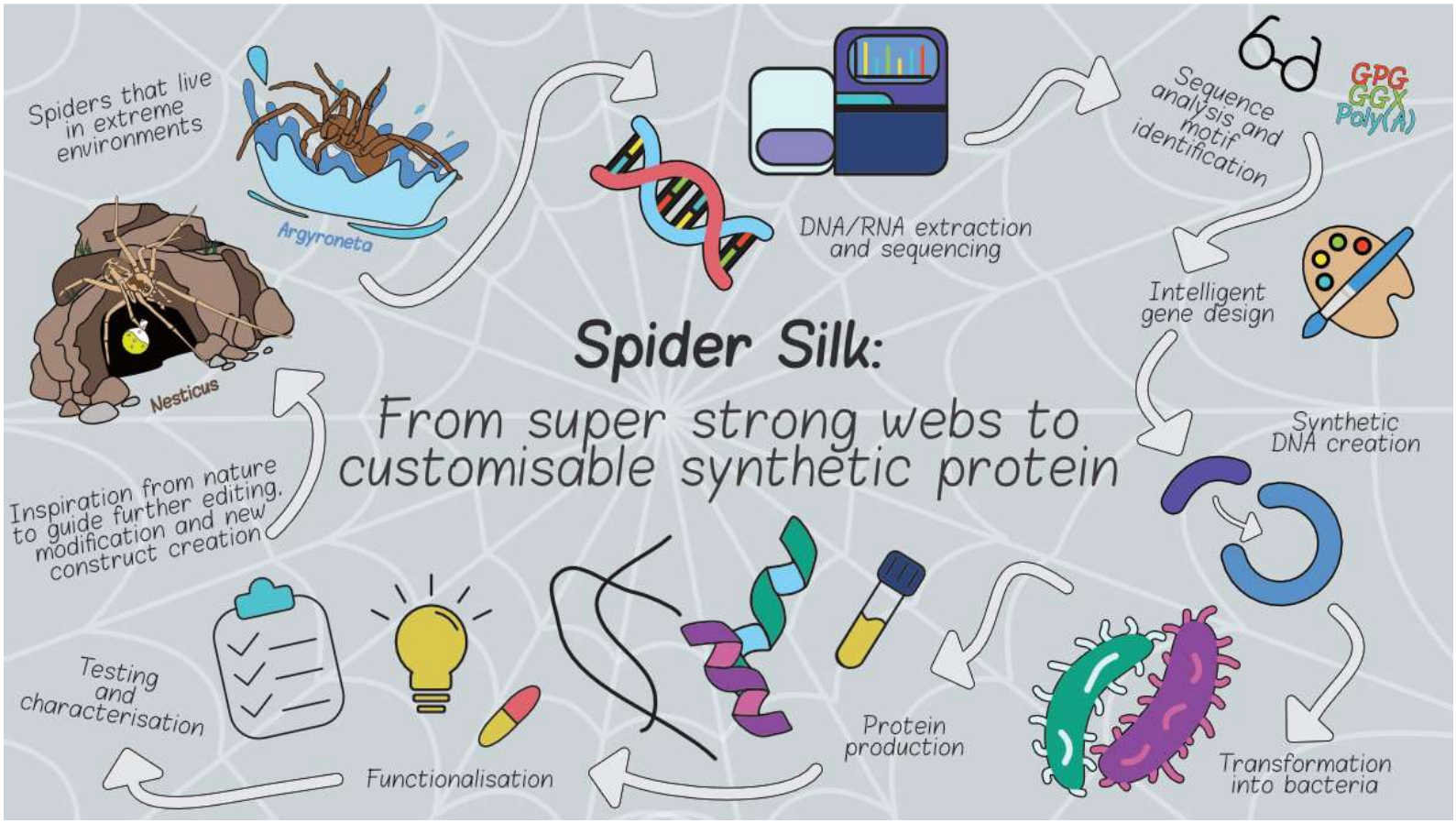


Table of Contents

Acknowledgements ..... **i**

COVID-19 Impact Statement ..... **ii**

Declaration ..... **ii**

Abstract ..... **iii**

Graphical Abstract ..... **v**

Table of Contents ..... **vi**

List of Figures ..... **xiii**

List of Tables ..... **xvi**

Commonly Used Abbreviations ..... **xvii**

Chapter One – An Introduction

1.1 An Introduction to Natural Silk **2**

1.2 Types of Spider Silk **4**

1.2.1 Major Ampullate (MaSp) 4

1.2.2 Flagelliform and Aggregate (FlaG and AgSp) 5

1.2.3 Minor Ampullate (MiSp) 5

1.2.4 Aciniform (AcSp) 5

1.2.5 Tubuliform (TuSp) 6

1.2.6 Pyriform (PySp) 6

1.3 From Protein Secretion to Fibre Formation 8

1.4 The Role of Repetitive Motifs in Different Silk Types 10

1.5 Next Generation Sequencing Methods 12

1.5.1 Illumina Sequencing 14

1.5.2 SMRT Long Read Sequencing 16

1.6 Spiders in Extreme Environments and *Argyroneta aquatica* 19

1.7 Silk and Biotechnology 26

1.8 Uses of Recombinant Spidroins 29

1.9 Aims and Objectives 33

Chapter Two:

An exploration of the first sequenced linyphiid and nesticid silk genes; represented by cave dwellers

2.1 Introduction 36

2.1.1 Spider Silk 36

2.2.2 Linyphiids and Nesticids 38

2.2 Materials and Methods 44

2.2.1 Spider Collection 44

2.2.2 RNA Extraction 44

2.2.3 Sequencing 44

2.2.4 Transcriptome Assembly and Quality Control 44

2.2.5 Identification of Novel Spidroins 45

2.2.6 Defining N- and C-Terminal Domains 49

2.3 Results 48

2.3.1 RNA Extraction, Sequencing and Transcriptome Assembly 48

2.3.2 *Lessertia barbara* Population Analysis 51

2.3.3 Identifying Spidroin Sequences 51

2.3.4 Amino Acid Designations and Properties 54

2.3.5 Major Ampullate Silk (MaSp) 54

2.3.6 Potential Novel MaSp 58

2.3.7 Minor Ampullate Silk (MiSp) 59

2.3.8 Pyriform Silk (PySp) 60

2.3.9 Tubuliform Silk (TuSp) 60

2.3.10 Aggregate Silk (AgSp) 61

2.3.11 Flagelliform Silk (FlaG) 62

2.4 Discussion 63

2.4.1 Linyphiid and Nesticid Silks Assemble in the Expected Way 63

2.4.2 Potential Novel Silk Type 65

2.4.3 Other Silk Types 65

2.4.4 Limited Influence of External pH Environment on Silk Assembly 70

2.4.5 Optimisation of Silks in Resource Limited Conditions 71

2.4.6 Future Studies 73
2.4.7 Conclusions 74

Chapter Three:

Using RNA sequencing technologies to uncover novel and full-length silk genes in *Argyroneta aquatica* and *Cyrtophora citricola*

3.1 Introduction 76
3.1.1 From the Depths of Lakes to Desert Expanses 76
3.1.2 Next Generation Sequencing (NGS) 80
3.2 Materials and Methods 83
3.2.1 Short Read Sequencing 83
3.2.1.1 RNA Extraction 83
3.2.1.2 Sequencing 83
3.2.1.3 Transcriptome Assembly and Quality Control 84
3.2.1.4 Identification of Novel Spidroins 84
3.2.2 Long Read Sequencing 85
3.2.2.1 RNA Extraction 85
3.2.2.2 Sequencing 85
3.2.2.3 Sequence QC and Assembly 86
3.2.2.4 High Quality Transcript Classification and Annotation 86
3.2.2.5 Sequence comparisons using phylogenetics 86
3.3 Results 87
3.3.1 Short Read Sequencing 87
3.3.1.1 RNA Extraction 87
3.3.1.2 Transcriptome Assembly and Quality Control 87
3.3.2 Long Read Sequencing 91
3.3.2.1 RNA Extraction 91
3.3.2.2 Sequencing 91
3.3.2.3 CCS Read Statistics 92
3.3.2.4 Protein Coding Sequence Prediction 93
3.3.2.5 High Quality Transcript Classification and Annotation 94

3.3.3 Assembly and Quality Control 94
3.3.4 Comparison of NGS Methods Used in This Study 98
3.3.5 Identification of Novel Spidroins 98
3.3.6 *A. aquatica* Major Ampullate Silk 101
3.3.6.1 MaSp Terminal Regions 101
3.3.6.2 Repetitive Regions 103
3.3.7 Putative Full-length *A. aquatica* Major Ampullate Spidroins 104
3.3.8 Other *A. aquatica* Spidroin Types 109
3.3.9 *C. citricola* Major Ampullate Silk 111
3.3.9.1 MaSp Terminal Regions 111
3.3.10 Putative Full-length *C. citricola* Major Ampullate Spidroin 112
3.3.11 Other *C. citricola* Spidroin Types 114
3.4 Discussion 116
3.4.1 The First Full-length Spidroins of *A. aquatica* and *C. citricola* 116
3.4.2 Species and Possible Environment Specific Adaptations 119
3.4.3 Frameshift Nonsense or Additional Diversity? 120
3.4.4 Using Next Generation Sequencing to Study Spidroins 121
3.4.5 Conclusions 123

Chapter Four:

Creation and analysis of a novel recombinant aquatic silk protein

4.1 Introduction 125
4.1.1 Aquatic Spider Silk 126
4.1.2 Recombinant Spidroin Production in *E. coli* 128
4.1.3 4RepCT 129
4.2 Materials and Methods 133
4.2.1 General Conditions 133
4.2.2 pJE401-AquaRepCT 134
4.2.2.1 pJE401-AquaRepCT Plasmid Retrieval 134
4.2.2.2 Transformation 134
4.2.2.3 Small-Scale Expression Test 134

4.2.2.4 SDS-PAGE analysis (standard protocol) 135
4.2.2.5 Litre- scale Expression Test in LB 136
4.2.2.6 Expression in Minimal Media 136
4.2.3 pET32b-AquaRepCT 136
4.2.3.1 pET32b-AquaRepCT Creation 136
4.2.3.2 Expression Test 139
4.2.3.3 Litre-scale LB AquaRepCT Expression and Purification 139
4.2.3.4 Protein Concentration and Cleavage 139
4.2.3.5 Solubility Tests 140
4.2.4 pET32b-AquaCT 140
4.2.4.1 pET32b-AquaCT Vector Design 140
4.2.4.2 AquaCT Expression 140
4.2.4.3 AquaCT Purification 141
4.2.4.4 Large Scale Expression, Cleavage and Tag Separation 141
4.2.4.5 Size Exclusion Chromatography and Analysis 142
4.2.4.6 Protein Crystallography 142
4.2.4.7 Mass Spectrometry 143
4.2.4.8 Analytical Ultra Centrifugation 143
4.2.5 Size Exclusion Chromatography Multi Angle Light Scattering 143
4.2.6 Additional Construct Generation 1463
4.2.6.1 TEV-AquaCT 143
4.2.6.2 NTMaSp/AcSp 144
4.2.6.3 NTXX-AquaRepCT 144
4.3 Results 145
4.3.1 pJE401-AquaRepCT 145
4.3.1.1 Small-scale Expression Test 145
4.3.1.2 Litre-scale Expression Test and Purification 145
4.3.1.3 AquaRepCT Expression in M9 Minimal Media 147
4.3.2 pet32b-AquaRepCT 148
4.3.2.1 pET32b-AquaRepCT Creation 148
4.3.2.2 Comparison of pET-32b-AquaRepCT and 4RepCT 148

4.3.2.3 Litre-scale Expression and Purification 152
4.3.2.4 Cleavage and Solubility Tests 152
4.3.3 pET32b-AquaCT 154
4.3.3.1 pET32b-AquaCT Vector Design and Creation 154
4.3.3.2 AquaCT Expression and Purification 154
4.3.3.3 Tag Cleavage and Separation 155
4.3.3.4 Large Scale Cleavage of Tags 157
4.3.3.5 Size Exclusion Chromatography and Analysis 157
4.3.3.6 Protein Crystallography 159
4.3.3.7 Mass Spectrometry 159
4.3.3.8 Analytical Ultra Centrifugation 162
4.3.3.9 SEC-MALS 162
4.4 Discussion 164
4.4.1 Future Work 168
4.4.2 Conclusions 169

Chapter Five:

General Discussion

Final summaries and recommendations for future work

5.1 Summary 172
5.2 Future Study of Silks in Extreme Environments 173
5.3 Future Recombinant Mini-Spidroin Expression 174
5.4 Implications of Results 175

List of References 177

PIP Reflective Statement 203

The Appendices 206

Appendices for Chapter 2

A.2.1 Summary of silk sequences from each transcriptome 206

A.2.2 Amino acid composition of MaSp repetitive regions 208

A.2.3 Spidroin sequences from each transcriptome 210

Appendices for Chapter 3

A.3.1 Sequences used for phylogenetic analysis 217

A.3.2 Alignments of *C. citricola* MaSp2 terminal domains 219

Appendices For Chapter 4

A.4.1 pJE401 AquaRepCT Information 222

A.4.2 pET32b AquaRepCT Information 224

A.4.3 pET32b AquaCT Information 226

A.4.4 pET32b TEV-AquaCT 228

A.4.5 N-terminal Sequence Information 230

Appendices for Chapter 5

A.5.1 Recombinant Silk 1 – Linyphiid Silk (MaSp1) 234

A.5.2 Recombinant Silk 2 – Nesticid Silk (Possible Acidic Specific MaSp) 234

A.5.3 Recombinant Silk 3 - Aquatic Silk (TuSp) 235

A.5.4 Recombinant Silk 4 - Aquatic Silk (AcSp) 236

List of Figures

Figure 1.1: General schematic structure of a major ampullate spidroin gene 3

Figure 1.2: Different silk types and their uses 7

Figure 1.3: Silk gland regions 8

Figure 1.4: The proposed structural model of a major ampullate spider silk fibre... 10

Figure 1.5: The process of Illumina sequencing 15

Figure 1.6: The generation of subreads and Circular Consensus Sequences (CCS) ... 18

Figure 1.7: *Argyroneta aquatica* 22

Figure 1.8: Structure and organisation of the hydrophobic hairs of *Argyroneta aquatica* 23

Figure 1.9: e-SEM images of the diving bell 24

Figure 1.10: Hydrogel and water on the silk threads of *A. aquatica* 25

Figure 2.1: General protein structure of a major ampullate spidroin 36

Figure 2.2: Different silk types 37

Figure 2.3: *Lessertia barbara* 38

Figure 2.4: *Nesticus eremita* 39

Figure 2.5: Summary of spider relationships based on phylogenetic analyses 41

Figure 2.6: Example webs produced by Araneids, Linyphiids and Nesticids 42

Figure 2.7: Schematic of spidroin NTD 46

Figure 2.8: Schematic of spidroin CTD 47

Figure 2.9: BUSCO completeness scores for each assembled transcriptome 49

Figure 2.10: Comparison of published BUSCO scores for a range of spider gene assemblies 50

Figure 2.11: Alignment of CO1 genes from *Lessertia barbara* transcriptomes 51

Figure 2.12: *L. barbara* MaSp NTD alignment with *E. australis* MaSp NTD 55

Figure 2.13: General protein structure of a *Lessertia barbara* major ampullate spidroin 55

Figure 2.14: Visual overview of amino acid composition of selected MaSp C-terminal domains 57

Figure 2.15: Composition of MaSp1, 2 and 3 silks compared to the novel silk type from this study 58

Figure 3.1: *A. aquatica* and *C. citricola* example habitats 77

Figure 3.2: BUSCO Assessment scores for short read transcriptomes 89

Figure 3.3: A comparison of published BUSCO scores for a range of spider gene assemblies 90

Figure 3.4: SMRT sequencing read length distribution..... 92

Figure 3.5: SMRT Sequencing subread length distribution..... 92

Figure 3.6: CCS read length distributions..... 93

Figure 3.7: Annotated Unigene ratios..... 94

Figure 3.8: BUSCO assessment scores for long-read assemblies and CCS reads..... 96

Figure 3.9: BUSCO assessment scores for a selection of assembly methods using the *C. citricola* CCS read dataset 97

Figure 3.10: *A. aquatica* MaSp N- terminal domain types..... 102

Figure 3.11: *A. aquatica* MaSp C- terminal domain types 102

Figure 3.12: Example MaSp1 repeat with structure prediction 104

Figure 3.13: Neighbour joining tree of a selection of MaSp and MiSp CTDs..... 106

Figure 3.14: Bayesian inference of phylogeny tree of a selection of MaSp and MiSp C- terminal domains 107

Figure 3.15: Hydrophobicity plot of terminal domains..... 108

Figure 3.16: Hydrophobicity plot of *A. aquatica* full-length MaSp1 repetitive region 108

Figure 3.17: Representative repeat regions of different types of *A. aquatica* silks. 110

Figure 3.18: Schematic of *C. citricola* NTD 111

Figure 3.19: *C. citricola* MaSp NTD types 111

Figure 3.20: *C. citricola* C-terminal domain types..... 112

Figure 3.21: *C. citricola* MaSp2 schematic 112

Figure 3.23: *C. citricola* MaSp2 repeat block consensus sequence 113

Figure 3.23: Hydrophobicity plot of *C. citricola* MaSp repetitive region 113

Figure 3.24: Hydrophobicity plot of terminal regions of *C. citricola* MaSp2 114

Figure 3.25: *C. citricola* MaSp2 NTD linker region with structural prediction 114

Figure 4.1: General protein structure of a spidroin 125

Figure 4.2: *Argyroneta aquatica* inside its diving bell..... 127

Figure 4.3: Schematic representation of 4RepCT 130

Figure 4.4: AquaRepCT schematic..... 145

Figure 4.5: SDS PAGE gel from small scale expression test of pJE401-AquaRepCT. 146

Figure 4.6: SDS PAGE gel from large scale expression test of pJE401-AquaRepCT . 146

Figure 4.7: SDS PAGE gel from large scale expression test of pJE401-AquaRepCT grown in 1L minimal media 147

Figure 4.8: Hydrophobicity plots of AquaRepCT and 4RepCT 149

Figure 4.9: 4RepCT and AquaRepCT sequence alignment 149

Figure 4.10: Phyre2 secondary structure prediction of 4RepCT 150

Figure 4.11: Phyre2 secondary structure prediction of AquaRepCT 150

Figure 4.12: SDS PAGE gel purification of Rosetta-pET32b-AquaRepCT grown at litre scale 152

Figure 4.13: SDS PAGE gel showing thrombin cleavage of AquaRepCT 156

Figure 4.14: AquaCT schematic 154

Figure 4.15: SDS PAGE gel from large scale expression of Rosetta-AquaCT cells grown in 1L LB 154

Figure 4.16: SDS-PAGE gel of TRX-HIS-AquaCT at various concentrations 155

Figure 4.17: SDS-PAGE gels of TRX-HIS-AquaCT thrombin cleavage 156

Figure 4.18: SDS-PAGE gel of purification, desalt, cleavage and rev-HIS tag separation of AquaCT 157

Figure 4.19: Size exclusion SDS-PAGE gel showing the purification of AquaCT 158

Figure 4.20: SDS-PAGE of reducing gel of Sample A and B 159

Figure 4.21: Mass spectra of Sample A (suspected dimer) in a native state 160

Figure 4.22: Mass spectra of Sample A reduced by β-mercaptoethanol 160

Figure 4.23: Mass spectra of Sample B in native state 161

Figure 4.24: Mass spectra of Sample B reduced by β-mercaptoethanol 161

Figure 4.25: Analytical Ultra- Centrifugation results 162

Figure 4.26: Size Exclusion Chromatography trace 163

List of Tables

Table 1.A: A comparison of properties from a selection of natural, recombinant silks and manmade materials 32

Table 2.A: Total RNA and RIN scores for each sample used for sequencing 48

Table 2.B: Assembly metrics for each transcriptome 49

Table 2.C: Number of identified spidroin sequences and terminal domains used in analysis 52

Table 2.D: A summary of silk types and their general amino acid characteristics 53

Table 2.E: Amino acids, their one letter code and general properties 54

Table 2.F: Summary of the amino acid make up of various MaSp CTDs 56

Table 3.A: Summary of total RNA extracted (in µg) and RIN scores 87

Table 3.B: Assembly metrics of short read transcriptomes using QUAST 88

Table 3.C: Extracted RNA concentrations and quality scores 91

Table 3.D: Polymerase Read and Subread Statistics 91

Table 3.E: CCS read statistics 93

Table 3.F: Predicted protein coding sequences in CCS reads 93

Table 3.G: QUAST and Fasta statistic assessment of CCS read datasets and the best performing assembly 95

Table 3.H: Comparison of NGS methods used in this study 98

Table 3.I: Number of identified spidroin sequence fragments from whole-abdomen short read transcriptomes 99

Table 3.J: Full-length Spidroin Identification 100

Table 3.K: Full-length Silk Sequence Lengths 100

Table 3.L: *A. aquatica* MaSp Repeat types 103

Table 4.A: Standard Buffers 133

Table 4.B: *E. coli* strains used in this study and their features 133

Table 4.C: Primers used in this study 138

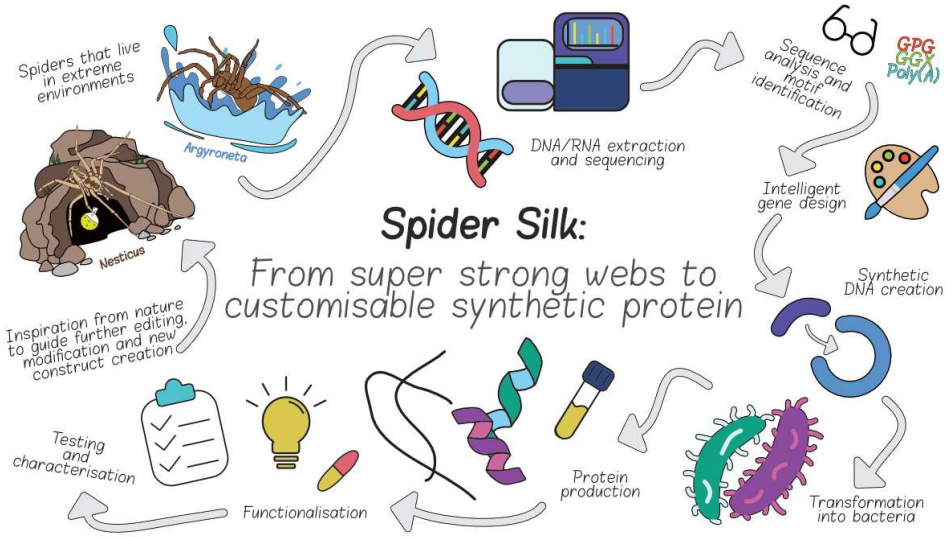
Table 4.D: PCR Programs A and B 138

Table 4.E: Comparison of properties of AquaRepCT and 4RepCT 151

Table 4.F: TRX-AquaRepCT in different buffers 153

Commonly used abbreviations

MaSp	Major ampullate silk	Dragline, the strongest silk type
MiSp	Minor ampullate silk	Used for the auxiliary spiral in an orb web
AcSp	Aciniform silk	Used for prey wrapping
AgSp	Aggregate silk	Sticky viscid silk, coats flagelliform fibres
TuSp	Tubuliform silk	Used to create the outer egg sac
FlaG	Flagelliform silk	The most extensible silk, coated with aggregate glue and used for the capture spiral in an orb web
PySp	Pyriform silk	Attachment glue silk
NTD	N-terminal domain	Non-repetitive amino terminal domain of a spidroin
CTD	C-terminal domain	Non-repetitive carboxyl terminal domain of a spidroin
IME	Intron Mediated Enhancement	The ability of an intron sequence to enhance the expression of a gene containing that intron



Chapter One

AN INTRODUCTION

1.1 An Introduction to Natural Silk

Silk is a protein-based, fibre forming structural protein, which often exhibits surprising and varied properties (McGrath and Kaplan, 1996). The ability to produce silk has evolved independently multiple times, but is only produced arthropods, specifically by individuals of the classes Insecta, Arachnida and Myriapoda (Craig, 1997, 2003; Sutherland *et al.,* 2010). Silken structures are often used to provide protection, assist with prey capture or help with locomotion (Nentwig and Heimer, 1987; Krafft and Cookson, 2012, Sutherland et al., 2011).

The larva of the domesticated silk moth, *Bombyx mori,* is probably the most common organism associated with silk production as it has been farmed and its silk harvested for human use for centuries. These silk fibres are called fibroins and are used to construct silken cocoons, inside which the larvae undergo metamorphosis (Chen, Porter and Vollrath, 2012). The fibroin is modular and made up of a heavy and a light chain connected by a disulphide bond. This is coated with sericin, a hydrophilic natural polymer, which is adhesive and can bind fibroins together, as well as act as a lubricant (Asakura, Okushita and Williamson, 2015; Sparkes and Holland, 2018). Sericin is often undesirable for downstream fibroin applications, but research has shown that it has potential to be used as a biomaterial in isolation from the fibroin itself (Wray *et al.*, 2011; Lamboni *et al.*, 2015; Suryawanshi *et al.*, 2020).

Another insect which makes silk, primarily for prey capture, is the glow-worm, *Arachnocampa spp.,* which lives in New Zealand caves. It uses silk fibres combined with an adhesive substance to create 'fishing lines and lures' which dangle from the walls (Von Byern *et al.*, 2016). This silk is mainly composed of crossed β-sheets which gives it a low tensile strength and could be an adaptation to protecting the nest in the case of snaring a larger prey item (von Byern *et al.*, 2019). These lines are highly extensible and well adapted to high relative humidity (Walker *et al.*, 2015).

Aquatic caddisfly larvae, *Trichoptera spp.*, use adhesive silks to construct a range of structures underwater, which often incorporate small rocks or sticks and act as protection during foraging or resting. These silks are often made up of repeated serine-X (where X is any other amino acid) motifs and the silks are also arranged in a

modular structure (Frandsen *et al.*, 2019). It is the phosphorylation of these serines and the accumulation of basic residues that have been postulated to allow underwater spinning (Stewart and Wang, 2010). The silks themselves are flat and ribbon-like and some progress has also been made towards using these as a medical biomaterial (Tszydel *et al.*, 2015; Kim, Sun and Moon, 2020).

Silk proteins produced by spiders are interesting because they have similar genetic organisations, but with a wide range of properties and behaviour (Römer and Scheibel, 2008). They are often tougher, stronger and less immunogenic than the silk produced by their insect counterparts (Kuhbier *et al.*, 2010). Natural spider silk proteins form several different types of materials, from strong fibres to sticky glue coatings. Spider silk proteins, called spidroins, are typically produced in specialised glands in the spider's opisthoma (abdomen). For example, orb weavers from the family Araneidae have seven different types of gland, which each produce differing amounts and types of specialised silk (Andersson *et al.*, 2013). Recently, it was shown that silk expression is not just restricted to these glands and up to 28 different spidroins produced by a single spider have been discovered, although this number varies between species studied (Babb *et al.*, 2017).

Spidroins are stored as a highly concentrated 'dope' in the glands and form filaments through changes in pH, altered salt composition and shear forces resulting from extrusion (Askarieh *et al.*, 2010). Generally, natural spidroins are made up of a non-repetitive N-terminal domain, a large region made up of repetitive amino acid units and a non-repetitive conserved C-terminal domain (Challis, Goodacre and Hewitt, 2006; Ayoub *et al.*, 2007). A schematic of the general gene structure of a major ampullate (MaSp) spidroin is shown in Figure 1.1.

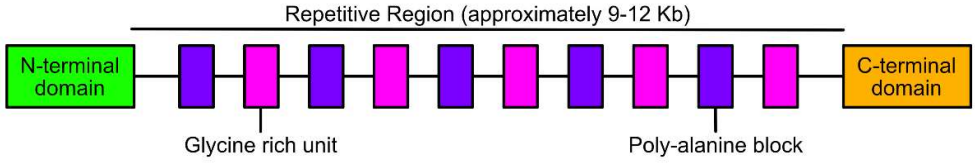


Figure 1.1: General schematic structure of a major ampullate spidroin gene.
The genes, much like their coded proteins, are large and are often mainly composed of alternating coding units. The N- and C-terminal domains are conserved between species and play important roles in protein folding.

Silks can be made from a combination of different spidroins, or from many repeats of the same one. The N-terminal domain of a silk protein controls solubility in the silk gland and undergoes pH dependent changes which triggers fibre formation (Hayashi and Lewis, 1998; Challis, Goodacre and Hewitt, 2006). The C-terminal domain is highly conserved between species, with up to 30% codon identity across the whole region (Strickland *et al.*, 2018). This remarkable degree of conservation allows it to carry out its important role in controlling the assembly of repetitive segments into fibres in all known silk types (Ittah *et al.*, 2006).

1.2 Types of Spider Silk

There are several different types of spider silk fibre, each specialised for a different use. Each silk type can be made up of one or more different spidroins, and spidroins of the same type are made in a specialised silk gland, which is connected to a set of spinnerets, from which the silk fibre emerges. Orb web weaving spiders, which include the genera *Trichonephila* and *Argiope*, produce seven different types, with different properties. These include major ampullate (MaSp), minor ampullate (MiSp), pyriform (PySp), tubuliform (TuSp), flagelliform (FlaG), aggregate (AgSp) and aciniform silk (AcSp). The way that each is used in web construction is detailed below and also shown in Figure 1.2.

1.2.1 Major Ampullate (MaSp)

One of the strongest natural silks is the major ampullate (MaSp), or dragline silk, which is used to create the main frame of a traditional orb-web. It has favourable tensile strength, toughness and elasticity when compared to man-made materials such as Kevlar (Gosline *et al.,* 1999). This is highlighted in Table 1.A (see page 32), which compares the material properties of some of the most studied types of silk with various types of synthetic materials. One of the toughest silks discovered to date is the major ampullate silk of the Darwin's bark spider (*Caerostris darwini*). This spider builds webs across rivers, spinning drag lines of up to 25.5 metres long (Agnarsson, Kuntner and Blackledge, 2010; Garb *et al.*, 2019). These fibres are twice as tough as silks from other species and over 10 times tougher than Kevlar, making it one of the toughest known biomaterials (Gregoriĉ *et al.*, 2011).

1.2.2 Flagelliform and Aggregate (FlaG and AgSp)

Flagelliform silk (FlaG) is used in the capture spiral of an orb web, and has a slightly lower tensile strength than MaSp, but over 200% extensibility (Adrianos *et al.*, 2013). This extensibility is enhanced through interaction with aqueous aggregate gland silk that coalesces to form sticky droplets and hydrates the silk fibre (Vollrath, 1999). The sticky droplets of the aggregate silk along the flagelliform fibres also adhere to prey, helping to keep it immobilised (Vollrath *et al.*, 1990).

Evolutionarily, in less derived spider species, adhesive force was provided, not by a glue, but by a special silk type that was hackled by comb-like appendages on the hind legs of the spider, termed cribellate silk (Joel *et al.*, 2015). This silk is combed in order to increase its surface area and increase intermolecular Van-der-Waals forces, which aid in prey adherence (Hawthorn and Opell, 2002). However, this process is both time and energy consuming and later ecribellate spiders developed the aggregate glue coating of FlaG silk (Blackledge and Hayashi, 2006; Blackledge *et al.*, 2009). Flagelliform glands are unique to members of the superfamily Araneoidea, and in some derived groups, such as theridiids, this silk can also be used to wrap prey (Eberhard, 2010; Mattina *et al*., 2008).

1.2.3 Minor Ampullate (MiSp)

Minor ampullate silk (MiSp) is used as the auxiliary spiral in orb web construction, stabilising the web and providing a template for the capture spiral. It has lower extensibility than flagelliform silk but retains a similar tensile strength (Vienneau-Hathaway *et al.*, 2017). When immersed in water, both major and minor ampullate silks have been shown to shrink by up to 50% (Work, 1985). This is called supercontraction, and can help the web to regenerate its shape in the morning dew after being deformed throughout the day by various mechanical stresses (Gosline, Denny and Demont, 1984).

1.2.4 Aciniform (AcSp)

Aciniform silk (AcSp) fibres are usually used to wrap prey and line the interior of egg sacs. It is almost twice as tough as major ampullate silk, but this is mainly due to its greater extensibility rather than the strength of the fibre (Hayashi, Blackledge and Lewis, 2004). The protein itself lacks the crystalline forming amino acid motifs found

in MaSp and FlaG (Chaw *et al.*, 2014), which could explain the reduction in strength and increase in extensibility (Rousseau, Lefèvre and Pézolet, 2009).

1.2.5 Tubuliform (TuSp)

Tubuliform silk (TuSp) is used for the tougher outer casing of the egg sac, and as such is only produced by female spiders. It is the only silk that is not consistently produced during a spider's lifetime, as it is exclusively made during the reproductive season. To facilitate this, the tubuliform glands undergo structural and morphological changes just before the eggs are laid (Tian and Lewis, 2006). The egg sac structure is made using two distinct fibre types (TuSp and AcSp), and strength varies depending on the way these are woven together (Alam *et al.*, 2016). It has been proposed that tubuliform is the most ancient of silks and other silk types evolved after this (Tian and Lewis, 2006).

1.2.6 Pyriform (PySp)

Finally, pyriform silk (PySp) is primarily involved in attachment, joining threads together and anchoring them to different substrates (Blasingame *et al.*, 2009). The pyriform silk fibres are coated with a gluey matrix and secreted into attachment disks, which contain nanofibrils and lipids for adhesion. Their overall strength can be controlled by spinneret movements and the location of silk depositions (Wolff *et al.*, 2015).

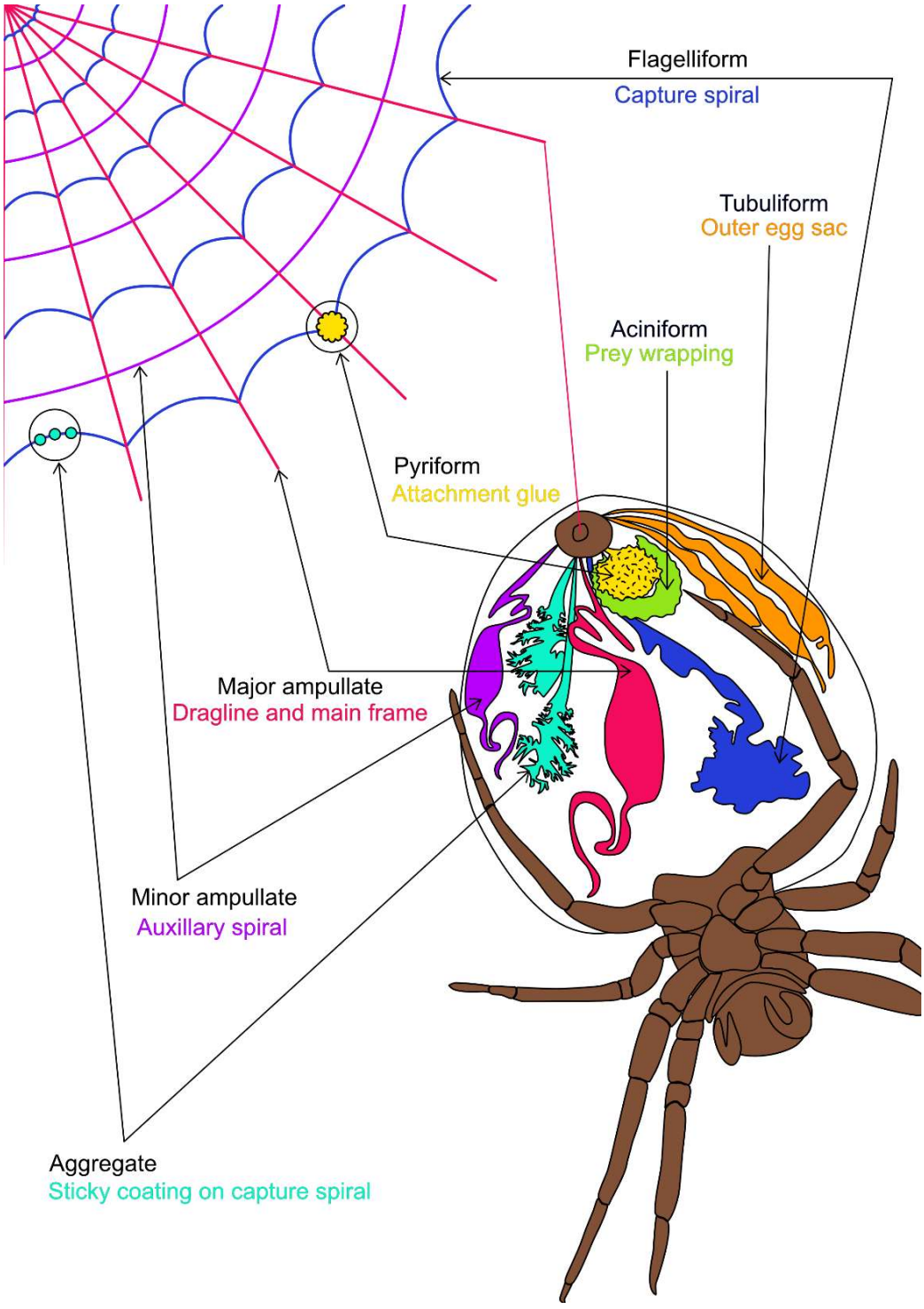


Figure 1.2: Different silk types and their uses.
The way that each silk type is used in the construction of an orb web and for other supporting activities is indicated, as well as general silk gland morphologies.

1.3 From Protein Secretion to Fibre Formation

Of the silk spinning process, most research has so far been conducted into major ampullate fibres, due to their attractiveness as a potential biomaterial as well as ease of identification and extraction. The major ampullate silk gland resides in the abdomen, with the other silk glands, and can be divided into three main parts: the tail, the sac and the duct, shown in Figure 1.3 (Andersson *et al.*, 2013). Natural silk production and spinning is a complex operation involving many processes, from chemical changes to physical forces (Scheibel, 2004; Vollrath, 2000).

Spidroins are secreted from three different types of tall columnar epithelial cells in the anterior end of the silk gland (the tail and proximal region of the sac, see Figure 1.3), which are then stored in the central sac region, before extrusion through the duct and out of the spinneret spigots (Andersson *et al.*, 2013; Rising and Johansson, 2015).

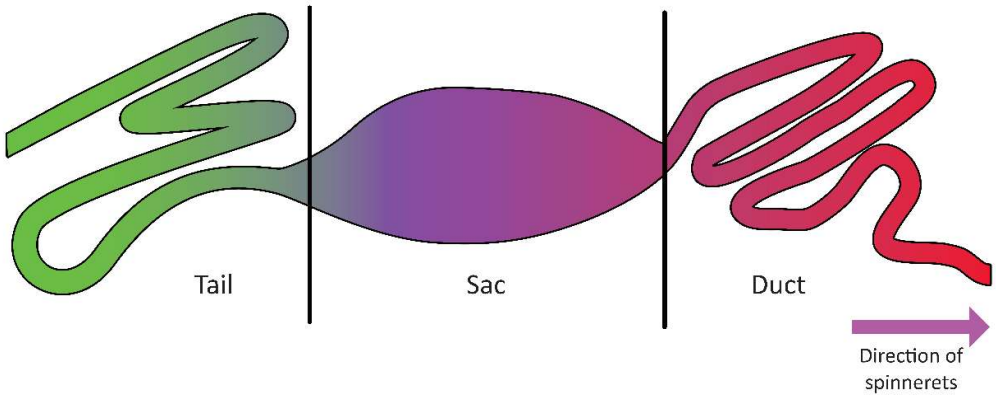


Figure 1.3: Silk gland regions.
The major ampullate silk gland is made up of the Tail (green), Sac (purple) and Duct (red), shown from left to right. Silk proteins are secreted in the tail and stored in the sac as a liquid dope. Further processing takes place as the silk proteins move towards the duct and are extruded as a fibre from the spinneret exterior spigots.

It has been proposed, that within the sac, spidroins are arranged in a micelle-like conformation, with the repetitive regions shielded in the centre by a hydrophilic outer shell made by the terminal domains (Jin and Kaplan, 2003). However, it is now generally accepted as being more likely that within the sac the proteins maintain an 'aquamelt'-like conformation, which aids processing and storage (Vollrath and

Knight, 2001; Holland *et al.*, 2012). This is also often referred to as a protein 'dope'. Solidification of the liquid dope is governed by changes in pH gradients, ion concentrations and shear forces exerted when the silk is pulled from the spigot (Andersson, Johansson and Rising, 2016). The key to the strength and toughness of silk is in the way that the proteins fold; the composition of amino acids that make up a silk fibre has dramatic consequences on its resulting strength, flexibility and other properties.

The molecular structure of fibrous MaSp consists of semi-amorphous and nanocrystalline β-sheet protein domains, which are made up of repeating amino acid motifs. It has been shown that spidroins in silk fibres are strongly oriented, and protein chains run predominantly parallel to the fibre (Van Beek *et al.*, 2002). In MaSp, poly-alanine regions form ordered β-sheet crystal domains, while glycine rich repeats form β- spirals and helices (Keten and Buehler, 2010). The relative quantity and composition of these repeats can confer strength (Poly(A)) or elasticity (glycine-rich regions) to the fibre. The crystalline β-sheet bundles align parallel to the fibril axis and are embedded in an amorphous matrix. During deformation, the cross links between these regions create a backbone structure to the fibre, and it is the stability of these beta sheet regions that is key to spider silk's strength (Hayashi, Shipley and Lewis, 1999; Gosline *et al.*, 2002). Tens to hundreds of these fibrils come together to create a high performance fibre (Giesa *et al.*, 2011). An overview of this can be seen in Figure 1.4.

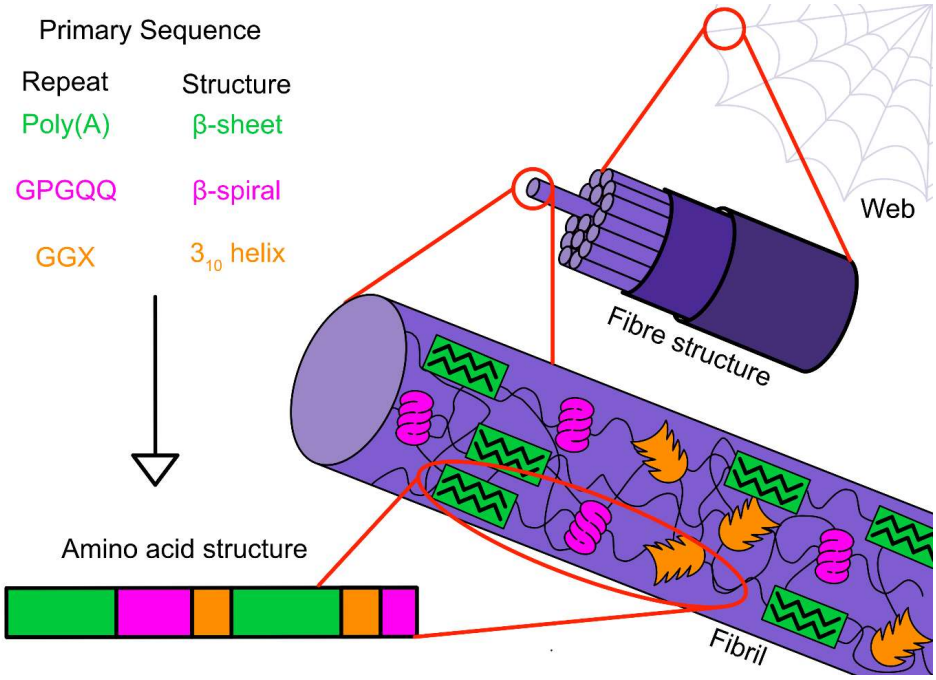


Figure 1.4: The proposed structural model of a major ampullate spider silk fibre.
Repeating amino acid motifs fold to create different secondary structures, which can result in crystalline β-sheets (green) becoming embedded in an amorphous matrix of β-spirals and alpha helices (pink, orange structures and black lines): creating a fibril. Hundreds of these fibrils come together to create a fibre, which is then used by the spider to create structures such as webs.

1.4 The Role of Repetitive Motifs in Different Silk Types

As the presence, absence and quantity of short amino acid motifs can have such an impact on the fibre properties, much research has focused on these repeating units. Some emphasis in this thesis is placed on the differing proportions of amino acids and their repetitive motifs in silks from different species and the potential impact these may have on the fibrous properties.

The presence of crystalline linked β-sheet regions can bind protein molecules together and increase tensile strength. These are often made up of poly-alanine repeat regions in major ampullate silk, which are also able to form many hydrophobic interactions, contributing to the increased tensile strength (Tokareva *et al.*, 2014). This motif is similar in minor ampullate silk, but here the β-sheets are generally made up of poly glycine-alanine (Poly(GA)) repeats (Vienneau-Hathaway et al., 2017). These cannot make as many hydrophobic interactions, which leads to a reduction in tensile strength compared to MaSp (Hayashi, Shipley and Lewis, 1999).

In the amorphous region of FlaG and MaSp fibres, a glycine-proline-glycine-X-X (GPGXX, where X is any amino acid) motif creates a 3D structure that mirrors that of elastin and confers elasticity to the fibre (Urry, Luan and Peng, 1995; Liu *et al.*, 2008; Liu, Shao and Vollrath, 2008). Flagelliform silk contains more of these uninterrupted repeats than major ampullate silk, which goes some way to explaining the difference in fibre extensibility from ~35% in major ampullate silk to ~200% in flagelliform (Hayashi and Lewis, 1998; Hayashi, Shipley and Lewis, 1999).

There can also be variation in sequence within individual silk types. For example, the major ampullate fibre of *Trichonephila clavipes* is composed of a mixture of MaSp1 and MaSp2 spidroins. MaSp1 is present in large quantities along the silk fibre, whereas MaSp2 is clustered in certain core areas of the fibre and absent from the ends (Motriuk-Smith *et al.*, 2005; Askarieh *et al.*, 2010). MaSp2 contains a higher proportion of proline residues which gives the silk elasticity, while MaSp1 gives the fibre its tensile strength (Santos-Pinto *et al.*, 2016). It is probably this mix of proteins that allow the silk to be flexible, elastic and remarkably strong. It has been suggested that due to extreme Poisson's ratios for the deformation of silks in various directions, silk should be viewed as a complex hierarchical structure rather than a continuum material (Brown *et al.*, 2014).

Non-conserved 'spacer regions' can also be found in MaSp and FlaG silk proteins, often flanking the β-sheet and amorphous regions. Spacer regions in both silks are relatively long, complex and charged (Tokareva *et al.*, 2014). No evidence has yet been provided for the role of these spacer regions, but some proposed roles include: preventing premature silk formation, promoting correct protein alignment into fibres or providing surface regions to allow association of silk with other molecules (Hayashi, Shipley and Lewis, 1999).

Interestingly, the non-fibre forming pyriform silk shows marked differences in amino acid content to other types of spidroin, as it lacks both glycine and proline residues in internal repeat blocks (Blasingame *et al.*, 2009). However, all silks to date contain a conserved C-terminal domain (CTD), including pyriform, which also retains the modular protein structure containing conserved terminal domains and intervening repetitive region (Perry *et al.*, 2010; Chaw, Saski and Hayashi, 2017). This suggests

that pyriform is indeed a true spidroin and has evolved differences in structure to fulfil a specialised role as a liquid glue substance, rather than as an adherent fibre.

Little work so far has been carried out on the post translational modifications of specific amino acids in silk and their effect. However, these may play an important role in forming the silk fibre structure and maintaining specific properties. It has been shown that the major ampullate silk of *Trichonephila* is post translationally modified, often in specific structural motifs. In both MaSp1 and MaSp2, up to 24 phosphorylation sites were identified (dos Santos-Pinto *et al.*, 2014). Also, the formation of β-sheet assembly is controlled by enzymatic phosphorylation in recombinant spidroins, showing that this is likely an important way to control conformation and material properties of both natural and synthetic silks (Winkler, Wilson and Kaplan, 2000).

The introduction of phosphate groups could change the way that the protein folds and, indeed, this has been found in major ampullate silk. In addition, phosphate anions in solution are known to promote hydrophobic interactions, which promote silk assembly and could increase fibre strength (Huemmerich *et al.*, 2004). Interestingly, the phosphorylation of caddisfly silk has shown to help this organism adapt to living in an aquatic environment (Stewart and Wang, 2010). Phosphorylated residues may also help to store the aqueous silk at high concentration in the glands and promote interactions upon the folding of silk fibres (Heim, Keerl and Scheibel, 2009; dos Santos-Pinto *et al.*, 2014). Phosphorylation of silk may be most important to examine, however, when designing and producing recombinant spidroins, as hinted at in Chapter 4.

1.5 Next Generation Sequencing Methods

The main difficulty in the genetic study of spidroins results from their unique organisations; they are both very long and repetitive. This means they are generally difficult to sequence using traditional methods, and to *de novo* assemble. Over time, a range of sequencing technologies have been used to explore the silk genes of spiders.

Generally, DNA sequencing technologies are classified into three generations. First generation methods enabled the very first sequencing of clonal DNA populations (Heather and Chain, 2016). The later sequencing methodologies (Next Generation Sequencing, NGS) comprises second generation sequencing, which greatly increased throughput by parallelising many reactions, and third generation sequencing, which can directly sequence single DNA molecules (Slatko *et al.,* 2018). NGS emerged in 2004, with the launch of pyrosequencing by 454 Life Sciences (Balzer *et al.,* 2010). This technology was able to produce 25 million bases with > 99 % accuracy per run by clonally amplifying DNA fragments on beads in the droplet of an emulsion, before sequencing in the wells of a specially designed fibreoptic slide (Margulies *et al.*, 2005). This represented a paradigm shift in sequencing, as sequencing reactions could be parallelised en-mass, greatly increasing the amount of DNA that could be sequenced in one run and reducing the need for much of the previously associated time and costs with first generation sequencing approaches (Heather and Chain, 2016).

Short read NGS approaches are low cost and highly accurate, useful for mass gene identification and producing high quality deep coverage of small genomes. They are also supported by a wide range of analysis tools and pipelines. However, this technique is susceptible to preferential amplification of repetitive DNA, and often the overlap between sequences is too short to enable the efficient *de novo* assembly of genomes or transcriptomes. For these approaches, genetic material is broken down, amplified and then sequenced. Long read sequencing is often called third generation sequencing because it moves away from this approach by directly sequencing single DNA molecules. While short read sequencers typically produce reads of up to 600 bases, long read technologies can generate reads in excess of 10kb (Pollard *et al.*, 2018; Amarasinghe *et al.*, 2020).

Long read sequencing is currently dominated by two technologies: Nanopore and Single-Molecule Real-Time (SMRT) sequencing. Nanopore sequencers measure the ionic current fluctuations that occur when single stranded nucleic acids pass through biological nanopores. Different nucleotides confer different resistance to the stretch of the nucleic acid within the pore and so the sequences of bases can be inferred by the specific patterns of fluctuation. SMRT sequencing detects fluorescence events

that correspond to the addition of one specific nucleotide by a polymerase tethered to the bottom of a tiny well. Long read technology is useful for full length isoform sequencing and *de novo* genome assembly applications, but these technologies are still overcoming early limitations in accuracy, throughput, cost effectiveness and analysis (Pollard *et al.*, 2018; Amarasinghe *et al.*, 2020).

Over the past 15 years, NGS technologies have been evolving, leading to substantial improvements in yield and quality over a relatively short timeframe. Improvements in chemistry, costs, throughput and accessibility are driving the emergence of new and varied technologies, as well as novel applications for these (Goodwin, McPherson and McCombie, 2016; Logsdon *et al.*, 2021). Chapters 2 and 3 of this thesis use Illumina and SMRT sequencing to identify and examine the transcripts of specific spider silk genes, and Chapter 3 suggests applications for which either of these would be most appropriate. In order to understand the difference between sequences derived from the different technologies used in these chapters, a brief description of how the sequences are generated is included here.

1.5.1 Illumina Sequencing

Illumina sequencing is based on a technique called bridge amplification, which allows paired end sequencing, something of which only Illumina NGS platforms are capable. This results in higher coverage, more reads and more data compared to single end sequencing, and increased probability of alignment to a reference (Ambardar *et al.*, 2016).

During Illumina sequencing, DNA is first fragmented and adapter sequences added to the ends. The DNA fragments are then made single stranded and added to a flow cell. The surface of the flow cell is covered in primers, complementary to the adapter sequences on the ends of the DNA. These primers bind to the adapters and then a DNA polymerase creates a complement of the hybridised fragment. The dsDNA is denatured and the original template washed away. The strands are then clonally amplified using bridge amplification, upon which the strand folds over and the adapter region hybridises to another primer on the surface of the chip. Polymerases form a double stranded 'bridge' which is then denatured, resulting in two single

stranded copies of the molecule tethered to the flow cell. This process is repeated millions of times, resulting in clonal amplification and clustering of the same DNA sequences. After this, the reverse strands are cleaved and washed away, leaving only the forward strands. Sequencing begins with the extension of the first sequencing primer to produce the first read. With each cycle, fluorescently tagged nucleotides compete for addition to the growing chain. After the addition of each nucleotide, the clusters are excited by a light source and a characteristic fluorescent signal is emitted, indicating which base has been added. The number of cycles determines the length of the read and each strand in a cluster is read simultaneously. This can be seen in Figure 1.5. Short read sequencing using this method is cost effective and allows the production of a large amount of data in a relatively short time. However, short reads make the generation of *de novo* transcriptomes and genomes difficult, as they do not provide sufficient coverage over repetitive sections (Schuster, 2008; Conesa *et al.*, 2016).

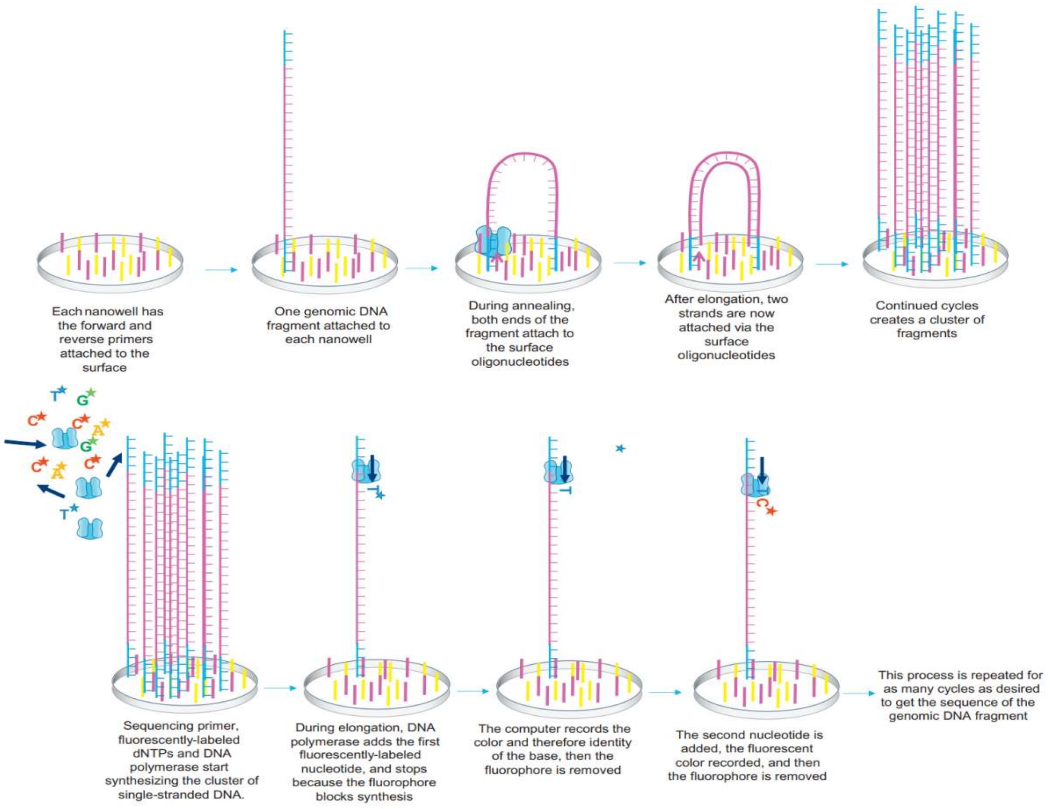


Figure 1.5: The process of Illumina sequencing.
Image reprinted from Clark et al. *(2019) with permission from Elsevier.*

1.5.2 SMRT Long Read Sequencing

Single-Molecule Real-Time (SMRT) Sequencing, developed by Pacific Biosciences, allows the real time sequencing of full-length transcripts of over 10kb. To create sequences, first, hairpin adapters are ligated to both ends of a target dsDNA molecule to create a closed, single stranded 'SMRTbell' template. The SMRTbell templates are then loaded onto a chip called a SMRT cell, which contains millions of tiny wells called zero-mode waveguides (ZMW). A single polymerase is immobilised in the bottom of each ZMW. One SMRTbell template diffuses into each ZMW. As the polymerase incorporates fluorescently labelled nucleotides, a specific emission spectrum of light is emitted and recorded. In this manner, sequencing is recorded in real time and, because the templates are circular, the polymerase can continue incorporating bases until it has reached the end of its lifetime. If the polymerase lasts long enough, both strands can be sequenced multiple times, called 'passes'. If this happens, the read (CLR) can be split into subreads by recognising the adapter sequences. The consensus sequence of multiple subreads from a single ZMW yields a Circular Consensus Sequence (CCS) with higher accuracy (Rhoads and Au, 2015). CCS reads are often referred to as HiFi reads. An overview of the generation of subreads and highly accurate CCS in Figure 1.6.

Importantly, SMRT sequencing allows the sequencing of extremely long single pieces of DNA. CCS reads also have an advantage of high accuracy, which allows them to be used directly for analysis, eliminating the need for assembly or read correction. Several studies have shown that genes of interest can be identified in high quality CCS data without assembly (Westbrook *et al.*, 2015; Shortreed *et al.*, 2020; Zurn *et al.*, 2020). One of the benefits of long reads is that repetitive regions can be sequenced with more confidence as errors are not introduced during assembly, which could prove to be a beneficial approach when examining often highly repetitive spider silk genes and transcripts.

One of the drawbacks to long read sequencing is the fact that throughput is a lot lower for long read sequencing than short reads, as SMRT sequencing is limited by the number of ZMWs per cell (Rhoads and Au, 2015). Also, the error rate of a raw read (CLR) is relatively high (10-15%). However, this error rate is reduced by

generating CCS reads with sufficient sequencing passes, which can yield 99% accuracy with 15 passes (Eid *et al.*, 2009). The CCS reads are limited in length by the lifetime of the polymerase, so very long transcripts (>10kb) may not be represented (Weirather *et al.*, 2015; Nakano *et al.*, 2017). This has led to several *de novo* assembly and hybrid assembly strategies being developed to create a more complete picture of a genome or transcriptome (Rhoads and Au, 2015; De Maio *et al.*, 2019; Gavrielatos *et al.*, 2021; Murigneux *et al.*, 2021).

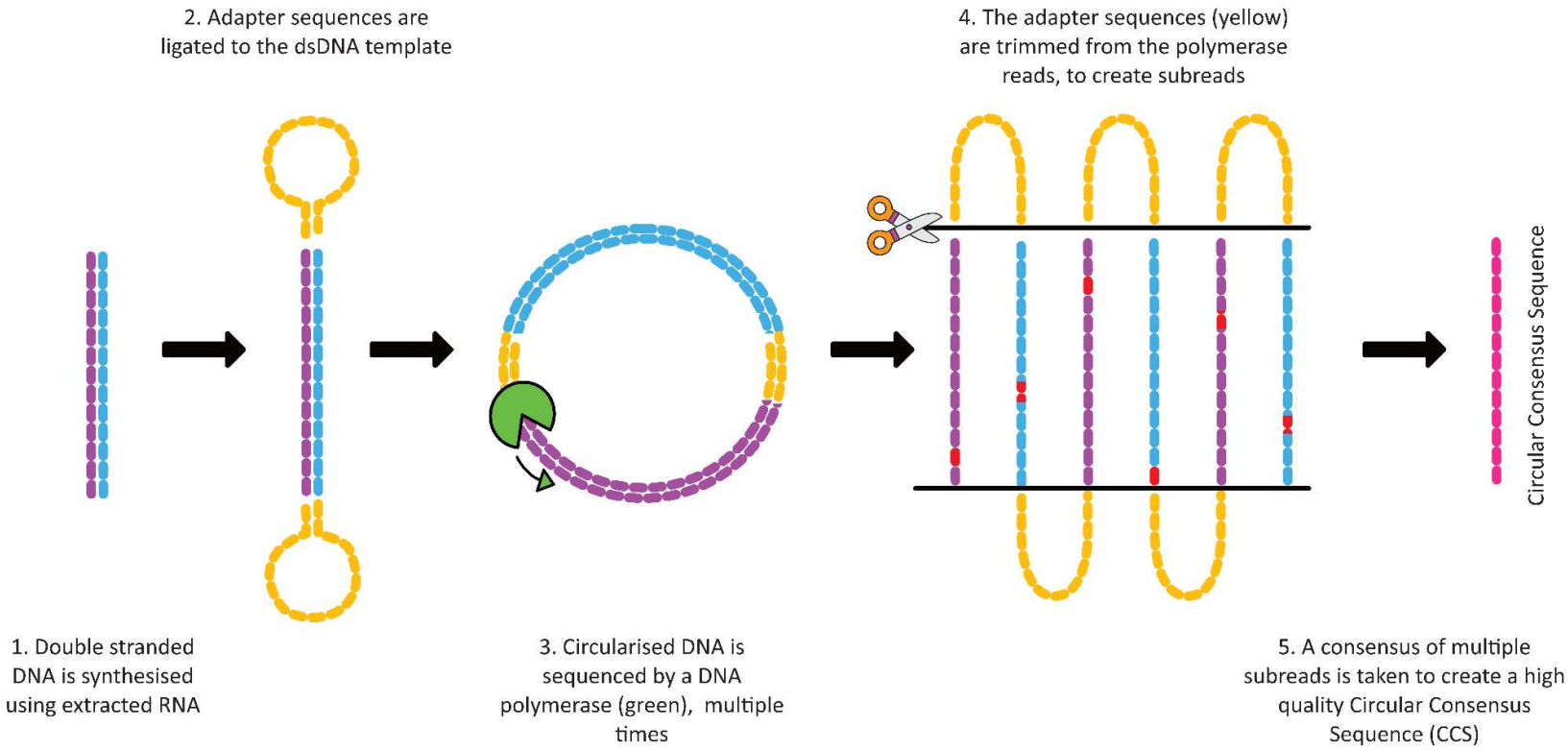


Figure 1.6: The generation of subreads and Circular Consensus Sequences.

1.6 Spiders in Extreme Environments and *Argyroneta aquatica*

The idea of bioprospecting, a way of using ecology to predict which species are likely to spin silks exhibiting extraordinary properties which can then be further investigated, is not a new one. This approach has been used to find the toughest silk discovered to date as well as a uniquely rigid silk used to create freestanding webs with high lateral stability, and so could also be used as a way to explore a greater range of possibly unexpected silk types (Agnarsson, Kuntner and Blackledge, 2010; Haynl *et al.*, 2020). In this thesis, the idea of bioprospecting is used to examine the silk and silk genes of several spiders which live in extreme or unusual environments, with a particular focus on the underwater spider, *Argyroneta aquatica*. This is done with a view to employing the information learnt to improve or modify the synthetic production of aquatic spidroins in the laboratory.

Spiders can adaptively optimise their energy and silk use to enable habitation of a range of environments (Canals, Veloso and Solís, 2015; Blamires *et al.*, 2016; Hernández Duran *et al.*, 2021). In order to inhabit areas of low prey abundance, some spiders, including *Lycosa lenta* and *Filistata hibernalis*, can lower their metabolic rate when starving (Anderson, 1974). Others display phenotypic plasticity to enable survival in environments with high temperature fluctuations, for example *Paraphysa parvula* which lives at high altitude in the Chilean Andes (Alfaro *et al.*, 2013). Caves have also become homes to many species, which are able to adapt to both humid and often hostile low resource environments. In the sulfuric acid springs in the caves of southern Italy, for example, reside representatives of both the Linyphiid and Nesticid families, which raises some important questions as to how well their silk is adapted for such challenging colonisation (Messina *et al.*, 2015; Mammola *et al.*, 2018; Pantini and Isaia, 2019). This is explored more in Chapter 2.

Spiders have also adapted to xeric environments, those of very low atmospheric humidity or precipitation, by lowering rates of evaporation (Canals, Veloso and Solís, 2015). This strategy could have been used by *Latrodectus hesperus* to invade desert habitats (Hadley and Quinlan, 1989). *Cyrtophora citricola* is a tent weaving species which colonises dry, desert-like environments ranging from Israel to Spain, creating large colony webs which often persist for weeks (Lubin, 1974). *Eratigena atrica* is a

commonly found house spider species in the UK, which usually lives in dry forests, but is thought to have originally inhabited caves, adapted to a range of temperatures and humidity (Rachwał, Rybak and Rogula-Kozłowska, 2018). The silk of some of these spiders is investigated further in Chapter 3.

Contrastingly, some spiders have made their home in areas of up to 100% humidity. The pirate wolf spider, *Pirata piraticus*, lives in a variety of semi-aquatic habitats in the borders of bodies of water. It relies on hydrophobic leg hairs to be able to walk on water and uses silk to create a retreat and keep its eggs safe and dry (Nørgaard, 1951; Hénaut *et al.*, 2018). The fishing spiders of the genus *Dolomedes* also live in semi-aquatic habitats, relying on vibrations on the surface of the water to sense prey and use silk to create dragline supports, which can be extensive (Bleckmann and Barth, 1984; Correa-Garhwal *et al.*, 2019). They are also able to dive underwater to catch aquatic invertebrates and small fish (Zimmermann and Spenee, 1989; Nyffeler and Pusey, 2014). *Desis marina* lives in the intertidal zone between sea and shore, creating silken retreats which store air in which they can survive for up to 19 days (McQueen, Pannell and McLay, 1983; Mclay and Hayward, 1987). While it has a lower respiration rate than other spiders, its survival underwater relies entirely on the ability of the silken structure to retain air and resist water (McQueen, Pannell and McLay, 1983).

Argyroneta aquatica is a unique spider in that it spends most of its life underwater. It is a member of the family Dictynidae and creates dome shaped webs, 'diving bells', which are filled with air and used underwater; functioning as a physical gill (Figure 1.8) (Selden, 2002; Seymour and Hetz, 2011). It is the hydrophilic property of this silk that makes *A. aquatica* particular interesting in this investigation. These spiders prefer areas of slow or non-moving water, with plenty of shade and water plants to which they anchor their webs, and can be found throughout Europe and Asia (Mammola, Cavalcante and Isaia, 2016). *A. aquatica* are commonly found living in areas of low partial pressure of oxygen and near *Sphagnum* moss, which can lower water pH and provides a habitat for a range of species (Henrikson, 1993; Glime and Lissner, 2013; Vicherová *et al.*, 2017). They consume aquatic invertebrates and small fish, and are equipped with a rather potent venom that affects both vertebrates and

invertebrates (Figure 1.9A) (Uzenbaev and Lyabzina, 2009; Nyffeler and Pusey, 2014). *A. aquatica* have a low resistance to the cold, and lower their oxygen consumption so they can stay underwater for longer during winter (Bromhall, 1988). Interestingly, this is one of the few spider species in which males are larger than females. This sexual dimorphism probably gives a mobility advantage to larger male spiders underwater when finding a mate, while females can create larger diving bells in relation to their size in which to lay and raise their eggs (Schütz and Taborsky, 2003, 2005). The spiders mate, eat and rest inside their diving bells, which can also act as the foundation for the egg sac (Figure 1.7).

A. aquatica have book lungs and a very elaborate, efficient tracheal system that permits submersion for the longest possible time. It does this by using almost all of the available oxygen in the diving bell before it has to replenish its supply (Levi, 1967). As surfacing increases the chance of predation, this system probably evolved as an adaptation to *A. aquatica*'s unusual aquatic habitat. To support this lifestyle, they use two types of air containment systems underwater: the diving bell and the plastron, seen in Figure 1.7B and C. The plastron is a thin layer of air which surrounds the opisthoma (abdomen). This air covers the stigmata (openings to the tracheal system) through which the spider breathes, and is a strategy employed by many other types of aquatic invertebrate (Thorpe, 1950; Flynn and Bush, 2008). In *A. aquatica*, the air is held on the surface of its body by hundreds of tiny, hydrophobic hairs. These hairs densely cover the spinnerets and abdomen, but are less dense on the legs and head, as the requirement for air there is minimised. The hairs on the legs probably help to stabilise the air bubble when it is taken underwater. These hairs and their arrangement can be seen in Figure 1.8.



Figure 1.7: **Argyroneta aquatica.**
A. aquatica *are able to both hunt and eat small prey underwater, such as brine shrimp (A). The plastron is a thin layer of air that covers the abdomen and is indicated by an arrow in (B). (C) shows an individual in a resting pose, half in and half outside a small diving bell. (D) shows an adult* A. aquatica *fully inside its diving bell, eating a small cricket. The threads that make up the bell are visible, as is their constraint on the air bubble inside it. Photographs taken by Robert Morton and used with permission.*

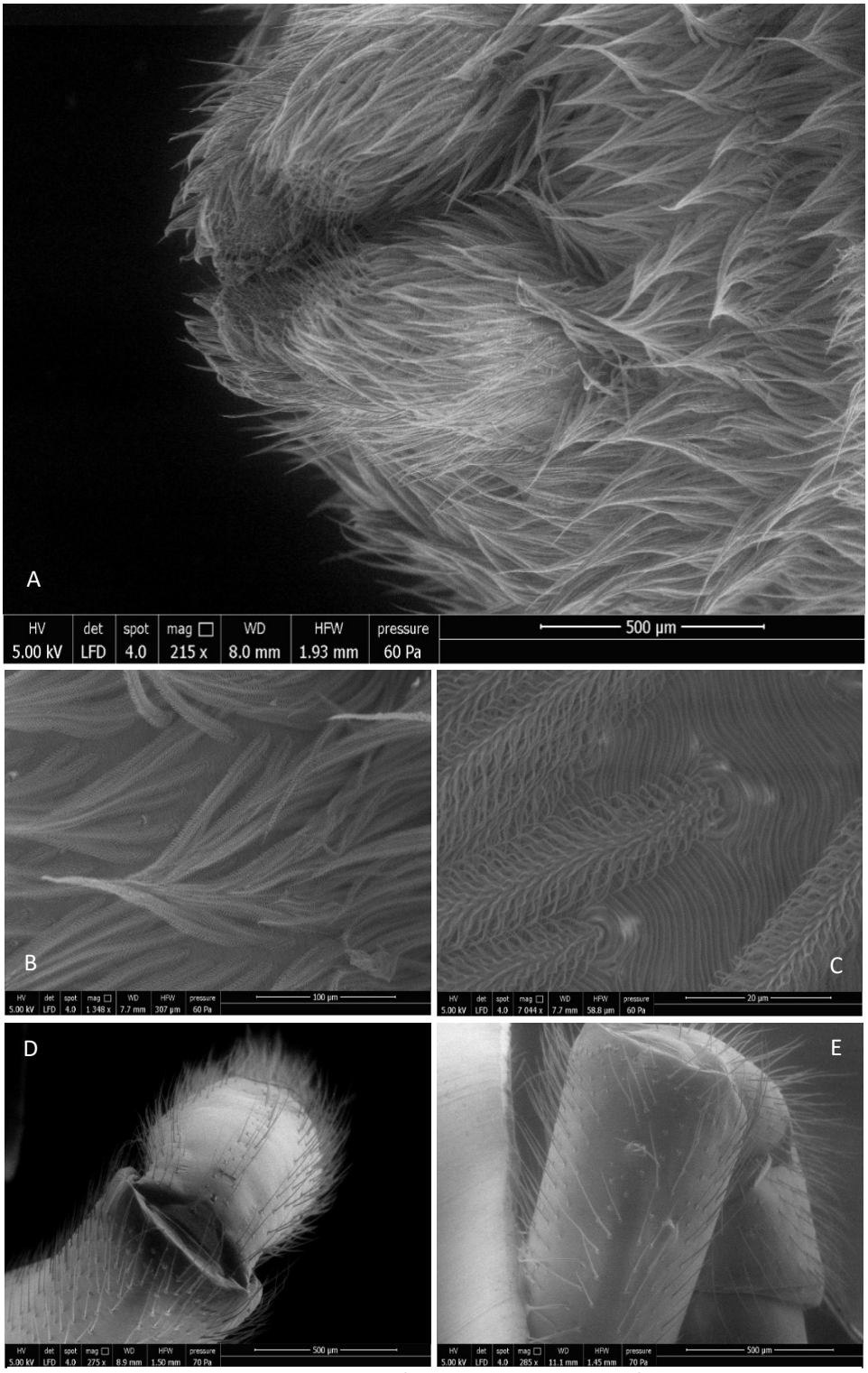


***Figure 1.8: Structure and organisation of the hydrophobic hairs of* Argyroneta aquatica.**
(A) shows the dense hairs covering the spinnerets and rear underside of the abdomen of an adult female Argyroneta aquatica*. (B) and (C) are magnified clumps of hairs on the abdomen, showing the fine microstructure and appearance of 'barbs'. (D) and (E) show the organisation of hairs, in stripes, on the rear (D) and front (E) legs of two individuals. Images taken by the author and staff at the nmRC (UoN) using environmental SEM in low vacuum and low pressure during the course of this project.*

A. aquatica make four types of silk underwater, which are used to construct the diving bell, egg sac, attachment threads and walking threads (for moving towards prey) (De Bakker *et al.*, 2006). There is some genetic data available, but the exact spidroins used to make up these silks have not yet been precisely determined (Strickland *et al.*, 2018; Correa-Garhwal Sandra *et al.*, 2019). The diving bell itself is composed of a mix of thick, thin and aggregated threads with what was thought to be no clear spatial ordination (De Bakker *et al.*, 2006).

The diving bell is constructed in a series of steps. First, a horizontal sheet is constructed, using anchoring points on the surrounding vegetation or substrate. e-SEM images taken during the course of this study suggest that the web is first woven as a rough guide, with thick threads, and reinforced using a more regular weaving pattern of thinner threads on the inside (Figure 1.9).

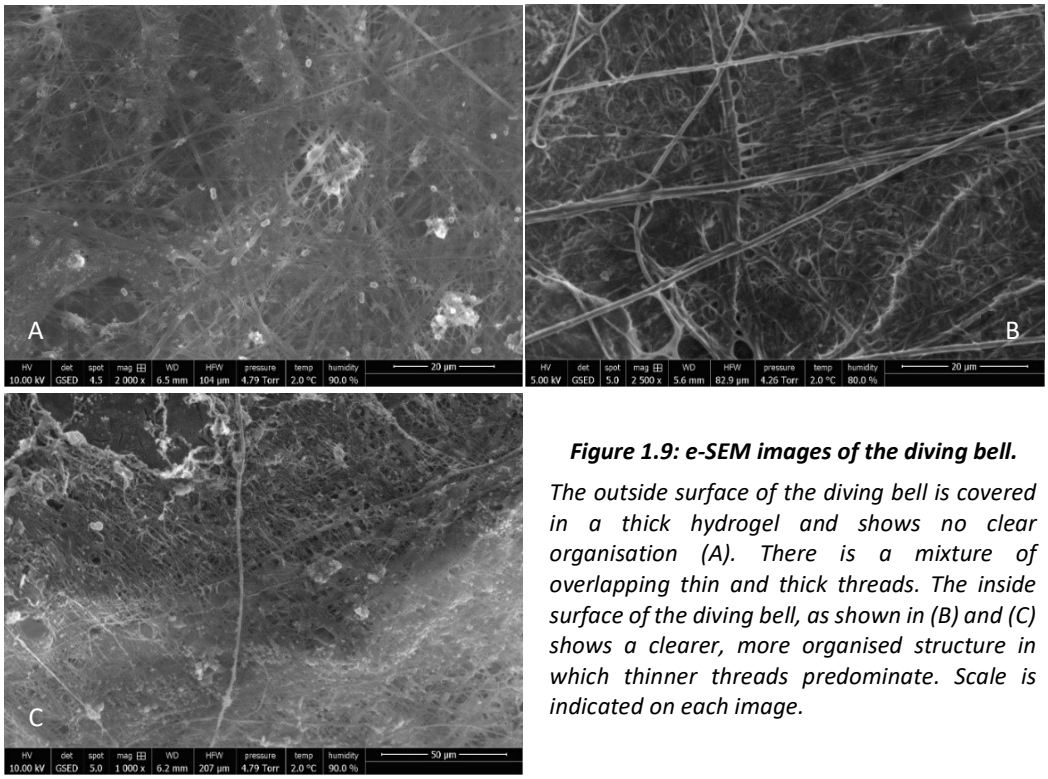


Figure 1.9: e-SEM images of the diving bell.

The outside surface of the diving bell is covered in a thick hydrogel and shows no clear organisation (A). There is a mixture of overlapping thin and thick threads. The inside surface of the diving bell, as shown in (B) and (C) shows a clearer, more organised structure in which thinner threads predominate. Scale is indicated on each image.

The spider returns to the surface of the water, collecting an air bubble which is carried between its two back legs and is then deposited under the sheet. The air is held by

surface tension between the silk fibres and creates an air filled, water free diving bell. This is repeated until the bell is large enough. The spider will spin a variety of sizes of diving bell; some will only accommodate the abdomen and some are enlarged to allow feeding or mating. This silken structure is able to effectively trap the air underwater, and can also take up some dissolved oxygen from the water, allowing for longer periods underwater (Seymour and Hetz, 2011). This also means that it can be said to be functioning as a physical gill, as well as providing a water free habitat in which to mate, hunt and reproduce. The spider inside its diving bell is shown in Figure 1.7.

Uniquely, the diving bell structure is coated with a proteinaceous hydrogel, which makes the structure very hydrophilic (Neumann and Kureck, 2013). It is yet unknown whether this coating is a form of liquid silk or whether it is a glycoprotein covering, but it is thought to be linked to the gas permeable properties of the web. In air, diving bell structures become dehydrated and the threads become very brittle (Neumann and Kureck, 2013; Neumann and Woermann, 2013). Figure 1.10 shows water and hydrogel coating the silk and its recedence when the humidity of the viewing chamber is lowered, taken during the course of this investigation. These properties suggest that the silk has evolved for an aquatic environment and is not simply the thread of terrestrial spiders used underwater.

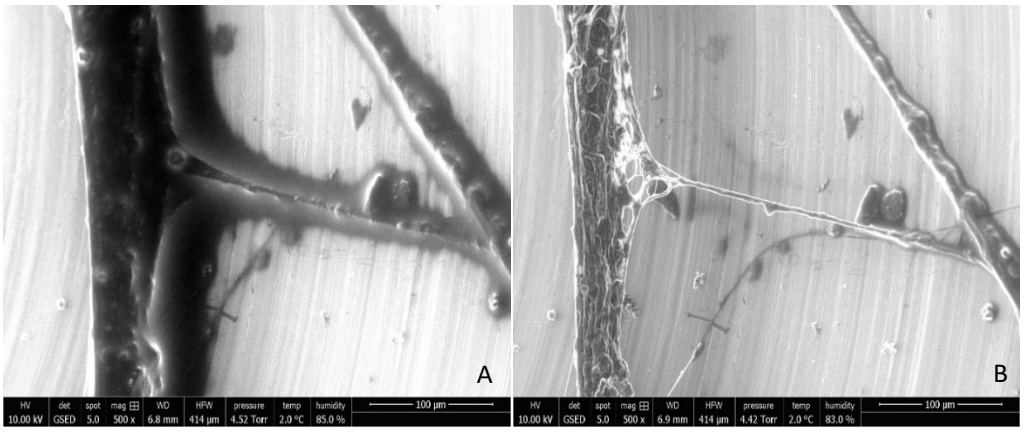


***Figure 1.10: Hydrogel and water on the silk threads of* A. aquatica.**
This image shows the hydrogel and water which coat the silken threads under relatively high humidity, and as the humidity is lowered to 83% (B) the water is removed, and the fine structure of the threads and globular hydrogel coating is revealed.

There is a relatively small amount of research into the genetic make-up of *A. aquatica* silk genes (see Chapter 3). However, the small number of available sequences suggest that the amino acid structure and modular make up is not too diverged from terrestrial silks. As they still retain the conserved C-terminus, modular gene structure and external spinning apparatus seen in other silks and spiders, it is suggested that fibre formation is the same as seen in other spiders, and that this silk would be a good candidate to recreate using existing miniaturised spidroin expression systems. This silk is also particularly interesting because of its potential uses in aquatic or humid environments in the engineering and medical industries, which would potentially make it more suited than normal terrestrial silks, which often exhibit unfavourable properties, such as supercontraction, upon contact with water.

1.7 Silk and Biotechnology

Diverse and unique biochemical properties make spider silk well suited for use as a biomaterial (Lewis, 2006). Its extraordinary strength and toughness, low immunogenicity and slow degradation makes it promising for use in the medical, aerospace and automotive industries, to name but a few. Most recombinant silks to date have been expressed in *Escherichia coli* as it is small, has a short generation time, is easy to manipulate and is relatively low cost to maintain (Scheibel, 2004). However, spider silks have also been expressed successfully in yeast, transgenic animals, insects and plants, as is discussed below.

A native sized dragline silk protein of greater than 280 kDa was produced in *E. coli* by Xia *et al.* (2010). This protein was expressed in low levels, but upon artificial spinning showed comparative mechanical properties to natural silk. The mechanical spinning process, used for fibre formation of many recombinant spidroins, consists of applying shear forces to the spidroin solution in the presence of a physiological buffer, known as wet spinning into a coagulation bath (Teulé *et al.*, 2007).

The repetitive regions of the two proteins that make up the dragline silk of the orb weaver, *Araneus diadematus,* have been used to create several recombinant spidroins in *E. coli*, with very different properties (Huemmerich *et al.*, 2004). Optimisation of this process, such as the fermentation of cells which increased

protein yield significantly, and the codon optimisation during cloning to reduce errors in translation, could be employed to optimise creation of other recombinant spidroins. A microfluidic device was used to combine these two recombinant spidroins into a fibre and using this it was possible to fine-tune the properties of the material (Rammensee *et al.*, 2008).

Artificial spinning is a time consuming and delicate process, unsuitable for reproduction on a large scale. Self-assembling silk fibres would provide a great benefit to recombinant silk fibre production. The recombinant protein 4RepCT is a miniaturised MaSp spidroin which is based on the silk of The African nursery web spider, *Euprosthenops australis,* and has been stably expressed in *E. coli.* This can form 3D structures independently of any additional mechanical spinning processes or manipulations (Stark *et al.*, 2007). It contains a thioredoxin tag for protein solubility, a poly-histidine tag for purification and a repetitive region. This protein also includes a non-repetitive C-terminal domain which allows dimerisation and fibre formation, much like in natural silks, upon thrombin or enterokinase mediated removal of thioredoxin, eliminating the need for costly and time-consuming artificial spinning processes. The role of the CTD has also been shown to be essential in fibre formation in other recombinant silk systems (Ittah *et al.*, 2006).

The addition of the conserved N-terminal domain sequence into the 4RepCT spidroin enables the rate of fibre formation to be controlled through changing the pH (Askarieh *et al.*, 2010). At higher pH values the NTD inhibits fibre formation, but fibres rapidly form when this is lowered. By enabling control of the timing of fibre formation, this study shows potential to allow the scaling up of the recombinant spidroin formation process. Genetic engineering has been used to increase the strength of 4RepCT by introducing additional covalent linkages in the fibre, showing that there are still steps that can be taken to improve this process (Grip, Johansson and Hedhammar, 2009). This expression system could prove to be a reliable way to produce specialised silk fibres with a range of possible uses.

Aciniform silk, used to wrap prey, has also been produced recombinantly in *E. coli* (Xu *et al.*, 2012). Two of the 200 amino acid repeats that make up the natural silk were sufficient to form silk-like fibres when exposed to shear forces but the mechanical

properties were not comparable to natural silk, probably due to the small spidroin size. Teulé *et al* (2007) combined the repeat unit of flagelliform silk with a MaSp repeat unit of *T. clavipes* in order to try and produce a silk that exhibited both strength and elasticity. A proteinaceous film formed spontaneously, which was then hand pulled, or subjected to wet spinning to create fibres (Teulé *et al.*, 2007). It was shown that the mechanical properties of these fibres did indeed vary according to the type and number of repetitive modules they contained. These studies show that it is possible to make not only major ampullate spidroins recombinantly, but also other silk types that go some way to recapitulating the properties of their natural counterparts.

Other prokaryotic systems have also been used to produce spidroins synthetically. *Salmonella typhimurium* has been used as an expression host in an effort to increase protein purity and yield, as it has more efficient protein secretion pathways than those of the Gram negative bacteria *E. coli* (Widmaier *et al.*, 2009; Widmaier and Voigt, 2010). However, secretion in this case relied heavily on the size of the protein, and was not particularly successful for proteins of large molecular weight, showing this system is not yet effective for large silk protein expression.

Spidroin expression has also been attempted using eukaryotic systems. For example, the methylotrophic yeast, *Pichia pastoris*, was used as an expression host for recombinant spidroins, and was able to produce large proteins of up to 3000 amino acids at high yields, of up to 1000 mg/l of culture (Fahnestock and Bedzyk, 1997). There was some heterogeneity of protein sizes produced and large losses in yield occurred during purification, yet there was no evidence of truncated synthesis as sometimes seen in *E. coli* production (Zhang *et al.*, 2015). As a producer of long, repetitive proteins, yeast seems to be a viable alternative to *E. coli* expression systems as it is stable in culture and can cope with the production of large proteins. This process has been used to create the first commercially available synthetic silk garment, by the company Bolt Threads (*Bolt Threads – Technology*, 2018).

Synthetic silk proteins have also been produced in insect cells using the baculovirus expression system, with some success (Huemmerich *et al.*, 2004; Ittah *et al.*, 2007). Silkworms transfected with a partial fragment of a *Trichonephila* MaSp1 gene were

able to incorporate this silk into their cocoons, but it did not perform as well as natural spider silk in mechanical tests (Wen *et al.*, 2010). Recombinant silk proteins have been produced in transgenic plants (Scheller *et al.*, 2001) at an agricultural scale, though at lower levels than the industry standard for recombinant proteins and these usually exhibit properties that are quite divergent from natural silk (Menassa *et al.*, 2004; Yang *et al.*, 2005). Plant-produced multimeric spider silk variants were produced and spun into fibres, but these also had properties that were less mechanically desirable than natural silk (Weichert *et al.*, 2014; Peng *et al.*, 2017).

Because of their ability to modify proteins post-translationally, mammalian cells have also been used to express recombinant spidroins. *In vitro*, mammalian cell lines have secreted proteins of up to 140 kDa, but with varying and often lower yields when compared to bacterial expression (Lazaris *et al.*, 2002; Grip *et al.*, 2006; Römer and Scheibel, 2008). Transgenic goats have also been used to produce spidroins in their milk in notable quantities (Karatzas, Turner and Lazaris-Karatzas, 2007; Copeland *et al.*, 2015; Munro *et al.*, 2017). This is a long process and not yet economically viable due to the costs involved, but highlights the durability and diversity of recombinant spidroins and the way they can be produced.

1.8 Uses of Recombinant Spidroins

Recombinant proteins like spidroins can be assembled into various morphologies such as films, capsules, beads and hydrogels, which makes them more useful for practical applications (Huemmerich, Slotta and Scheibel, 2006; Geckil *et al.*, 2010; Humenik, Smith and Scheibel, 2011). As a material for use in medicine, recombinant spider silk has shown much promise.

Hydrogels are 3D polymers that absorb water and swell, which is useful for various tissue engineering applications. These can be formed from many materials, including collagen or agarose, and have also been made from recombinant spidroins. *Araneus diadematus* MaSp has been used as the template for various recombinant silks (Huemmerich *et al*., 2004). One such recombinant spidroin, eADF4 (C16), is based on 16 repeats of the MaSp of *A. diadematus* and can self-assemble into stable hydrogels (Rammensee *et al.*, 2008), which can then be modified to make specific structures for

use in tissue engineering (Schacht and Scheibel, 2011). Microspheres can also be produced through the aggregation of recombinant silk (Slotta *et al.*, 2008; Rabotyagova, Cebe and Kaplan, 2009). These have the possibility to be used for drug delivery systems, by encapsulating water insoluble drugs, or by becoming impregnated by water soluble drugs (Liebmann *et al.*, 2008; Lammel *et al.*, 2011).

Natural composite materials are already extensively used in the aerospace industry (Balakrishnan *et al.*, 2016). Recombinant spidroins have been combined with synthetic materials in order to create semi-natural composites, such as spider silk epoxy. By depositing recombinant MaSp spidroins onto an epoxy matrix, a thin, tough, strong and light material was created which could be suitable for use on an aircraft fuselage (Archana *et al.*, 2018). Silk has been used to reinforce composite materials with results that are comparable to glass fibre reinforced composites and outperformed plant fibre reinforced composites during testing (Shah, Porter and Vollrath, 2014). Recombinant silk could recapitulate this strengthening, making it useful in aerospace and the automotive areas of research, while being more cost effective to produce than natural silk.

The use of different buffers, spinning methods and genetic or chemical modifications all allow the adjustment of material properties to meet different demands. For example, recombinant spidroins can be engineered to change the rate of degradation when *in vivo* and post-spinning conditions can significantly affect fibre mechanical properties (Albertson *et al.*, 2014; Müller-Herrmann and Scheibel, 2015). By genetically or chemically modifying recombinant spidroins, additional functionality for bioengineering applications can be conferred. Electrospinning is often used for recombinant fibre formation, but has also been used to construct wound dressings using this material, which can enhance wound healing *in vivo* (Zhao *et al.*, 2017; Liu *et al.*, 2021). In addition, this technique was used to create an air filter device using the recombinant spidroin eADF4(C16), with high filter efficiency (Lang, Jokisch and Scheibel, 2013).

Genetic fusion has been used to produce recombinant spidroins with various affinity domains, such as fibronectin, to enhance cell adhesion and growth, IgG as a way to present biomolecules or enzymes for use as biosensors (Jansson *et al.*, 2015; Zhao *et*

al., 2017; Tasiopoulos, Widhe and Hedhammar, 2018). This approach allows the successful functionalisation of silk; however, it is a process that takes time and needs to be repeated for each new desired fusion peptide. In addition, large fusion peptides could disrupt the mechanical properties of the fibre. Harvey *et al.* (2017) were able to 'decorate' spider silk with antibiotic molecules post-fibre formation, in a rapid reaction that could be applied to other silk types.

It has been shown that the size of a recombinant silk protein is a key factor in controlling mechanical properties (Xia *et al.*, 2010). This is because larger proteins can make more inter- and intra-molecular interactions which strengthen the fibre, showing that it is important to try and produce recombinant silks that are as close to their native size as possible, in the hope of mimicking their natural properties. A comparison of the mechanical properties of some natural and recombinant silks alongside selected man-made materials is shown in Table 1.A.

Full-length spidroins are difficult to produce in microorganisms, due to their large size and repetitive nature. The highly repetitive domains of silk genes lead to homologous recombination, deletions, transcription pauses and errors, the production of inclusion bodies and low yields. Codon optimisation of the designed vector to the host expression system can help to increase yield and reduce errors in both yeast and bacteria (Elena *et al.*, 2014; Kim, Yoo and Kang, 2015). Additionally, because the spidroins are so large, depletion of the microorganismal tRNA pools may result in errors and ribosome stalling (Rosano and Ceccarelli, 2014). By artificially increasing the tRNA availability, Xia *et al*. (2010) produced full length silk proteins, comparable in size and mechanical properties to natural silks.

The applications of recombinant silks are currently limited by a lack of tuneable functionality. As research progresses, chemical and genetic modification of recombinant spidroins is becoming more common and materials are beginning to resemble those which could be directly applied in a practical setting. By using the knowledge gained from understanding natural silks, especially those with unusual properties, it will be possible to expand the repertoire of functionally useful recombinant silks.

Table 1.A: A comparison of properties from a selection of natural, recombinant silks and manmade materials.

Material		Silk Type	Species	Strength (GPa)	Elasticity (%)	Toughness (MJ/m^{-3})	Ref
Natural Spider Silk		Major ampullate	*A. diadematus*	1.1	27	180	(Gosline *et al.*, 1999)
			T. clavipes	1.2	17.2	111	(Agnarsson, Boutry and Blackledge, 2008)
			G. cancriformis	1.1	33.1	193	
			D. spinosa	1.3	19.1	124	
			L. geometricus	0.7	31	117	
			L. hesperus	1.4	30.3	181	(Swanson, Blackledge and Hayashi, 2007)
			C. darwini	1.6	52	354	
		Minor ampullate	*A. argentata*	0.92	22	137	(Blackledge and Hayashi, 2006)
			A. gemmoides	1.4	33	137	(Stauffer, Coguill and Lewis, 1994)
		Flagelliform	*A. diadematis*	0.5	270	150	(Römer and Scheibel, 2008)
		Tubuliform	*A. argentata*	0.48	19	95	(Blackledge and Hayashi, 2006)
		Aciniform	*A. trifasciata*	0.68	86	376	(Hayashi, Blackledge and Lewis, 2004)
			A. argentata	1.1	40	230	(Blackledge and Hayashi, 2006)
Natural Insect Silk		Fibroin	*B. mori*	0.6	18	70	(Gosline *et al.*, 1999)
Recombinant Spider Silk	4RepCT	Major ampullate	*E. australis*	0.2	10	-	(Stark *et al.*, 2007)
	CC1	Major ampullate	*E. australis*	0.11	12	-	(Grip *et al.*, 2006)
	96mer	Major ampullate	*T. clavipes*	0.51	15	-	(Xia *et al.*, 2010)
	W4	Aciniform	*A. trifasciata*	0.12	37	-	(Xu *et al.*, 2012)
Kevlar 49		-	-	3.6	2.7	50	(Gosline *et al.*, 1999)
Carbon fibre		-	-	4	1.3	25	
Steel		-	-	1.5	0.8	6	
Nylon 6.6		-	-	0.95	18	80	

1.9 Aims and Objectives

Chapter 1 (current) aims to present the pertinent background information needed to understand the context of the following research, as well as the key areas for consideration when creating a recombinant silk creation 'pipeline'. It also presents some new information on the morphology of *A. aquatica*, a spider which lives almost entirely underwater and explores the structure of the silk it produces.

Chapter 2 explores the silk of two remarkable spiders; *Lessertia barbara* and *Nesticus eremita*, which live in or around the acid caves of Italy. These represent an understudied group of spiders that live in an extreme environment. This investigation aims to demonstrate a robust method to identify a number of silk termini sequence transcripts and present some possible environment and species-specific silk motifs present in different silk types. To do this, Illumina sequencing is used, an accessible, commonly used method of identifying spidroin sequence fragments, and *de novo* transcriptome assembly. The adaptability of this approach is also highlighted, as it can successfully amplify transcripts even from very degraded specimen samples. Using an RNA approach allows only the analysis of actively transcribed genes and ablate the problems often encountered with DNA analysis, such as 'dead' genes or transposable elements. Specific differences are found between this silk and that of related spiders such as araneids, which have been amongst the most thoroughly studies to date, and as such are well-represented by sequence data. These changes likely influence the physical properties of the resulting silk, but conclude that the basic mechanism of production and assembly is the same. Some sequence features are also highlighted that may be useful to include when trying to fine tune the properties of recombinant silks for use in different environments.

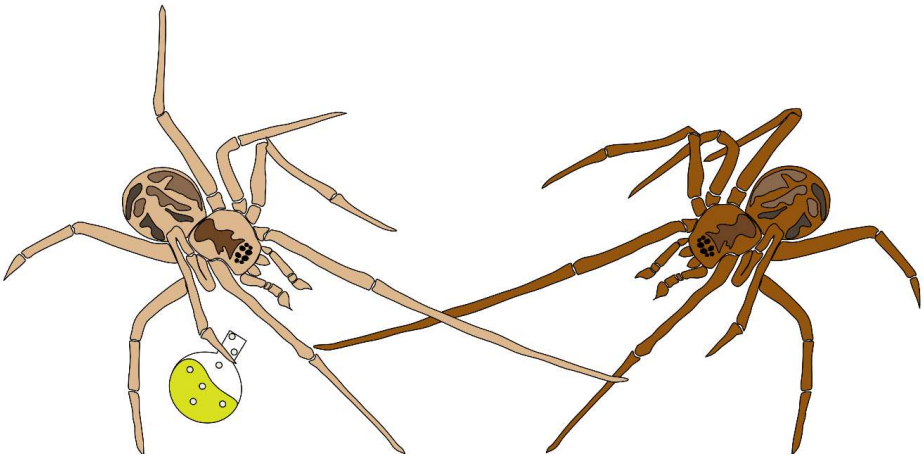
Chapter 3 looks further into the methods available for silk transcript sequencing and identification, and aims to compare and combine approaches, while exploring diversity of the silk of several species; including the water spider *Argyroneta aquatica* and *Cyrtophora citricola*. Five short-read transcriptomes were constructed, using a range of starting tissue types and spider species. Additionally, long read sequencing is used to present the first full length silk sequences from *A. aquatica* and *C. citricola*. It is shown that transcriptomics using short read data is still a useful and more

affordable way to learn about specific regions of genes of interest. While long read sequencing demands more time, money and ultimately resources, it can unearth a wealth of information. By comparing these methods, it is possible in this chapter to recommend suitable approaches for future silk investigations and suggest some important features for recombinant spidroin design. The sequences generated and bioinformatic pipelines created are planned to be deposited to EMBL-EBI and GitHub for public use.

Chapter 4 utilises the knowledge gained from RNA sequencing, and the structure and sequence of spider silk genes, to create the first artificial miniaturised aquatic silk spidroin, AquaRepCT, in *E. coli*. In order to assess the polymerisation and dimerisation properties of this molecule, both the full length protein and the C-terminal domain in isolation were expressed and several structural studies undertaken. These results suggest that the repetitive regions have more importance than previously thought in producing fibres. This has implications for controlling solubility or rate of fibre formation in future constructs. Improvements for future constructs are also suggested. This chapter highlights that the modular 4RepCT system is flexible and adaptable, allowing the production of a diverse range of different silk protein types.

Chapter 5, the discussion, gives a brief overview of the findings, and recommends further direction and routes of investigation, as well as presents several novel construct designs for future recombinant spidroin expression.

Overall, this thesis aims to draw a link between natural and synthetic silk proteins, highlighting tools, techniques and key motifs that should be appreciated or included in the study of these molecules, whether they be at the genetic or protein level.



Chapter Two

An exploration of the first sequenced linyphiid and nesticid silk genes; represented by cave dwellers

Journey to the Centre of the Earth

2.1 Introduction

2.1.1 Spider Silk

Most spiders produce silk, made up of proteins called spidroins, and use this to help support prey capture, feeding and reproduction in an array of different environments. These proteins are diverse in both form and function and are used to create several different types of materials, from strong fibres to sticky glue coatings (See Chapter 1 for more information). Generally, natural spidroins are made up of a non-repetitive amino terminal domain (N-terminal domain, NTD), a large region made up of repetitive amino acid units, and a non-repetitive conserved carboxyl terminal domain (C-terminal domain, CTD) (Challis, Goodacre and Hewitt, 2006; Ayoub *et al.*, 2007). There are also often linker regions between the repetitive region and non-repetitive termini (Figure 2.1). Both non-repetitive terminal domains are conserved between species and the NTD is responsible for protein transport and pH dependent changes during fibre formation (Hayashi and Lewis, 1998). The CTD is highly conserved between species with up to 30% codon identity across the whole region, and every spidroin has the same CTD structure (Strickland *et al.*, 2018). This remarkable degree of conservation likely allows it to carry out its important role in controlling the assembly of repetitive segments into fibres in all known silk types (Challis, Goodacre and Hewitt, 2006; Ittah *et al.*, 2006).



Figure 2.1: General protein structure of a major ampullate spidroin.

Much of our understanding of silk morphology and material behaviour has historically come from work based on spiders from the family Araneidae, a group that includes the orb-weavers (Garrison *et al.,* 2016). A brief explanation of the types and uses of the silk produced by these spiders follows, in order to provide a framework for understanding the way that these silks are used by spiders in general. Some spiders can produce up to seven different types of silk fibre, each specialised for a different use. These include major ampullate (MaSp), minor ampullate (MiSp), pyriform (PySp),

tubuliform (TuSp), flagelliform (FlaG), aggregate (AgSp) and aciniform silk (AcSp). A summary of these silk types, as well as how they are used in a conventional orb web are shown in Figure 2.2.

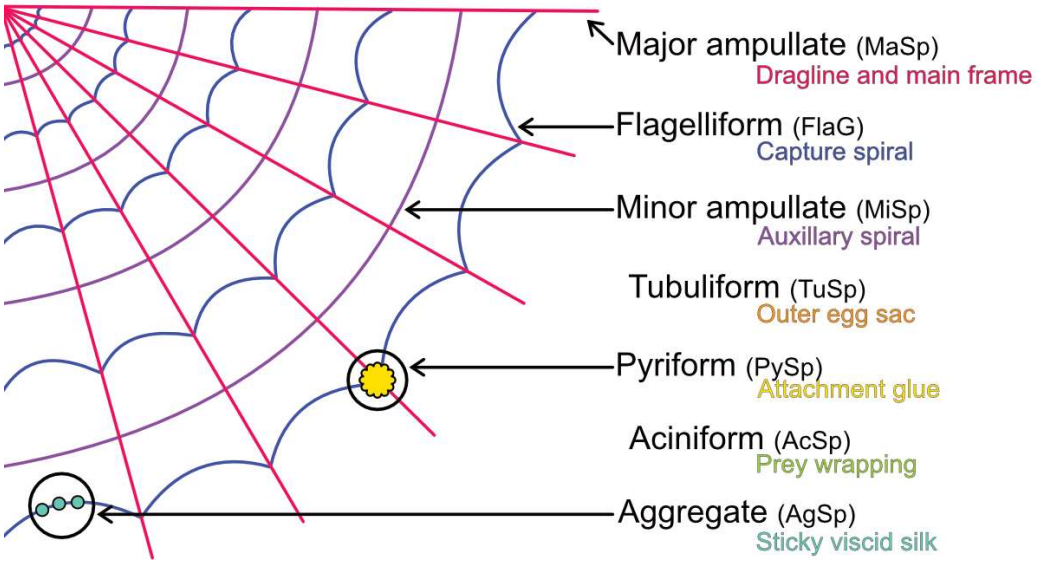


Figure 2.2: Different silk types.
This figure shows a typical orb web, showing where each silk type is used in web construction or for other supporting activities. This is based on the seven silk types produced by orb-weaving species, from the Araneidae family.

One of the strongest natural silks is the major ampullate, or dragline, silk which is used to create the main frame of a traditional orb-web. It is the strongest and toughest silk type; with strength comparable to man-made materials such as Kevlar (Gosline *et al.*, 1999). Minor ampullate silk is used to stabilise the orb web, providing a template for the capture spiral (Römer and Scheibel, 2008). This silk is thinner, less tough and less strong than major ampullate silk (Stauffer, Coguill and Lewis, 1994; Papadopoulos *et al.*, 2009). Pyriform silk is used in the attachment of fibres to solid substrates by forming sticky attachment disks (Blasingame *et al.*, 2009). Tubuliform silk is the only silk that is not produced consistently during a spider's lifetime and is used to create the tough outer casing of the egg sac (Tokareva *et al.*, 2014). Flagelliform silk is used in the capture spiral of an orb web, and has a slightly lower tensile strength than major ampullate, but has over 200% extensibility (Adrianos *et al.*, 2013). Aggregate silk forms sticky droplets along flagelliform fibres and act to adhere to prey and hydrate the FlaG fibre (Vollrath *et al.*, 1990). Aciniform silk fibres

are almost twice as tough as major ampullate silk and is used to wrap prey, as well as form the inner layer of egg sacs (Hayashi, Blackledge and Lewis, 2004).

2.1.2 Linyphiids and Nesticids

The Linyphiidae is the second largest family of spiders with over 4,300 species (World Spider Catalog, 2020). Linyphiids are found worldwide in all terrestrial biotopes and make up perhaps the most widely distributed spider family. However, this group of spiders is peculiarly understudied, and research is possibly hindered by the fact that many linyphiids are small, hard to identify even when adults, well-dispersed and often live in hard to reach or extreme environments. Linyphiids are small to medium spiders that live in a range of habitats; from hot, arid deserts to alpine tundra and bogs. They have even been found living in continental Antarctica (Pugh, 2004; Gavish-Regev and Bosmans, 2012; Růžička, Vaněk and Šmilauer, 2012). One further and perhaps less immediately conspicuous extreme environment in which linyphiids are found is in extremes of pH. One such example of this is *Lessertia barbara* (Simon, 1884), a linyphiid that resides in the acidic caves of Italy (Figure 2.3). This troglophile species has a body length of 2.3-3.3 mm and has also been found in Algeria, Italy, Morocco and Spain, mostly in caves but also in moist areas in the open air, which suggests that humidity is important for their lifestyle (Wunderlich, 1995; Bosmans, 2006; Mammola *et al.*, 2018; Pantini and Isaia, 2019; World Spider Catalog, 2020).



Figure 2.3: **Lessertia barbara.**
Male (left) and female (right) specimens of Lessertia barbara*. Scale bar shows 1.0 mm. Photographs of specimens collected in Morocco by Pierre Oger* (Oger, 2019)*, included with permission.*

The Nesticidae is a small family of spiders, with 280 species spanning 16 genera and little is known about their silks (World Spider Catalog, 2020). The species that reside in temperate zones are mainly restricted to cave-like environments, while those that live in the tropics or sub-tropics rely less on cave environments and can be found in leaf litter or grass (Lehtinen and Saaristo, 1980; Lin, Ballarin and Shuqiang, 2016). One example of a cave dwelling nesticid that inhabits environments similar to *L. barbara* is *Nesticus eremita* (Figure 2.4). *N. eremita* is also a small spider (4-7 mm) and can be found in caves, tunnels and cave-like synanthropic habitats throughout Europe, and also New Zealand (Brignoli, 1971; Vink and Dupérré, 2011).



Figure 2.4: **Nesticus eremita.**
Nesticus eremita photographed in Grotta dei Bambocci cave, central Italy, by Marcelli Massimiliano (Massimiliano, 2010), licensed under CC BY-SA 3.0.

Both *L. barbara* and *N. eremita* can be found living in the sulphuric acid caves of Monte Conca, Italy, in and around the acid springs themselves. Here, the relative humidity is consistently around 70-90% and the temperature ranges from 16-18^{o}C (Messina *et al.*, 2015). The acid droplets in the cave have a pH as low as 1, yet this environment can support a complex food web, including microbes, isopods, beetles and spiders. In both *Nesticus* and linyphiids, troglophilism seems to have arisen independently multiple times (Hedin, 1997b, 1997a; Miller, 2005). The presence of these small spiders in this unusual habitat raises important questions regarding the

way that their silk is able to persist and, presumably, have function in a humid, often low pH, environment. Currently, there is very little genomic information available for these species and to date no published silk genes, but they are phylogenetically placed closely to the araneid clade (Figure 2.5), for which there exists more extensive data.

While linyphiids, such as *L. barbara*, are thought to be derived from orb weaving ancestors (Figure 2.5), the webs produced by these spiders are starkly different. Most linyphiids use silk to form sheet webs, constructed of mainly structural and some sticky threads. These are dense, irregular and usually horizontal sheets, supported from above and below by loose threads, shown in Figure 2.6. Flying insects tend to become entangled and then trapped between these 'knock down' threads above the web, which the spider can then shake so that the prey falls onto the sheet web itself (Foelix, 1983). It is proposed that aciniform and pyriform silks are used to wrap prey and minor and major ampullate silk is used for the majority of web building (Foelix, 1983). The importance of the capture thread silks (aggregate, flagelliform) has been debated as these are often economised by ecribellate spiders, a group which includes linyphiids (Peters and Kovoor, 1991; Blackledge *et al.*, 2009). This economisation of silk glue use may be important as these spiders do not consume their webs and so cannot re-coup these spent resources. Linyphiids characteristically have poorly developed aciniform glands, and tend to have less well-developed flagelliform and aggregate glands than araneids, due to their lower importance in web-building (Peters and Kovoor, 1991). Unfortunately the full diversity of linyphiid web architecture remains poorly understood (Benjamin and Zschokke, 2004).

Linyphiids are also well known for their ability to use their silken threads to facilitate aerial travel throughout their lifecycle, in a process called ballooning (Duffey, 1956, 1998). To do this, they expel one long piece of silk that is caught by the wind, which can then carry them metres to kilometres away from their starting point (Thomas, Brain and Jepson, 2003; Woolley *et al.*, 2016). The silk used for this is thought to be major ampullate (Sheldon *et al.*, 2017; Cho *et al.*, 2018). While juveniles of other species have been reported to balloon, in many cases adults become too large to stay

airborne. This is in contrast to adult linyphiids, which retain and exhibit this behaviour frequently.

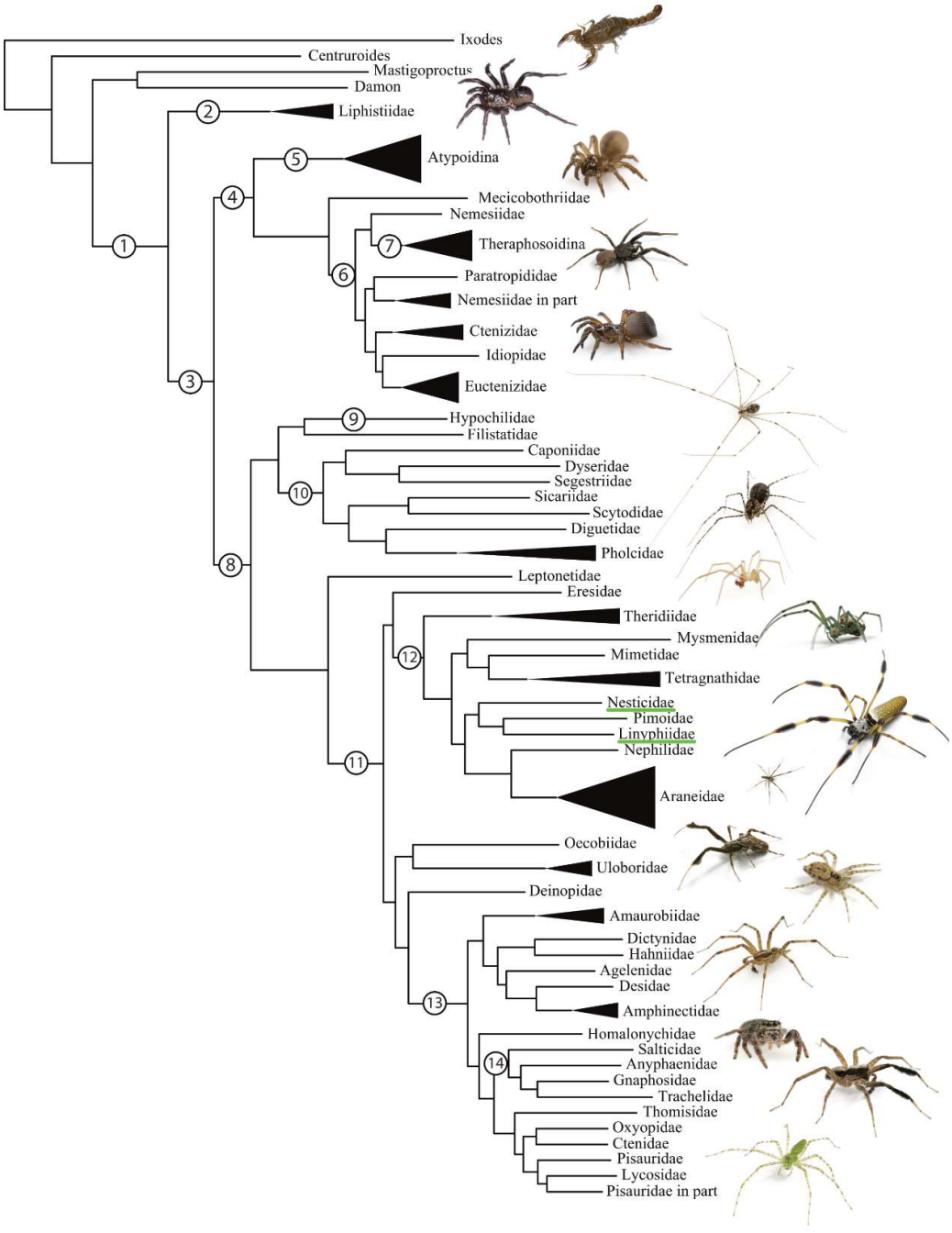


Figure 2.5: Summary of spider relationships based on phylogenetic analyses.
Green lines indicate the family of the spiders examined in this chapter. Figure modified from Garrison et al. (2016) under CC BY 4.0.

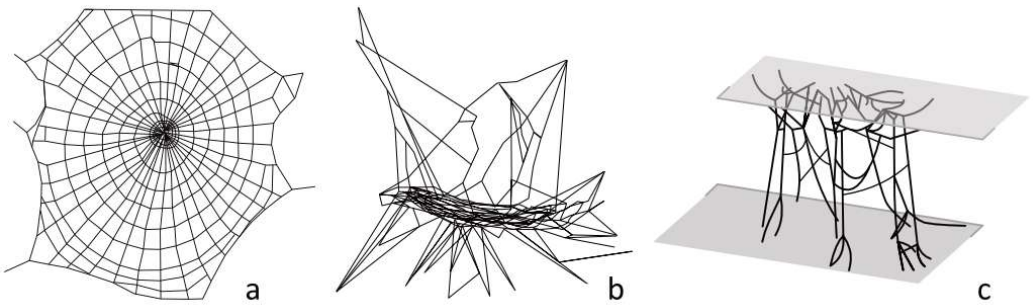


Figure 2.6: Example webs produced by Araneids, Linyphiids and Nesticids.
The typical structure of an orb web (a) produced by members of the Araneid family, features a regular circular web structure, adapted to intercepting and trapping flying insects. An example sheet web produced by some types of linyphiid (b), consists of a dense, irregularly shaped horizontal sheet, surrounded by threads designed to secure the web and intercept prey. Nesticid webs (c) consist of threads forming a loose sheet or tangle, often anchored between two cave walls (grey). The lower threads are often coated with sticky viscid silk, used to trap prey.

Species of the family Nesticidae are sedentary silk weavers which share some basic characteristics to other spider groups such as Linyphiidae and Theridiidae. They create scaffold webs, sometimes with a loose sheet, predominantly made of dry threads, with some longer dry lines with a series of small viscid silk droplets near their end (Figure 2.6) (Gertsch, 1958). These threads are assumed to be major or minor ampullate silk, coated with aggregate glue (Boutry and Blackledge, 2008). Their web is perhaps most similar to tangle webs made by theridiids, and nesticids share many common features with this family (Arnedo *et al.*, 2004; Adams, 2019).

These small spiders do not stray far from their webs and tend to carry their egg sac with them until the spiderlings hatch (Gertsch, 1958; Cokendolpher and Reddell, 2001). To immobilise their prey, nesticids use sticky lines connected to the web, which are coated with viscid silk, and swathe prey using aciniform fibres (Barrantes and Eberhard, 2007). Some individuals have also been observed to attack isopods underwater and it has been suggested that *N. eremita* can disperse using silken parachutes, ballooning much like linyphiids (Hedin, 1997a; Cokendolpher and Reddell, 2001).

This chapter shows, for the first time, the genetic sequence of the silk of small, acid cave-dwelling spiders from the families Linyphiidae and Nesticidae and their amino acid sequences. This chapter aims to investigate the composition of these silks and to identify, by analysis of silk sequences from two different species, any key changes in amino acid make-up that may be due to environmental adaptation. These spiders are good study organisms to use when investigating natural variation and evolution of spider silk genes because each family produces silk and uses it for a wide range of activities, while also persisting at environmental extremes.

Two species are examined in this study that have adapted to environmental extremes: *Lessertia barbara* and *Nesticus eremita.* These species share an overlapping geographical range in and around the acidic springs of the Monte Conca Cave system in Italy and were collected in the same date range. In addition, two transcriptomes are *de novo* assembled: one for each *L. barbara* and *N. eremita* using specimens taken from areas of extreme pH (pH 1.5-1.7) inside the cave. Additionally constructed, is a *de novo* assembled transcriptome using genetic material from *L. barbara* individuals taken from the same environment, but which were found in areas of neutral pH away from the acid springs, in order to see if any difference in silk expression was identifiable.

2.2 Materials and Methods

2.2.1 Spider Collection

Specimens of *Lessertia barbara* and *Nesticus eremita* were collected from the acid springs of the Monte Conca Cave, Italy, by Marco Isaia and Giuseppe Nicolosi (University of Turin) from April - October 2017 and immediately placed in RNALater (Ambion). At the same time, specimens of *L. barbara* were collected from areas of neutral pH away from the acid springs in the same cave as a non-acid environment comparison. The recorded pH in the acid springs at time of sample collection was 1.5-1.8. Samples were stored at room temperature (RT) in RNALater for over one year until RNA extraction in 2019. Manufacturers recommend that it should be only stored at RT for up to one week, after which it should be frozen, therefore, the RNA was expected to have degraded significantly.

2.2.2 RNA Extraction

For whole body RNA extraction of *L. barbara* and *N. eremita*, up to nine individuals were removed from RNALater (Ambion) before immediate submersion in liquid nitrogen. At this point, all tissue samples were disrupted using a pestle and mortar and RNA extracted using the GeneJET RNA purification kit (ThermoFisher) in under 45 minutes. The concentration of RNA was then recorded using a spectrophotometer (Nanodrop 2000), before sequencing of the three separate samples (*L. barbara* (acid), *L. barbara* (non-acid) and *N. eremita*).

2.2.3 Sequencing

RNA sequencing was carried out by Macrogen. Total RNA integrity was tested using an Agilent Technologies 2100 Bioanalyser or 2200 TapeStation. mRNA was purified using the Illumina TruSeq stranded mRNA LT sample prep kit (library protocol part # 15031047 Rev E) and sequencing carried out using NovaSeq6000 s4 Sequencing (150 PE, 1 Gb), to generate over 20 million paired end reads of 101 base pairs.

2.2.4 Transcriptome Assembly and Quality Control

Trimming and quality control of the sequenced reads before assembly were essential to highlight and remove any erroneous sequence variants introduced during library

preparation, sequencing and imaging. Filtered reads were then assembled into contigs, sets of overlapping DNA segments that together represent a consensus region of DNA, using Trinity, before translation into protein coding sequences using TransDecoder.

Barcodes added during the sequencing process and low-quality sequences were removed using Trimmomatic v.0.36.6 with the following settings: Leading: 20, Trailing: 20, Sliding window: 4:20, Minimum length: 36 (Bolger, Lohse and Usadel, 2014). FastQC v.0.11.8 was used to manually check the quality of the trimmed sequences (Andrews, 2010). Trimmed and filtered paired end sequencing data was then assembled into contigs using Trinity v.2.2.0 (Grabherr *et al.*, 2011). The quality and completeness of each transcriptome was assessed using QUAST v4.1 (Gurevich *et al.*, 2013) and BUSCO v.3, which provides a quantitative measure of transcriptome completeness based on evolutionarily-informed expectations of gene content from near-universal single-copy orthologs, selected from OrthoDB v9 (Simão *et al.*, 2015). Some tools were run using the public server at usegalaxy.org (Afgan *et al.*, 2016).

2.2.5 Identification of Novel Spidroins

Transdecoder v.1.0.3.0 was used to identify coding regions within the transcript sequences and translate them into their corresponding amino acids (Haas *et al.*, 2013). The output files were then used in a BLAST (Basic Local Alignment Search Tool) search against a custom database of spidroin (silk) DNA or amino acid sequences. This custom database was created using all spider silk protein sequences available on the NCBI database (downloaded on 23-June-19). This BLAST search was used with an E-value cut off of $1e^{-5}$ to $1e^{-4}$ (Altschul *et al.*, 1990). Sequences identified were then BLAST searched again against the entire NCBI non-redundant protein database to confirm identity. Sequences that were identified using BLASTp and BLASTn were kept as the most likely silk sequences, and those that were highlighted by BLASTx, tBLASTx and tBLASTn kept as spidroin-like, for future analysis. Alignments and sequence analysis were carried out using Geneious or Clustal (Geneious, 2018; Madeira *et al.*, 2019). Published spidroin sequences were retrieved from GenBank and used to help identify termini regions and confirm classification of silk type. Silk identification checks were done by eye and alignments refined manually. SignalP was used to

predict secretory signal peptides in protein sequences, as signal peptides have been shown to be a key element in the transport of spider silk transport across the endoplasmic reticulum during secretion (Wen, Liu and Meng, 2017). Multiple sequence alignments were carried out using CLUSTAL Omega (Madeira *et al.,* 2019) implemented in the Geneious software (Geneious, 2018).

2.2.6 Defining N- and C-Terminal Domains

The N-terminal domain (NTD) of the silk sequences was defined using a set of criteria based on major ampullate spidroins and is approximately 125 residues long. The start of this region was taken from the conserved first PWXN motif, where X stands for any amino acid (but is most commonly S or T and less commonly A or K or L or G). Approximately 29 amino acids (aa) downstream of the starting amino acid is a well conserved DXXXD motif. At positions 57 and 58 there is most often a HK (or XK) motif which leads into a highly conserved 'LQALCMAMATSMAEIAASE' sequence. This is referred to as a region of high conservation. Eight amino acids later there is a conserved K and the NT ends on a conserved E residue approximately 27 amino acids later. This is based on analysis of the MaSp NTD and only used as a general guide for analysis of other silk types, which are diverged from this. An overview is shown in Figure 2.7.

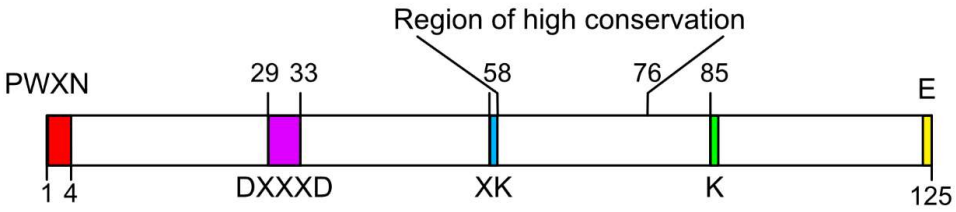


Figure 2.7: Schematic of spidroin NTD.
Specific residues and their position in the N- terminal domain are indicated. Numbering convention is taken from the first residue in the conserved PWXN-like motif.

The C-terminal domain (CTD) of the silk sequences in this study are approximately 72 amino acids long. The start of this region is defined using the nearest arginine (R) residue upstream from the 'Q' in the conserved QALLEVITAL motif (approximately 45

residues). There is also an arginine residue approximately nine residues before this in most MaSp and MiSp C-terminal regions. As all C-terminal domains contain some version of the QALLEVITAL motif, the 'Q' was used as marker, and the CTD ends approximately 26 amino acids after this residue, when the conservation of residues between species and silk types drops dramatically; this is generally a serine (Figure 2.8).



Figure 2.8: Schematic of spidroin CTD.
Although few specific residues are indicated here, C-terminal domains tend to show a high degree of conservation between silk types and between species. The hydrophobicity profile of this region can also be used to indicate its presence. Numbering convention is taken from the conserved arginine (R) residue approximately 45 residues upstream from the conserved QALLE motif.

2.3 Results

2.3.1 RNA Extraction, Sequencing and Transcriptome Assembly

Three transcriptomes were assembled from short read Illumina sequencing reads that covered the expressed RNA from two species: *Lessertia barbara* and *Nesticus eremita*. The total RNA and RIN (RNA Integrity Number) scores for each sample used for sequencing can be seen in Table 2.A. Quality assessment revealed that transcriptome assembly scores were fairly low (Figures 2.9, 2.10 and Table 2.B), but putative spidroin sequences were identified in each.

Table 2.A: Total RNA and RIN scores for each sample used for sequencing.

	L. barbara (acid)	*L. barbara* (non-acid)	*N. eremita* (acid)
RNA (μg)	10.78	8.27	17.92
RIN	2.2	2.7	2

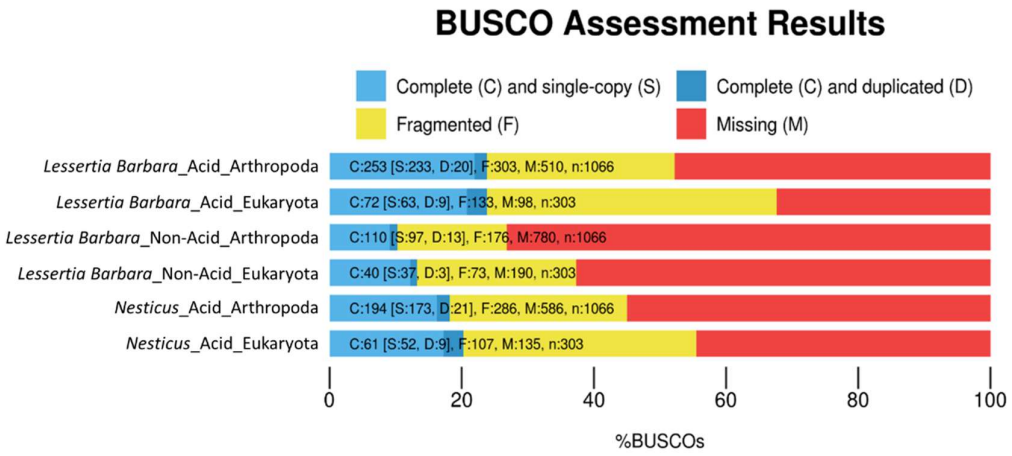


Figure 2.9: BUSCO completeness scores for each assembled transcriptome.
This was assessed against the published Eukaryota or Arthropoda BUSCO set (v9.0), as indicated in the figure.

Table 2.B: Assembly metrics for each transcriptome

	Lessertia barbara (acid)	*Lessertia barbara* (non-acid)	*Nesticus eremita* (acid)
# contigs (>= 0 bp)	24476	8186	17861
# contigs (>= 1000 bp)	732	173	502
Largest contig	11166	8942	6159
Total length (#bases in assembly)	3936808	1146450	3125308
Total length (>= 0 bp)	9673628	3131954	7249072
Total length (>= 1000 bp)	1038077	250258	646762
N50	749	730	725
N75	607	586	598
L50	1826	558	1597
L75	3293	1001	2787
GC (%)	40.09	42.7	39.26

Assembly metrics generated using QUAST. The N50 is defined as the minimum contig length needed to cover 50% of the assembly, and the N75 the same but to cover 75%. The L50/75 is the number of contigs of at least length = N50/75.

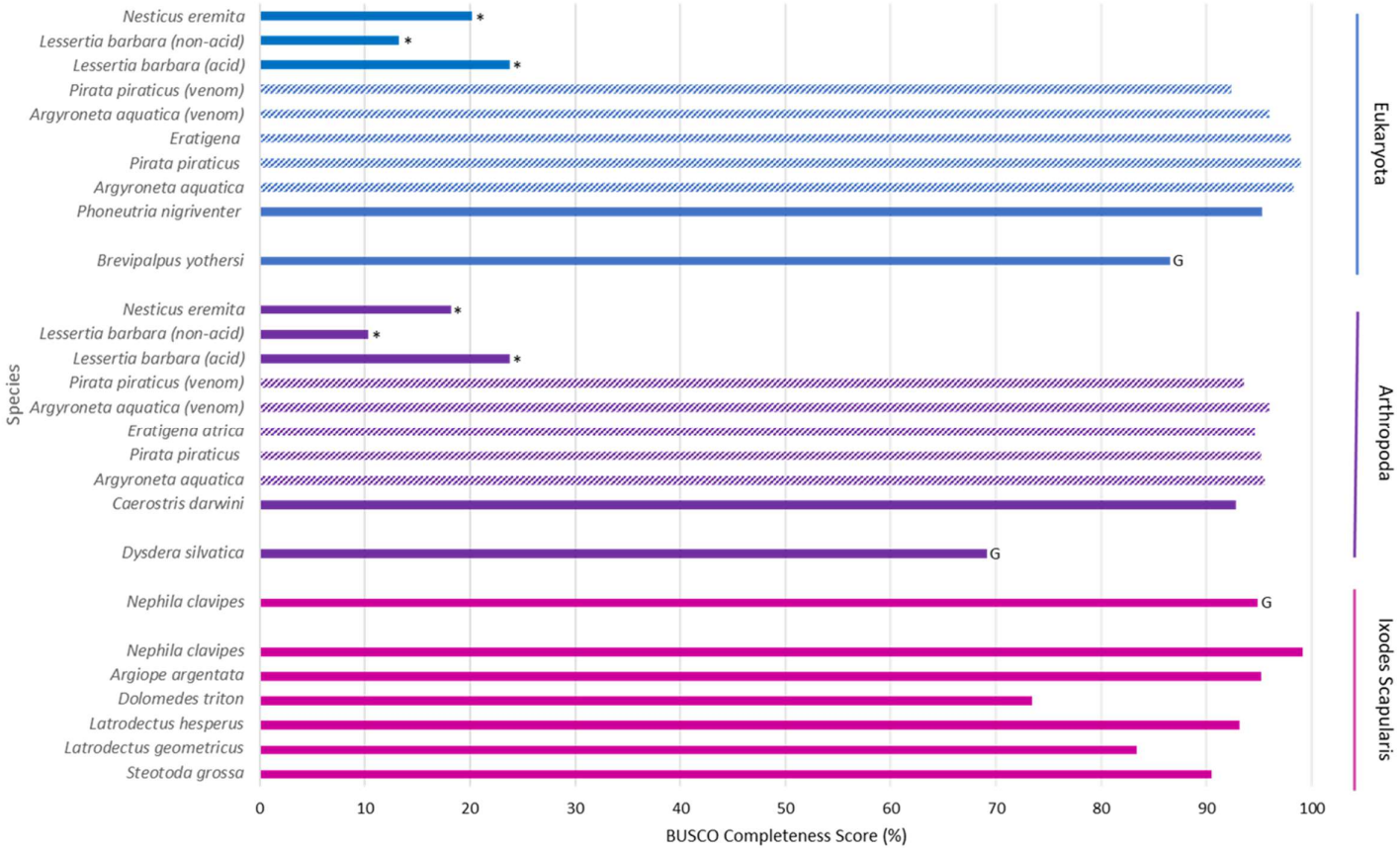


Figure 2.10: Comparison of published BUSCO scores for a range of spider gene assemblies.

*A variety of BUSCO sets can be used to assess transcriptome completeness, which are designated by the various colours; blue (Eukaryota), purple (Arthropoda), pink (the deer tick, Ixodes scapularis). Bars marked with 'G' represent a BUSCO score for a genome, scores for the transcriptomes generated here are designated with * and patterned bars are scores from assemblies generated for parallel studies at the same time as those of the acid spiders (see also Chapter 3).*

2.3.2 Lessertia barbara *Population Analysis*

To determine whether the two assembled transcriptomes from *Lessertia barbara* represented two distinct sub populations, the sequenced CO1 gene fragment was isolated from the transcriptomes of both the acid and non-acid *L. barbara* population and compared (Figure 2.11), with 100% identity.

```
Consensus                       ATGTATCCACCTCTTGCTTCGTTAGAAGGGCATTCTGGGAGATCTGTCGATTTTGCTATT   60
LBA_COl_TRINITY_DN42045_c0_gl_il ATGTATCCACCTCTTGCTTCGTTAGAAGGGCATTCTGGGAGATCTGTCGATTTTGCTATT   60
LBN_COl_TRINITY_DN6252_c0_gl_il  ATGTATCCACCTCTTGCTTCGTTAGAAGGGCATTCTGGGAGATCTGTCGATTTTGCTATT   60

Consensus                       TTTTCCTTGCACTTAGCTGGTGTTTCTTCAATTATGGGGGCTATTAATTTTATTTCTACT  120
LBA_COl_TRINITY_DN42045_c0_gl_il TTTTCCTTGCACTTAGCTGGTGTTTCTTCAATTATGGGGGCTATTAATTTTATTTCTACT  120
LBN_COl_TRINITY_DN6252_c0_gl_il  TTTTCCTTGCACTTAGCTGGTGTTTCTTCAATTATGGGGGCTATTAATTTTATTTCTACT  120

Consensus                       ATTTTAAATATACGTGGGTATGGTATAACTTTAGAAAAAGTTCCATTATTTGTATGGTCG  180
LBA_COl_TRINITY_DN42045_c0_gl_il ATTTTAAATATACGTGGGTATGGTATAACTTTAGAAAAAGTTCCATTATTTGTATGGTCG  180
LBN_COl_TRINITY_DN6252_c0_gl_il  ATTTTAAATATACGTGGGTATGGTATAACTTTAGAAAAAGTTCCATTATTTGTATGGTCG  180

Consensus                       GTATTAATTACAGCAATTTTATTATTATTATCGTTACCTGTGTTAGCAGGTGCTATTACT  240
LBA_COl_TRINITY_DN42045_c0_gl_il GTATTAATTACAGCAATTTTATTATTATTATCGTTACCTGTGTTAGCAGGTGCTATTACT  240
LBN_COl_TRINITY_DN6252_c0_gl_il  GTATTAATTACAGCAATTTTATTATTATTATCGTTACCTGTGTTAGCAGGTGCTATTACT  240

Consensus                       ATACTTTTGACTGATCGAAATTTTAATACATCATTTTTTGACCCATCGGGGGGGGGGAGAT  300
LBA_COl_TRINITY_DN42045_c0_gl_il ATACTTTTGACTGATCGAAATTTTAATACATCATTTTTTGACCCATCGGGGGGGGGGAGAT  300
LBN_COl_TRINITY_DN6252_c0_gl_il  ATACTTTTGACTGATCGAAATTTTAATACATCATTTTTTGACCCATCGGGGGGGGGGAGAT  300

Consensus                       CCTATTTTATTCCAGCATTTATTT                                      324
LBA_COl_TRINITY_DN42045_c0_gl_il CCTATTTTATTCCAGCATTTATTT                                      324
LBN_COl_TRINITY_DN6252_c0_gl_il  CCTATTTTATTCCAGCATTTATTT                                      324
```

Figure 2.11: Alignment of CO1 genes from* Lessertia barbara *transcriptomes.
Figure created using Clustal Omega (EBI).

2.3.3 Identifying Spidroin Sequences

A selection of BLAST tools (BLASTn, BLASTp, BLASTx, tBLASTx) were used to identify any potential spidroin sequences by searching against a custom spidroin database to an E-value of at least $1x10^{-4}$, then specific regions annotated by eye. The number of sequences found, separated into silk types and used for further analysis is shown in Table 2.C. Silk types were determined by the identity of two BLAST hits with the highest bit score to the sequence of interest. If these did not agree or there were not two hits, the sequence was analysed for the presence of the defining repetitive motif of each silk type and named 'silk-like', for future analysis. The N- and C-terminal domains were identified according to the convention described in section 2.2.6.

Annotation in the CTD region varies as to whether it is defined from the first or second upstream arginine resdiue. It is proposed here that the region between these two residues is too variable when comparing between species. Therefore, for continuity, the CTD here is always defined as starting at the second arginine, to make classification and comparison of the CTD easier between silk types. For deeper analysis, this region can be extended and compared, especially when using closely related species or groups. All spidroins used in analysis can be found in Appendix A.2.3.

Table 2.C: Number of identified putative spidroin sequences and terminal domains used in analysis.

Silk Type		*Lessertia barbara*	*Nesticus eremita*
Major Ampullate (MaSp)	Unique Sequences	2	3
	CTD sequences	1	3
	NTD sequences	1	0
Minor Ampullate (MiSp)	Unique Sequences	1	1
	CTD sequences	1	1
	NTD sequences	0	0
Aciniform (AcSp)	Unique Sequences	0	0
	CTD sequences	0	0
	NTD sequences	0	0
Pyriform (PySp)	Unique Sequences	3	0
	CTD sequences	1	0
	NTD sequences	0	0
Tubuliform (TuSp)	Unique Sequences	2	1
	CTD sequences	0	1
	NTD sequences	2	0
Aggregate (AgSp)	Unique Sequences	8	27
	CTD sequences	1	0
	NTD sequences	0	3
Flagelliform (FlaG)	Unique Sequences	0	1
	CTD sequences	0	0
	NTD sequences	0	0

Silk sequences are summarised in Appendix A.2.1. Table 2.D shows a summary of silk types found and their general characteristics. Individual silks are analysed in further detail in the following sections.

Table 2.D: A summary of silk types and their general amino acid characteristics.
No/missing data is indicated by grey shaded squares. The orange shaded square indicates that this information is taken from a putative sequence.

		Elastic	Crystalline	Amorphous	Crystalline	Coil	Lack proline?	Non-repetitive terminal domains	
	Silk type	β-turn spiral (GPGXX)	β-sheets (alanine rich)	Helix (GGX)	β-sheet (GAG)	Serine rich		NT	CT
Lessertia barbara	MaSp	×	✓	✓	✓	✓	✓	✓	✓
	MiSp	×	✓	✓	✓	✓	✓		✓
	PySp	×	×	✓	×	✓	×		✓
	TuSp	×	✓	✓	×	✓	✓	✓	
	AgSp	✓	✓	✓	×	✓	×		✓
	FlaG								
	AcSp								
Nesticus eremita	MaSp	✓	✓	✓	×	✓	×		✓
	MiSp	×	×	✓	✓	✓	✓		✓
	PySp								
	TuSp	×	✓	×	×	✓	✓		✓
	AgSp	✓	✓	✓	×	✓	×	✓	
	FlaG	✓	×	✓	×	✓	×		
	AcSp								

2.3.4 Amino Acid Designations and Properties

Table 2.E shows show the amino acids, with their one letter code and general properties, useful when analysing silk sequences in the following sections (Morris *et al*., 2013).

Table 2.E: Amino acids, their one letter code and general properties

Amino acid	1 letter abbreviation	Side chain properties
Alanine	A	Small, hydrophobic
Arginine	R	Hydrophilic, basic
Asparagine	N	Small, polar, hydrophilic
Aspartic acid	D	Small, hydrophilic, acidic
Cysteine	C	Small, polar, hydrophobic
Glutamine	Q	Polar, hydrophilic
Glutamic acid	E	Hydrophilic, acidic above pH 4.2
Glycine	G	Small, non-polar
Histidine	H	Hydrophilic, basic
Isoleucine	I	Non-polar, hydrophobic.
Leucine	L	Non-polar, hydrophobic.
Lysine	K	Hydrophilic, basic
Methionine	M	Non-polar, hydrophobic
Phenylalanine	F	Aromatic, hydrophobic
Proline	P	Small, non-polar, hydrophobic
Serine	S	Small, polar, hydrophilic
Threonine	T	Small, polar, hydrophilic
Tryptophan	W	Non-polar, aromatic, hydrophobic
Tyrosine	Y	Polar, aromatic, hydrophobic
Valine	V	Small, hydrophobic

2.3.5 Major Ampullate Silk (MaSp)

2.3.5.1 Lessertia barbara

One MaSp N-terminal domain sequence was retrieved from the *L. barbara* transcriptome which showed high homology to NTDs of other published MaSp genes, with the closest match being to a spider outside the Araneidae, to the pisaurid *Euprosthenops australis* MaSp1 domain (E-value of 4.76×10^{-27}) (Figure 2.12).

```
                          10         20         30         40         50         60
Consensus                 PWXNPXXAXXFMNXFXXXXSXXGXXFTXXQLXDMSXIXXSXXXXJZXLXXXGXXSXXKXQXXB   64
E. australis MaSp1 NT     PWTNPGLAENFMNSFMQGLSSMPG--FTASQLDDMSTIAQSMVQSIQSLAAQGRTSPNKLQALN   62
L. barbara (acid) NT      PWANPQTADAFMNNFLSSVTSNSGGIFTGDQLSDMSDIGFSIMNQLERLNRGGHSSARKMQMMD   64

                               70         80         90        100        110        120
Consensus                 MAFASXMAEXAXXEXXXXXXXXXXXXXIXXAXSXXFXQTTGXXNQXXINXIXXLXSMFAQXXXXN  128
E. australis MaSp1 NT     MAFASSMAEIAASEEGGGSLSTKTSSIASAMSNAFLQTTGVVNQPFINEITQLVSMFAQAGMN-   125
L. barbara (acid) NT      MAFASTMAEVALAEADNPQ--AALGAISNALSQSFMQTTGSSNQASINSIMSLISMFAQNAGEN   126
```

Figure 2.12: **L. barbara *MaSp NTD alignment with* E. australis *MaSp NTD.***
This shows the NTD of the newly identified L. barbara MaSp *aligned with the NTD region from an* Euprosthenops australis *MaSp (accession number CAJ90517.1).*

After the N-terminal region there is a linker region of approximately 50 amino acids before the repetitive poly(A) tracts. This region contains GPV and QTQ motifs, but the repetitive region that follows it lacks proline, making it a putative MaSp1. SignalP reports a 50% likelihood of a Sec signal protein at the beginning of the *L. barbara* fragment.

The repetitive regions of the *L. barbara* MaSp silk fragments contains the widely accepted repeating motifs found in major ampullate silks; Poly(A), GGX, GAG. However, these lack proline, so probably represent MaSp1. These sequences were found in both acid and non-acid transcriptomes.

Two sequences were found, one from the acid and from the non-acid transcriptome, that included a MaSp C-terminal domain. The sequences were identical, so were treated as one, and contained both the conserved QALLE motif and the canonical cysteine at position 40 expected in a MaSp CTD. A general *Lessertia barbara* MaSp1 gene structure is proposed in Figure 2.13. Figure 2.14 and Table 2.F show how the MaSp sequences compare with published CTDs.

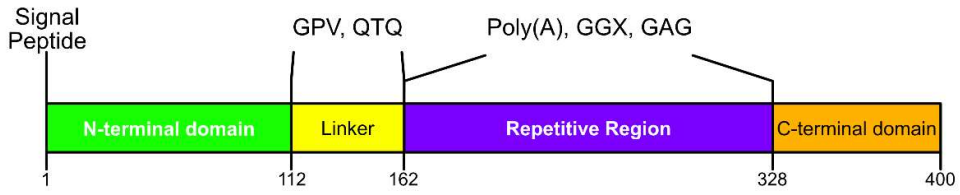


Figure 2.13: General protein structure of a* Lessertia barbara *major ampullate spidroin.

2.3.5.2 Nesticus eremita

No *Nesticus eremita* major ampullate N-terminal regions were identified, but three putative sequences were identified that contained a portion of the repetitive region and the C-terminal domain. Three different *N. eremita* putative CTDs were identified. Each one contained the canonical QALLE motif, the canonical cysteine, required for spidroin dimerisation, and comparable proportions of hydrophobic and hydrophilic amino acids to those published (Table 2.F, Figure 2.14).

Table 2.F: Summary of the amino acid make up of various major ampullate C-terminal domains.
Sequences from this study are underlined.

Major ampullate C-terminal domain	pI (%)	Percentage of amino acids					
		Acidic	Hydrophobic	Basic	Polar (uncharged)	Hydrophilic	Charged
Araneus ventricosus MaSp2	6.7	2.6	60.4	3.9	32.9	43.4	6.5
Argiope amoena	5.3	2.6	53.3	2.6	41.6	48.1	5.2
Agelenopsis aperta	5.3	2.7	56.8	2.8	37.9	48.7	5.5
Trichonephila clavipes MaSp2	4.0	3.9	55.9	1.3	39.0	48.1	5.2
Latrodectus hesperus	5.2	3.9	48.1	3.9	44.2	49.4	7.8
Peucetia viridans	6.1	2.4	52.9	2.4	42.1	46.9	4.8
Araneus ventricosus	4.3	3.2	59.0	1.6	36.0	45.8	4.8
Argiope aurantia	4.3	3.0	56.9	1.5	38.4	46.1	4.5
Argiope trifasciata	5.3	2.6	55.2	2.6	39.7	50.0	5.2
Euprosthenops australis	6.1	2.4	54.2	2.4	41.3	46.0	4.8
Lessertia barbara (acid)	6.7	3.1	54.6	4.1	38.1	45.3	7.2
Lessertia barbara (non-acid)	6.7	3.1	54.6	4.1	38.1	45.3	7.2
Nesticus eremita (DN24404)	11.3	2.0	50.4	4.1	43.3	51.5	6.1
Nesticus eremita (DN24404_c1)	9.1	2.2	51.0	2.2	44.3	52.3	4.4
Nesticus eremita (DN27466)	9.3	4.1	50.6	4.1	41.2	47.4	8.2

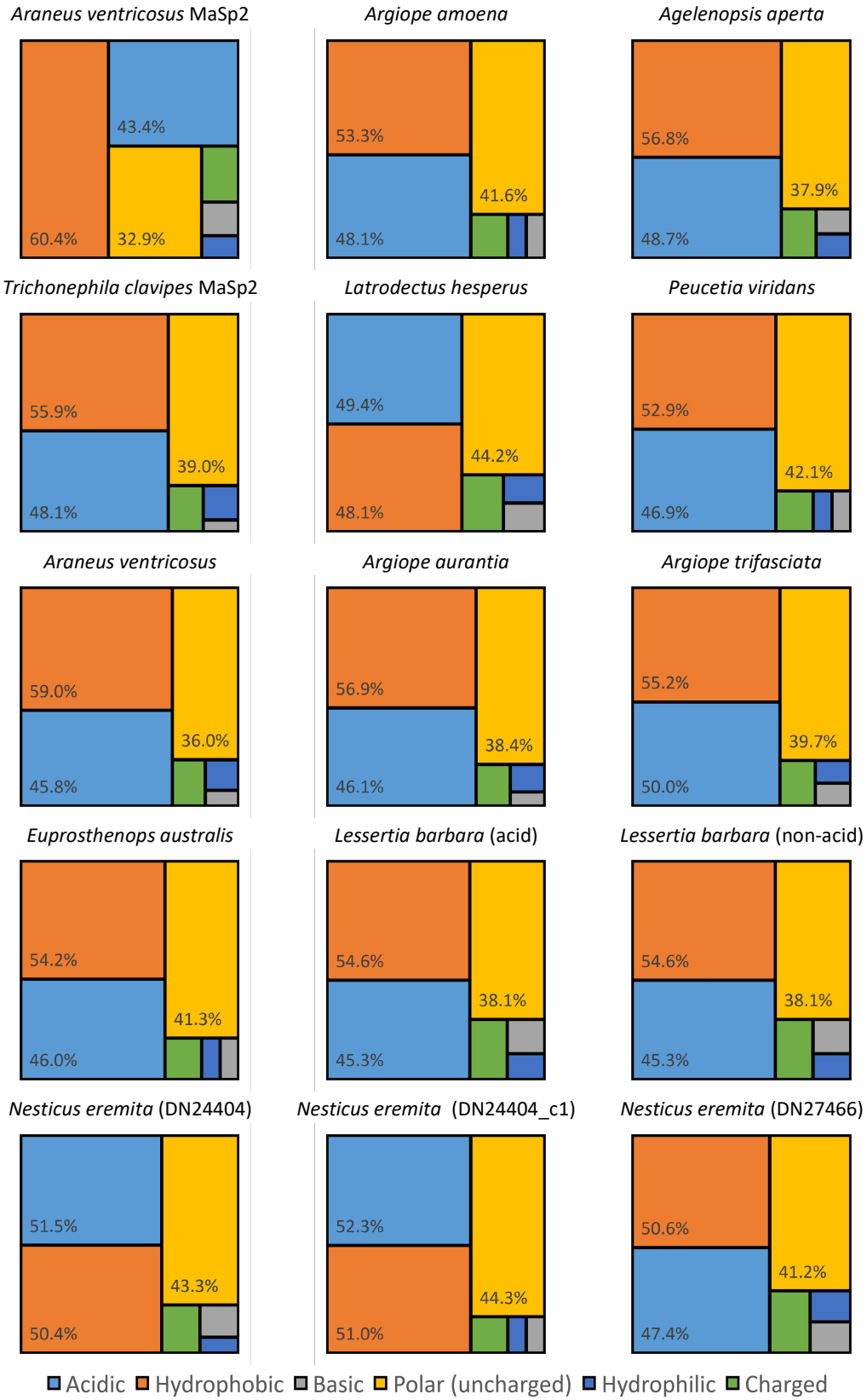


Figure 2.14: Visual overview of amino acid composition of selected MaSp C-terminal domains.

2.3.6 Potential Novel MaSp

The *N. eremita* RNA sequences comprising the MaSp repetitive region contained the repeating motifs poly(A), GG(X) and GPG. The presence of proline marks this silk as a probable MaSp2, but it does contain rather more glutamic acid and aspartate (polar, acidic residues) than seen in normal terrestrial MaSp2 silks (1 in every 24 residues on average in the non-termini region), as shown in Figure 2.15. Aspartic acid (D) residues, which are polar and acidic, are also seen in *L. barbara*, but at a lower frequency of three every 159 residues. There is also a high prevalence of the repeating motif SS, which is polar and could promote hydrogen bonding to water molecules.

It is rare for polar, acidic residues to be found in MaSp2 silks, and while they are found sometimes in MaSp1, this silk type does not follow the other motif conventions. A representative sample of the amino acid makeup of different MaSps can be seen in Figure 2.15. Sequence identifiers and values are shown in Appendix A.2.2.

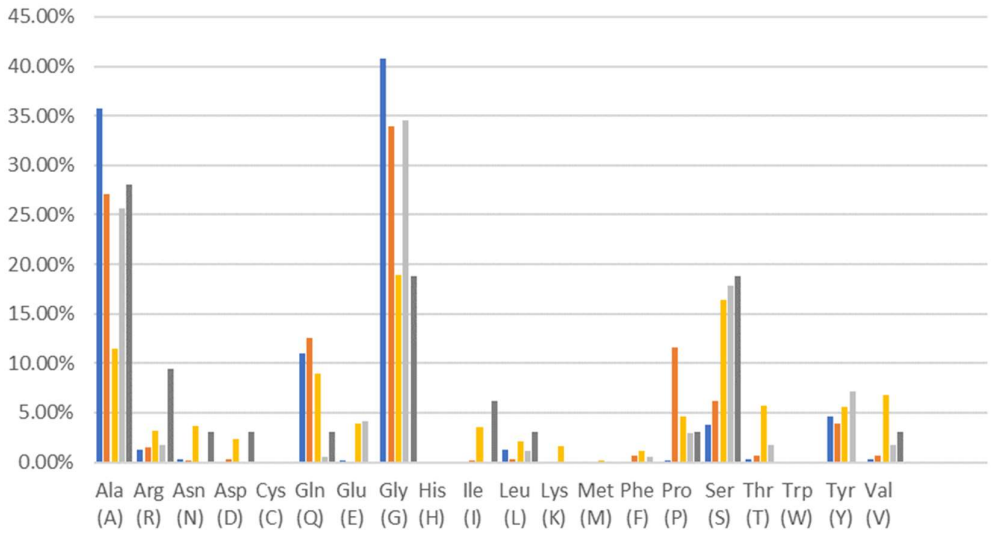


Figure 2.15: Composition of MaSp1, 2 and 3 silks compared to the novel silk type from this study. *Blue bars represent the averaged percentage that each amino acid appears in the repetitive regions of selected MaSp1 sequences from the literature, orange bars are averaged MaSp2 values and yellow bars are averaged MaSp3 values. The light and dark grey represent the amino acid content in the two potential novel MaSp types found in the* Nesticus eremita *transcriptome.*

2.3.7 Minor Ampullate Silk (MiSp)

Two representative sequences of the minor ampullate silk type were identified, one from each of the *L. barbara* and *N. eremita* transcriptomes, each of which contain the C-terminal region and a portion of repetitive sequence.

Only one other sequence was found that could represent a MiSp CTD fragment, in the *N. eremita* transcriptome, but this sequence was found to have more residues diverged from the canonical residue sequences than the others. When a BLAST search was conducted against the NCBI nucleotide collection, it was found to be more similar to fibroin heavy chain-like isoforms found in *Stegodyphus dumicola*, the African social spider. This more diverged sequence has a variation on the QALLEVITAL motif: QSLTEVVSSLV. This sequence was omitted from further analysis as a MiSp but could warrant further investigation in the future.

The repetitive region of the *L. barbara* MiSp fragment is comprised of a similar amount of alanine and glycine with tyrosine and glutamine; lacking the more often observed glycine domination of MiSp silks. Only one R (arginine) appears in the repetitive region, but this region is not long enough to determine individual repeat architectures or indeed if this is a MiSp 1 or 2. The sequence lacks proline, as is expected in MiSps and also has a very low TPM (transcript per million) count compared to other spidroins, but the conclusions that can be drawn from this are limited due to the quality of the transcriptome assemblies and starting RNA.

The repetitive region of the *N. eremita* MiSp fragment again has an equitable proportion of G and A residues, and the repeat $G(A)_n$ predominates, separated by areas of T and S rich regions. This fragment contains two repeats of GAGA-TS repeat blocks before the CT. Only one proline residue appears in the repetitive region. The repeats found in both of these sequences can be categorised as GX GGX $G(A)_n$ or SS. Glycine, alanine and serine are the most abundant amino acids in both sequences. Both MiSps generally utilise GC rich codons and are hydrophobic.

The CTD of both MiSp sequences align well to published sequences and contain the canonical QALLE motif (QVLLEIVSSLIH; QMLLEIVSSLIH). All MiSp CTDs tend to have a conserved histidine after this motif, which is conserved in some silk types in other

spiders. There seems to be no linker region between the repetitive region and the CTD.

2.3.8 Pyriform Silk (PySp)

Three non-isoform putative pyriform sequences were identified in the *L. barbara* transcriptomes, two from the acid and one from the non-acid environment. One unique CTD was identified and used for analysis, which was characteristic of pyriform CTD, and there were no differences in amino acid makeup between the non-acid and acid sequences. No pyriform sequences were identified in the *N. eremita* transcriptome. The repetitive region is composed mainly of serine and glutamine, with low glycine content. There is more glutamine than proline and there are no alanine repeats more than two amino acids long. One proline rich motif was found in each repetitive region, comprising RRPRPAPR, but interspersed proline residues were lacking. There were several repeats comprised of Q, V, S and A.

There is a region rich in charged amino acids (arginine, glutamic acid and aspartic acid) just before the CTD, but the fragment is not long enough to determine whether this is a true linker or a novel repeat type. The C-terminal domain contains the conserved region 'EV(V/L)VEALLET(M/L)AALL' which is found in other PySp silks and seems to be characteristic of PySp CTDs.

2.3.9 Tubuliform Silk (TuSp)

Four putative tubuliform sequences were found in the *L. barbara* transcriptome, two of which covered the NTD and some repetitive region, which were used in analysis.

The N-terminal region contains the conserved regions of the NTD, including the QEQEDLDAISSVILS motif. However, they are notably missing a D/E (aspartic acid/glutamic acid) residue at position 68, which is replaced with an N (asparagine). No signal peptide was predicted via SignalP. There are two key cysteines present in positions 14 and 91. While the sequenced repetitive regions are quite short, they consist mainly of alanine, serine and glutamine residues, with very few glycine residues. The sequenced repetitive regions (outside of the conserved termini) completely lack proline.

One putative TuSp fragment was identified in the *N. eremita* transcriptome, containing the repetitive region and C-terminal region. The repetitive region is again comprised of mainly alanine and serine, but lacks the high content of glutamine, found in the *L. barbara* fragments. Two prolines are present, in the 15 amino acid region directly before the CTD. The repeating motifs are most commonly AX, SA and SQ, as seen in other tubuliform silks.

The CTD is 77 amino acids long and contains almost all of the expected conserved TuSp CTD residues but does not contain any cysteines. The expected cysteine is instead replaced by a proline, with similar hydrophobicity and the hydrophobic profile of the surrounding area does not appear to be changed significantly.

2.3.10 Aggregate Silk (AgSp)

From the *L. barbara* transcriptomes, eight putative aggregate spidroin fragments were identified that contained at least one defined repeat and one of these contained a CTD. No NTD sequences were found. As the sequences extracted from these transcriptomes are quite short, analysis is fairly limited.

The *L. barbara* aggregate compare well with other published aggregate repeats, so even though most of the transcripts lack a non-repetitive region, it is proposed that they are from aggregate spidroins. Repeats are well defined and tend to be separated by GPVNS tracts. The C-terminal domain shows good alignment in the latter half to published termini regions, while the beginning of the CTD contains more R, K and H residues than seen in other AgSp CTDs.

There were twenty-seven putative aggregate spidroins identified in the *N. eremita* transcriptome, and three contained putative NT regions. The proposed NT regions are quite diverged from published aggregate NTD and tend to contain more uncharged or non-polar (V, A, P, G) amino acids than would be expected. However, the lack of availability of published aggregate spidroin N-terminal regions does not lend itself to further analysis here.

The repetitive sequences are generally longer than found in the *L. barbara* transcriptome, allowing some more structural analysis of the transcripts. When

sequences are aligned, the general structure contains two repeats, then two half repeats, then two more repeats, and then this whole structure repeats.

Whereas in *L. barbara* transcripts, repeats are separated by tracts of GPVNS repeats, the *N. eremita* repeats generally are back-to-back and well conserved from residues 1-28. There is a region of difference in the middle of the repeat tracts ranging from 1-17 residues, whereupon the last 14 residues become fairly well conserved again.

There appears to be two types of internal AgSp repeat. Those that are mainly non-polar or uncharged and those that have mainly polar basic and some polar acidic inserted amino acids in the second quarter of the repeat region. This is predicted to reduce the chance of having a beta strand. Each repeat starts with GPGTTPG and the repeats themselves are mainly nonpolar or uncharged.

2.3.11 Flagelliform silk (FlaG)

No unequivocal flagelliform fragments were found for either species, however a potential flagelliform-like sequence was found in the *N. eremita* transcriptome, comprising the repetitive region only.

This region contains the GPGFG motif found in flagelliform silks, with multiple GPG motifs. BLAST searches against the wider online protein databases only returned hits to either flagelliform, major ampullate silk or collagen. There are six repeats of the GG(X) motif and also GGX(n) and GPG motifs, typical of flagelliform silks, but no silk terminal regions.

2.4 Discussion

In this study, three transcriptomes were reconstructed from sequenced genetic material of *Lessertia barbara* and *Nesticus eremita*. The assembled transcriptomes represent a total of over 10,000 high quality sequenced base pairs which include over 100 silk gene sequences. Even using degraded RNA samples, these transcriptomes yielded several silk sequence fragments which were suitable for further analysis. The assembled silk gene fragments shed light on the genetic and protein makeup of spider silk used in extreme acid environments. These spiders are clearly able to adapt to diverse external environments, either through phenotypic plasticity or because of localised selection for different traits within populations that occupy divergent environments in and outside a cave system. These findings help develop understanding of the way that this family of proteins has diversified while maintaining important key properties in order to fulfil all the necessary functional requirements and also persist under extreme environmental conditions.

2.4.1 Linyphiid and Nesticid Silks Assemble in the Expected Way

Of the seven major silk types, five were identified in this study, as used by two different acid-cave dwelling spider species; and many fragments contain the expected N-and C-termini regions (Table 2.C).

The NTD MaSp1 sequence isolated from the *L. barbara* transcriptome showed high homology to N-terminal regions of other published major ampullate silk genes and four unique MaSp CTD sequences were isolated from the linyphiid transcriptomes, which showed remarkable homology to published C-terminal domains. The C-terminal regions are remarkably conserved to other species, suggesting that the silk is assembled in the same way regardless of the environment it is used in. In the CTDs identified, the pH sensitive 'switch' region is conserved and the regional makeup constrained to the percentages of hydrophobic and hydrophilic residues found by Strickland *et al.* (2018) (Table 2.F, Figure 2.14). The terminal regions of both of these small spider species were homologous to both theridiid and araneid silks, which could suggest that this silk represents an evolutionary step from one derived silk type to

another, which is well suited to environmental extremes. The high degree of conservation of the N- and C-terminal regions in all silk types underpins the research suggesting the importance of these regions in dimer formation; especially that the assembly of all known types of spider silk are governed by a lock and trigger mechanism involving both the N- and C-terminal domains (Andersson *et al.,* 2014).

No MaSp2 sequences were found in the *L. barbara* transcriptome, but only MaSp1, which lacks proline. It has been shown that amino acids derived from citric acid metabolites (proline, glutamine), are more metabolically expensive to produce, and spiders can reduce the expression of silks containing these amino acids (e.g. MaSp2) when nutrient intake is limited (Craig *et al.*, 1999; Guehrs *et al.*, 2008; Blamires, Wu and Tso, 2012). These results, if not an artifact due to the completeness of the transcriptome, could indicate that the silk type used by *L. barbara* is dependent on external environment, or the availability of resources in the environment, as resources are often scarcer inside caves.

MaSp2 was, however, found in *Nesticus eremita*, which were found in regions of low pH. There are several explanations for this: this could mean that these individuals are better adapted to a cave dwelling lifestyle, and consequently more successful at catching prey, so are able to invest more resources into the silk they produce. Proline may also be more important for maintaining silk function in areas of low pH. Conversely, it may simply be that the attempt to amplify and sequence MaSp2 from *L. barbara* here was unsuccessful because this silk type was not being expressed in high copy numbers at the time, or the fragments that were retrieved from sequencing are just not long enough to distinguish differences between the major ampullate silks in order to highlight more subtle amino acid differences. Either way there is a clear need for more investigation in this area, using longer read sequencing approaches coupled with physical characterization of silks.

All identified major ampullate silks contained the motifs expected: Poly(A), GGX and GAG. The repetitive region of both *L. barbara* MaSp fragments contains the widely accepted repeating motifs found in major ampullate silks; Poly(A), GGX, GAG. However, both lack proline, so were designated as putative MaSp1 sequences. This sequence was found in both the acid and non-acid transcriptomes. MaSp1 silk is

stronger and less stretchy than other silks, and so could be more suited for use in aerial travel. The *Nesticus* major ampullate silks contained more glutamic acid and other polar, acidic residues than seen in other MaSp2 silks, which could mark it as a candidate novel silk type adapted for use in acidic environments.

2.4.2 Potential Novel Silk Type

The *Nesticus eremita* RNA sequences comprising the MaSp repetitive region contained the repeating motifs poly(A), GG(X) and GPG. The presence of proline marks this silk as a probable MaSp2. However, it does contain rather more glutamic acid and other polar, acidic residues than seen in normal terrestrial MaSp2 silks (Figure 2.15). This is rarely seen in published MaSp2 sequences, but the pattern is common in MaSp1 sequences. Polar aspartic acid (D) residues are also seen in *L. barbara*, but at a lower frequency; three every 159 residues. It is very rare for polar, acidic residues (such as aspartic or glutamic acid) to be found in the repetitive region of MaSp2 silks, as they are mainly comprised of uncharged or non-polar amino acids, such as serine, threonine or alanine. Rarer amino acids are polar and basic (K, R and H). More often, polar residues are found in MaSp1 sequences, although they are again predominated by smaller, uncharged and non-polar amino acids. This can be seen in Appendix A.2.2.

The fairly recently described MaSp3 lacks the Poly(A) and GPG motifs found in other major ampullate silks, but do contain larger, more polar amino acids (usually more serine or threonine). MaSp3 also has more acidic residues than MaSp1 or 2 (Collin *et al.*, 2018) which is similar to that found in our MaSp silks. It could be pertinent to suggest that the *N. eremita* MaSp2 is more reminiscent of the newly proposed MaSp3, as it a high proportion of S and T and acidic residues, but also contains the poly(A) tracts typical of MaSp1 and 2, making it a new silk-type hybrid. The strange silk found in *Nesticus* could be a nesticid-only silk, suited to life in a resource scarce environment, or extremes of pH.

2.4.3 Other Silk Types

Representatives of four of the other main silk types were isolated from the transcriptomes. This is the first study to support the view that spiders from the family

Linyphiidae and Nesticidae use a range of silks, comprised of at least four types; MaSp, MiSp, AgSp, TuSp. PySp was found only in the *L. barbara* transcriptomes and the repeats are more similar to those found in the Araneidae family, rather than those identified from the theridiids, which is to be expected considering their evolutionary relationship. No aciniform spidroins were identified from either transcriptome. This can be expected in one case, as linyphiids tend to have both reduced silk glands and reduced numbers of aciniform spigots; probably due to the fact that they do not attack prey by swathing in silk, but by biting (Peters and Kovoor, 1991; Yoshida, 1999). However, aciniform spigots have been observed on members of the nesticid family, so it is proposed that here the transcripts were too low copy number or too degraded to survive amplification for sequencing (Coddington, 1989). They may also be so diverged from other silk sequences that they are not easily identified using homology searches to known sequences.

In the silk types identified, there was a surprising amount of homology to published sequences when the differences in silk use and ecology is considered, as well as a tendency to conserve some of the key sequence motifs that dictate the fibrous properties of the various silk types. This is summarised in Table 2.D, and lends evidence to the idea that the silk used by these spiders is not specially evolved for cave dwelling, even at extreme pH, but can still function adequately. It could also suggest that the ability to withstand acidic environments is something that many other silks may already possess because this trait is pleiotropic and could confer other physical benefits, an area which warrants further investigation.

The C-terminal domains of the minor ampullate sequences showed high homology to published sequences, and all contained similar amino acid makeup and the canonical QALLE motif. The *N. eremita* MiSp was the most reminiscent of published sequences in terms of repetitive region amino acid composition, with the motif 'GAGA' predominating. The repetitive region of the *L. barbara* MiSp fragment is made up predominantly of alanine, rather than the more often observed glycine-domination of MiSp silks. This region is composed of glycine and alanine, with tyrosine and glutamine. Only one R (arginine) in the whole sequence, but the repeats are not long enough to determine MiSp1, MiSp2 or individual repeat architectures. Proline is

found to be lacking in both species in this silk type, which is to be expected as is usually rare in other minor ampullate silks, but the presence of this silk in both transcriptomes suggests that it is an important resource in the silk repertoire.

No *N. eremita* PySps were found, which could be explained by the fact that nesticids reside in the 'reduced pyriform clade' (Coddington, 1989). However, pyriform silk spigots are still present in nesticids, if reduced, and it is proposed that further sequencing interrogation may uncover these sequences (Arnedo *et al.*, 2004; Álvarez-Padilla and Hormiga, 2011).

The repetitive region of the *L. barbara* PySp is composed mainly of serine and glutamine, with low glycine content, which mirrors that found in other pyriform silks (Chaw, Saski and Hayashi, 2017; Correa-Garhwal *et al.*, 2019). There is more glutamine than proline and there are no alanine repeats more than two amino acids long. The PX motif found in other pyriform silks was absent (Correa-Garhwal *et al.*, 2018). The QQ motif in particular showed a degree of periodicity between the polar glutamine and serine and non-polar proceeding residues, as has been described before (Perry *et al.*, 2010). This could create a surface suitable for adherence in both hydrophobic and hydrophilic environments. The identified repeats are more similar to those found in the Araneidae family than those identified from the theridiids, which is to be expected based on their evolutionary placement. No clear adaptations to adherence in an acidic environment were identified.

Tubuliform sequences were found in all transcriptome assemblies, which showed homology to published sequences. In the *Lessertia barbara* tubuliform NTD sequences, there are 2 cysteines present in positions 14 and 91, which are conserved across other tubuliform NTDs. It has been speculated that these play a key role in fibre formation in other tubuliform spidroins (Wen, Liu and Meng, 2017; Collin *et al.*, 2018). The sequenced repetitive regions consist mainly of alanine, serine and glutamine residues, as found in other tubuliform silks. However, these regions also completely lack proline. While proline is not expected to be present in TuSp repeats, it is variable as to whether proline appears in tubuliform silks in other species, as it is often present in linker regions or areas between repeats. No poly-threonine tracts are seen, which are seen in aquatic spiders (Correa-Garhwal *et al.*, 2019). The *N.*

eremita TuSp more closely resembles terrestrial, orb-weaving sequences in terms of amino acid composition, which is expected as TuSp is not used in web construction but for the conserved function of creating the egg sac.

The *N. eremita* TuSp CTD does not contain any cysteines, which were proposed to be key in allowing dimerization in other silk types. It has been shown that the C-terminal domain is involved in initiation of fibre assembly and has a role in controlling the final properties of the fibre (Ittah *et al.*, 2006; Heidebrecht *et al.*, 2015). It does this through helping the orientation of fibres, enhancing the natural structural transition during fibre formation (Xu *et al.*, 2017). It has been proposed that the conserved cysteine situated just before the QALLE motif has a key role in facilitating this process (Guerette *et al.*, 1996; Sponner *et al.*, 2004, 2005; Garb, Ayoub and Hayashi, 2010). However, experimental evidence focussing on recombinant spidroins has suggested that fibre formation is largely independent of covalent interactions (and so disulphide bridges) between the CTDs, and that stiffness of fibres can rely mainly on intermolecular interactions between poly-alanine blocks in the repetitive region (Grip, Johansson and Hedhammar, 2009). Mutational studies have shown that removing the cysteine and replacing it with a similar amino acid still allows fibre formation, suggesting that it is the secondary structure of the CTD and not the disulphide bridge that is critical (Ittah *et al.*, 2007). NTD situated cysteines could instead be involved in disulphide linkage in the case of a recombinant TuSp that lacks the CTD (Hu *et al.*, 2005). The lack of a conserved cysteine in the CTD may be because this domain may be less important for fibre formation and silk protein solubility for TuSp as it is for MaSp, which may also be the case for aciniform silk (Xu *et al.*, 2012). Phyre2 protein structure prediction places this residue just before the conserved alpha-helical, highly hydrophobic conserved QALLE motif, suggesting that a change at this residue would still allow the formation of the necessary secondary structure of the CTD (data not shown) (Sponner *et al.*, 2004; Kelley *et al.*, 2015). In all, this suggests that a lack of cysteine in this region does not necessarily mean that the silk folds in a different way to the proposed model, lending credence to the idea that linyphiid silk is assembled in a manner similar to other published spidroins.

The aggregate spidroin sequences are rich in glycine, proline, serine and threonine. Notably, serine and threonine can undergo glycosylation and are abundant, which can help to promote adhesion. Many β-sheets, turns and random coils are predicted (using EMBOSS) which are proposed to aid in target adhesion via suspension bridge formation (Collin *et al.*, 2016). The fact that so many representatives of the aggregate repeat region in these transcriptomes are identified, but very few (CTD) or no (NTD) representatives suggests that these regions are so diverged that they are no longer uncovered by the stringent BLAST search used in this study, or that they may fold very differently to other silk types. It is interesting that even though there is a variable region within the aggregate repeats, this seems to affect the (predicted) secondary structure minimally, suggesting that the structure here is still important for the function of the protein. High levels of inter-gene homogenisation are seen in the repeat modules of the aggregate spidroins. High levels of repeat homogeneity are seen throughout the silk gene family, and are often attributed to gene conversion or unequal crossing over, which leads to intragenic concerted evolution (Hayashi, Blackledge and Lewis, 2004; Ayoub *et al.*, 2007; Garb *et al.*, 2007; Chaw, Saski and Hayashi, 2017). Stabilising selection could also contribute to the protein sequence maintenance seen in the aggregate sequences (Chaw *et al.*, 2014).

One putative flagelliform fragment was identified in the *N. eremita* transcriptome, and none in *L. barbara,* which could indicate a key evolutionary loss of silk type. Flagelliform silk spigots have been observed in members of the nesticid family (Coddington, 1989), and while molecular evidence of this silk has not been presented before, it could be expected to be the silk coated with aggregate glue, used as sticky gum-foot trap lines attached to their webs, rather than MaSp as is proposed (Boutry and Blackledge, 2008; Coddington, 1989).

This fragment has six repeats of the GG(X) motif which is known to contribute to extensibility (Hayashi and Lewis, 2001; dos Santos-Pinto *et al.*, 2018). There are also GGX(n) and GPG motifs, typical of flagelliform silks. However, the glycine rich GPGXX motif is not found, which usually contributes β-turns and allows the silk to be spring-like (Li *et al.*, 2017). The lack of GPGXX motif and conserved terminal regions makes it impossible to determine whether this is a true flagelliform silk protein. However, if

this could be sequenced in future studies, it could be used to track the evolutionary loss or conservation of this silk gene as spiders evolved.

2.4.4 Limited Influence of External pH Environment on Silk Assembly

Spiders were collected inside caves that exhibited either neutral pH or in pH of down to 1.5. While this same area is subject to spatial and temporal variations in pH, so the same spider may experience both conditions during the year, the approach used in this study, RNA transcriptomics, can hope to uncover any differences in protein use at the time which the spiders were collected. This means that while it is expected that the gene set in the same species to be the same, the proteins transcribed to support survival in response to a particular environment may vary.

Analysis of the CO1 gene, one of the mitochondrial regions commonly used to distinguish between species (Hedin and Maddison, 2001), extracted from the two *L. barbara* transcriptomes suggests that the two populations belonged to the same species, as they aligned with 100% identity (Figure 2.11). No motifs or specific regions of the silk RNAs examined differed between the two populations and often the same silk gene was found in both transcriptomes. This data, and analysis by eye, suggests that the silk used by these spiders is the same, irrespective of their environments. For the main bulk of analysis these two transcriptomes were treated as one.

By examining silk sequences retrieved from transcriptomes of spiders that were found in very different pH environments (1.5-neutral), no extreme differences between their amino acid make up were found. This suggests that silk assembly is unlikely to be highly specialised to acid environments and is already adapted to function in a wide range of pH habitats. Low pH interferes with residue charge and at lower pH many aspartate, glutamate and histidine (polar, charged) residues become protonated, and their charges change to neutral, neutral and positive respectively. Therefore, most proteins adapted for use in low pH tend to have mainly acidic (negatively charged at a neutral pH) amino acids on the surface (Baker-Austin and Dopson, 2007; Reed *et al.*, 2013). As a higher proportion of glutamic and aspartic acid residues were identified in the pyriform, tubuliform and MaSp NTD of *L. barbara*, and the MaSp of *N. eremita*, it is proposed that these may be adaptations to the lower

pH, but as these are subtle changes, they may exist in more linyphiids or nesticids to equip their silks for use in a wide range of environments. To show whether these residues are indeed important, the proteins must be modelled to see where they reside in the protein structure, and also the sequenced silk repertoire of the linyphiid family should be expanded to allow further comparisons.

2.4.5 Optimisation of Silks in Resource Limited Conditions

The repetitive motifs that make up the major ampullate silks contained most of the expected repeating motifs. Only MaSp1 sequences were retrieved from the *Lessertia barbara* transcriptomes, which importantly do not contain proline in the repetitive region. This silk could be more resistant to water and pH changes, as in terrestrial silks proline acts to allow disorder and unfolding of the silk fibre between crystalline regions, which often leads to supercontraction in the presence of water and fibres with little proline contract very little, if at all (Liu *et al.*, 2008; Savage and Gosline, 2008; Pérez-Rigueiro *et al.*, 2019). In addition, the classical supercontraction motif (YGGLGS(N)QGAGR) is absent from all MaSp silks found, suggesting that the silks are not as susceptible to structural changes due to water addition than silks of araneids (Yang *et al.*, 2000).

Two unique pyriform spidroins were identified in the *L. barbara transcriptomes*, which have a higher glutamine content than proline. While high glutamine mirrors that found in other pyriform silks, the low proportion of proline is unexpected. Proline is thought to provide extensibility and flexibility to pyriform spidroins in other species (Chaw, Saski and Hayashi, 2017), and the lack of these may suggest that the fibres are less extensible than others. The tubuliform silks resemble that of other published sequences and are mainly comprised of alanine, serine and glutamine residues. The conserved regions are as expected, but the repetitive regions completely lack proline. The identified minor ampullate spidroin sequences also lack proline, but this is common in other published MiSps as they usually lack proline. These properties, and the lack of proline incorporation, may be especially important to allow survival and maintenance of a functioning MaSp silk when living in caves,

which are both low in pH, high in humidity and, one could expect, resource limited and scarce in prey.

Synthesising amino acids for use in silks is metabolically expensive, and some amino acids are more costly than others. For example, amino acids that are synthesised in a single step (for example alanine, aspartate) or from simple precursors (like serine, leucine and glycine) are produced at a lower energetic cost than those derived from multi-step reactions, or from citric acid-cycle metabolites (such as asparagine, methionine, glutamine and proline) (Dufton, 1997; Gutiérrez-Preciado, Romero and Peimbert, 2010). Studies have found that when dietary protein is limited, for example when food is scarce, spiders are able to alter their silk chemical composition to include more lower-cost amino acids. Some spiders, including orb and tent-weavers (such as *Argiope aetherea, Leucauge blanda* and *Cyrtophora moluccensis)* preferentially shift towards producing more MaSp1 in times of starvation, thereby reducing their proline and glutamine output (Blamires, Wu and Tso, 2012), however this effect is not universal (Blamires, Chao and Tso, 2010).

The lack of proline in the silk of the linyphiid *L. barbara* may also make it more able to travel aerially, as the silks could be more rigid and possibly more able to withstand wind forces, propelling the spider further than an elastic action would. Further investigation into the use of MaSp1 and MaSp2 silk in the linyphiid family would be useful to shed more light on this. It is also possible that the lack of proline found in these transcriptomes may simply be due to incomplete sequencing coverage, or degradation of the RNA before the libraries were constructed. As glutamine and proline containing MaSp2 silks were identified in the *N. eremita* transcriptome, this could suggest that the exclusive use of MaSp1 is more dependent on the spider family, rather than its environment.

As proline is hard to retrieve evolutionarily, if they truly don't produce MaSp2, this means that linyphiids may never be able to re-create silks with the properties of araneid proline-rich silk. It is interesting to postulate what might happen to the silks of linyphiids which arrive in 'new' extreme environments and evolve separately to other spider species. It could be proposed that a lack of proline in silk may limit linyphiids to weaker silk, limiting them to being small (yet extraordinary) creatures.

2.4.6 Future Studies

Due to the degradation of the RNA and improper sample storage before extraction, assessment of these transcriptomes revealed them to be of low quality, as expected. However, due to the difficulties in obtaining these particular specimens, analysis was continued. BUSCO scores were often less than half that seen in the literature, or achieved by the transcriptomes assembled later in this thesis (see Chapter 3). Therefore, it was necessary to treat all sequence fragments with caution, and so much further analysis and deeper gene sequencing was impossible. However, the sequences gene fragments do provide the first insight into the silks and their use by these small, cave dwelling spiders. Future studies should aim to obtain and use less degraded specimens for RNA extraction, which should drastically improve the length and quality of sequenced spidroin genes, and combine this with a more comprehensive population analysis to investigate the adaptation of silks used in different pH environments.

Only one potential flagelliform silk fragment was identified, which contained the repeating units GPG and PQGPGFGG, as well as the secondary structure prediction which suggests that it may be a flagelliform. If a true flagelliform silk is found in this species it could have implications on the development and evolution of silks throughout the Araneae family tree. However, as a conserved terminal domain was not identified in conjunction with this fragment, it is impossible to say that it is a flagelliform fragment for certain, but provides some hope that a future study may be able to amplify isolate and characterise these genes in a nesticid.

Further studies could attempt to search for silk genes in this, and in more improved datasets using an increased E-value during BLAST searches. This could allow for the discovery of more diverged silk types, which may be expected in the case of pyriform, for example, if it were to be more different in order to allow adherence in a more acidic environment than that encountered by most spiders.

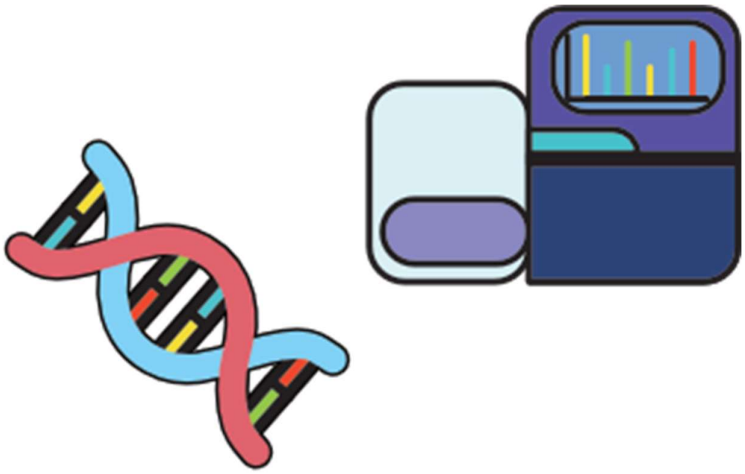
In the future, spidroin sequencing should be updated to reflect the advancements in technology, as newer technologies are able to cover long, repetitive and complex sequencing. Using nanopore long read or Single Molecule Real Time Sequencing

(SMRT) is looking more promising as a realistic time and cost-effective solution to uncovering full silk gene sequences from genetic material (See Chapter 3). By combining approaches, and using multiple platforms, it has been shown that the number and length of unique transcripts can be increased (Posnien *et al.*, 2014), which would be invaluable to further investigating these remarkable group of spiders.

Future studies should also investigate the physical adaptations of *N. eremita* silk and venom proteins for use in a low pH environment. Venoms could prove useful when investigating novel acid resistant toxins. Special attention should be taken to test the physical properties of silks in the presence and absence of acid, and mass spectrometry approaches used to determine exactly which silk types are used in the presence of low pH. This dataset provides a good starting point for both silk and venom protein analysis, which could help to guide future studies and highlight routes to acid resistance when creating nature-inspired recombinant biomaterials.

2.4.7 Conclusions

To date, there has been no other exploration into the silk genes of linyphiids or nesticids. The presence and composition of genes from the major silk types in two species living in the same environment is shown here, and the first sequenced silk genes from *Lessertia barbara* and *Nesticus eremita* are presented. The flexibility of a next generation sequencing approach is also demonstrated, to successfully amplify transcripts of interest even with sub-optimal sample quality. While there are some distinct differences in protein composition when compared with those of other spiders, these data indicate that the silks of these extreme pH cave-dwelling species most likely assemble and fold in the same way as more well-researched silks used as the basis for biomaterial formation and depend on a pH sensitive mechanism for correct folding. These findings are true for the spiders in this study that live both in and outside of extreme pH environments and could be vital in understanding the adaptation of silks through the spider family.



Chapter Three

Using RNA sequencing technologies to uncover novel and full-length silk genes

The Silken Code

3.1 Introduction

The repertoire of published silk genes is increasing rapidly with time and is proving invaluable in the rational design and creation of recombinant materials in the laboratory. By examining silks from a wider range than has been studied to date, including those from spiders than live in extreme environments, we can hope to expand the knowledge of these natural silks and the ways that they differ for each environment. Research into silk genes has in the past relied on sequencing methods that are time consuming, rely on out-of-date technologies or only yield short gene fragments. This study employs a combined approach of Single Molecule Real-Time (SMRT) sequencing, alongside Illumina sequencing, in order to attempt to identify full-length silk genes. This combination of approaches is quickly becoming a more cost and time effective method for *de novo* gene assembly. Full-length silk genes offer an obvious advantage to researchers trying to miniaturise spidroins for recombinant expression; by putting a repetitive fragment in context of a whole gene its importance can more easily be rationalised, and key motifs included in synthetic constructs.

3.1.1 From the Depths of Lakes to Desert Expanses

Spiders have been found in nearly every habitat and across all ecological niches; from the depths of lakes to desert expanses. The species examined here: *Argyroneta aquatica, Cyrtophora citricola, Pirata piraticus* and *Eratigena atrica,* are organisms which inhabit very different environments and create very different silken structures in order to do so.

The water spider, *Argyroneta aquatica* is the only spider that can live entirely underwater. It lives in marshes and wetlands, and has a wide distribution across Europe and Japan, see Figure 3.1 and Chapter 1 for more information (Seyyar and Demir, 2009; Seymour and Hetz, 2011; Nyffeler and Pusey, 2014). They vary from light brown to almost black in colour, depending on their distribution and prefer slow or non-moving water with vegetation which allows for the anchoring of a silken 'diving bell' (Mammola, Cavalcante and Isaia, 2016). This diving bell starts as a silk sheet, which is spun underwater and then filled with an air bubble, transported from

the surface using the surface tension created by thousands of tiny hydrophobic hairs that coat *A. aquatica*'s legs and abdomen.



Figure 3.1:* A. aquatica *and* C. citricola *example habitats
(A) shows an example marsh habitat of Argyroneta aquatica*, taken in Doksy, Czechia. The water is slow moving and densely covered in reeds and sphagnum moss, providing shade and protection for the spiders, which spend most of their time submerged. (B) shows an example dry and arid habitat of* Cyrtophora citricola*, taken in Rota, Spain. These spiders use cacti plants as anchoring points for their webs and can cover areas of up to several square metres. Photographs taken by the author.*

The diving bell is made up of a range of silk fibres that show no clear spatial ordination on the outside and range in thickness from <1 µM to greater than 15 µM. However, investigations carried out during this study have found that the orientation of the threads is more ordered on the inner surface of the bell, which suggests that thick threads are first used to construct a rough guide for the bell, and then thinner, more ordered structural threads are used to reinforce it. This structure is coated with a proteinaceous hydrogel, which makes it very hydrophilic, but this hydrogel dries and cracks as humidity is lowered. The identity of this hydrogel coating is as yet unknown. More information about the structure and makeup of the physical diving bell can be found in the introductory chapter to this thesis, Chapter 1.

Morphological studies *of A. aquatica* have found silk spigots and glands that indicate the presence of major ampullate (MaSp), minor ampullate (MiSp), pyriform (PySp), aciniform (AcSp) and tubuliform (TuSp) silks, but not flagelliform (FlaG) or aggregate (AgSp) (Strickland M, unpublished). Previous genetic studies have found evidence of major ampullate, pyriform, aciniform and tubuliform silks produced by *Argyroneta aquatica* (Strickland *et al.*, 2018; Correa-Garhwal *et al.*, 2019).

Early studies that tried to elucidate the identity of the silks that make up the diving bell used resin to cover up spinnerets and observed resulting web morphologies. This suggested that the diving bell is mainly produced by glands connected to the median and posterior spigots (minor ampullate and aciniform) while threads attached to the surrounding vegetation were mainly made up of major ampullate threads (Wasowska, 1977). Further proteomic analysis of the diving bell have identified AcSp, PySp and MaSp spidroins in this structure, along with other, unidentified elements (Correa-Garhwal *et al.*, 2019).

In contrast to the aquatic environment inhabited by *A. aquatica*, the colonial tent web building spider species, *Cyrtophora citricola*, has a wide distribution across tropical and subtropical areas of the world, and is native to the Mediterranean basin and also parts of Asia and Africa (Lubin, 1974). They are colonial species and can be found in aggregations of up to thousands of individuals (Mestre and Lubin, 2011). However, each spider maintains an individual territory by building and maintaining their own web within the colonial matrix, and they do not cooperate in prey capture or care of spiderlings (Yip *et al.*, 2016). Their typical environment is hot and arid, and these spiders cannot survive in temperatures below -1^{o}C (Blanke, 1972). An example habitat, from which the spiders in this study were taken, is composed of dense cacti in the Province of Cádiz, Spain, and can be seen in Figure 3.1.

Each individual *C. citricola* web is composed of a horizontal or slightly domed fine meshed sheet, with an irregular barrier web above and below. These barrier or supporting threads function as knock-down traps, with which insects collide and fall onto the sheet below, where the spider can grasp and subdue the prey (Lubin, 1974). This whole structure is non-sticky and is often referred to as a tent web. In a colony, these irregular webs are often interconnected using these supporting threads, to

create dense framework of webs. Examination of the silk glands of *C. citricola* suggest that they have vestiges of the triad spigots; they can no longer produce flagelliform or aggregate silk, and this is an evolved loss. They do, however, possess silk glands for major ampulate, minor ampullate, tubuliform, aciniform and pyriform silks, and it has been suggested that the webs are made primarily of major and minor ampullate threads (Peters, 1993).

This colony formation behaviour allows *C. citricola* to take advantage of wide, open spaces with an abundance of flying insects, and can provide increased protection for egg sacs (Lubin, 1974). This may have some implication on the strength of tubuliform silk, as it may not have to be as inherently protective to ensure egg survival. Unlike some orb weavers, these webs are semi-permanent, in that they are not recycled, eaten by the spider or rebuilt daily, which may also have implications for the amount of resources invested into these more permanent threads.

The small wolf spider *Pirata piraticus* lives in a wide variety of semi-aquatic habitats, such as pond or stream margins. It is widespread throughout Western and Central Europe and is an ambush predator; it does not rely on a web to catch prey. This spider instead uses its silk to create a silken tube-shaped retreat, in which it will moult and lay eggs (Locket and Millidge, 1951; NMBE, 2021). Some *P. piraticus* use silken trip lines around the burrow, presumably used to sense the approach of prey. Silk is also used to wrap the eggs, rolled into a cocoon shape and is then carried by the mother until the eggs hatch (Nørgaard, 1951). *P. piraticus* are very sensitive to low humidity and they prefer to hide and live amongst vegetation, such as sphagnum moss. They can run on the surface of water or dive under the surface when startled or to chase prey, remaining submerged for up to 24 hours due to the presence of hydrophobic hairs on their legs which can trap air (Nørgaard, 1951; Hénaut *et al.*, 2018).

Eratigena atrica, the house spider, is the representative terrestrial spider chosen for use in this study due to its ease of collection and love of dry indoor environments. It is found mainly in Europe, Central Asia and northern Africa and is usually nocturnal, found in dark corners in homes or outbuildings. It was probably originally a cave dweller or lived in dry forests (Rachwał, Rybak and Rogula-Kozłowska, 2018; World Spider Catalog, 2020). This species weaves funnel webs, which are flat sheets created

by irregularly arranged dense threads connected to a funnel-shaped tunnel, in which it waits to ambush its prey.

By examining in detail the silk of two very different spiders, *Argyroneta aquatica* and *Cyrtophora citricola*, as well as constructing transcriptomes of *Pirata piraticus* and *Eratigena atrica*, this study is able to produce a wealth of information concerning silk fibres that are used in a range of environments. It is also shown that these bioinformatic approaches are robust enough to be used for a range of species and tissues, as well as uncover novel spidroin sequences to allow for direct comparisons between environments and evolutionary timescales.

3.1.2 Next Generation Sequencing (NGS)

The main difficulty in the study of spidroins is due to their unique organisations; they are both very long, as their genes often exceed 10kb, and are composed of several repeating blocks. This means they are expected to be difficult to *de novo* assemble, and genes cannot be reconstructed using a reference because these are often not available (Treangen and Salzberg, 2012; Liao *et al.*, 2019). However, the author has successfully used short read *de novo* assembly approaches to identify novel spidroin sequence fragments in different spider species, even with sub-optimal RNA quality (see Chapter 2).

Short read approaches are low cost and highly accurate, useful for mass gene identification and producing high quality deep coverage of small genomes. They are also supported by a wide range of analysis tools and pipelines. However, this is susceptible to preferential amplification of repetitive DNA, and often the overlap between sequences is too short to enable the efficient *de novo* assembly of genomes or transcriptomes. During Next Generation Sequencing (NGS), genetic material is broken down, amplified and then sequenced. Long-read sequencing is often called third generation sequencing, because it moves away from this approach by directly sequencing single DNA molecules. While short read sequencers typically produce reads of up to 600 bases, long-read technologies can generate reads in excess of 10kb (Amarasinghe *et al.*, 2020). The differences between these next generation sequencing technologies and approaches are expanded upon in Chapter 1.

This investigation expands upon the knowledge of transcriptome construction and interrogation using short read data and assemble the transcriptomes of four different spider species, using significantly higher quality starting material than was used in Chapter 2. A relatively new technology, Single Molecule Real Time (SMRT) sequencing, is used to uncover, for the first time, full-length silk genes in *Argyroneta aquatica* and *Cyrtophora citricola*.

This study employs both Illumina and SMRT sequencing technologies to probe the genetic make-up of spider silk and create sequence resources for future analysis of silk proteins.

Firstly, using reads derived from Illumina short read technology, transcriptomes are *de novo* assembled, representing three different spider species: *A. aquatica, P. piraticus* and *E. atrica*. The assembled transcriptomes represent a total of over 290 million high quality sequenced base pairs which include over 200 partial silk gene sequences.

The aim of these transcriptomic assemblies was to unearth novel proteins and investigate the genetic and proteomic makeup of silk transcripts in the sphere of known terrestrial spidroins. Specimens of *A. aquatica* and *P. piraticus* were chosen as representatives of spider species that either live exclusively in, or predominantly around water. *E. atrica* was chosen as a non-water dwelling comparison species due to its abundance and ease of collection. The transcriptome assembly scores for all assembled transcriptomes have been included to demonstrate the variation in assembly quality and completeness depending on the size and type of tissue used. It was found that the amount of starting tissue, for example extracted chelicerae and associated glands or whole abdomen, does not have a dramatic effect on transcriptome quality.

Secondly, long-read sequencing technology was used to investigate the silk genes of *Argyroneta aquatica* and a colonial tent-web spider, *Cyrtophora citricola*, in more detail. Long-read sequencing represents a huge leap forward in sequencing technology, allowing the relatively easy identification of full-length genes or transcripts as a result of not having assemble many shorter, repetitive fragments. *C.*

citricola was chosen as a comparison species to highlight differences between silks used in very dry (*C. citricola*) and very wet (*A. aquatica*) environments.

The information learned here about aquatic silk sequences will be used to design and suggest improvements to an existing recombinant spidroin system, in the hopes of recapitulating the properties of silk that functions underwater in a recombinant protein. These transcriptomic assemblies may also prove useful for future studies which aim to more completely characterise the transcriptome of the species examined. Also, different sequencing methods are compared and contrasted to provide a signpost for future studies as to which technology is more appropriate. This study proposes that long-read sequencing is an essential tool necessary for characterising the full-length silk gene, but short reads are also valuable for smaller studies that wish to identify key features in non-repetitive spidroin terminal regions. Finally, the first published full-length silk genes of *Argyroneta aquatica* and *Cyrtophora citricola* are presented.

3.2 Materials and methods

3.2.1 Short Read Sequencing

3.2.1.1 RNA Extraction

Specimens were collected from Nottingham, UK (*Eratigena atrica, Pirata piraticus*) and Doksy, Czech Republic (*Argyroneta aquatica*) and housed in the lab before RNA extraction in 2019. Two adult female specimens of each species were euthanised in ethanol and the abdomen separated from the cephalothorax before immersion in liquid nitrogen. At this point, all tissue samples were disrupted using a pestle and mortar and RNA extracted using the GeneJET RNA purification kit (ThermoFisher) in under 45 minutes. RNA was eluted in 100 µL RNase free water and the concentration recorded using a spectrophotometer (Nanodrop 2000), before storage on ice for QC and library preparation.

3.2.1.2 Sequencing

Whole abdomen RNA sequencing was carried out by DeepSeq (University of Nottingham). RNA concentrations were measured using the Qubit Fluorometer and the Qubit RNA BR Assay Kit (ThermoFisher Scientific; Q10211) and RNA integrity assessed using the Agilent TapeStation 4200 and the Agilent RNA ScreenTape Assay Kit (Agilent; 5067-5576 and 5067-5577). mRNA was purified from 1 mg of total RNA using the NEBNext Poly(A) mRNA Magnetic Isolation Module (E7490). Indexed sequencing libraries were then prepared using the NEBNext Ultra Directional RNA Library Preparation Kit for Illumina (NEB; E7420) and NEBNext Multiplex Oligos for Illumina, Index Primers Set 1 (NEBNext; E7335). Libraries were quantified using the Qubit Fluorometer and the Qubit dsDNA HS Kit (ThermoFisher Scientific; Q32854). Library fragment-length distributions were analysed using the Agilent TapeStation 4200 and the Agilent High Sensitivity D1000 ScreenTape Assay (Agilent; 5067-5584 and 5067-5585). Libraries were pooled in equimolar amounts and final library quantification performed using the KAPA Library Quantification Kit for Illumina (Roche; KK4824). The library pool was sequenced on the Illumina NextSeq 500 on a NextSeq 500 Mid Output Kit v2.5 300 cycle kit (Illumina; 20024905), to generate over 20 million pairs of 150-bp paired-end reads per sample.

3.2.1.3 Transcriptome Assembly and Quality Control

Barcodes and low-quality sequences were removed using Trimmomatic v.0.36.6 with the following settings: Leading: 20, Trailing: 20, Sliding window: 4:20 and a minimum length of 36 (Bolger, Lohse and Usadel, 2014), carried out by DeepSeq. FastQC v.0.11.8 was used to manually check the quality of the trimmed sequences (Andrews, 2010). Trimmed and filtered paired end sequencing data was then assembled into contigs using Trinity v.2.2.0 (Grabherr *et al.*, 2011). The quality and completeness of each transcriptome was assessed using QUAST v4 (Gurevich *et al.*, 2013) and BUSCO v.4 (Simão *et al.*, 2015)

3.2.1.4 Identification of Novel Spidroins

Transdecoder v.1.0.3.0 was used to identify coding regions within the transcript sequences and translate them into their corresponding amino acids (Haas *et al.*, 2013).

The output files were then used in a Basic Local Alignment Search Tool (BLAST) search against a custom database of spidroin DNA or amino acid sequences (created using all spider silk protein sequences available on the NCBI protein database as of 26-Jan-21, with hits manually curated to certify these were correctly designated as spidroins) to identify potential novel silks, with an E-value cut off of $1e^{-5}$ to $1e^{-4}$ (Altschul *et al.*, 1990). These were then BLAST searched again against the NCBI non redundant protein database to confirm identity as a silk sequence, by comparison with the BLAST hit that returned the smallest E-value and by-eye identification of signature silk-type repetitive motifs, an approach used by other groups for spidroin identification. (Sangaard *et al.,* 2014; Prodoscomi *et al.,* 2011; Babb *et al.,* 2017; Collin *et al.,* 2018; Na *et al.,* 2019; Garb *et al.,* 2019).

Spidroin sequences identified by BLASTp and BLASTn were extracted and alignments and sequence analysis were carried out using Geneious or Clustal Omega v1.2.2 (Geneious, 2018; Madeira *et al.*, 2019). In some cases, alignment images were prepared using ESPript v3.0 (Robert and Gouet, 2014). Published spidroin sequences were retrieved from GenBank and used to help identify terminal regions and confirm classification of silk type. SignalP was used to identify secretory signal peptides in protein sequences, as signal peptides have been shown to be a key element of in the

transport of spider silk transport across the endoplasmic reticulum during secretion (Wen, Liu and Meng, 2017). N- and C-terminal regions were defined as stated in Chapter 2.

3.2.2 Long Read Sequencing

3.2.2.1 RNA Extraction

Specimens were collected from one of two sites in the UK: Thursley nature reserve or the Potteric Carr nature reserve (*Argyroneta aquatica*) or from the Province of Cádiz, Andalusia, in Spain (*Cyrtophora citricola)*. Individuals were housed in the laboratory before RNA extracted in 2020. One adult female *Argyroneta aquatica* and one subadult female *Cyrtophora citricola* was euthanised in ethanol and the abdomen separated from the cephalothorax before immersion in liquid nitrogen. RNA was then extracted from the abdomens using the GeneJet RNA extraction kit in the manner described previously. RNA was eluted in 100 µL RNase free water and quality assessed using the Nanodrop before samples were shipped on dry ice for sequencing.

3.2.2.2 Sequencing

All library preparation and sequencing was carried out by Macrogen. RNA libraries were prepared using the SMRTbell® Express Template Prep Kit 2.0 and templates sequenced using the Sequel® Binding Kit 3.0 and Sequel® Sequencing Plate 3.0 using one SMRT cell. Sequencing and read processing was performed on the PacBio Sequel II. First, raw polymerase reads, sequences of nucleotides incorporated by the polymerase while reading a template, which reflect the movie length and include high-quality regions and adapter sequences, were generated. Each polymerase read was partitioned to form one or more subreads, which contain a sequence derived from a single pass of a polymerase on a single strand of insert within a SMRTbell template, with no adapter sequences. Then, Circular Consensus Sequence (CCS) reads were determined using subreads taken from a single ZMW, which requires at least two full pass subreads from the insert.

3.2.2.3 Sequence QC and Assembly

A combination of web-based and locally installed tools was used for sample QC and assembly. RNASpades (Bushmanova *et al.*, 2019), Geneious Assembler (Geneious, 2018) and Canu (Koren *et al.*, 2017) were used with CCS reads to create several different assemblies, which were then assessed using QUAST (Gurevich *et al.*, 2013), Fasta statistics (Seemann and Gladman, 2012) and BUSCO v5 (Simão *et al.*, 2015). Illumina data were included to see if quality was increased when using RNASpades and Geneious Assembler. Short read data could not be used for Canu, as it can homogenise repeats and mix up haplotypes. The Galaxy server was used for some assembly and analysis tools (Afgan *et al.*, 2016). Spidroins were identified from CCS or assembly data using the methods described in section 3.2.1.4.

3.2.2.4 High Quality Transcript Classification and Annotation

The IsoSeq3 pipeline was used to classify CCS reads into low and high quality transcripts, which were then clustered into non-redundant transcripts using CD-HIT-EST (Fu *et al.*, 2012). Transcripts were then functionally annotated using seven different public databases; GO, UniProt, NCBI NR, Pfam, EggNOG, NT and KEGG. These steps were carried out by Macrogen.

3.2.2.5 Sequence comparisons using phylogenetics

Sequence alignments were made using CLUSTAL Omega v1.2.2 (Madeira *et al.,* 2019). Trees were constructed to highlight similarity between sequences rather than their relatedness in evolutionary time. To this end, trees were created with the Geneious tree builder (Geneious Prime v2021.0.3) using the Neighbour-Joining method without rooting in order to explore grouping of sequences rather than estimating phylogenetic relationships. Trees were also created using a Bayesian inference method in MrBayes v3.2.6, through the MrBayes Geneious plugin (Huelsenbeck and Ronquist, 2001), with a 500,000 chain length and 10,000 burn-in length. Two methods of tree construction were used to show that relationships between sequences were maintained regardless of algorithm used and to avoid artefact incorporation during interpretation. These trees used all available MiSp/MaSp sequences from the RTA clade, a clade of araneomorph spiders which are united by the possession of a Retrolateral Tibial Apophysis (RTA) (a projection on the tibia of

male pedipalp), a sister group to orb-weaving spider which contains half of all spider diversity and of which *A. aquatica* is a member (Coddington and Levi, 1991; Blackledge *et al.*, 2009). The sequences used are shown in Appendix A.3.1.

3.3 Results

3.3.1 Short Read Sequencing

3.3.1.1 RNA Extraction

For Illumina sequencing, RNA was extracted from the samples and the total amount of RNA and its RNA Integrity Number (RIN), which indicates the quality of the RNA on a scale of 1-10, recorded in Table 3.A.

Table 3.A: Summary of total RNA extracted and RIN scores.

Sample	*A. aquatica*	*P. piraticus*	*E. atrica*
Tissue type	Abdomen	Abdomen	Abdomen
RNA (µg)	12.36	27.06	54.15
RIN	9.8	8.8	9.5

3.3.1.2 Transcriptome Assembly and Quality Control

Sequences assembles using Trinity were assessed using QUAST (Table 3.B). The transcriptomes of *A. aquatica*, *P. piraticus* and *E. atrica* abdomens included complete homologs to >92 % of the core eukaryotic genes and arthropod benchmarking genes (BUSCO) (Figure 3.2). This is well within the limits of other published gene sets; generated through both genomic and transcriptomic sequencing. A summary of all scores is shown in Figure 3.3.

Table 3.B: Assembly metrics of short read transcriptomes using QUAST.

The N50 is the minimum contig length needed to cover 50% of the assembly, and the N75 the same but to cover 75%. The L50/75 is the number of contigs of at least length = N50/75.

	A. aquatica	*P. piraticus*	*E. atrica*
# contigs	34803	39346	36071
# contigs (>= 0 bp)	79087	100504	98969
# contigs (>= 1000 bp)	19972	19746	17480
Largest contig	26257	15721	20003
Total length (# bases in assembly)	59648882	54126902	50903089
Total length (>= 0 bp)	73228562	72847519	70143365
Total length (>= 1000 bp)	49338274	40547425	38047738
N50	2385	1739	1822
N75	1292	997	993
L50	7497	9484	8133
L75	15925	19794	17611
GC (%)	34.51	33.93	34.21

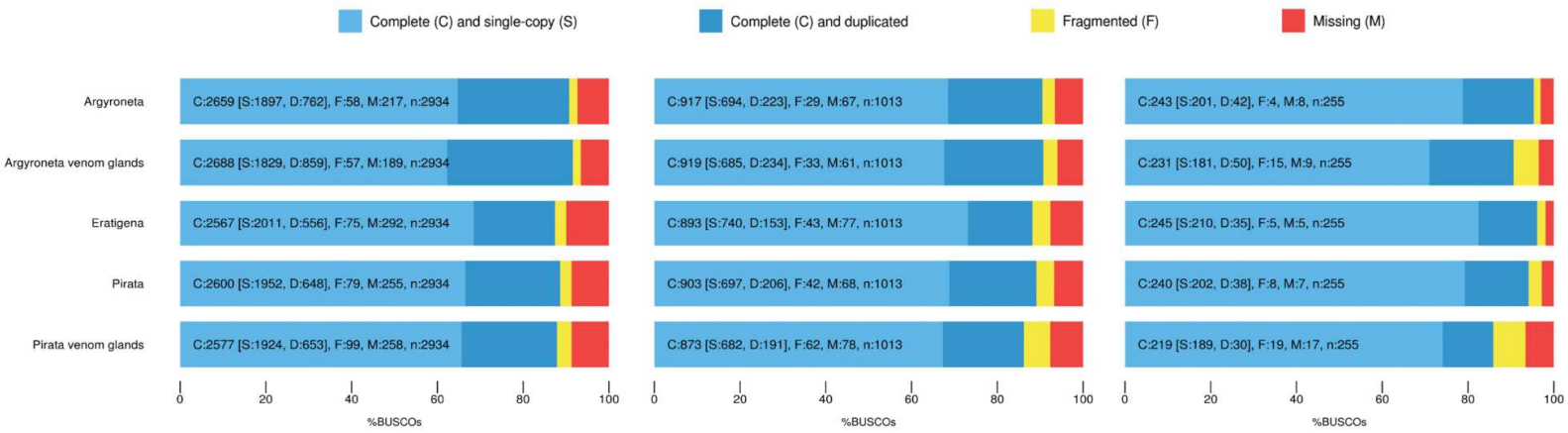


Figure 3.2: BUSCO Assessment scores for short read transcriptomes.
Assessed against the Arachnida (left), Arthropoda (middle) and Eukaryota (right) BUSCO (v10) datasets.

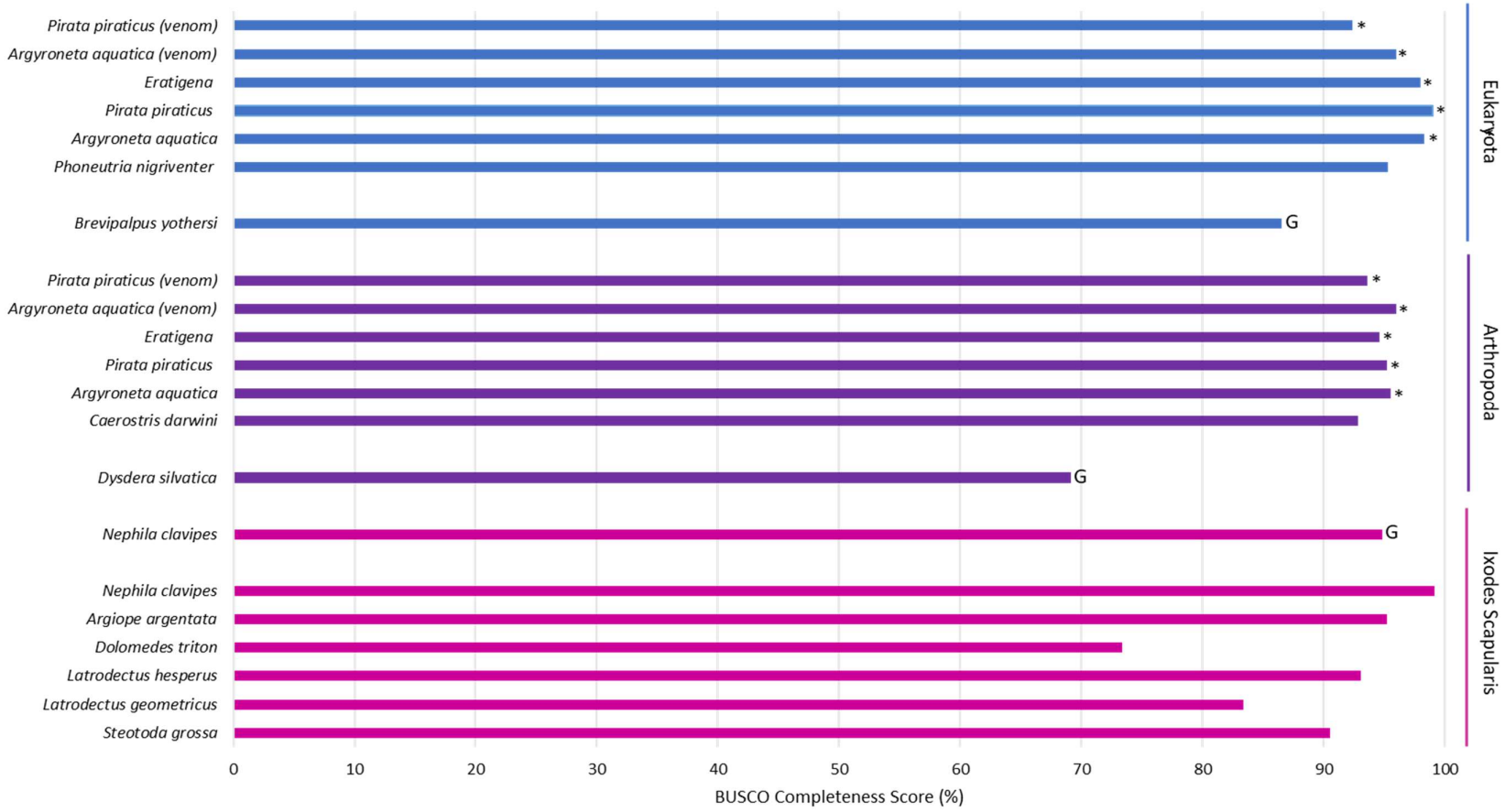


Figure 3.3: A comparison of published BUSCO scores for a range of spider gene assemblies.

Bars marked with 'G' represent a BUSCO score for a genome, while scores for the transcriptomes generated for this study are designated with a (*). A variety of BUSCO sets can be used to assess transcriptome completeness, which are designated by the various colours; blue (Eukaryota), purple (Arthropoda), pink (the deer tick, Ixodes scapularis*).*

3.3.2 Long Read Sequencing

3.3.2.1 RNA Extraction

The total amount of extracted RNA for long-read sequencing and its purity was assessed using a spectrophotometer (Table 3.C).

Table 3.C: Extracted RNA concentrations and quality scores.

Sample	ng/µL	260/280	260/230	260
A. aquatica	297.12	2.16	2.42	7.42
C. citricola	219.96	2.13	2.157	5.51

3.3.2.2 Sequencing

The polymerase read and subread statistics generated by Macrogen are shown in Table 3.D, Figure 3.4 and Figure 3.5.

Table 3.D: Polymerase Read and Subread Statistics.

	Total number of bases	Number of reads	N50*	Average read length
Polymerase Reads				
A. aquatica	37,470,358,607	537,505	132,849	69,711
C. citricola	33,758,637,377	550,354	115,805,	61,339
Subreads				
A. aquatica	36,660,762,854	18,816,821	2,133	1,948
C. citricola	33,004,391,375	17,375,015	2,208	1,899

**The N50 indicates that 50% of all bases come from reads longer than this value.*

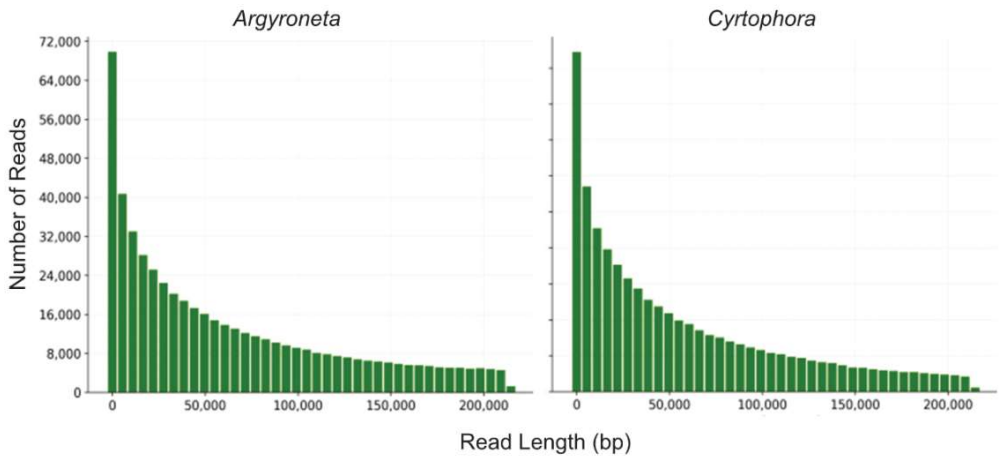


Figure 3.4: SMRT sequencing read length distribution.

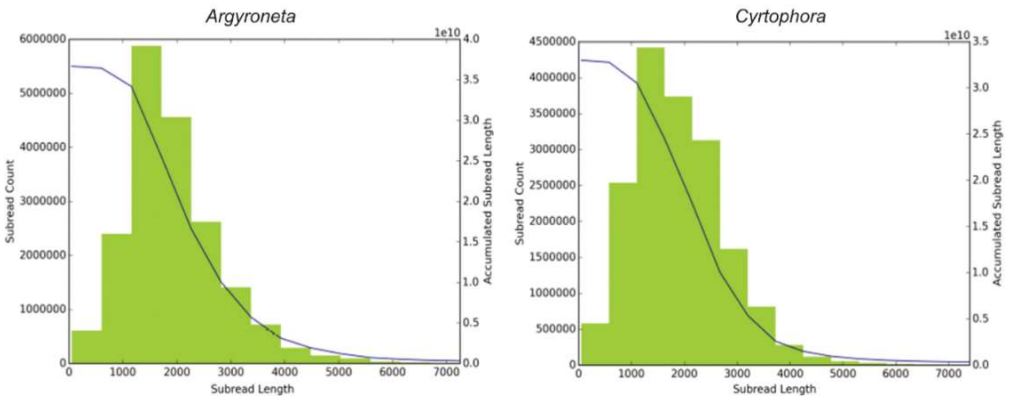


Figure 3.5: SMRT Sequencing subread length distribution.

3.3.2.3 CCS Read Statistics

CCS read statistics for both samples are shown in Table 3.E and the read length distributions in Figure 3.6. The quality of a CCS read is indicated by a Phred quality score. A Q score of 10 (Q10) indicates that there is a 1 in 10 chance of the base pair being misidentified (90% accuracy), Q20 is a 1 in 100 chance (99% accuracy), Q30 is 99.9% accuracy and Q40 indicates a 99.99% accuracy rate. The average quality scores of the CCS reads are also indicated in Table 3.E.

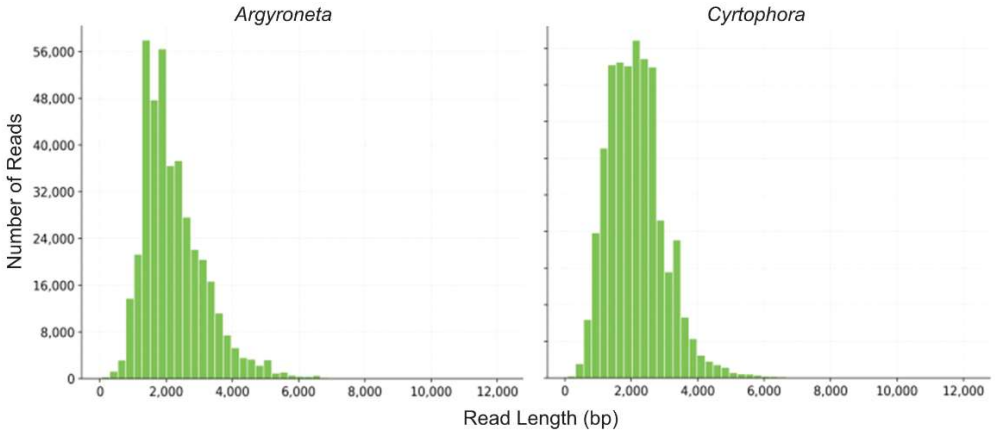


Figure 3.6: CCS read length distributions.

Table 3.E: CCS read statistics.

Sample	CCS Read bases	CCS reads	Average read length	Average quality	Minimum sequence length	Maximum sequence length	GC%
A. aquatica	896,522,253	401,869	2,230	Q39	46	12,291	40.7
C. citricola	862,531,371	400,841	2,151	Q38	75	12,280	40.3

3.3.2.4 Protein Coding Sequence Prediction

Transdecoder was used to predict likely open reading frames in the CCS read files, and the number of predicted proteins is shown in Table 3.F.

Table 3.F: Predicted protein coding sequences in CCS reads.

	Number of predicted protein coding sequences	Maximum sequence length
A. aquatica	394,546	3,252
C. citricola	476,848	3,773

3.3.2.5 High Quality Transcript Classification and Annotation

Transcripts were classified and returned 21,139 non-redundant, high-quality transcripts for *A. aquatica* and 21,871 for *C. citricola*. Several publicly available databases were then used to functionally annotate each transcript, creating robust annotated databases available for future use. The amount of successfully annotated transcripts varied between databases, which is as expected as each database will have a different proportion of genes shared with the species studied here, and no database is constructed on spider genes, so low scores are expected. The annotated Unigene ratio for each sample using each database is shown in Figure 3.7.

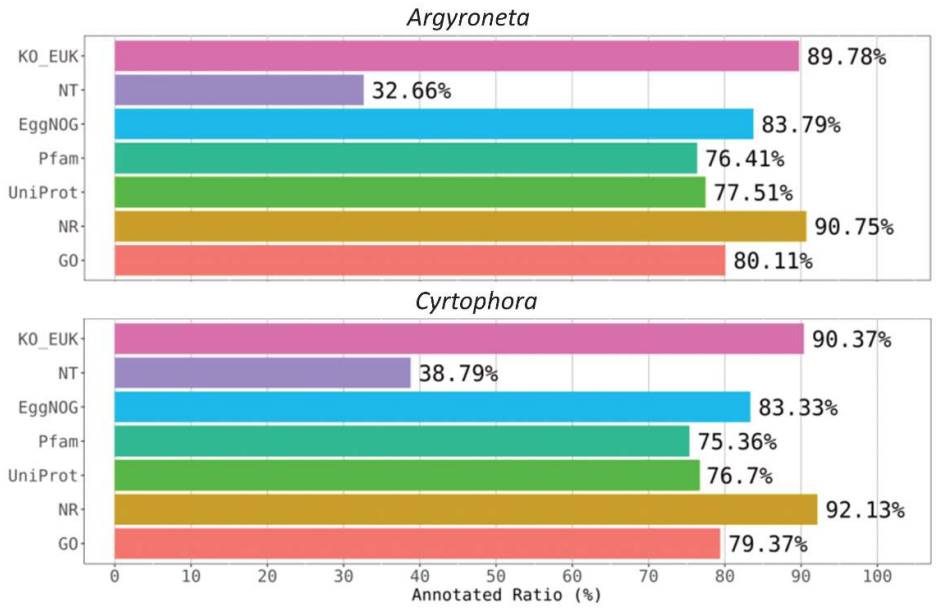


Figure 3.7: Annotated Unigene Ratios.
The percentage of transcripts annotated by each database is indicated for both samples. Databases used are: KEGG Orthology Eukaryotic (KO_EUK), NCBI Nucleotide (NT), EggNOG 4.0 (EggNOG), EMBL_EBI Pfam (Pfam), Uniprot, NCBI protein (NR) and the Gene ontology resource (GO).

3.3.3 Assembly and Quality Control

Several tools were used to assemble CCS reads in the hope of assembling full-length spidroin sequences that exceeded the maximum CCS read length. Subreads were not assembled due to restrictions on computing resources. The descriptive statistics of the CCS reads and best performing assemblies for *A. aquatica* and *C. citricola* are shown in Table 3.G. Figure 3.8 shows the BUSCO scores for each sample of both the

CCS read files and the assembly for each sample that returned the most convincing spidroin sequences upon manual comparison with original CCS reads, as well as the highest BUSCO coverage score. Figure 3.9 shows an example of the variation in BUSCO scores for the different assemblies attempted for the *C. citricola* dataset.

Table 3.G: QUAST and Fasta statistic assessment of the CCS read datasets and the best performing assembly.

		A. aquatica		*C. citricola*	
		CCS reads only (unassembled)	Geneious Assembler	CCS reads only (unassembled)	Geneious Assembler
QUAST	# contigs (>= 0 bp)	401869	53416	400841	52788
	# contigs (>= 1000 bp)	385258	52400	376421	50695
	Total length (>= 0 bp)	896522253	155359863	862531371	141045734
	Total length (>= 1000 bp)	883109534	154517965	842860445	139313069
	# contigs	400688	53372	399525	52729
	Largest contig	12291	18118	12280	17605
	Total length	896084865	155342227	862002278	141020081
	GC (%)	40.66	37.69	40.25	37.86
	N50	2430	3218	2372	2983
	N75	1818	2400	1849	2224
	L50	135856	17721	140634	17252
	L75	242888	31678	241897	30919
	# N's per 100 kbp	0	659.57	0	598.71
Fasta Statistics	GC content	40.7	37.7	40.3	37.9
	Len N50	2430	3217	2371	2983
	Len max	12291	18818	12280	17605
	Len mean	2230	2908	2151	2671
	Len median	2003	2658	2099	2449
	Len min	46	74	75	291
	Num_bp	896522253	37388428	862531371	141045734
	Num bp not N	896522253	37372931	862531371	140200981
	Num_Seq	401869	31979	400841	52788

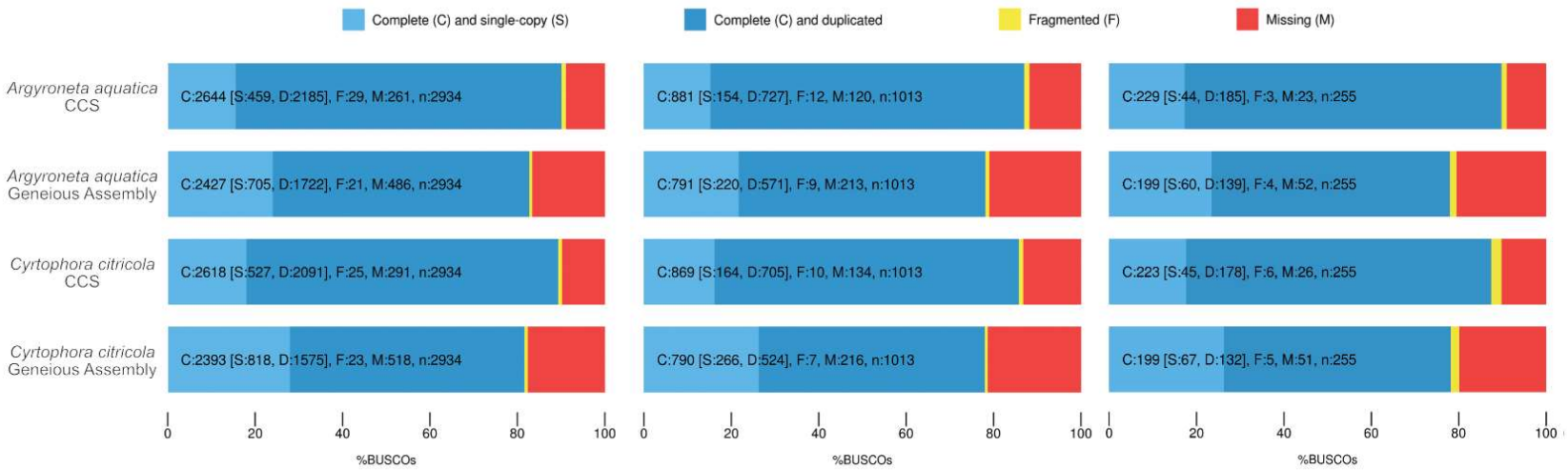


Figure 3.8: BUSCO assessment scores for long-read assemblies and CCS reads.
Assessed against the Arachnida (left), Arthropoda (middle) and Eukaryota (right) BUSCO (v10) datasets.

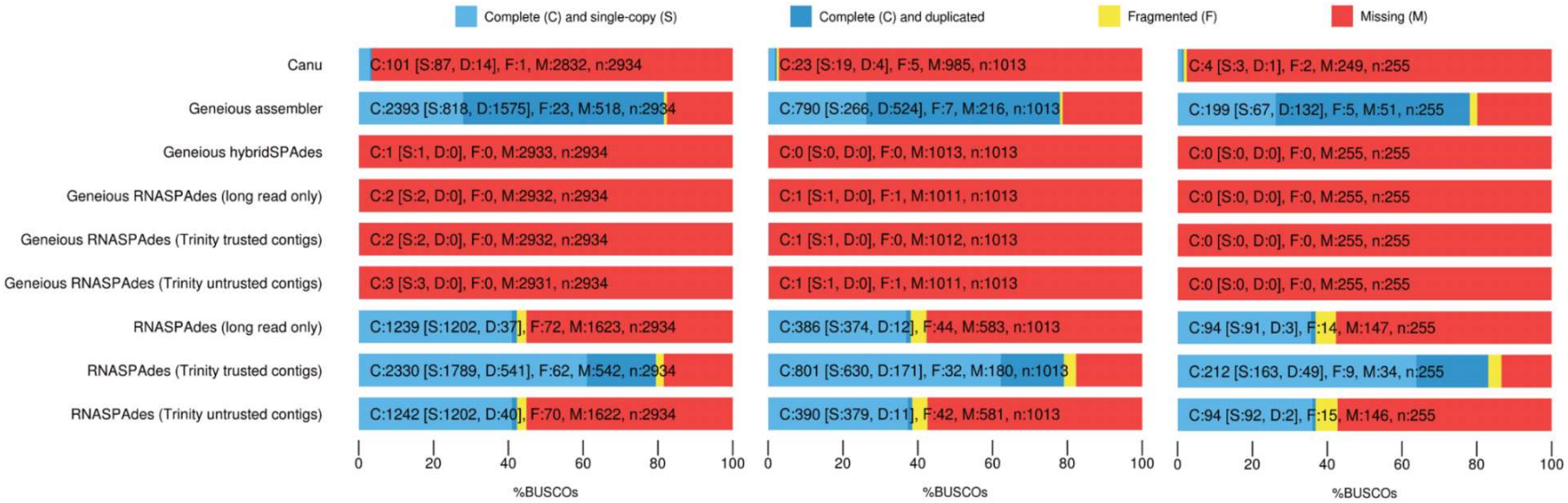


Figure 3.9: BUSCO assessment scores for a selection of assembly methods using the* C. citricola *CCS read dataset.
Assessed against the Arachnida (left), Arthropoda (middle) and Eukaryota (right) BUSCO (v10) datasets.

3.3.4 Comparison of NGS Methods Used in This Study

Several parameters were used to compare the data return and cost of the sequencing methods used in this study, which is shown in Table 3.H. Error rates and instrument specifications were taken from the literature (Fox and Reid-Bayliss, 2014; Rhoads and Au, 2015; PacBio, 2020; Illumina, 2021).

Table 3.H: Comparison of NGS methods used in this study.

	Illumina NextSeq (Mid output kit)	Illumina NovaSeq (Flow cell s4)	PacBio Sequel
Output range	20-120 Gb	65-6000 Gb	Up to 20 Gb
Run time	26 hours	36 hours	20 hours
Reads per run	130 - 400 million	Up to 20 billion	500,000
Max read length	2 x 150 bp	2 x 101 bp	> 20 kb
Samples per run	2 - 8	12 - 400	1
Price per sample	£604	£220	£2,100
Year Sequenced	2018	2019	2020
Error rate	0.1-1 %	0.1-1 %	10-15 %

3.3.5 Identification of Novel Spidroins

A selection of BLAST tools (BLASTn, BLASTp, BLASTx, tBLASTx) were used to identify potential spidroin sequences as described in section 3.2.1.4, and the number of sequences found shown in Table 3.I. These sequences are explored in further detail in the following sections. Spidroin fragments from these assemblies were used to elucidate the identity of certain amino acids more accurately or used to map to reference when the quality of long-read data was called into question. Sequences from *P. piraticus* and *E. atrica* are not examined here in any more detail as these were used as indicators of transcriptome quality variation from different spiders, and more focus is placed on the sequences of *A. aquatica* and *C. citricola.*

Table 3.I: Number of identified spidroin sequence fragments from whole-abdomen short read transcriptomes.

		Argyroneta aquatica	*Pirata piraticus*	*Eratigena atrica*
Major Ampullate	# Sequences	7	16	9
	C-terminal domain sequences	2	8	2
	N-terminal domain sequences	2	5	5
	Non-Isoforms	5	11	8
Minor Ampullate	# Sequences	0	0	1
	C-terminal domain sequences	0	0	0
	N-terminal domain sequences	0	0	0
	Non-Isoforms	0	0	1
Aciniform	# Sequences	8	8	6
	C-terminal domain sequences	0	1	0
	N-terminal domain sequences	2	1	2
	Non-Isoforms	7	7	5
Aggregate	# Sequences	0	0	0
	C-terminal domain sequences	0	0	0
	N-terminal domain sequences	0	0	0
	Non-Isoforms	0	0	0
Tubuliform	# Sequences	2	0	18
	C-terminal domain sequences	1	0	3
	N-terminal domain sequences	1	0	4
	Non-Isoforms	2	0	6
Flagelliform	# Sequences	0	0	0
	C-terminal domain sequences	0	0	0
	N-terminal domain sequences	0	0	0
	Non-Isoforms	0	0	0
Pyriform	# Sequences	4	14	0
	C-terminal domain sequences	1	0	0
	N-terminal domain sequences	0	1	0
	Non-Isoforms	2	3	0

Spidroins from the long-read dataset were identified and annotated according to type. The number of putative full length spidroins found for each silk type is indicated in Table 3.J. This table shows which silk types were identified in the long-read datasets, as well as how many long-read isoforms, if any, were identified. Table 3.K shows the CCS and CDS lengths of the complete silks found in each sample.

Table 3.J: Full-length Spidroin Identification.
Unexpected results are underlined.

	Argyroneta aquatica		*Cyrtophora citricola*	
Silk Type	*Spidroin type identified?*	*Number of full-length sequences*	*Spidroin type identified?*	*Number of full-length sequences*
Major Ampullate	Y	3	Y	1
Minor Ampullate	N	0	Y	0
Aciniform	Y	0	Y	0
Aggregate	N	0	N	0
Tubuliform	Y	3	N	0
Flagelliform	N	0	N	0
Pyriform	Y	1	Y	0

Table 3.K: Full-length Silk Sequence Lengths.

	Silk Type	CCS read length (bp)	Predicted CDS length (aa)
Argyroneta aquatica	MaSp	4818	1606
	MaSp	4804	1600
	MaSp	4081	1360
	TuSp	9048	1532
	TuSp	9087	2962
	TuSp	9040	2790
	PySp	3469	1115
C. citricola	MaSp	5755	1853

The following analysis focuses only on sequences identified in the CCS data files, and it is possible that the assembly of the CCS reads would yield more full-length sequences. Repeats were predicted, in cases where repeats units were complex, using RADAR, which identifies approximate repeats, as well as complex repeat architectures involving many different types of repeats (Heger and Holm, 2000). In the following analysis, major ampullate silk receives more focus, as it is the most well characterised silk in the literature, and is often viewed as the most promising to recapitulate as a synthetic biomaterial. The other silk types identified are described in brief, in relation to their probable importance in the construction of the diving bell; and the respective *C. citricola* silks annotated to provide a comparison point between silks in another environment.

3.3.6 *A. aquatica Major Ampullate Silk*

From the short read transcriptomes, none unique MaSp fragments were identified, containing four C- and four N-terminal domains. Only repetitive region fragments connected to a terminal region were used for analysis. In the long-read dataset, only protein sequences over 1000 amino acids arising from the BLAST search were used for further analysis; which left 59 putative MaSp sequences, including two C-terminal types and six N-terminal region types.

3.3.6.1 MaSp Terminal Regions

Terminal regions were defined using the general criteria set out in Chapter 2. Two different MaSp N-Terminal domains (NTD) were identified in our datasets, Type A and Type B (Figure 3.10). Both contain a predicted five helical secondary structure and a similar percentage of acidic and polar uncharged residues, while NT type A (which starts with a PWXN motif) is slightly more hydrophobic (52% compared to 44.8% in type B). Type A was also found to be associated with a region of potential frameshift (repeat Type 2, see Table 3.L).

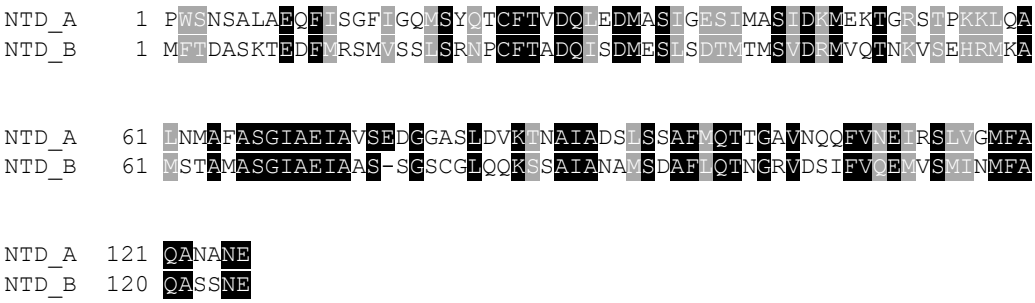


Figure 3.10: **A. aquatica *MaSp N-terminal domain types.***
NTD Type A (NTD_A) and NTD Type B (NTD_B) alignment showing similarities and differences between sequences found in MaSp silks.

Two different MaSp C-Terminal Domains (CTD) were also identified, which were represented in both datasets. The start of this region is defined using the nearest arginine (R) residue approximately 45 residues upstream from the 'Q' in the conserved QALLEVITAL-type motif. Both CTD types contain an arginine (R) residue approximately 32 residues upstream of the Q from the QALLE motif, and CTD type 2 contains an aspartic acid (D) 28 residues upstream, which don't appear in any other published CTD to date.

In the *A. aquatica* MaSp CTD, the QALLE motif takes the following form: QVLME(I/V)(I/A)SALVH. The two underlined residues are most commonly A/L or V/M in other species' C-terminal domains. Both CTD types contain similar proportions of amino acids and similar structures (Figure 3.11).

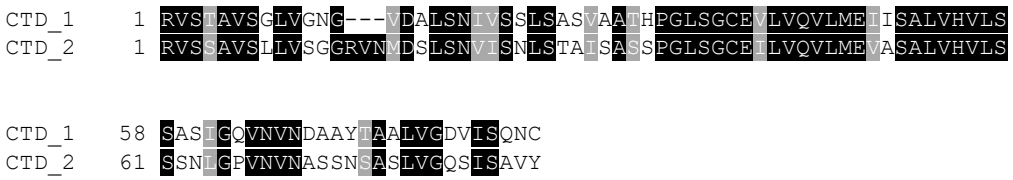


Figure 3.11: **A. aquatica *MaSp C-terminal domain types.***
CTD Type 1 (CTD_1) and CTD Type 2 (CTD_2) alignment showing similarities and differences between sequences found in MaSp silks.

3.3.6.2 Repetitive Regions

The repetitive regions of the major ampullate silks are predominantly made up of Poly(A), GAG, GGY, and GQG motifs, as expected. Generally, repeat blocks can be separated into two types (Table 3.L). Type 1 repeats are classical representatives of MAsp1, although they lack extended (more than four in a row) Poly(A) motifs. Conspicuously there are several R and D resides in these repeats and no prolines. Type 2 repeats display QQ, Poly(P) and Poly(A) motifs, but always appear with repeat type 1 blocks. Further analysis shows that these repeat blocks arise due to a frameshift in the sequence, with either one base pair addition or subtraction.

Table 3.L:* A. aquatica *MaSp Repeat types.

Repeat Type	Motifs	General properties
Type 1	GAGA, GGY, RR, EE	Classical MaSp1 motifs, very hydrophobic
Type 2	Poly(Q), Poly(A), PEMVD	More acidic, charged and polar uncharged residues. Suspected frameshifted motifs. PolyQ may have implications on protein interaction and structure.

Major ampullate repetitive regions that are connected to CTD Type 1 have both repeats, sequences containing CTD Type 2 have only repeat Type 1. NT Type A is only found with CT Type B and vice versa, which suggests that these genes are not recombining.

The hydrophobic motif 'GV' as reported by Correa-Garhwal *et al.,* 2019, is only found in silks with type 1 repeats. The published 'AmSp' fragment, which was not designated MaSp1 or 2 is present in both MaSp1 full-length sequences. It was probably designated AmSp due to the STY motif, which it is suggested here to be an NT-linker region, before the MaSp1 sequence, and it is here proposed that it represents a classical MaSp1 sequence.

The proposed unique *A. aquatica* spidroin, rich in DVAG motifs discovered by Correa-Garwhal *et al.,* (2019), was also identified in our transcriptome as a sequence

fragment, however, it was not present in any long-read spidroin sequence. Homology to the terminal region was not deemed high enough, so this may be a peptide which makes up the coating of the diving bell but is probably not a spidroin. Additionally, no MaSp2 silks (which contain proline) were identified in our datasets.

3.3.7 Putative Full-length A. aquatica Major Ampullate Spidroins

Two major ampullate (MaSp1) spidroin sequences which contained both a CTD and an NTD, as well as expected repetitive region motifs were identified.

The DNA sequences are comprised of 4,818 (Argyroneta_59572611_ccs MaSp1_1a) and 4,081 (Argyroneta_68419713_ccs p1 MaSp1_2a) base pairs, and their predicting coding sequences are 1498 and 1257 amino acids, respectively. Short, interspersed repeats constitute the differences between the two sequences, and both share the same CT and NT, so are probably isoforms of the same gene; thus, the longest was used for further analysis.

This silk contains only repeat type 1, with classical GAG and GGY motifs, and lacks proline and di-serine motifs. It does not contain elongated regions rich in DVEG residues, but does have a slightly higher proportion of R, E or D repeats than seen in terrestrial silks and the motif 'RRGGY' is common. It also lacks proline and Poly(A) tracts of more than three residues. The repetitive sequence 'GAGAAAGAGAGADASAGATGAGRRGGY' appears 11 times in the sequence (Figure 3.12), but larger repeats are not well defined or identical repeats to each other within the sequence.



Figure 3.12: Example MaSp1 repeat.
Structure prediction using Garnier is shown above the sequence, and shows a coil, alpha helix, coil, possible short beta strand, turn, coil structure.

Phylogenetic analysis was carried out on the C-terminal region of this full-length silk, as the repetitive region shows some similarities with MiSp this enabled the silk type to be determined based on phylogenetic similarity. The CTD was used in this analysis

as it shows more conservation between silk types from different species than the NTD. Within the groupings shown in this analysis, sequences generally reflect the species phylogenetic relationship of the spiders that they are from, for example *A. Argyroneta* CTD sequences are more similar to those from *Agelenopsis aperta*, an agelenid, than they are to Araneidae CTDs (*Argiope, Trichonephila* etc.). C-terminal domain sequences form two main groups; MaSp and MiSp and the sequences of interest fall within the MaSp group, leading us to confirm it as a MaSp (Figure 3.13). This is also reflected in a constructed Bayesian inference of phylogeny tree (Figure 3.14).

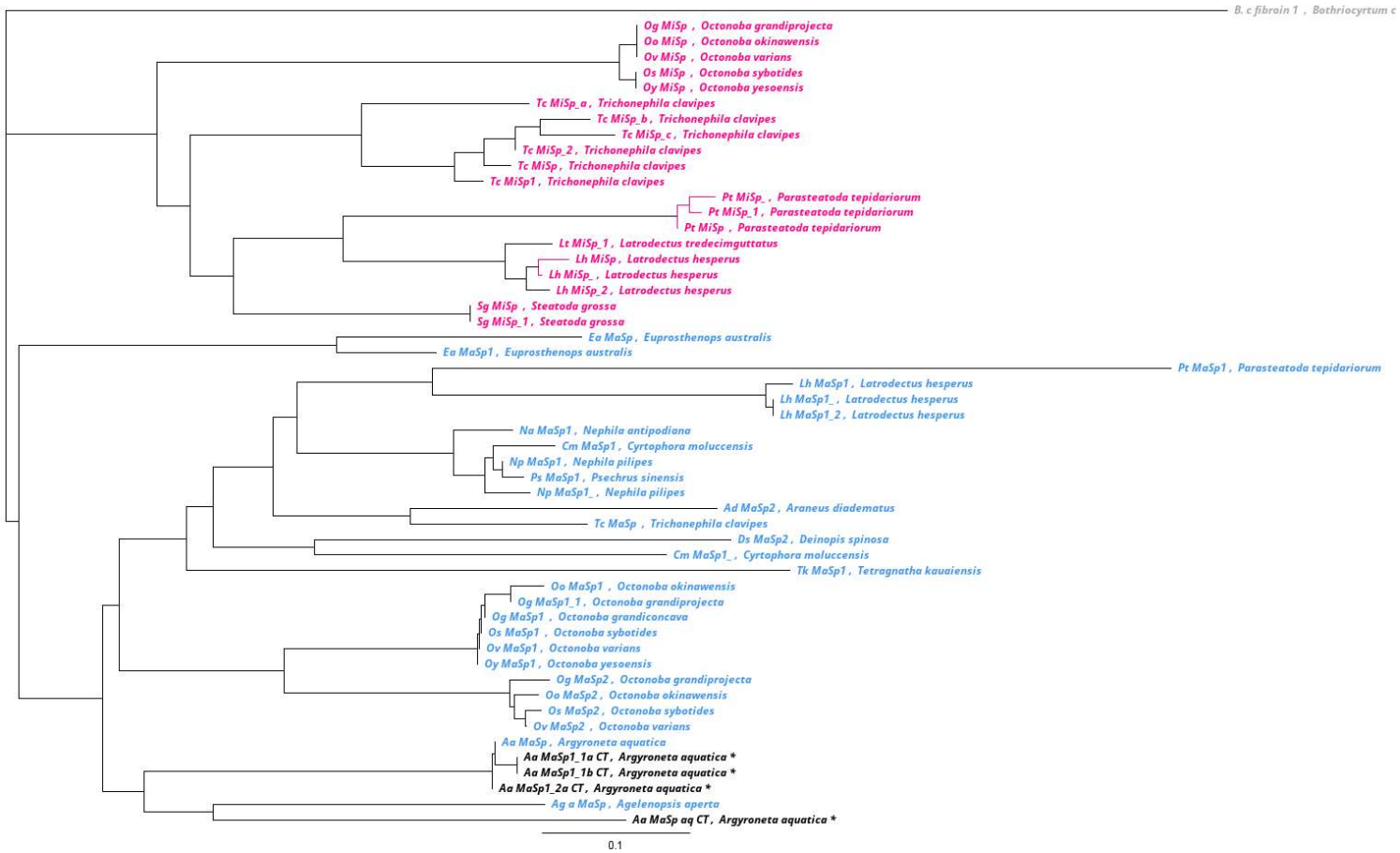


Figure 3.13: Neighbour joining tree of a selection of MaSp and MiSp C-terminal domains.

All available MaSp and MiSp sequences from the RTA clade were included, as well as a selection of sequences from Araneidae spiders, as these represent the bulk of available sequences. Sequence abbreviations are defined in Appendix A.3.1. MaSp CTDs are indicated in blue, while MiSp CTDs indicated in pink. The CTD sequences taken from the full-length A. aquatica silk in this study are in bold, outgroup in grey.

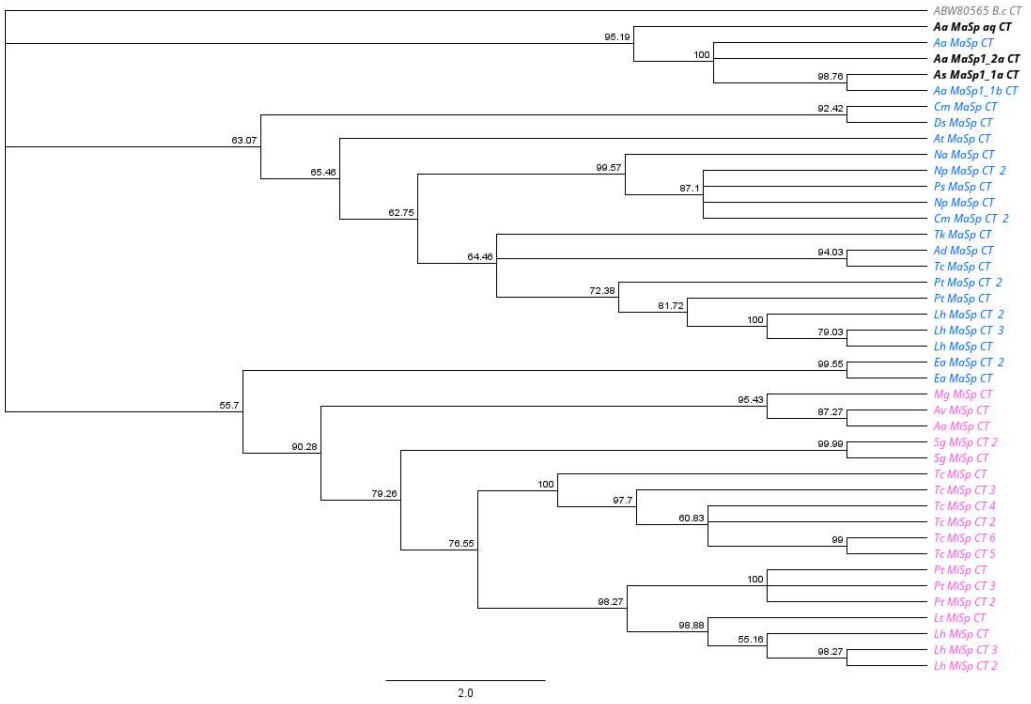


Figure 3.14: Bayesian inference of phylogeny tree of a selection of MaSp and MiSp C-terminal domains.

Sequence abbreviations are defined in Appendix A.3.1. MaSp grouping is indicated in blue, while MiSp grouping indicated in pink, outgroup is in grey. Sequences of interest (full-length A. aquatica silk sequence CTD from this study) are in bold.

The hydrophobicity plots of the NTD, repetitive and CTD are shown in Figure 3.15 and Figure 3.16. The repetitive region is made up of approximately 46 loosely defined repeat units. Within repeats, some variation in sequence and hydrophobicity is seen, which is more than exhibited by the *C. citricola* silk (discussed in later sections). Repeats are slightly less hydrophobic than those found in *C. citricola* MaSp and there appears to be no obvious N-terminal linker sequence.

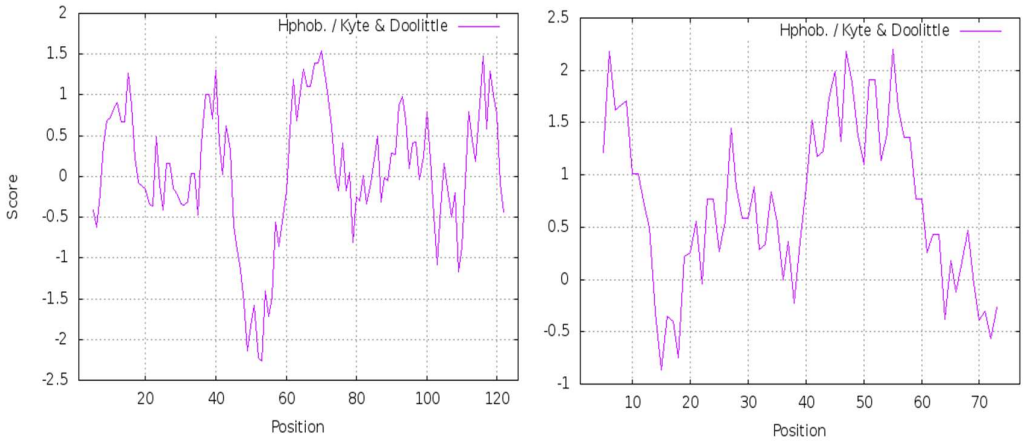


Figure 3.15: Hydrophobicity plot of terminal domains.
The full-length A. aquatica *MaSp1 NTD (left) and CTD (right) were assessed using the Kyte-Doolittle hydrophobicity scales.*

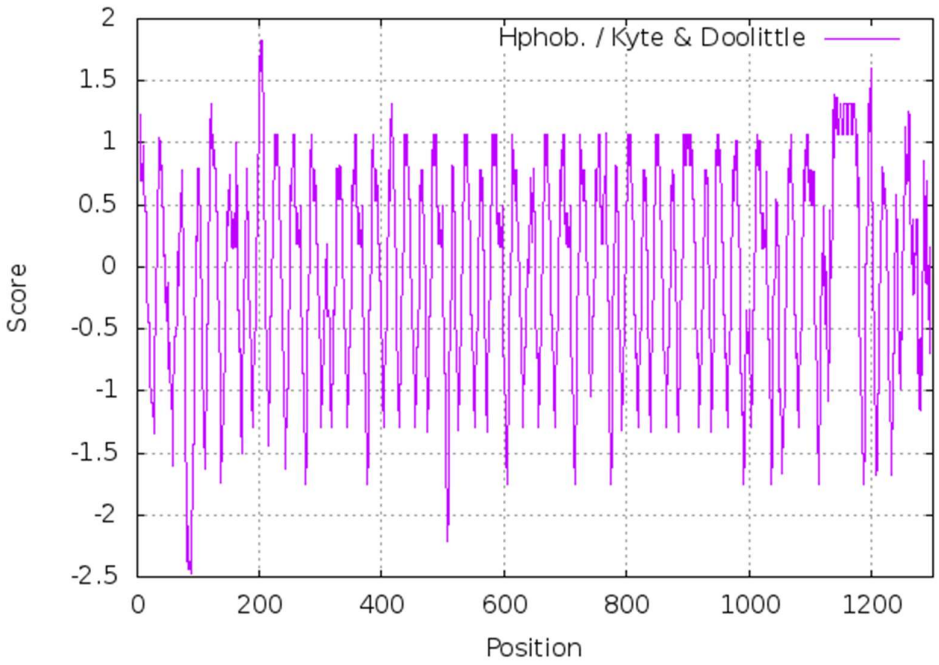


Figure 3.16: Hydrophobicity plot of* A. aquatica *full-length MaSp1 repetitive region.
Kyte-Doolittle hydrophobicity scale was used, and approximately 46 peaks of hydrophobicity can be seen.

3.3.8 Other A. aquatica Spidroin Types

Aciniform, tubuliform and pyriform spidroin sequences were found in this dataset, including two full-length PySp and eight putative full-length TuSp transcripts. A representative repeat alignment can be seen in Figure 3.17.

No minor ampullate or flagelliform silk type sequences were found in either the Illumina transcriptome or long-read sequencing datasets. In addition, no true aggregate silks (AgSp2) were identified in any of the *A. aquatica* sequences. However, a number of non-spidroin AgSp1 sequences were identified.

A full-length aciniform silk gene was not sequenced; however 34 fragmented sequences were available for analysis, which contained the motifs AAA, SSS, FV, QT, VVV, AEAIA. The hydrophobic GV motif, which is also found in found in MaSp, was only enriched in one sequence (64619193). This silk type does not contain any glycine rich repeats, but alanine and serine predominate at 20-30% of amino acid content each. There are very few (if any) prolines seen in the repetitive region, and a lower glycine content than seen in other terrestrial aciniforms. Arginine punctuates the sequence every 14 and 150 residues.

Three unique full-length tubuliforms were identified, each with the same NTD and CTD, which suggest one locus for this gene. Serine and alanine again predominate, but more arginine residues than compared to other tubuliform silks. The repetitive motifs are composed of TTTTT, AAAA (no more than 4), AA, SA, SQ, SSSS (no more than four) and are arranged in fairly large blocks (~210).

One full-length pyriform sequence was also identified, and two NTD types found. The full-length sequence was not the longest sequenced fragment. The repetitive motifs are made up of SS, TT, PP, as expected from this silk type.

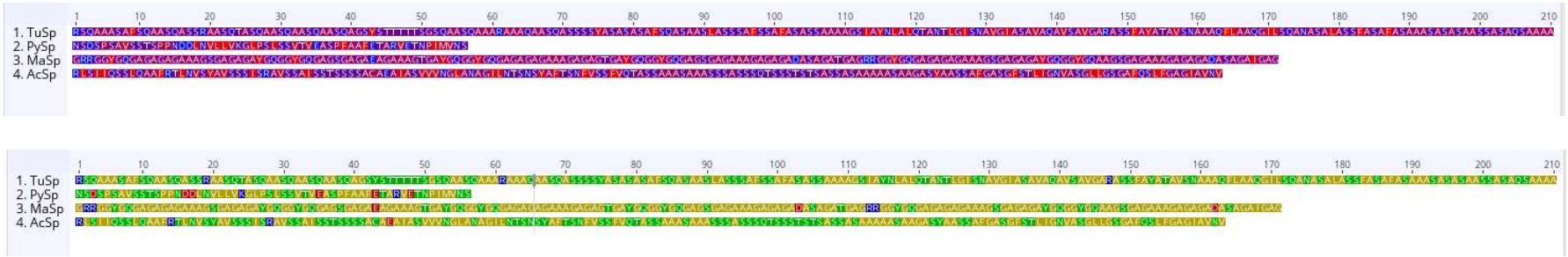


Figure 3.17: Representative repeat regions of different types of* A. aquatica *silk.

This figure shows a representative repeat of each silk type identified in A. aquatica, highlighted using two different colour schemes. The repeats, generally, are predominated by non-polar or polar uncharged, hydrophobic residues, broken up with occasional polar residues. The top panel shows amino acids coloured in terms of hydrophobicity (from red for more hydrophobic though purple to blue indicating not hydrophobic). The lower panel is coloured in terms of polarity (green – polar, uncharged, yellow- non-polar, red – polar, acidic, blue – polar, basic.

3.3.9 *C. citricola Major Ampullate Silk*

Over 200 sequences were returned from an initial BLAST search for major ampullate silks in the *C. citricola* CCS files, and 159 were confirmed to represent MaSp. Only sequences over 700 amino acids were used for analysis, which represented both MaSp1 and MaSp2, and one unique full-length major ampullate sequence was identified.

3.3.9.1 MaSp Terminal Regions

Two types of NTD were identified in the full-length sequence dataset, both 124 residues long. Both identified NTD start with the well conserved PW(S/D)(N/S) motif, where X in this case is S or D. Approximately 27 aa downstream of the starting amino acid, the usually well conserved DXXXD motif is XQX(E/D)D, and the XK motif at positions 57 and 58 is present, which leads into a the very well conserved, predicted alpha helical structure LQAMNMAFASSMAEIAATE, which is seen in nearly all other published NTD sequences – the region of high conservation. There is a conserved K eight amino acids later, and the NTD generally ends on a conserved E residue. A schematic of a general *C. citricola* MaSp NTD is shown in Figure 3.18, and an alignment of NTD types found in the full-length sequences shown in Figure 3.19.

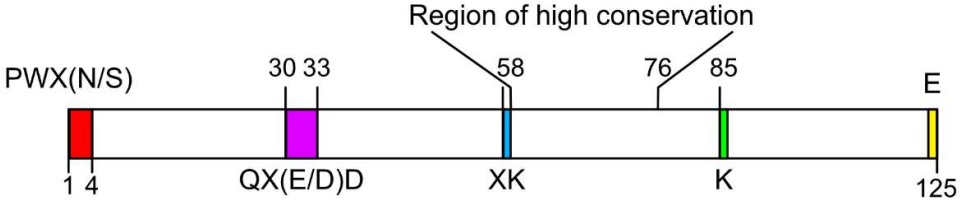


Figure 3.18: Schematic of* C. citricola *MaSp NTD.

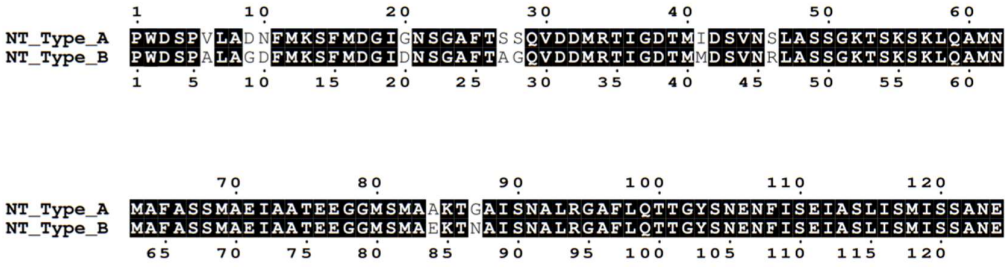


Figure 3.19:* C. citricola *MaSp NTD types.

Three CTD isoforms were identified in the MaSp sequences, which all contained the canonical QALLEVITAL-like motif, as defined previously (Figure 3.20). CT type A appears with MaSp1 repeat regions, while CT type B and C appear with MaSp2 repeats, suggesting there are at least 2 loci for MaSp2 in *C. citricola*.

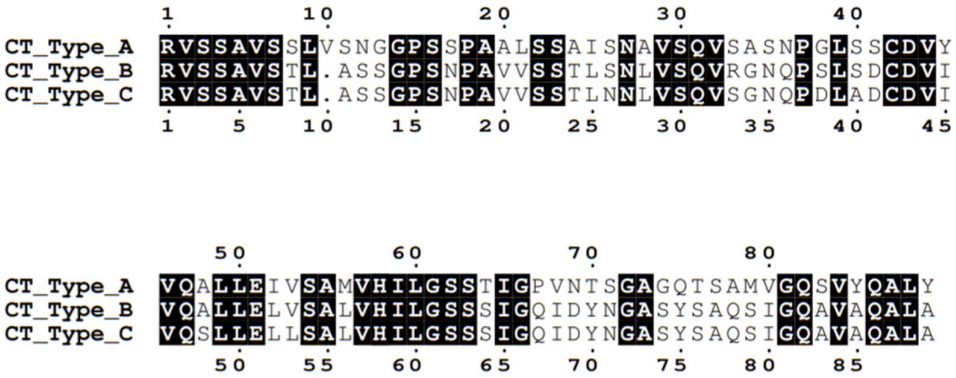


Figure 3.20: C. citricola CTD types.
The Canonical 'QALLE' motif, present in all CTDs, can be seen starting at position 47..

3.3.10 Putative Full-length C. citricola Major Ampullate Spidroin

One full-length major ampullate silk gene was identified in the *C. citricola* CCS reads. No signal peptide was predicted, and the presence of proline marks this as a MaSp2.

The DNA sequence of Cc_FL_MaSp2_1 is comprised of 5755 base pairs and the predicted coding sequence is 1,825 amino acids long. It contains an NTD (type 2), repetitive region and CTD (Type C). The repeat region contains GPG, GAY and shorter Poly(A) motifs. A schematic can be seen in Figure 3.21.

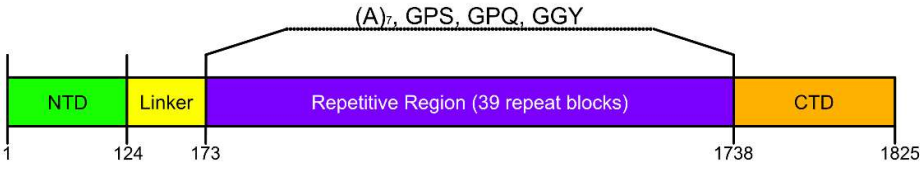


Figure 3.21:* C. citricola *MaSp2 schematic.

Each repeat block has one trough of hydrophobicity (~20aa long), and there are 39 repeat units of approximately 32 amino acids. The repetitive region is made up of

$(A)_7$, GPS, GPQ and GGY motifs. Glycine, alanine, proline and serine predominate, and this region is mainly made up of hydrophobic (69 %) and polar uncharged (30.2 %) residues. A consensus repeat can be seen in Figure 3.22, and hydrophobicity plot of the repetitive region is shown in Figure 3.23.

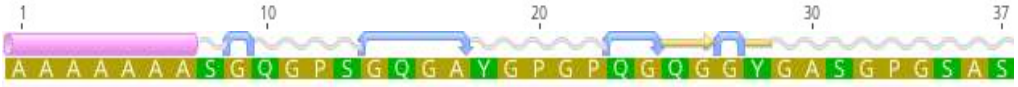


Figure 3.22: C. citricola MaSp2 repeat block consensus sequence.
Structure prediction using EMBOSS can be seen above the sequence, amino acids are coloured using the Geneious polarity scale (Yellow: Non-polar, Green: Polar, uncharged).

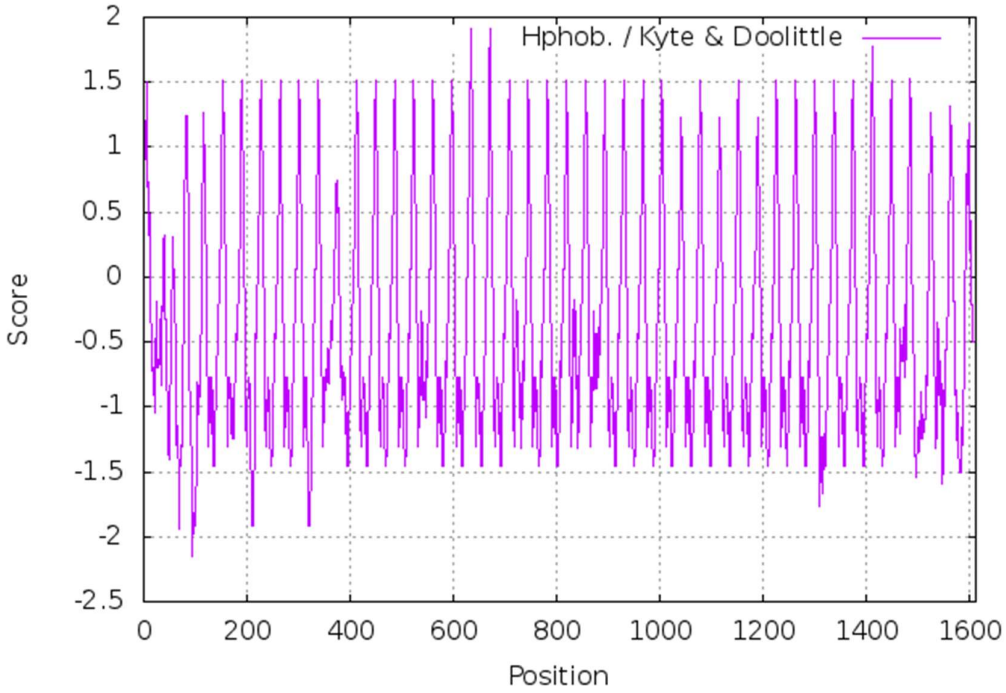


Figure 3.23: Hydrophobicity plot of* C. citricola *MaSp repetitive region.

Hydrophobicity plots of the repetitive region and NTD show a probable linker region after the NTD of 50-70 amino acids (Figures 3.23-3.24). This proposed linker sequence is shown in Figure 3.25, and the predicted structure shown is mainly composed of turns, beta strands and coils. Importantly, this region lacks the middle alpha helix that characterises subsequent repeats. The CTD is characteristic of MaSp2 spidroins from other species and shows the expected hydrophobicity profile (Figure 3.24). Appendix

A.3.2 shows the NTD and CTD of the full length MaSp2 aligned with a selection of terminal domains from other species, to highlight the homology between these regions.

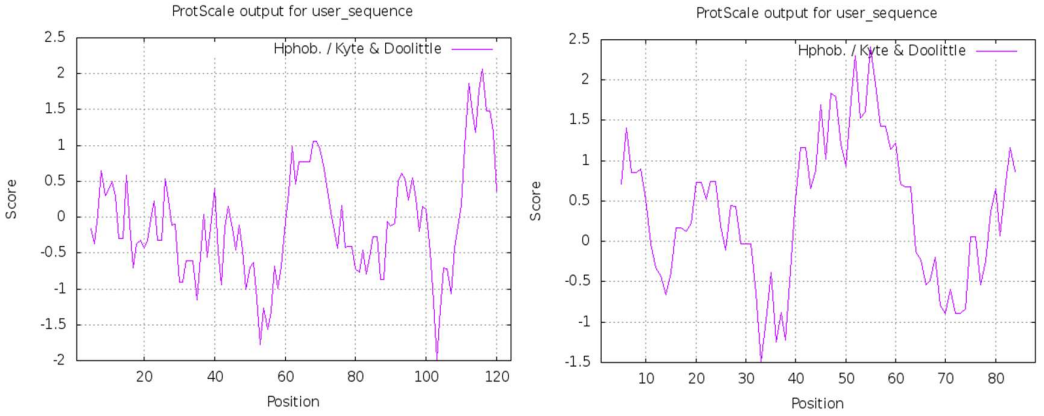


Figure 3.24: Hydrophobicity plot of terminal regions of C. citricola MaSp2.
N-terminal (left) and C-terminal (right) region of the full-length C. citricola MaSp2, assessed using the Kyte-Doolittle hydrophobicity scale.



Figure 3.25:* C. citricola *MaSp2 NTD linker region.
Structure prediction using EMBOSS can be seen above the sequence, amino acids are coloured using the Geneious polarity scale (Yellow: Non-polar, Green: Polar, uncharged).

3.3.11 Other C. citricola *Spidroin Types*

No other full-length spidroin sequences were identified in this dataset. Several fragmented MiSp, AcSp, TuSp and PySp were identified which contained either N- or C-terminal domains. All identified silk types contained the classical, expected amino acid motifs of their silk type. The starting material for this sample was slightly lower quality than that of *A. aquatica*, and it is suggested that this may have a contributing factor to the lower numbers and shorter CCS reads. In addition, the longer full-length sequences in the literature are often well above the 10kb achieved in this sequencing run, which suggests that the spidroins used by *C. citricola* are longer than this.

No AgSp, Flag or TuSp sequences were retrieved from this dataset, which was not wholly unexpected, as these spiders have not been noted to produce AgSp or Flag. In addition, TuSp is only produced by adult female spiders and the genetic material for this study was taken from a sub-adult as this was the only specimen available.

3.4 Discussion

This study presents five high quality transcriptomes from a variety of starting tissues, constructed using data from Illumina sequencing, which represent over 131,000 predicted protein sequences. This demonstrates a robust pipeline for the generation and identification of short read spidroin transcripts from different tissues and spider species. Also, the first full-length sequenced silk genes of both *Argyroneta aquatica* and *Cyrtophora citricola* are presented, as well as a functionally annotated database of non-redundant high quality long-read transcripts. By using two technologies to generate sequencing data, it is possible to compare and contrast these, making suggestions for future investigations. This data, taken together, will be valuable for future studies into the genetic and proteomic makeup of spider silks in the chosen species.

3.4.1 The First Full-length Spidroins of A. aquatica *and* C. citricola

This study presents eight full-length silk genes, the first described for *Cyrtophora citricola* and *Argyroneta aquatica*. These full-length genes show that the accepted modular nature of silk genes is shared in both these species, as well as highlighted the diversity of silks between species, and will pave the way for the more rational design of targeted miniaturised silks. As the whole gene has now been sequenced, it can be more accurately determined which regions are repeated and probably more important for silk function, whereas previous studies have relied solely on short, fragmented sequence information.

The predicted full-length *A. aquatica* MaSp is 1498 amino acids long, and contains the classical MaSp1 repetitive motifs, but lacks the proposed aquatic-specific hydrophobic GV motif, as reported by Correa-Garwhal (2019). The CTD of this silk shows high homology to MaSp1 CTD of other spidroins, while the NTD is more diverged. When compared to the *C. citricola* NTD, it is shown to be more hydrophobic and shows less pronounced peaks of hydrophilicity (Figure 3.16, Figure 3.24).

This full-length silk has been designated as a major ampullate silks, due to motif and terminal domain similarities to this silk type. However, it shows similarities to both MaSp and MiSp, lacking Poly(A) tracts as would be expected in MaSp, and only

exhibiting a maximum of three alanine residues together. It also lacks the serine rich 'spacer' regions expected in MiSps, and homology and phylogenetic analysis of the terminal domains determines that it is a major ampullate (Figures 3.13, 3.14). The lack of Poly(A) tracts could suggest that this silk has a lower tensile strength than other MaSps, as the GAGA motif is able to form β-sheets, but with a lower binding energy, leading to a lower tensile strength (Hayashi, Shipley and Lewis, 1999; Van Beek *et al.*, 2002; Asakura *et al.*, 2004; Holland *et al.*, 2008; Zhao *et al.*, 2021). However, this sequence does contain the GAGAGA motif, and its alanine rich regions flanked by 'GA', as seen in other minor ampullate silks (Colgin and Lewis, 1998). The presence of GGY is as expected, which furthers the suggestion that this is important to modulate intermolecular self-assembly and crosslinking in silks, and that tyrosine (Y) is an important amino acid for self-assembly of fibres (Partlow *et al.*, 2016; Malay, Arakawa and Numata, 2017). There is no evidence of the SX_4 motif, used for calcium binding in other aquatic silks (Ashton and Stewart, 2015).

The repetitive region is made up of 46 loosely defined repeat units, which show more variation in length and hydrophobicity than those that make up the *C. citricola* silks. This could suggest that there is less selection pressure on this silk to maintain rigidly defined and controlled repeat units, as the silk needs to be more amenable to activity in 100% humidity. Additionally, as the repeat regions in this silk are loosely defined, this may suggest that it is a more recently evolved silk, as sufficient evolutionary time has not passed to allow the concerted repeat evolution and extreme homogeneity as seen in other silks (Johannesson *et al.*, 2005; Chaw *et al.*, 2014; Persi, Wolf and Koonin, 2016)

In the *A. aquatica* transcriptome, four full-length tubuliform silk and one full-length pyriform silk sequences were identified. The pyriform silk lacks long QST rich regions and has more basic and acidic residues than expected. One full-length pyriform silk sequence was also identified, and two NTD types found. The full-length sequence was not the longest sequenced fragment, which suggests that either two pyriform silk genes exist or that the RNA is being alternatively spliced. This was a lot shorter than many published pyriform silks, which is interesting as it could suggest that this glue

has had to evolve more rapidly than other silk types to maintain its function under water.

No full-length *C. citricola* aciniform silk sequences were found, which suggests that the transcripts are too long to have multiple sequencing polymerase passes, as AcSp fragments have been seen both in this dataset and other short-read transcriptomes (data unpublished). This is as expected, as published full-length AcSps are generally greater than 3000 amino acids and other full-length silks from this study were generally longer than this (Table 3.K).

All full-length sequences have a stretch of 10 lysine residues (K) approximately 32 amino acids downstream of the stop codon, giving further weight to the fact that this is probably the true stopping point of the protein. Lysine tracts can help to control translation by inducing ribosomal pausing in other species, from yeast to humans (Brandman *et al.*, 2012; Tsuboi *et al.*, 2012), and evidence of these tracts suggests that spiders use the same mechanism (Arthur *et al.*, 2015; Koutmou *et al.*, 2015).

The data presented here suggests that there exists two genetic loci for MaSp2 silks in *C. citricola*, and one for MaSp1, due to the presence of multiple terminal domains. The MaSps contain shorter Poly(A) stretches than has been found previously, which could suggest that the silks are less crystalline and weaker, which is backed up by anecdotal evidence suggesting that this silk is less strong than silk of, for example, the golden orb weaver, *Trichonephila spp*. Proline is very often seen in these silks, which could help contribute to toughness and allow the silk to be used for long periods of time.

The identified full-length silk genes in this dataset are at the lower end and shorter than those described previously (1741-3779 aa), which could be a quirk of *A. aquatica* silks or due to the fact that this data will naturally only reveal silks under 10 kb (see Table 3.K) (Han *et al.*, 2013; Chaw, Saski and Hayashi, 2017; Stellwagen and Renberg, 2019). The lack of full-length *C. citricola* silks found suggest that these silk transcripts are longer than 10 kb, which is more reminiscent of terrestrial orb weaving genes than of *A. aquatica* (Chen *et al.*, 2012; Ayoub *et al.*, 2013).

Some silks are notable by their absence in our dataset, for example, *A. aquatica* minor ampullate silk was not identified, but this species has been shown to possess minor ampullate glands and spigots (Strickland, M, unpublished). This could mean that MiSp is used in too low levels to be identified by these sequencing methods, or that its sequences have diverged enough from that of terrestrial MiSp that it can no longer be identified by these methods. Other missing silks, however, are expected. No flagelliform or aggregate silks were found in either species, which is to be expected as these spiders do not produce traditional orb-webs. In addition, *C. citricola* is known to lack the necessary glands and spigots usually associated with these silk types (Coddington, 1989).

Tubuliform silk was not identified in the *C. citricola* dataset, which was also expected as a subadult specimen was used and only adults are expected to produce this (Tian and Lewis, 2006). This suggests that this use of silk types is conserved between these species. In addition, no true aggregate silks (AgSp2) were identified in any of the *A. aquatica* sequences. However, a number of AgSp1 sequences were identified, which could be used as reinforcing in silk glands, as previously hypothesised (Collin *et al.*, 2016; Correa-Garhwal *et al.*, 2019).

3.4.2 Species and Possible Environment Specific Adaptations

In the identified putative *A. aquatica* silk genes, it is generally found that a higher proportion of acidic and basic residues (including D and R) are present than would be expected, especially when compared to *C. citricola* silk, used in predominantly dry environments. The NTD and CTD of silks used by both these species follow the expected amino acid patterns to published silks, suggesting that they retain the same assembly mechanisms (Strickland *et al.,* 2018).

Interestingly, no representative MaSp2 (proline containing silks) or MaSp2-like silks were found in either the long-read or short read sequence datasets. As far as is currently known, no MaSp2 evidence has been found in any studies of *Argyroneta aquatica*. This may be because proline disorders silk, making it more prone to supercontraction (Liu *et al.,* 2008). The hydrophobic GV motif is found in MaSp1 silks with type 1 repeats which may be important for maintenance of fibre structure

underwater (S. Correa-Garhwal *et al.*, 2019). This author did not agree with the assertions made by this group that the 'AmSp or 'Sp' sequence fragments were unique to *A. aquatica*, but posit that they are smaller portions of MaSp1 or not true spidroins.

3.*4.3 Frameshift Nonsense or Additional Diversity?*

Several silks in these CCS datasets contained KED, Poly(Q), Poly(P) and Poly(A) motifs, which are often charged and highly hydrophobic. Sometimes these appear with the classic MaSp1 type motifs, and sometimes they appear on their own.

These may be a result of a frameshift during translation, which has sometimes happened more than once in a sequence, leaving the CTD recognisable. The NTD may not be present because it is not translated or because the sequence was lost during library preparation. However, this frameshifted protein must be produced by the spider's RNA polymerase, as it derives from an RNA CCS read. This is not completely unexpected, as frameshifts often happen in repetitive regions of DNA (Ripley, 1990; Lovett, 2004). The frameshift point, where the sequence slips back into in frame is usually at a glycine or glutamine. The protein that is produced may have novel properties, such as promotion of aggregation or adherence to prey due to increased QQQ motifs, or it may be non-functional. The high proportion of D and E residues may also promote calcium ion bonding and help to make the silk more visco-elastic, or may also be a way for the fibre to retain its crystalline properties and inter-molecular bonds, which are presumably lost in the MaSp1 silk that lacks these motifs. This may be a true MaSp1/2 hybrid, produced by the spider in lieu of producing two types of fibre and using both to create a dragline fibre. This phenomena has been seen before in spider silks, but in those cases brought the silk to be much more similar to other silk types (Beckwitt and Arcidiacono, 1994).

Recent research has suggested that frameshift mutations, where the hydrophobicity profile remains unchanged, could be a powerful evolutionary mechanism for creating new proteins with vastly different sequences, yet similar physicochemical properties to the proteins from which they originate (Bartonek, Braun and Zagrovic, 2020).

Frame shifted proteins can also sometimes retain the same function, and this has previously been found in tubuliform silks (Chaw *et al.*, 2018; Huang *et al.*, 2021).

It is proposed here that frameshift mutations which occur more than once inside a spidroin are a powerful way to increase diversity, and they are maintained in the transcript repertoire because the contribute directly to the silk property. However, more experimental evidence is needed before this hypothesis can be proven, both on the genetic and material property levels. This would improve the confidence in the fidelity of the RNA sequencing by proving that it is present in the genome or resulting protein.

3.4.4 Using Next Generation Sequencing to Study Spidroins

This investigation has successfully used two approaches to identify silk transcripts from spider RNA. The relatively low-cost Illumina sequencing presents a convenient snapshot to identifying transcripts and RNA fragments, especially when the focus is on non-repetitive spidroin terminal domains or similar. However, as shown in Chapter 2, low quality RNA can lead to very short fragments being identified and assembled, and the true value of these sequences may be eclipsed by the wealth of information provided by a full-length transcript. Full-length sequences can lend more confidence to the resolving of repetitive regions and allow a deeper exploration into the variation in these regions between species.

A reference genome is not currently available for many spider species, including those used in this study, which meant that it was necessary to reconstruct each transcriptome *de novo*. While traditionally, the N50 value (defined by Miller *et al.,* (2010, as the longest contig, such that all contigs of at least that length compose at least 50% of the bases of the assembly), is used to assess the quality of *de novo* transcriptomes, this value measures the continuity of an assembly, but not its accuracy. The actual aim of a transcriptome is to get complete coverage of all transcribed genes, as one contig per transcript, rather than the long contig lengths as is the aim with genomic sequencing. Therefore, BUSCO was also used to assess transcriptomic completeness, to provided a quantitative measure for the assessment of transcriptome completeness based on evolutionarily informed expectations of

gene content from these near-universal single-copy orthologs. The assemblies presented here achieved a range of contig numbers and N50 values (see Table 3B), a range which can be explained by the difference in specimen quality and freshness before RNA extraction. Additionally, BUSCO completeness scores were of <85% in all cases, which compares to and, in some cases, outperforms other high quality published transcriptomes (Figure 3.3).

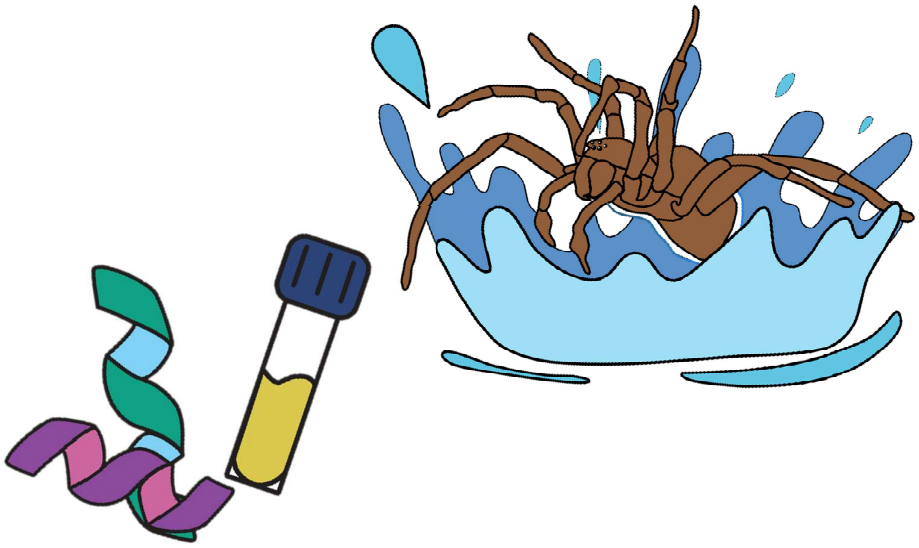
These scores, coupled with the number of spidroin sequences retrieved, that represent nearly all silk types currently known to be produced by spiders, show that the generated data is a robust starting point for future and more in-depth studies about the proteomic make up of these extraordinary spider species. The long-read sequencing data provided over 800,000 high quality CCS reads and more than 870,000 predicted protein coding sequences. Assembly was not necessary for these long reads for genes under 10kb in length, but this would have been attempted if less restrictions on computing power and time were in place, which could possibly have uncovered all silk genes in their entirety.

While SMRT Sequencing can provide full-length, high quality transcripts, its cost is fairly prohibitive, as one SMRT sequencing sample can be almost 10x more expensive than a run on a short read sequencer, and the bioinformatic tools and computing resources required in its analysis are often much greater. A comparison of the sequencing technologies used in this study is shown in Table 3.H. While some variation is expected dependent on the time in which sequencing was ordered and service provider, there still exists a gap in attainment between these technologies on several fronts.

3.4.5 Conclusions

A broad range of silk sequences have been presented in this investigation, from several species which inhabit environments of varying humidity. A range of bioinformatic approaches have been used and this study has gone some way towards showing the versatility of using silks as a study system, highlighting several areas which are in great need of further study, expanded upon in Chapter 5. Further investigation should be carried out to correlate sequence and physical properties of these silks in order to prove a solid base for biomaterial expression in the laboratory.

The current and future advances in the field of NGS make it important to critically evaluate the available tools for each study. Tools should be chosen carefully at the outset, taking into account cost, time and analysis needed. As these technologies continue to improve, more can be learnt about spidroins and their properties can be modelled on a larger scale before attempts are made to produce them in the laboratory, which should lead to a dramatic increase in quality of recombinant spidroins tailored for use in specific environments.



Chapter Four

Creation and analysis of a novel recombinant aquatic silk protein

The Lost Silk of Atlantis

4.1 Introduction

Spider silk has many of the desired properties needed for a biomaterial; it is tough, biocompatible, provokes a minimal immune response in mammals and has remarkable tensile strength; it can be as strong and tough as high performance man made materials such as Kevlar and steel (Rising *et al.*, 2005). Spider silk is often more desirable for use as a biomaterial than the more readily available and easier to harvest silkworm silk for two main reasons. Firstly, spider silk is much stronger and tougher than silkworm silk (Gosline *et al.,* 1999, Andersson *et al.,* 2016). Secondly, silkworm silk is coated with sericin, an immunogenic protein coating that is difficult to remove from the silk fibres themselves, which spider silk lacks (Sparkes and Holland, 2018; MacIntosh *et al.,* 2018). By considering these properties and the relative difficulty of farming spiders due to their cannibalistic nature, much interest has focussed on the recombinant production of synthetic spider silk as a scalable and animal free biomaterial process (Römer and Scheibel, 2008). Due to their versatility and range of biocompatible, biodegradable and physical properties, recombinant spider silk proteins are promising candidates for use in scaffolds for tissue engineering, drug delivery systems and nanodevices, to name just a few (Kiseleva *et al.,* 2020; Spiess *et al.,* 2010).

Natural silk spidroins vary in size between individuals and species. Generally, spidroins are made up of a non-repetitive N-terminal domain (NTD), a large region made of repetitive glycine/alanine rich units and a non-repetitive conserved C-terminal domain (CTD) (Gaines IV and Marcotte, 2008), an example of which is shown in Figure 4.1. The N- and C-terminal domains are evolutionarily conserved and have been shown to have important roles in the formation of silk fibres (Rising and Johansson, 2015).



Figure 4.1: General protein structure of a spidroin.

4.1.1 Aquatic Spider Silk

The idea of bioprospecting, the exploration of biodiversity for new resources of social or commercial value, is not a new one. It has in the past been used to predict which spider species are likely to spin silks exhibiting extraordinary performance properties (Coddington *et al.*, 2002; Agnarsson, Kuntner and Blackledge, 2010; Beattie *et al.*, 2011; Miller, 2015). *Argyroneta aquatica* is unique, as it is the only spider to spend most of its life underwater. It is a member of the family Dictynidae (Wheeler *et al.*, 2017) and creates dome shaped webs, 'diving bells', which are filled with air and used underwater: functioning as a physical gill (Figure 4.2) (Selden, 2002; Seymour and Hetz, 2011). *A. aquatica* consume aquatic invertebrates and small fish and these spiders mate, eat and rest inside their diving bells, which can also act as the foundation for an egg sac (Uzenbaev and Lyabzina, 2009; Nyffeler and Pusey, 2014). The silk used to create this diving bell is able to effectively trap the air as it is taken under water and can also take up some dissolved oxygen from the water, allowing for longer underwater periods (Seymour and Hetz, 2011). This structure is also gas permeable, allowing carbon dioxide to diffuse out over time, which makes the silk an interesting candidate for further investigation as a gas exchange membrane. More information can be found in Chapter 1.



Figure 4.2: **Argyroneta aquatica *inside its diving bell.***
The silken threads of the diving bell can be seen constraining an air bubble. The spider sits inside the bell, eating a small cricket. Photograph taken by Robert Morton and used with permission.

There is a relatively small amount of published research into the genetic make-up of *A. aquatica* silk genes (see Chapter 3 of this thesis). The available sequences suggest that the amino acid structure and modular make up is minimally diverged from terrestrial silks, while the number of expressed spidroin types vary (Strickland *et al.*, 2018). As they retain the conserved CTD, modular gene structure and external spinning apparatus seen in other silks and spiders, it is suggested that fibre formation is the same as seen in other spiders, making this a good candidate to express in an existing recombinant spidroin expression system. The fibres have the potential to be strong, flexible, non-immunogenic and form structures that could be useful for diverse uses such as liquid filters or 3D porous scaffolds suitable for tissue engineering (Debabov and Bogush, 2020; Gustafsson *et al.*, 2020; Salehi, Koeck and Scheibel, 2020).

4.1.2 *Recombinant Spidroin Production in* E. coli

Spidroins are very large (usually greater than 250 kDa), very repetitive proteins which normally contain high proportions of glycine and alanine, all of which makes them very difficult proteins to express recombinantly. Most studies to date describe the expression of miniaturised spidroins containing repeats of 26-40 amino acids repeated from 4-32 times (Debabov and Bogush, 2020), and most commonly expressed in bacteria (Venkatesan, Chen and Hu, 2019; Whittall *et al.*, 2020) due to its ease of genetic manipulation, low cost, short generation time and potential for industrial scaleup (Chen, 2012). However, recombinant spidroin yields are usually low (2-15 mg/L) and these sharply reduce with a molecular mass increase (Debabov and Bogush, 2020). Strategies to increase yields have included greatly increasing the available pool of glycyl tRNA (by 10-30 times) which enabled the production of proteins up to 500 kDa (Xia *et al.*, 2010; Bowen *et al.*, 2018). This aimed to recapitulate the natural system of increased tRNA in the epithelial cells of the silk glands which support the extreme need of certain amino acids such as alanine and glycine, showing that mimicking natural systems can be favourable when improving recombinant protein expression (Candelas *et al.*, 1990). Other studies have included both globular terminal regions to increase solubility and expression levels, with some success (Finnigan *et al.*, 2020).

There are several examples of spidroin production in *E. coli.* The repetitive regions of the two proteins that make up the major ampullate silk (MaSp) of *Araneus diadematus* have been used to create recombinant spidroins, which each exhibit very different properties due to differences in the amino acid makeup of the repetitive region (Huemmerich *et al.*, 2004). A native sized MaSp of greater than 280 kDa was produced in *E. coli* and expressed in low levels, but upon artificial spinning showed comparative mechanical properties to natural silk (Xia *et al.*, 2010). Aciniform silk (AcSp) has also been produced recombinantly in *E. coli* and two 200 amino acid repeats were sufficient to form silk-like fibres when exposed to shear forces, albeit with lower mechanical strengths compared to natural silk (Xu *et al.*, 2012). By combining the repeat unit of flagelliform silk (FlaG) with a MaSp repeat unit of *Trichonephila clavipes*, a silk was produced that exhibited both strength and

elasticity. Wet spinning or hand-pulling produced fibres, with varying mechanical properties that depended on the number and combination of repeats (Teulé, Addison, *et al.*, 2012). Another recombinant protein, NT2Rep was expressed in *E.coli* with relatively high yields (125 mg/L) and then spun into fibres in a highly efficient process, which shows promise for increasing expression levels and fibre yields of recombinant silks (Andersson *et al.*, 2017). Very large (>556 kDa) recombinant spidroins were also produced using a combination of synthetic biology and chemical tools and the resulting fibres showed comparative strength, modulus, extensibility and toughness to natural silks (Bowen *et al.*, 2018). These studies show that it is possible to make a range of recombinant silks that can in some way recapitulate the properties of their natural counterparts.

Recombinant spidroins are often subject to post-production processing in order to turn them into functional materials. This includes artificial spinning, which can combine recombinant spidroins into a fibre and enables fine-tuning of the resultant properties (Rammensee *et al.*, 2008). However, artificial spinning is a time consuming and delicate process, unsuitable for reproduction on a large scale. The mechanical spinning process, used for fibre formation of many recombinant spidroins, consists of applying shear forces to the spidroin solution in the presence of a physiological buffer into a coagulation bath (Teulé *et al.*, 2007), which is another time consuming process. Self-assembling silk fibres would provide a great benefit to recombinant silk fibre production. The use of other expression systems to create recombinant silks has also been attempted, for example using yeast, insect cells, other bacteria, animal cells and transgenic plants, but production at a larger scale remains challenging (Fahnestock and Bedzyk, 1997; Scheller *et al.*, 2001; Huemmerich, Scheibel, *et al.*, 2004; Yang *et al.*, 2005; Widmaier *et al.*, 2009; Teulé, Miao, *et al.*, 2012; Jones *et al.*, 2015).

4.1.3 4RepCT

The recombinant protein 4RepCT is a miniaturised major ampullate spidroin that has been stably expressed in *E.coli* and can form 3D structures independently of any additional mechanical spinning processes or manipulations (Stark *et al.*, 2007) (Figure 4.3). It contains a thioredoxin tag for protein solubility, a poly-histidine tag for

purification and a repetitive region, as well as a non-repetitive C-terminal domain that allows dimerisation and fibre formation, strengthened with a disulphide bond between two cysteine residues in the CTD much like in natural silks, upon the thrombin mediated removal of thioredoxin.

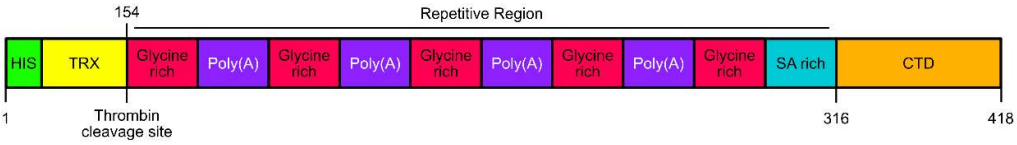


Figure 4.3: Schematic representation of 4RepCT.

The poly-histidine tag (HIS, shown in green), is followed by a thioredoxin solubility tag (TRX, yellow) and then a thrombin cleavage site. The repetitive region of 4RepCT is made up of 4 Poly(A) blocks (purple), five glycine rich regions (pink) and a serine and alanine rich linker region before the C-terminal domain (CTD, orange).

The role of the CTD has been shown to be essential in fibre formation in recombinant silk systems (Ittah *et al.*, 2006). Spontaneous fibre formation eliminates the need for costly and time-consuming artificial spinning processes, although it does not enable control of fibre length and thickness. This silk has also been functionalised using site specific chemical conjugation with antibiotic and fluorescent molecules post-fibre formation (Harvey *et al.*, 2017). The repetitive region and CTD is based on the silk of the African nursery web spider, *Euprosthenops australis*. This spider uses major ampullate silk to construct sheet webs from which it forages and prey items are caught by becoming entangled by these threads (Madsen, Shao and Vollrath, 1999). These spiders are widely distributed throughout Africa (Cruz Da Silva and Sierwald, 2014). While there is not a wealth of available literature on the ecology or natural silk use of this spider, it has been shown that this silk is stiffer, stronger and less extensible than silk from the more well studied orb-weaving species, such as *Araneus diadematus* or *Trichonephila clavipes*. This natural silk also has a higher poly alanine content than most other published major ampullate proteins, which leads to a greater proportion of crystalline β-sheet structures (Shao and Vollrath, 1999; Vollrath and Knight, 2001; Pouchkina-Stantcheva and McQueen-Mason, 2004; Stark *et al.*, 2007).

The addition of a conserved NTD sequence, based on a natural *E. australis* MaSp fragment, into the 4RepCT spidroin has enabled the rate of fibre formation to be controlled through changing the pH (Askarieh *et al.*, 2010). At higher pH values the NTD inhibits fibre formation, but fibres rapidly form when this is lowered. By enabling control of the timing of fibre formation, this shows potential to allow the scaling up of the recombinant spidroin formation process. Genetic engineering, by substituting specific amino acids, has been used to increase the strength of 4RepCT by introducing additional covalent linkages in the fibre, showing that there are still steps that can be taken to improve fibre formation (Grip, Johansson and Hedhammar, 2009).

This system, if it could be reconstructed with elements based on the silk of a different spider species, may prove to be a reliable way to produce specialised silk fibres with a range of possible uses.

By using knowledge of the silk produced by *Argyroneta aquatica*, this study shows the creation of a miniaturised major ampullate aquatic spidroin (AquaRepCT), expressed in *E. coli*. Initial expression levels were low in pJE401, but when the construct was moved to a different expression vector, pET32, a marked improvement was seen, due to the presence of the more powerful T7 expression system than the T5 promotor present in the previous pJE401 construct. This produced a soluble protein which does not form fibres spontaneously upon removal of the solubility tags, proposed to be due to differences in the repetitive region from 4RepCT. This protein can be produced at yields of up to 150 mg/L and demonstrates the versatility and robustness of the new AquaRepCT expression system. Next, the C-terminal domain alone was expressed and analysed in order to investigate its dimerisation properties. It was shown that the CTD was indeed capable of dimerization, which suggests that it is the repetitive region of this construct that hinders fibre formation. Insights can be gained from this study that are useful when designing other recombinant silks, which need to display certain properties, such as a lack of, or more controllable fibre formation. Further constructs have been designed in order to investigate this protein further, and improve its solubility and fibre forming properties. This study has demonstrated that the principle of recombinant production of miniaturised spidroins can be applied to different silk types using the same expression model to create

different proteinaceous materials. With further refining, this method could be used to create a range of functionalised recombinant spidroins with properties tailored for use as silk-based biomaterials, using natural spidroins as inspiration.

4.2 Materials and methods

4.2.1 General Conditions

All kanamycin sulfate concentrations were used at 50 µg/mLand all ampicillin concentrations 100 µg/ml. All overnight incubation steps for liquid cultures were left for 16-18 hours at 37°C with shaking at 180 RPM. These standard conditions were used throughout, unless otherwise indicated. The standard buffers used for this set of experiments and their function are shown in Table 4.A. The *Escherichia coli* strains used in this study and their features are shown in Table 4.B.

Table 4.A: Standard Buffers.

Buffer name	Contains	pH	Function
Buffer A	20 mM Tris, 300 mM NaCl, 15 mM Imidazole	7.5	Equilibration Buffer
Buffer A1	20 mM Tris, 500 mM NaCl, 50 mM Imidazole	7.5	Equilibration Buffer
Buffer B	20 mM Tris, 300 mM NaCl, 700 mM Imidazole	7.5	Elution Buffer
Buffer C	20 mM Tris	8.0	Dialysis Buffer
Buffer C1	20 mM Tris, 300 mM NaCl	8.0	Dialysis Buffer

Table 4.B:* E. coli *strains used in this study and their features.

E. coli strain	Description	Used for	Reference
NEB5α	T1 phage resistant and *endA* deficient	High quality plasmid preparations and plasmid storage	(Anton and Raleigh, 2016)
BL21	Designed for protein expression	Protein expression using T5 promotor system	(Jeong, Kim and Lee, 2016)
DL41	Methionine auxotroph	Protein expression allowing downstream labelling with ^{35}S-methionine and selenomethionine	(Hendrickson, Horton and LeMaster, 1990)
Rosetta (DE3)	Supply of tRNAs for rarely used codons in *E. coli*, providing universal translation, not limited by the codon usage of *E. coli*	Enhancing protein expression in T7 expression systems	(Fathi-Roudsari, Akhavian-Tehrani and Maghsoudi, 2016)

4.2.2 pJE401-AquaRepCT

4.2.2.1 pJE401-AquaRepCT Plasmid Retrieval

Escherichia coli NEB5α cells containing the recombinant plasmid pJE401-AquaRepCT were provided by Dr David Harvey. This encoded the *Argyroneta* miniaturised spidroin protein (4 repeats) fused with an N-terminal thioredoxin tag and an *A. aquatica* MaSp CTD. Cells were revived from glycerol stocks and grown overnight at 37^{o}C on an LB agar plate containing kanamycin sulfate (Sigma Aldrich), before one colony was used to inoculate 5 mL LB supplemented with kanamycin and incubated overnight. The plasmid pJE401-AquaRepCT was then purified from this culture using the Monarch plasmid miniprep kit and standard manufacturer's protocol (NEB) then eluted into 50 µL nuclease free water.

4.2.2.2 Transformation

Chemical transformation was used to insert 75 ng of pJE401-AquaRepCT into 40 µL chemically competent BL21 *E. coli* cells (NEB). Cells were defrosted on ice before being incubated with DNA on ice for 30 minutes. Cells were then placed at 42^{o}C for 30 seconds before being placed on ice for 5 minutes. Pre-warmed LB (37^{o}C) was added (1 ml), then transferred to a 37^{o}C shaking incubator, shaking at 250 RPM for 1 hour.

In tandem, 75 ng of pJE401-AquaRepCT was also transformed into 40 µL electrocompetent *E. coli* DL41 *E. coli* cells (CGSC, Yale) using the Bio-Rad MicroPulser Electroporator at 2.5 kV using a single pulse. Cells were electroporated at 1700 V, and 1 mL pre-warmed LB added before culture was incubated at 37^{o}C, shaking at 250 RPM for one hour. Transformation reactions were then removed from the incubator and 50 µL cells plated onto LB agar + kanamycin and incubated overnight at 37^{o}C.

4.2.2.3 Small-Scale Expression Test

For both transformed cell lines, the expression test was carried out in the same way. Five colonies were randomly picked from overnight plates and used to inoculate 5 mL LB + kanamycin. This was incubated overnight. The cultures were then removed and 250 µL placed into 50 mL fresh LB broth + kanamycin and incubated until OD_{600} = 0.7. Cultures were then split into two 25 mL flasks and one of these induced with 1 mL of

1 M Isopropyl β-D-1-thiogalactopyranoside (IPTG) to create a final concentration of 1 mM. Control cultures were grown in the same way, but IPTG was not added at this step. BL21 cultures were then incubated at 20^{o}C overnight, with shaking at 250 RPM. DL41 cultures were incubated at 30^{o}C for four hours, shaking at 250 RPM. After this step, culture samples were OD corrected to 1 and centrifuged at 13000 RPM for 10 minutes on a benchtop centrifuge, before being frozen at -20^{o}C.

4.2.2.4 SDS-PAGE analysis (standard protocol)

SDS-PAGE gels were made according to the standard methods and were comprised of a stacking gel layer (120 mM Tris pH 6.8, 0.1 % w/v SDS, 5 % acrylamide polymerised with 0.1 % (w/v) ammonium persulfate (APS) and 5 µL TEMED) set on top of a resolving gel (375 mM Tris pH 8.8, 0.1 % w/v SDS, 12 % acrylamide polymerised with 0.1 % (w/v) APS and 4 µL TEMED). For analysis of induced cultures, sample pellets were thawed and 200 µL CellLytic (Sigma), 1 µL lysozyme (50 mg/ml), 0.5 mL benzonase and 3 µL $MgCl_2$ (1M) added and resuspended. Samples were centrifuged at 13000 *g* for 10 minutes before removal of the supernatant, constituting the soluble fraction, and resuspension of the remaining pellet in 20mM Tris(hydroxymethyl)aminomethane, pH 8 (200 µL). Reducing loading buffer (100 µL) was added to all samples, incubated at 95^{o}C for 10 minutes and 12 µL loaded onto the prepared SDS-PAGE gel. To analyse the un-induced sample, cells were lysed and resuspended as before. Reducing loading buffer was added as before and tubes incubated at 95^{o}C for 10 minutes, then loaded onto the SDS-PAGE gel. Gels were electrophoresed at 180 V for 10 minutes and then 200 V for approximately 45 minutes, or until the gel front reached the bottom of the gel, using a PowerPac 200 (BIORAD). Gels were stained by washing with deionised water and warmed gently in the microwave for 2 minutes, water was then removed and the process repeated 3 times. Simply Blue GelSafe gel stain (ThermoFisher) was then added and heated gently in the microwave for 3-4 minutes before stain was removed. The gel was washed with deionised water, heated gently in a microwave for 2 minutes and visualised using a GelDoc system (BioRad).

4.2.2.5 Litre- scale Expression Test in LB

One colony of DL41 or BL21 containing pJE401-AquaRepCT was used to inoculate 5 mL LB + kanamycin and grown overnight. Each starter culture was added to 1 litre LB + kanamycin and incubated at 37^{o}C, shaking at 180 RPM until OD_{600} reached 0.7.

Cultures were induced with a final concentration of 1 mM IPTG and incubated overnight at 20°C, shaking at 250 RPM. Cultures were centrifuged at 4500 *g* (Beckmann J25 using JA10 rotor) for 12 minutes and pellets suspended in Buffer A before being frozen at -80°C. Pellets were thawed and sonicated at 15 µM for 5 minutes on ice in 30 second bursts of power and rest (MSE Soniprep 150 Sonicator (Sanyo)). These were then centrifuged at 35000 *g* for 30 minutes (Beckmann J25 using JA25.50 rotor), and the supernatant analysed using Immobilized Metal Ion Affinity Chromatography (IMAC) (using the AKTA Pure and UNICORN Software). Supernatants were loaded onto column packed with Ni^{2+} Sepharose and equilibrated with Buffer A. Buffer A was used to wash the column for approximately 20 column volumes (CV) before bound proteins were eluted with Buffer B in a step-up manner of 5, 10, 15, 40 and 100 %. The insoluble protein fraction was dissolved in 10 mL 8 M urea and diluted 1 in 10 before being analysed with the column wash and IMAC fractions by SDS-PAGE, using conditions described previously.

4.2.2.6 Expression in Minimal Media

Overnight DL41 cultures were prepared as described previously and used to inoculate 1 litre of M9 minimal media supplemented with kanamycin, methionine (100 µg/ml) biotin (100 µg/ml) and thiamine (100 µg/ml) (Geerlof, 2010). This was grown shaking at 37°C until the OD_{600} reached 1.0. Cultures were induced with a final concentration of 1 mM IPTG and incubated overnight at 37°C, shaking at 250 RPM. Cultures were then processed and the protein purified as previously described before analysis using SDS PAGE.

4.2.3 pET32b-AquaRepCT

4.2.3.1 pET32b-AquaRepCT Creation

E. coli BL21 pJE401-AquaRepCT cells were grown overnight in 5 mL LB + 5µL kanamycin and the plasmid purified using the Sigma GenElute plasmid miniprep kit

(manufacturer's protocol). Cells were also used to make 40 % glycerol stocks and stored at -80^{o}C. Purified plasmid was then digested using NdeI and XhoI for 30 minutes at 37^{o}C and heat inactivated at 65^{o}C for 20 minutes. The empty cloning vector pET32b (Novagen) was cut using XhoI for 60 minutes at 37^{o}C and heat inactivated at 60^{o}C for 20 minutes. BamHI (1 µL) was added to both the digested pJE401-AquaRepCT plasmid and digested pET32b plasmid and incubated for 30 minutes at 37^{o}C. Digested products were run on a 1 % agarose gel for 45 minutes and then viewed using a transilluminator. The correct bands (814 bp AquaRepCT fragment and 5860 bp pET32b) were excised from the gel using a scalpel and purified using the Genelute gel extraction kit (Sigma) and eluted in deionised water.

The gel-excised pET32b vector and AquaRepCT insert were ligated by adding 1 µL DNA ligase and incubated at room temperature (RT) for 20 minutes, then chemically transformed into Rosetta (DE3) cells. The standard chemical transformation protocol used was as follows: 2-5 µL DNA was added to 50 µL Rosetta (DE3) (Novagen) cells and incubated on ice for 30 minutes. The Rosetta *E. coli* strains are designed to enhance expression of proteins by providing tRNAs for rarely used *E. coli* codons, allowing universal translation that is not limited by *E. coli* tRNA usage. This was then heated at 42^{o}C for 30 seconds before being cooled on ice for 2 minutes. Pre-warmed SOC media (250 µL) was then added and the reaction mixture incubated at 37^{o}C, shaking at 180 RPM for 1 hour before plating onto LB+100 µg/mL ampicillin. To check the ligation was a success, 5 µL ligation mix was also used to run on a 1 % agarose gel for 45 minutes. Three colonies from the transformation were then picked, and either made into glycerol stocks or miniprepped and the sequence validated using Sanger sequencing using primers 'T7_Promotor' and 'T7_Terminator' (Table 4.C).

Sequencing revealed that the translated gene sequence was out of frame in relation to the promotor, so site directed mutagenesis was carried out using primers Q5SDM_INS_F and Q5SDM_INS_R; designed to insert 1 bp (adenine) after the BamHI site using the NEBase changer protocol and Q5 mutagenesis kit (NEB). Briefly, primers, purified pET32b-AquaRepCT and Q5 reagents were mixed, and PCR program A carried out (Table 4.D). For phosphorylation, ligation and template removal, 1 µL PCR product was added to 1 µL KLD enzymes and incubated at Room Temperature

(RT) for 10 minutes. This was then incubated on ice for 30 minutes before chemical transformation into 50 µL NEB5α cells, chosen as they are T1 phage resistant and *endA* deficient, which allows high quality plasmid preparation and storage. Transformed cells were plated onto LB agar + 100 µg/mL ampicillin and incubated at 37°C overnight. Four colonies were picked, grown in 5 mL LB + ampicillin at 37°C overnight, shaking at 180 RPM before the plasmid was purified and was sequenced by Macrogen using primers 'T7_Promotor' and 'T7_Terminator' (Table 4.C). For storage, 40 % glycerol stocks were also made and stored at -80°C.

Table 4.C: Primers used in this study.

Primer name	Primer Sequence
T7_Promotor	TAATACGACTCACTATAGGG
T7_Terminator	GCTAGTTATTGCTCAGCGG
Q5SDM_INS_F	aGGTGGCAACGGTGGCATC
Q5SDM_INS_R	GGATCCGATATCGCCATGG
Q5CT_3/8/19_F	ACCCGTAATACTGCGTAC
Q5CT_3/8/19_R	TGGATCCGATATCGCCAT
Q5SDM_8/21/2020_F	ttttcagggcACCCGTAATACTGCGTAC
Q5SDM_8/21/2020_R	tacaggttttcTGGATCCGATATCGCCAT

Table 4.D: PCR Programs A and B.

PCR program	Step	Temperature (°C)	Time (seconds)	
A	1	98	30	
	2	98	10	30 cycles
	3	66	20	
	4	72	180	
	5	72	120	
B	1	98	30	
	2	98	10	25 cycles
	3	64	20	
	4	72	210	
	5	72	120	

4.2.3.2 Expression Test

Purified plasmid clone I1 (in-frame pET32b-AquaRepCT) was chemically transformed into Rosetta cells (using the standard protocol, except cells were rescued with 950 µL RT SOC) and incubated overnight at 37°C. Six colonies were picked into 5mL LB + ampicillin and grown overnight at 37°C, shaking at 180 RPM. For the expression test, 1 mL of overnight culture was used to inoculate 100 mL LB + ampicillin and grown until OD_{600} ≈0.6 at 37°C, shaking at 280 RPM (approximately 3 hours). Cultures were split into two 50 mL flasks and one flask induced with 1 mM IPTG. Both flasks were incubated at 20°C for 4 hours, shaking at 180 RPM. Cultures were diluted to an OD_{600} of 1 in a final volume of 1 mL and used for analysis as described in section 4.2.2.4. Samples were centrifuged at 13000 RPM for 1 minute and supernatant removed before pellets were frozen at -20°C overnight. Samples were processed and analysed using SDS-PAGE as described in section 4.2.2.4.

4.2.3.3 Litre-scale LB AquaRepCT Expression and Purification

To create a starter culture, 1 colony of pET32b-AquaRepCT was used to inoculate 5 mL LB + ampicillin and grown overnight. For the main growth, 2.5 mL overnight culture was added to 1L LB + ampicillin and grown at 37°C until OD_{600} reached 0.7 (approximately 5 hours). Culture was then induced and purified using IMAC as previously described.

4.2.3.4 Protein Concentration and Cleavage

Fractions containing AquaRepCT from purification were pooled and dialysed against Buffer C overnight at 4°C in SpectraPor molecular porous membrane tubing (6-8 kDa MWCO). Sample concentration was measured using a Nanodrop and further concentrated using a 10,000 MWCO filtration column, centrifuged at 10-minute intervals at 4200 RPM. Total yield of protein was approximately 75 mg/l. Protein was then flash frozen in liquid nitrogen before storage at -80°C. For cleavage, 500 µL of 3 mg/mL AquaRepCT was added to 50 µL thrombin and incubated at 37°C, shaking at 110 RPM, for thrombin mediated removal of solubility tags. Complete cleavage was confirmed using SDS-PAGE analysis (Mw thioredoxin tag 14kDa, Mw AquaRepCT ≈27kDa).

4.2.3.5 Solubility Tests

To test solubility during heat exposure, 3.3 mg/mL aliquots of TRX-AquaRepCT were rapidly defrosted and centrifuged briefly to check for precipitation. Soluble protein was added 1 : 5 protein : buffer and pipetted 4 times to mix, carried out by Dr David Harvey, University of Nottingham. Samples were then centrifuged at 13000 RPM x 20 mins and supernatant analysed spectrophotometrically at 280 nm using the Nanodrop (Thermo), to see if the concentration of the protein in solution was decreased.

4.2.4 pET32b-AquaCT

4.2.4.1 pET32b-AquaCT Vector Design

Primers were designed to delete the repetitive region of pET32b-AquaRepCT using the Q5 mutagenesis kit (NEB). Primers Q5CT_3/8/19_F and Q5CT_3/8/19_R (Table 4.C) were mixed with the standard reagents and 1 µL of purified primer before PCR program B was carried out (Table 4.D). PCR product (1 µL) was KLD treated and incubated for 10 minutes at 25^{o}C and then 30 minutes on ice. The reaction mix (5 µL) was transformed into one 50 µL aliquot of *E. coli* NEB5α cells by incubating on ice for 30 minutes before placing it at 42^{o}C for 30 seconds. The standard transformation protocol was then followed before plating onto LB + ampicillin and overnight incubation.

One colony was picked from the transformation plates and grown in 5 mL LB + ampicillin overnight. Plasmid was then purified as described previously and the sequence verified using the primer T7_Promotor and Sanger Sequencing (Macrogen). The pET32b_AquaCT plasmid was used to transform *E. coli* Rosetta using the standard transformation protocol and stored as 40 % glycerol stocks at -80^{o}C.

4.2.4.2 AquaCT Expression

Rosetta *E. coli* containing pET32b-TRXAquaCT was used to inoculate 10mL LB + ampicillin, and grown overnight at 37°C, shaking at 180 RPM. This was then used to inoculate 1 litre of LB + ampicillin and incubated at 37°C, shaking at 180 RPM until OD_{600} was between 0.7-1 (approximately 4hrs). Protein expression was induced with 1 mM IPTG and the culture incubated at 20°C for 15hrs, shaking at 180 RPM. The cell

suspension was centrifuged at 4500 *g* (Beckmann J24 using JA10 rotor) for 12 minutes and supernatant discarded. Cell pellets were re-suspended in 15 mL Buffer A1 (see Table 4.A) and frozen at -80°C.

4.2.4.3 AquaCT Purification

Cell suspensions were defrosted and lysed by sonication at 15 µM amplitude for a total of 5 minutes at 4°C. Bacterial lysate was then centrifuged at 35000 *g* for 30 minutes at 4°C (Beckmann J25 using JA25.50 rotor). The soluble fraction (supernatant) was collected and loaded onto a pre-equilibrated Ni^{2+} affinity column (Hi-Trap chelating HP, GE) using an AKTA start with UNICORN software. Buffer A1 was used to wash the column and a gradient of Buffer B used to elute the purified protein. Purified AquaCT was transferred to dialysis membrane tubing (Spectrum Spectra/por 6-8 kDa MWCO) and dialysed in Buffer C1 (Table 4.A) overnight. Protein concentration was then determined by spectrophotometry (Nanodrop) and if necessary concentrated using a spin column (3000 kDa MWCO, VivaSpin 20). AquaCT was analysed by SDS-PAGE using pre-cast NuPAGE 10 % Bis-Tris protein gels (Invitrogen) using Tris-Glycine buffer.

Thrombin was added at a ratio of 1 : 1000 to 1 mL of 3 mg/mL protein in Buffer C1 supplemented with 1.4 mM $CaCl_2$ in order to cleave the solubility tag. Cleavage reactions were left at 37°C or RT for up to 2 hours, shaking gently at 20 RPM before analysis by SDS-PAGE.

4.2.4.4 Large Scale Expression, Cleavage and Tag Separation

Induced culture of *E. coli* Rosetta containing pET32b-TRXAquaCT (1 litre) was centrifuged and frozen according to the previous protocol (section 4.2.4.2). This pellet was defrosted and 1 tablet complete EDTA-free protease inhibitor added, before sonication using the standard protocol. Lysate was centrifuged at 35000 RPM for 30 minutes before the supernatant was loaded onto a 5 mL HiTrap column using a peristaltic pump at a flow rate of 2 ml/min. The column was washed with 10 Column Volumes (CV) Buffer A1 and protein eluted using 2 CV Buffer B. The eluate was then passed through a PD-10 desalt column (GE healthcare) to remove imidazole, and eluted into Buffer C. This was pooled and 1.4 mM $CaCl_2$ and 550 µL thrombin added, before incubation at 37°C for 45 minutes, shaking at 20 RPM. A reverse His- protocol

was performed using a HiTrap column equilibrated with Buffer C to remove the thioredoxin tag and the flowthrough was collected. One half of a protease inhibitor tablet (cOmplete EDTA-free protease inhibitor (Sigma Aldrich)) was added to inactivate thrombin and the results analysed on an SDS-PAGE gel. The Rev-HIS flowthrough (containing cleaved protein) and Buffer A wash was pooled and AquaCT concentrated using a spin column (VivaSpin 5000 MWCO) at 45000 RPM x30 minutes until the concentration was at least 5.5 mg/ml.

4.2.4.5 Size Exclusion Chromatography and Analysis

Protein sample was concentrated to 5 mL total volume and loaded onto a HiLoad 26/60 Superdex 75 preparation grade column for size exclusion overnight in Buffer C1, using the standard protocol. Samples were then analysed using SDS-PAGE under denaturing conditions. Briefly, a Mini-Protean TGX Pre-cast gel was electrophoresed for 10 minutes at 200 V (without samples) in Tris Glycine Running buffer (25 mM Tris, 192 mM glycine, 0.1 % SDS). Samples were then loaded and the gel re-electrophoresed at 300 V for 15 minutes, before the gel was washed in deionised water and stained using InstantBlue Coomassie stain (Abcam) for 15 minutes and then imaged using a Gel doc system (BioRad).

Protein was then split into two samples, A and B, before being concentrated using a Vivaspin 3,000 MWCO concentrating column. Thrombin was added to Sample A at 1:1000 and incubated for 1hr at room temperature. Samples were then analysed on a denaturing gel as previously described with and without a reducing agent (β-mercaptoethanol).

4.2.4.6 Protein Crystallography

Hampton Research Index HT (Batch ID 214448, 13/06/19) and Morpheus HT-96 (Batch 039-1-46, 13/06/19) crystallisation screens were used in a sitting drop format, using 10 nl of sample A or B per well, at 3 mg/mL and 5.4 mg/mL (Sample A) or 5.7 mg/mL (Sample B) concentrations. Plates were incubated at 20°C and checked periodically.

4.2.4.7 Mass Spectrometry

For mass spectrometry, 20 µL protein sample was mixed with 3 % methanol and 1 % formic acid before being dialysed with distilled H_2O through a membrane (SpectraPor) for 30 minutes. Sample was then mixed with 20 µL acetonitrile and 2 % formic acid before being loaded onto a Micromass Q-tof micro-Mass spectrometer for exact mass MS/MS, carried out by technical staff at the Research Complex at Harwell (RCaH).

4.2.4.8 Analytical Ultra Centrifugation

A Sedimentation Velocity (SV) Analytical Ultra-Centrifugation (AUC) run was carried out on a Beckman Optima XLI Ultracentrifuge. The density and viscosity of the running buffer, Buffer C1, was measured using an Anton Parr Density and Viscosity meter and then the cells were loaded into the AUC with assistance from Ryan Lithgo (University of Nottingham). Absorbance measurements were recorded and translated using ProteomLab v6.2 and the data analysed using SEDFIT.

4.2.5 *Size Exclusion Chromatography Multi Angle Light Scattering (SEC-MALS)*

Size exclusion chromatography was performed using a Superdex 200 10/300 GL column (GE healthcare). Column effluent flowed through an inline UV detector, Dawn Heleos Multi Angle Light Scattering (MALS) detector and a rEX refractive index detector. Data were analysed in ASTRA (Wyatt). This was carried out by technical staff at the Research Complex at Harwell (RCaH).

4.2.6 Additional Construct Generation

4.2.6.1 TEV-AquaCT

A TEV protease recognition site (ENLYFQG) was inserted into the pET32b-AquaCT construct using primers Q5SDM_8/21/2020_F and Q5SDM_8/21/2020_R (Table 4.C) and PCR program B was run (Table 4.D). The standard Q5 site-directed mutagenesis (NEB) protocol was followed for subsequent steps. Transformed *E. coli* were grown in 5 mL LB + ampicillin overnight, with shaking, and plasmid was retrieved using the Sigma Genelute plasmid miniprep kit and stored at -20^{o}C or chemically transformed into *E. coli* NEB5α cells and stored at -80^{o}C. The vector map and sequences can be found in Appendix A.4.4.

4.2.6.2 NTMaSp/AcSp

Using sequences from the *Argyroneta aquatica* transcriptome created during the course of this thesis (Chapter 3), two N-terminal regions were extracted and designed to be expressed in a pET32b plasmid, with an included TEV site. Two NTDs were selected, one from a MaSp sequence and another from an AcSp sequence, these sequences can be seen in Appendix A.4.5. Sequences were ordered from Aruru molecular, delivered in a modified pET32b vector and then used for further cloning. The vector map can be seen in Appendix A.4.5. Further discussion of the transcriptome and methods used to generate these sequences can be found in Chapter 3.

4.2.6.3 NTXX-AquaRepCT

The full-length construct pET32b-NTXX-AquaRepCT (NTXX can be either NTMaSp or NTAcSp) was created using the existing pET32b-AquaRepCT and the newly created pET32b-NTXX constructs in a ligation reaction.

1 µg total DNA (pET32b-AquaRepCT and pET32b-NTXX) was digested using XhoI and NcoI enzymes in Cutsmart buffer, for 30 minutes at 37^{o}C and then 15 minutes at 80^{o}C. Products were loaded onto a 1 % agarose gel and electrophoresed for 30 minutes at 150 V, before the appropriate bands were excised and purified using the Genelute gel purification kit. Purified products were mixed in a 5 : 1 insert: vector ratio and placed into a ligation reaction using T4 DNA ligase for 10 minutes at room temperature, before heat inactivation at 65^{o}C for 10 minutes. 5 µL Ligation mixture was then transformed into 50 µL NEB5α competent cells using the standard protocol. Plasmid was then miniprepped using the standard protocol and some stored in NEB5α cells at -80^{o}C. The full length NTXX-AquaRep-CT insert sequence is shown in Appendix A.4.5. A sample of plasmid was sent to RL at the Appleton Rutherford Laboratories where NTMaSp-AquaRep-CT was expressed and purified according to the protocols used for AquaCT.

4.3 Results

4.3.1 pJE401-AquaRepCT

The modified pJE401 vector (pJE401-AquaRepCT) encoding the fusion protein Thioredoxin-(Histidine)$_6$-AquaRepCT was transformed into two strains of *E. coli*, BL21 and DL41 (CGSC, Yale). The vector map and sequence information can be seen in Appendix A.4.1. A schematic of AquaRepCT is shown in Figure 4.4.

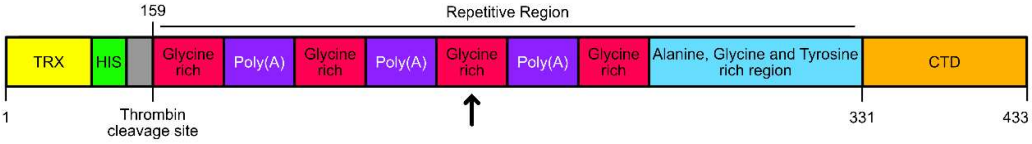


Figure 4.4: AquaRepCT schematic.
A thioredoxin tag (TRX) is followed by a Poly-histidine tag (HIS) before a linker region (grey) connecting the AquaRepCT repetitive region and the C-terminal domain (CTD, orange). The arrow indicates the region which contains more negatively charged, polar amino acids in comparison to 4RepCT; it is in this area, and the area coloured in turquoise that are most diverged when compared to the sequence of 4RepCT.

The fusion protein TRX-AquaRepCT has a molecular weight of 37 kDa, which was validated by sequencing of the vector insert and subsequent calculation of the molecular weight using an online calculator (EXpasy).

4.3.1.1 Small-scale Expression Test

Figure 4.5 shows an SDS-PAGE gel from the small-scale expression test. The protein of interest, AquaRepCT, has a molecular weight of approximately 37 kDa. There is no evident band at this molecular weight present exclusively in the induced culture lanes.

4.3.1.2 Litre-scale Expression Test and Purification

A large scale (one litre) expression test in both BL21 and DL41 strains of *E. coli* was then carried out, as anecdotal evidence suggested that protein production is accelerated in large scale cultures, which may be due to the increased oxygen which can be introduced to the cultures in baffled shake flasks, rather than in 50 mL tubes, as well as making protein purification more effective. The resulting protein was

purified using Immobilized Metal ion Affinity Chromatography (IMAC). The resulting fractions containing protein were analysed using SDS-PAGE, shown in Figure 4.6.

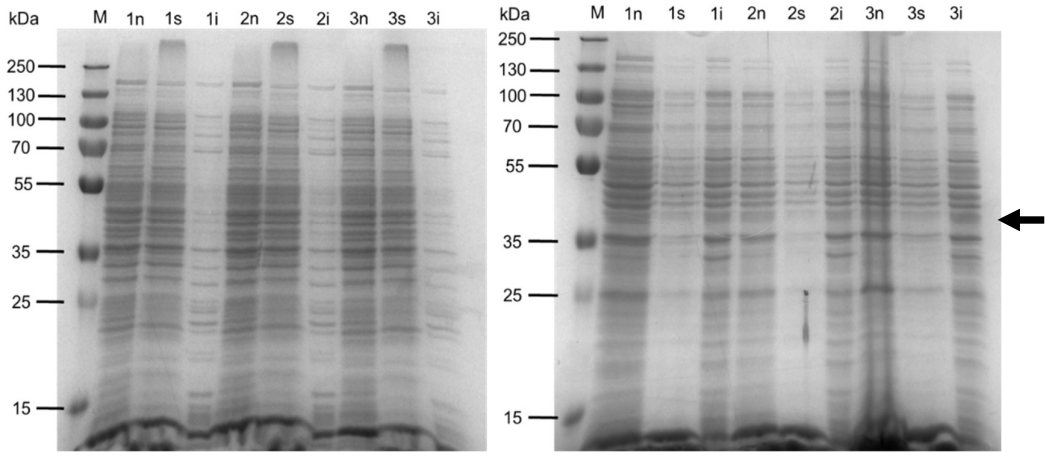


Figure 4.5: SDS PAGE gel from small scale expression test of pJE401-AquaRepCT.
This gel shows non-induced, soluble protein and insoluble protein fraction from 3 example colonies of BL21 (left panel) and DL41 (right panel) cells transformed with the plasmid pJE401-AquaRepCT. Numbers designate separate colonies, 'n' indicates the non-induced culture, while 'i' indicates the insoluble protein fraction and 's' indicates soluble protein fraction of the induced cultures. 'M' shows the marker, PageRuler Plus (NEB). The arrow indicates the area where a band would be expected for AquaRepCT in the induced lanes.

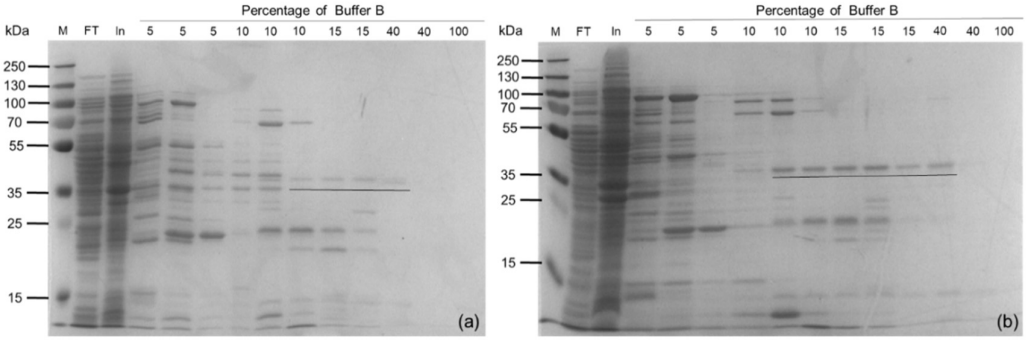


Figure 4.6: SDS PAGE gel from large scale expression test of pJE401-AquaRepCT.
This gel shows the fractions obtained from IMAC purification of induced BL21 (a) and DL41 (b) cultures. 'M' shows the marker, PageRuler Plus (NEB), 'FT' indicates a sample of the flow-through from the column as supernatant was added and 'In' indicates the insoluble fraction. The protein of interest, AquaRepCT, has a molecular weight of approximately 37kDa, at which size bands can be seen in both gels. Faint bands at this expected size are evident in (a) at 10-40% Buffer B elutions, which have been indicated by underlining with a solid black line. In (b) more evident bands can be seen at 0-40% Buffer B elutions. The bands seen in both gels at approximately 70kDa could represent AquaRepCT dimers.

In the purification of both induced BL21 and DL41 there are evident bands at the appropriate molecular weight for AquaRepCT, 37 kDa, eluted at concentrations of 70-

280 mM of Imidazole. The cell line DL41 produced slightly more protein than BL21. The bands are smaller than those obtained by expression of the miniature spidroin 4RepCT at this scale (data not shown), showing that low expression of AquaRepCT is obtained when the expressed in DL41 or BL21 cells grown in LB media.

4.3.1.3 AquaRepCT Expression in M9 Minimal Media

DL41-pJE401-AquaRepCT was then grown in one litre of defined minimal media, and the resulting protein purified and analysed using IMAC and SDS-PAGE. Successful expression in these cells, grown in minimal media, shows potential for downstream labelling of the protein with ^{35}S-methionine and selenomethionine, a click chemistry reaction which has been successfully employed with 4RepCT (Harvey *et al.*, 2017). Protein analysis gels are shown in Figure 4.7.

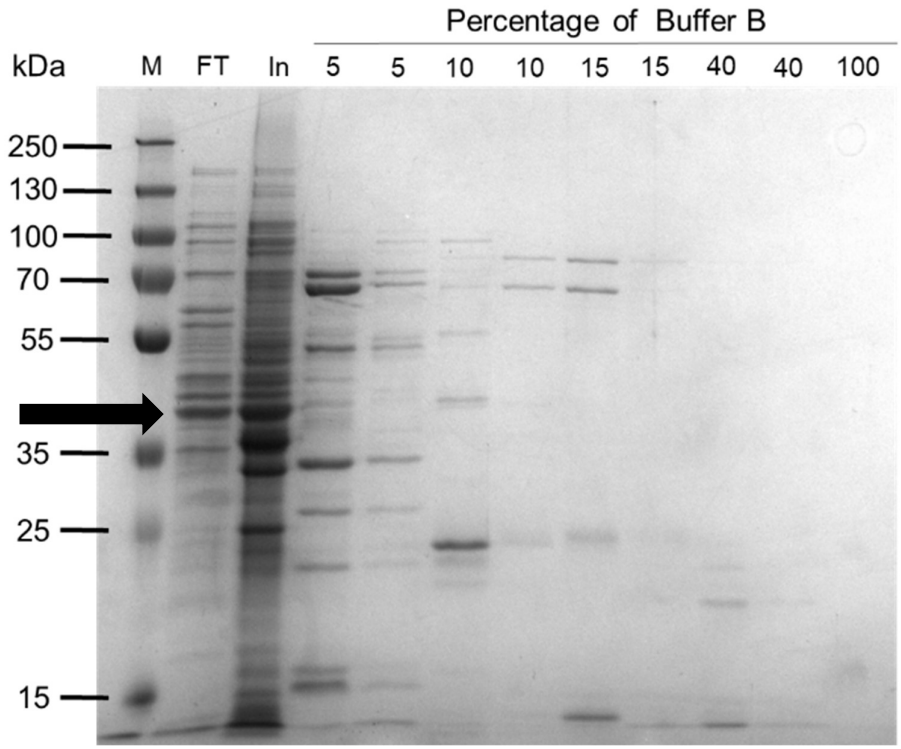


Figure 4.7: SDS PAGE gel from large scale expression test of pJE401-AquaRepCT grown in 1L minimal media.

'M' shows the marker, PageRuler Plus (NEB), 'FT' indicates a sample of the flow-through from the column as supernatant was added and 'In' indicates the insoluble fraction. AquaRepCT should give a band at 37 kDa, but this is not visible in the elution fractions. There is a band at this size in the flow-through and insoluble fractions, indicated with an arrow. There are also bands visible at the expected size for AquaRepCT dimers, approximately 70kDa, in the elution fractions of 0-15% Buffer B.

Some bands are visible at the expected molecular weight of AquaRepCT (37 kDa) in the elution fractions. However, a band is seen in both the insoluble and flowthrough fractions, which suggests that the protein produced was mainly insoluble and some expressed protein was soluble but did not bind to the purification column effectively.

4.3.2 pet32b-AquaRepCT

4.3.2.1 pET32b-AquaRepCT Creation

The plasmid pJE401-AquaRepCT and the empty vector pET32b were digested using the restriction enzymes NdeI and XhoI before ligation to create pET32b containing the insert AquaRepCT. The sequence encoding the in-frame DNA sequence of AquaRepCT was amplified with the primers 'T7_Promotor' and 'T7_Terminator' (Table 4.C) and the protein sequence verified. The vector map and sequence information are supplied in Appendix A.4.2.

The sequenced Clone I1 had the highest sequencing coverage so was taken forward for protein production. The protein is comprised of 433 amino acids, had a molecular weight of 41392.47 Da, a theoretical pI of 5.18 and an extinction coefficient of 45520 M^{-1} cm^{-1} before tag cleavage (Expasy, PROTPARAM).

4.3.2.2 Comparison of pET-32b-AquaRepCT and 4RepCT

The amino acid makeup, hydropathy plots and secondary structural predictions of 4RepCT and AquaRepCT were compared using protein alignments, EXpasy and Phyre2, to highlight similarities and differences between sequences (Figures 4.8-4.11 and Table 4.E) (Gasteiger *et al.*, 2003; Kelley *et al.*, 2015).

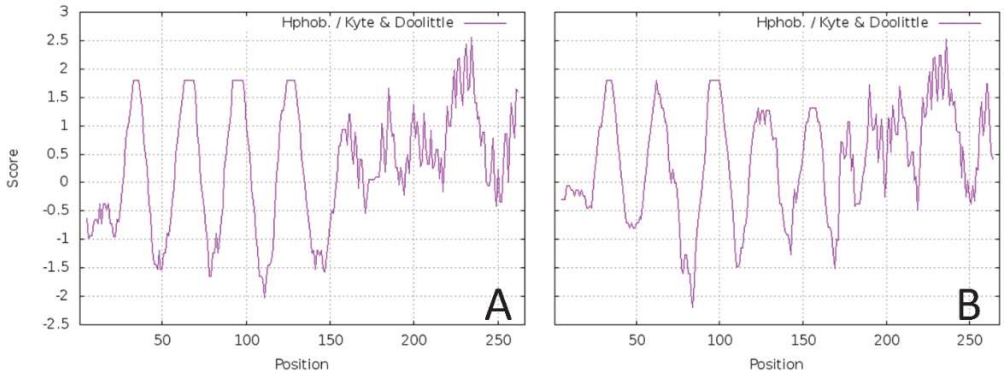


Figure 4.8: Hydrophobicity plots of AquaRepCT and 4RepCT.
AquaRepCT (A) and 4RepCT (B). Hydrophobicity plots calculated on ProtScale, using the Kyte & Doolittle scale (Gasteiger et al., 2003).

```
4RepCT        1 SGNSGIQGQGGYGGLGQGGYGQGAGSSAAAAAAAAAAAAGGQGGQGQGGYGQGSGGSAAA
AquaRepCT     1 GGNGGIGGYGGYGGLGYGGYGYGAGSSAAAAAAAAAAAAGGYGGYGYGGYGYGGGSAAA

4RepCT       61 AAAAAAAAAAAAAGRGQGGYGQGS-GGNAAAAAAAAAAAAAAGQGGQGGYGRQSQGAGSA
AquaRepCT    61 AAAAAAGAGAGAGEGAGGYRRYGGYGDDGAAAAAAAAAAAAAAAGGYGGYYYEEGGAGAG

4RepCT      120 AAAAAAAAAAAAAAGSG---Q-GGYGGQGQGGYGQSSA--SASAAASAASTVANSVSRLSS
AquaRepCT   121 AAGAAAASAAGAAATGGARRAGGYG-YGAGGAAGAAAAAGAAGGAGTRNTAYNSVSLLSS

4RepCT      174 PSAVSRVSSAVSSLVSNGQVNMAALPNIISNISSSVSASAPGASGCEVIVQALLEVITAL
AquaRepCT   180 SASNARVSTAVSGLVGNGV-D--ALSNIVSSLSASVAATHPGLSGCEVLVQVLMEIISAL

4RepCT      234 VQIVSSSSVGYINPSAVNQITNVVANALAQVL
AquaRepCT   237 VHVLSSASIGQVNVNDAAYTAALVGDVLSQNC
```

Figure 4.9: 4RepCT and AquaRepCT sequence alignment.
Protein alignment; highlighting shows sequence similarities between 4RepCT and AquaCT.

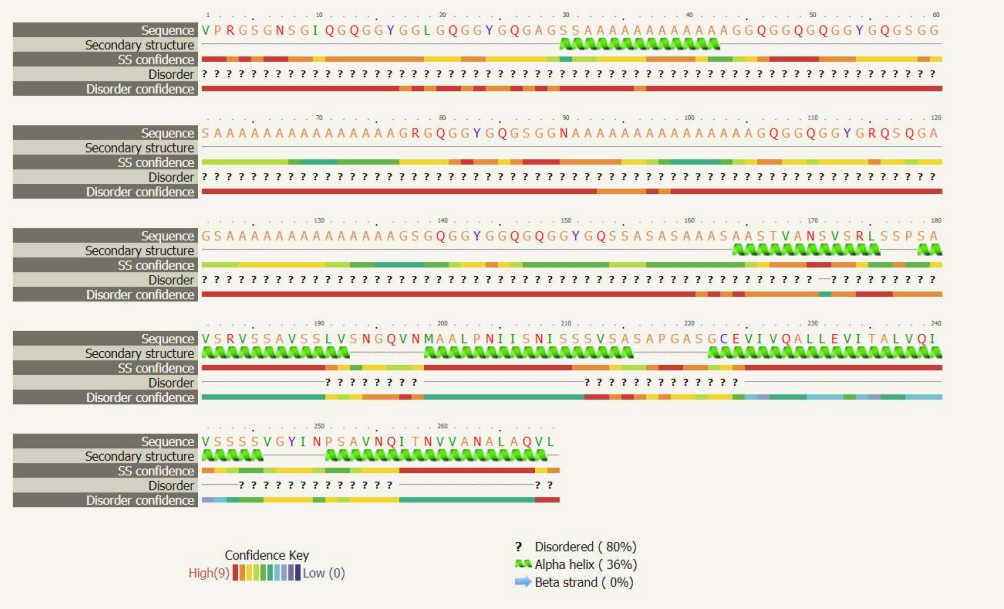


Figure 4.10: Phyre2 secondary structure prediction of 4RepCT.

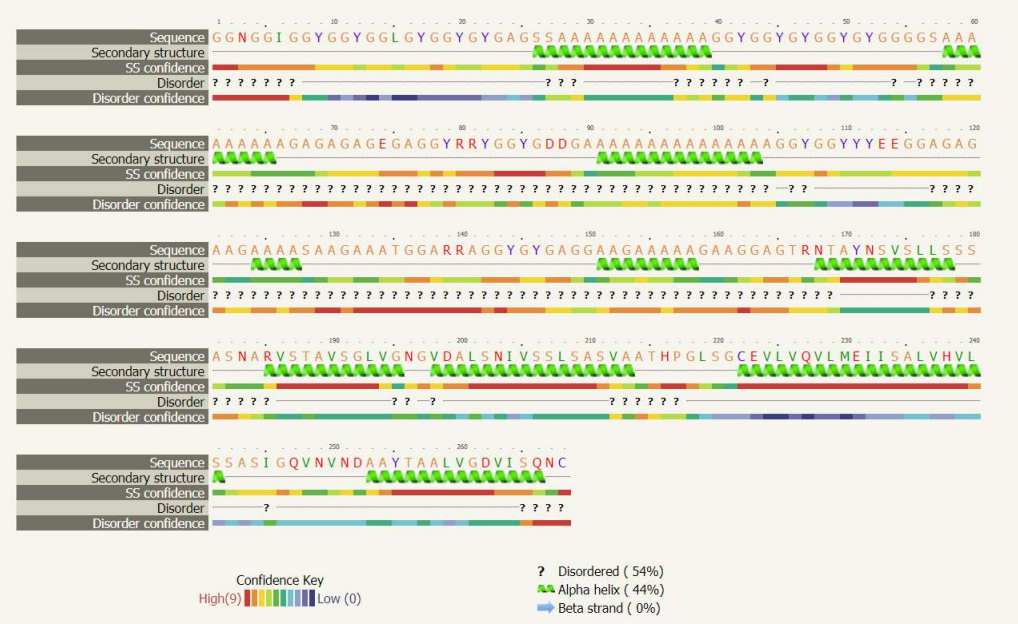


Figure 4.11: Phyre2 secondary structure prediction of AquaRepCT.

Table 4.E: Comparison of properties of AquaRepCT and 4RepCT.

Key differences are highlighted in bold and underlined.

	TRX-AquaRepCT	TRX-4RepCT	AquaRepCT (cleaved with thrombin)	4RepCT (cleaved with thrombin)
Number of amino acids	433	421	304	269
M_w (Da)	42392.47	40182.23	27470.53	23612.43
Theoretical pI	5.18	5.71	**4.55**	**8.07**
Extinction coefficient	45520	26025	31415	11920
Instability index	13.56	23.97	**13.07**	**28.09**
Aliphatic index	76.10	77.77	69.21	73.42
GRAVY	0.105	0.063	0.218	0.288
Ala (A)	22.60%	22.60%	28.3%	29.40%
Arg (R)	2.10%	1.40%	2.30%	1.1%
Asn (N)	3.00%	3.60%	3.00%	4.10%
Asp (D)	5.30%	3.80%	**3.90%**	**0.00%**
Cys (C)	0.90%	0.70%	0.70%	0.40%
Gln (Q)	1.60%	6.70%	**1.30%**	**8.60%**
Glu (E)	2.80%	2.40%	2.30%	1.1%
Gly (G)	19.6%	16.20%	23.7%	20.80%
His (H)	2.50%	1.90%	1.00%	0.00%
Ile (I)	3.70%	4.30%	2.30%	3.30%
Leu (L)	6.20%	5.50%	4.30%	3.30%
Lys (K)	3.00%	3.30%	1.00%	0.4%
Met (M)	1.60%	1.70%	1.30%	1.1%
Phe (F)	1.20%	1.20%	0.30%	0.00%
Pro (P)	2.10%	2.60%	1.00%	1.50%
Ser (S)	7.60%	10.90%	**8.60%**	**14.50%**
Thr (T)	3.20%	2.60%	2.60%	1.10%
Trp (W)	0.50%	0.50%	0.00%	0.00%
Tyr (Y)	5.30%	2.40%	**6.90%**	**3.00%**
Val (V)	5.10%	5.90%	5.30%	6.7%
Negative residue (%)	8.08	6.18	**6.25**	**1.12**
Positive residue (%)	5.08	4.75	**3.29**	**1.49**

4.3.2.3 Litre-scale Expression and Purification

A litre-scale expression of AquaRepCT using pET32b-AquaRepCT in *E. coli* Rosetta (DE3) was then carried out and showed a much higher expression level than was seen in the pJE401 vector, seen in Figure 4.12. Total yield of protein was approximately 75 mg/l and AquaRepCT was concentrated up to 40 mg/mL with no observable precipitation.

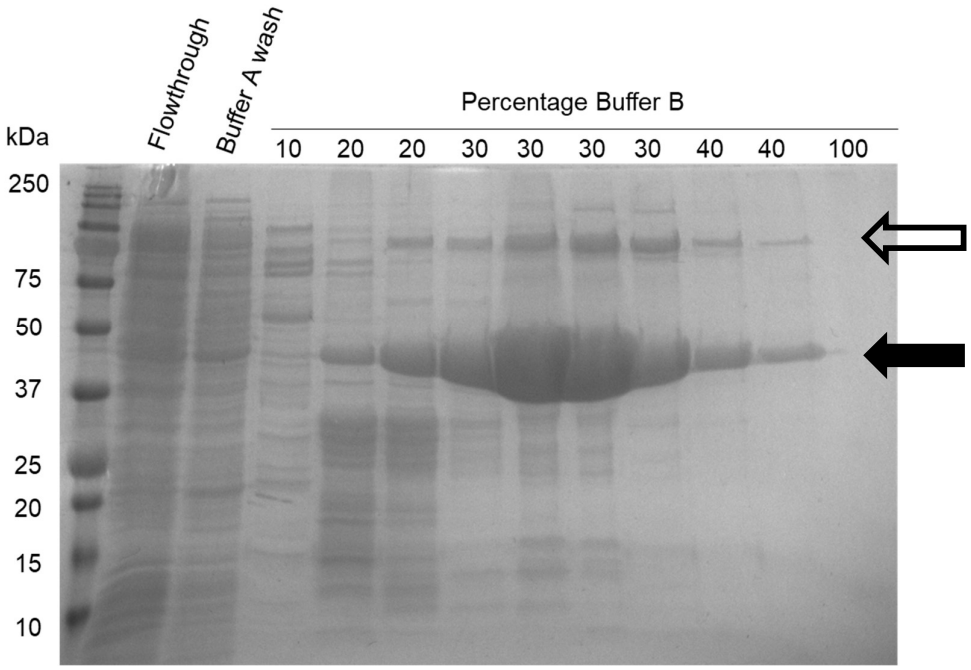


Figure 4.12: SDS PAGE gel from large scale purification of Rosetta-pET32b-AquaRepCT grown in 1L LB.

Samples of the flowthrough (effluent from column as cell supernatant was added) and when the column was washed initially with Buffer A) are indicated in lanes 2 and 3. The large band at approximately 42 kDa in the subsequent lanes represents the soluble AquaRepCT protein (solid arrow). There are also bands visible at the expected size for AquaRepCT dimers, approximately 84 kDa, in the elution fractions of Buffer B (clear arrow). This was done under non-reducing conditions.

4.3.2.4 Cleavage and Solubility Tests

Protein was then cleaved of the expression and solubility tag using thrombin (Figure 4.13). No observable fibres were formed.

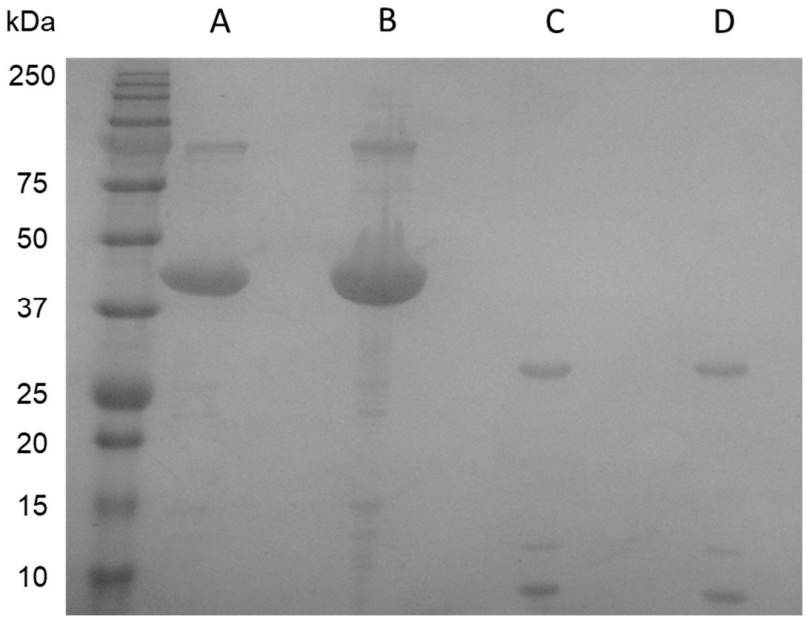


Figure 4.13: SDS PAGE gel of thrombin cleavage showing AquaRepCT.
Gel shows AquaRepCT before thrombin cleavage at 3.3 mg/mL (A) and 5.5 mg/mL (B). Lanes C and D show AquaRepCT after digestion with thrombin for 1 hour at 37°C at 3.3 mg/mL and 5.5 mg/mL respectively.

Non-cleaved protein was then assessed for its solubility in a range of buffers and the amount of portein lost during buffer transfer was assessed using microvolume spectrophotometry at 280 nm. The results are shown in Table 4.F. No precipitation was observed during defrosting of samples or when incubated in temperature ranging from 18-37°C in Buffer B, B1 or C (See Table 4.A).

Table 4.F: TRX-AquaRepCT in different buffers.

Buffer	pH range	Initial observation	After 20 mins at room temperature	Protein loss (%)
TRIS	7.0 - 9.0	Clear	Clear	0
HEPES	7.2-8.2	Clear	Clear	0
PIPES	6.1–7.5	Clear-cloudy	Cloudy	50
BIS TRIS	5.8-7.2	Clear-cloudy	Cloudy	100
MES	5.6 - 6.8	Cloudy	Cloudy	100

4.3.3 pET32b-AquaCT

4.3.3.1 pET32b-AquaCT Vector Design and Creation

The sequence encoding the in-frame DNA sequence of AquaCT was amplified with the primer T7 Promotor (Table 4.C) and sequenced to check fidelity (Macrogen). The sequence and vector map are detailed in Appendix A.4.3. The protein schematic can be seen in Figure 4.14.

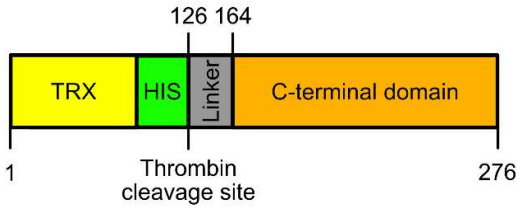


Figure 4.14: AquaCT schematic.
A thioredoxin tag (TRX) is followed by a Poly-histidine tag (HIS) before the thrombin cleavage site, connecting the C-terminal domain to the tags via a linker region (grey).

4.3.3.2 AquaCT Expression and Purification

A one litre expression of AquaCT was carried out using *E. coli* Rosetta as an expression host. Protein was purified using Ni^{2+} IMAC and all wash and elution steps were collected in 4 mL fractions. SDS-PAGE gel analysis can be seen in Figure 4.15.

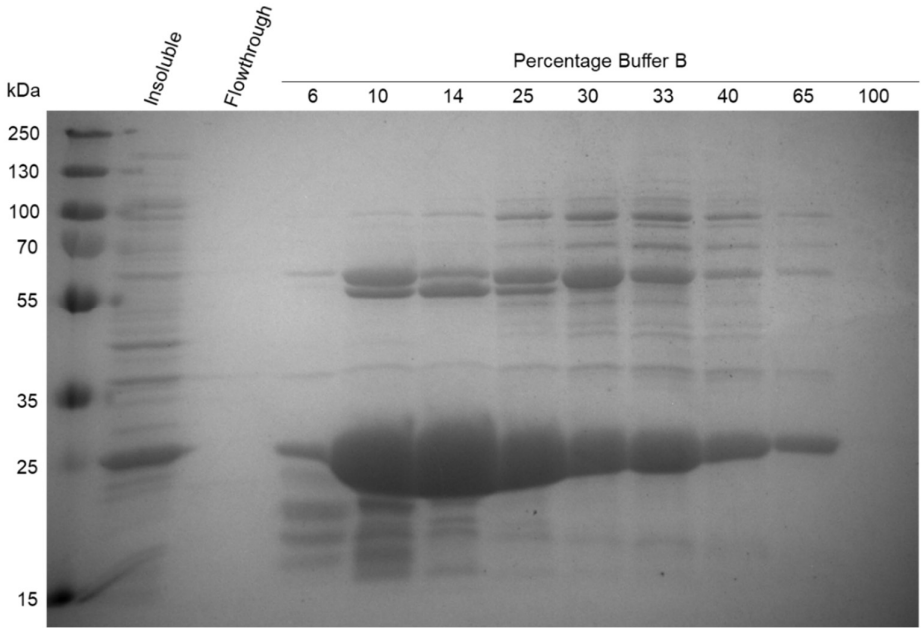


Figure 4.15: SDS PAGE gel from large scale expression of Rosetta-AquaCT cells grown in 1L LB.
Percentage of Buffer B is approximate, as samples were taken during a gradient run. TRX-HIS-AquaCT is visible as expected as a thick band at around 25 kDa in all Buffer B elutions, and a potential dimer band is also visible.

From this purification, fractions from approximately 5-40% Buffer B were pooled and dialysed against Buffer C1. Protein was concentrated to 16 mg/mL and a standard curve of AquaCT run on an SDS-PAGE gel (Figure 4.16). The protein was also freeze-thawed to check whether the band migration was affected when run on a gel and the spin concentrator flow through analysed to see if any protein was lost during concentration. Total yield of AquaCT was approximately 48 mg (concentrated to 12 mL of 4 mg/ml).

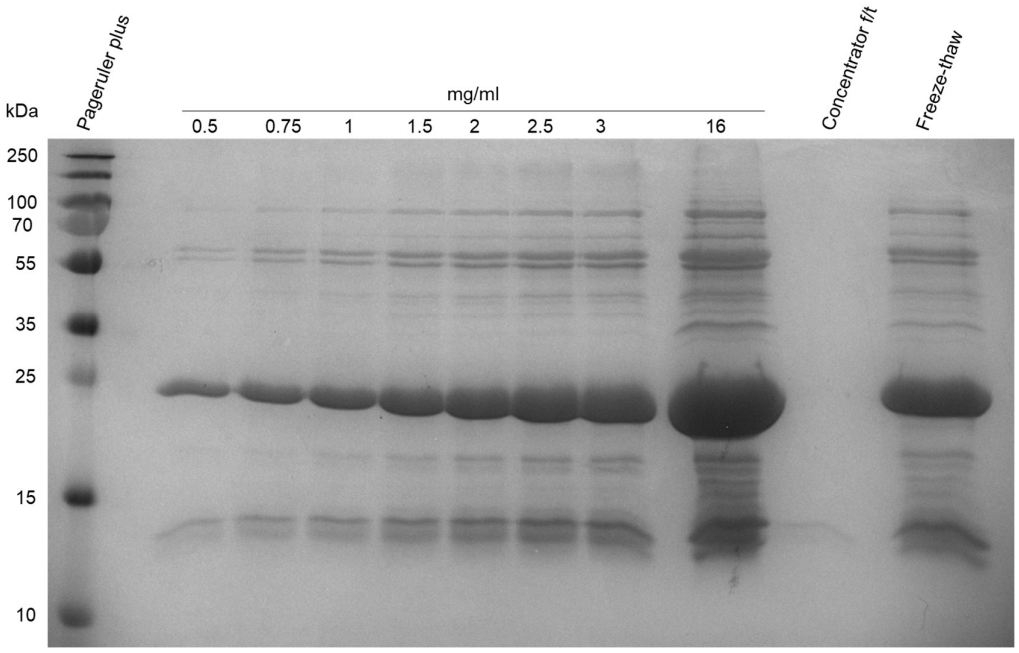


Figure 4.16: SDS-PAGE gel of TRX-HIS-AquaCT at various concentrations.
This gel shows TRX-HIS-AquaCT at various concentrations, as well as a sample of the flowthrough from the concentrator column to check if any protein was being lost through the membrane (concentrator f/t) and a sample of 16 mg/mL AquaCT frozen at -80°C and then thawed to see if the monomer/dimer state changed, shown in in the final lane.

4.3.3.3 Tag Cleavage and Separation

Thrombin was added at a ratio of 1:1000 to 1 mL of 3 mg/mL protein in Buffer C1 supplemented with 1.4 mM $CaCl_2$ to cleave the solubility tag. Cleavage reactions were left at 37°C or RT for 2 hours, shaking gently at 20 RPM and 10 µL samples taken regularly. Cleavage reactions were analysed using SDS-PAGE and was shown to be complete after 45 minutes at both temperatures. After this, the sample becomes over digested and non-specific cleavage and aggregation bands appear (Figure 4.17).

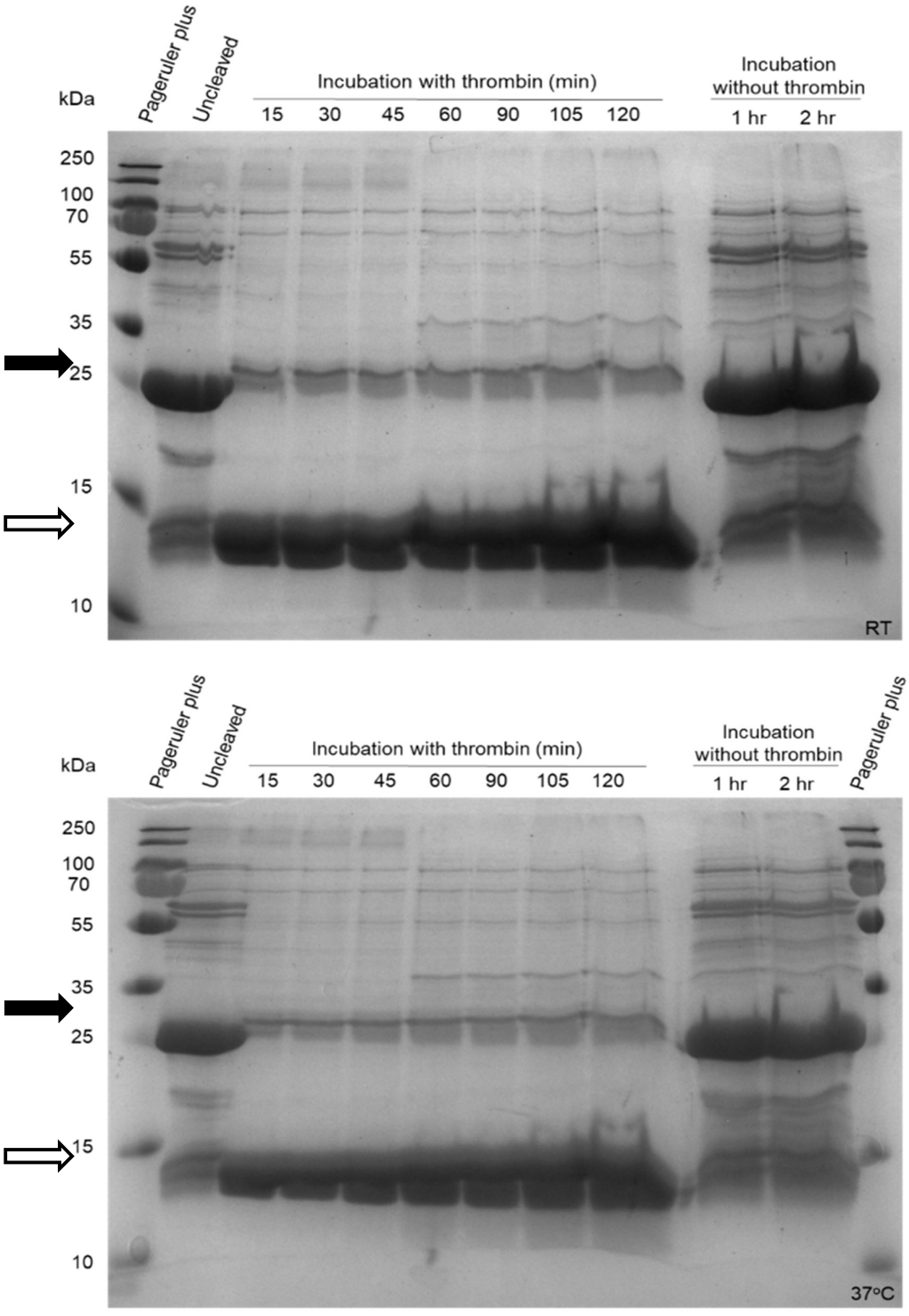


Figure 4.17: SDS-PAGE gels of TRX-HIS-AquaCT thrombin cleavage.

Cleavage was carried out at Room Temperature (RT, top panel) and 37°C (bottom panel), and a control sample without thrombin taken at 1 and 2 hours. The large band in the undigested sample shows un-cleaved TRX-AquaCT (27 kDa). In the digested sample lanes, the band present here probably represents a cleaved dimer of AquaCT (solid arrow, ~30 kDa) while the thick band around 15 kDa represents cleaved AquaCT monomer (15 kDa) and thioredoxin tag (13.7 kDa) (clear arrow). Non-specific cleavage and aggregation products can be seen from 60 minutes onwards in both gels.

4.3.3.4 Large Scale Cleavage of Tags

AquaCT was purified using a HiTrap column before thrombin cleavage and a reverse-HIS protocol carried out to remove the solubility tags. Figure 4.18 shows the purification and removal of the thrombin and HIS tags from AquaCT.

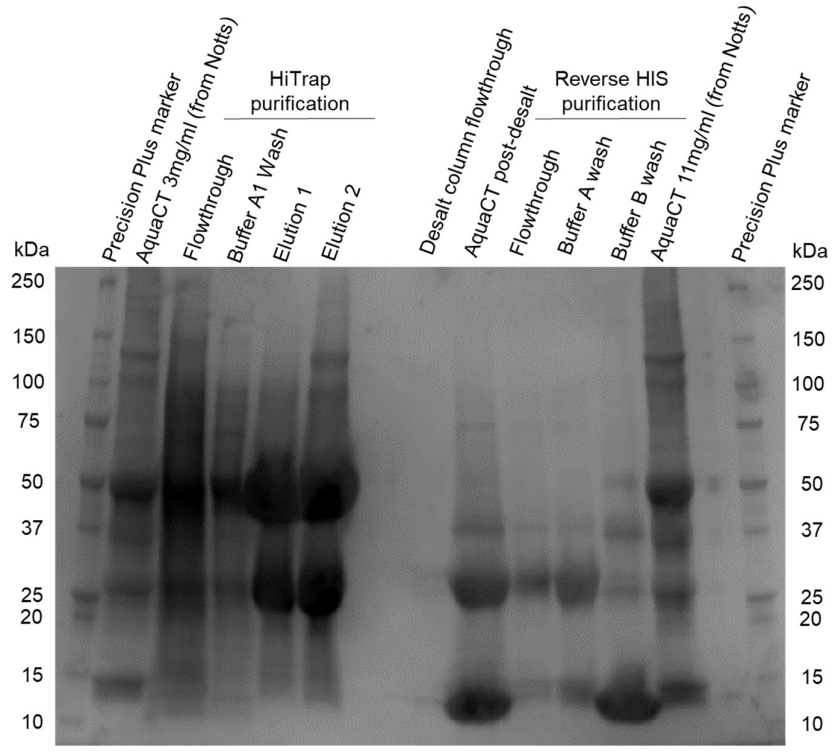


Figure 4.18: SDS-PAGE gel of purification, desalt, cleavage and rev-HIS tag separation of AquaCT. *This gel shows a sample of purified AquaCT at 3mg/ml, as well as a sample of non-specific protein (flowthrough), unbound AquaCT (Buffer A1 wash) and eluted AquaCT (Elution 1 and 2) resulting from the HiTrap purification. On the right hand side of the gel can be seen the elution fraction after processing using a desalt column (to remove imidazole) and the fractions after thrombin cleavage and reverse-HIS purification, which have enabled the separation of the thrombin tag.*

4.3.3.5 Size Exclusion Chromatography and Analysis

Protein sample was concentrated to 5 mL total volume and loaded onto a HiLoad 26/60 Superdex 75 preparation grade column for further purification using size exclusion overnight in Buffer C. Fractions were then analysed on a denaturing SDS-PAGE gel which showed two distinct bands at approximately 25 and 14 kDa (Figure 4.19).

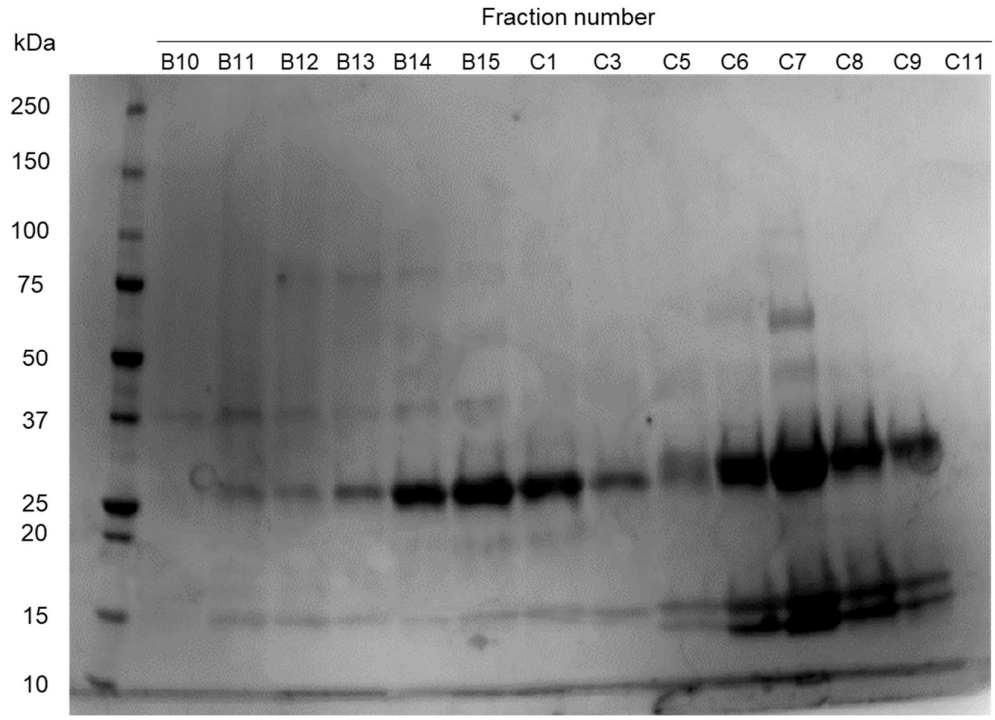


Figure 4.19: Size exclusion SDS-PAGE gel showing the purification of AquaCT.
A gradient of Buffer B was run, and samples taken from a range of concentrations. Two species are evident in fractions C2-9, while only one band is evident in fractions B15-C3, so these were split into two separate samples and analysed separately.

The bands were split into two samples for further analysis: Sample A containing the suspected dimer (pooled fractions B15, C1, C2, and C3) and Sample B containing both the suspected monomer and dimer (pooled fractions C5, C6, C8, and C9). Samples were concentrated and Sample A cleaved with thrombin before being analysed on a denaturing gel with and without the reducing agent β-mercaptoethanol (Figure 4.20).

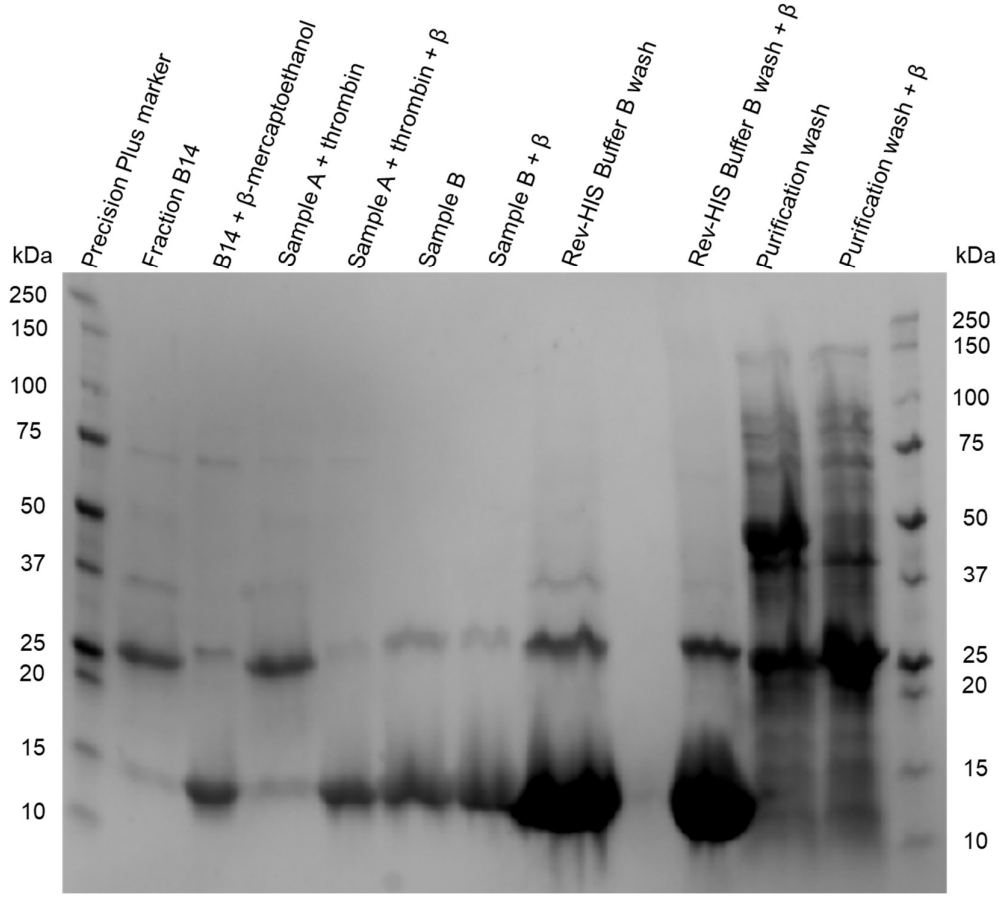


Figure 4.20: SDS-PAGE of reducing gel of Sample A and B.
Gel shows each sample (A and B) after concentrating with the addition of an extra thrombin digestion step, or with and without a reducing agent, β-mercaptoethanol.

Additional thrombin had no effect at reducing the size of the band at approximately 27 kDa, demonstrating that the samples were already fully digested. Sample A and Sample B were further concentrated to 5.4 mg/mL (A) and 5.7 mg/mL (B).

4.3.3.6 Protein Crystallography

Protein was set using the sitting drop method into two types of crystal screen trays at two different concentrations. No crystals formed within six months of monitoring.

4.3.3.7 Mass Spectrometry

The results from the mass spectrometry of suspected AquaCT dimers are shown in Figures 4.21-4.24.

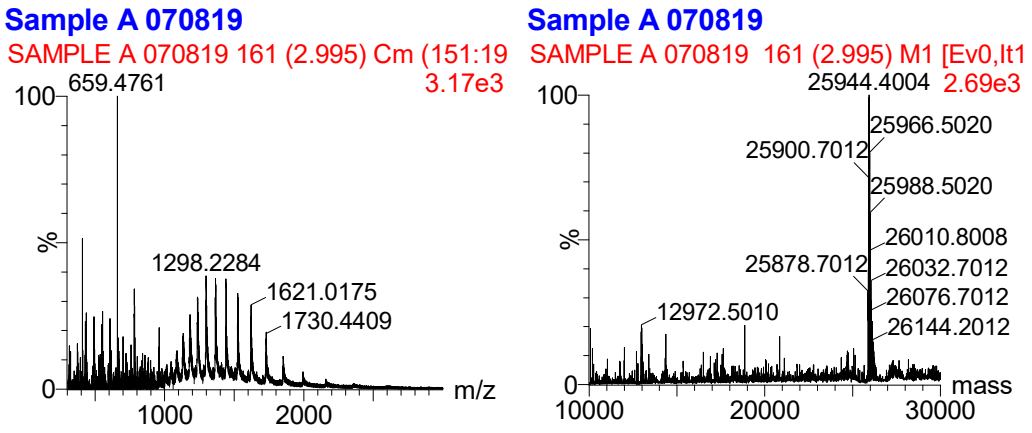


Figure 4.21: Mass spectra of Sample A (suspected dimer) in a native state.
The mass to charge ratio (m/z) is shown in the left-hand panel, and that has been converted to molecular mass in the right-hand panel.

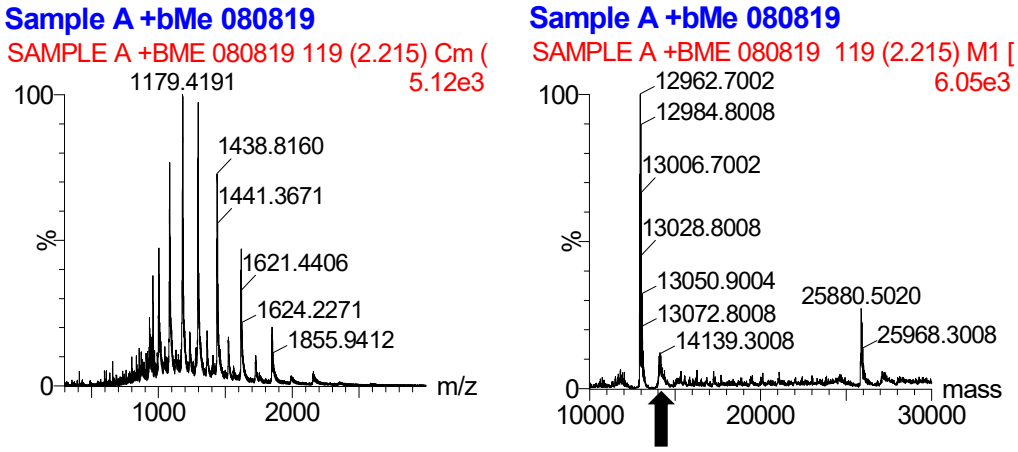


Figure 4.22: Mass spectra of Sample A reduced by β-mercaptoethanol
Solid arrow shows suspected AquaCT monomer peak. The mass to charge ratio (m/z) is shown in the left-hand panel, and that has been converted to molecular mass in the right-hand panel.

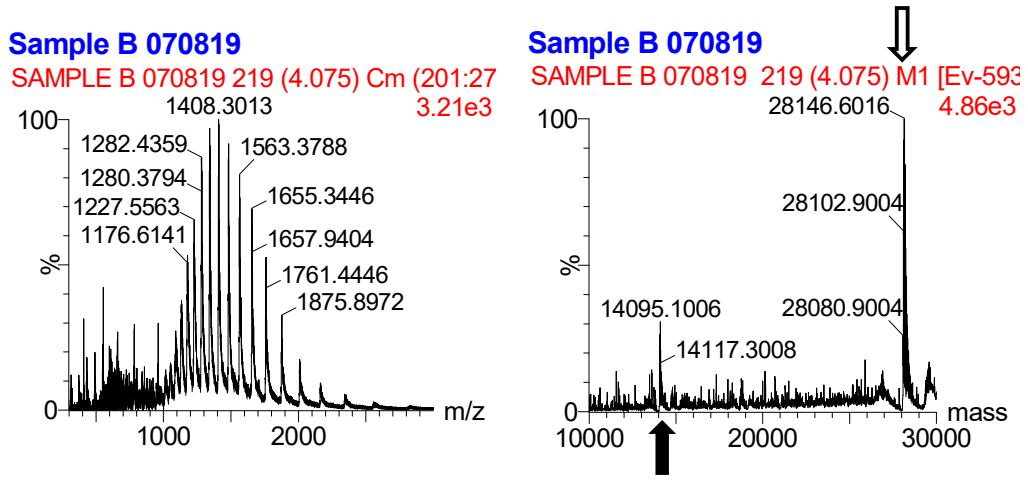


Figure 4.23: Mass spectra of Sample B in native state

Mw peak of suspected AquaCT monomer is indicated by solid arrow, clear arrow indicates suspected dimer. The mass to charge ratio (m/z) is shown in the left-hand panel, and that has been converted to molecular mass in the right-hand panel.

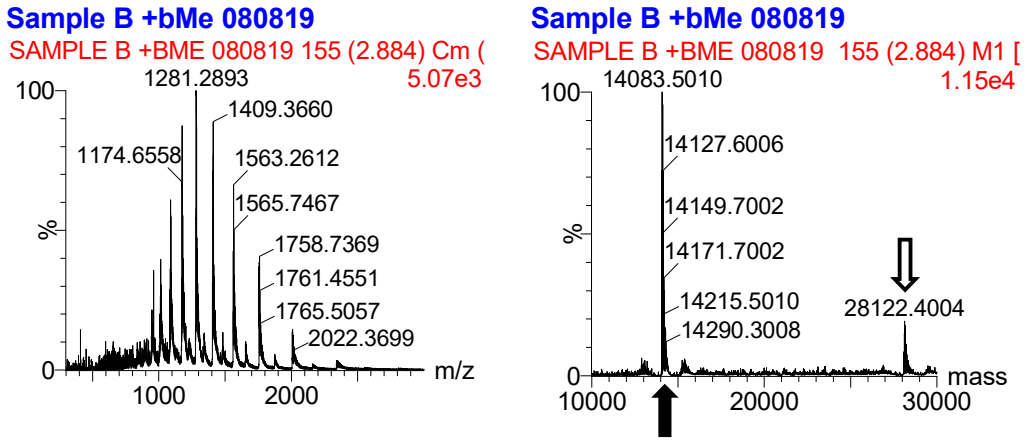


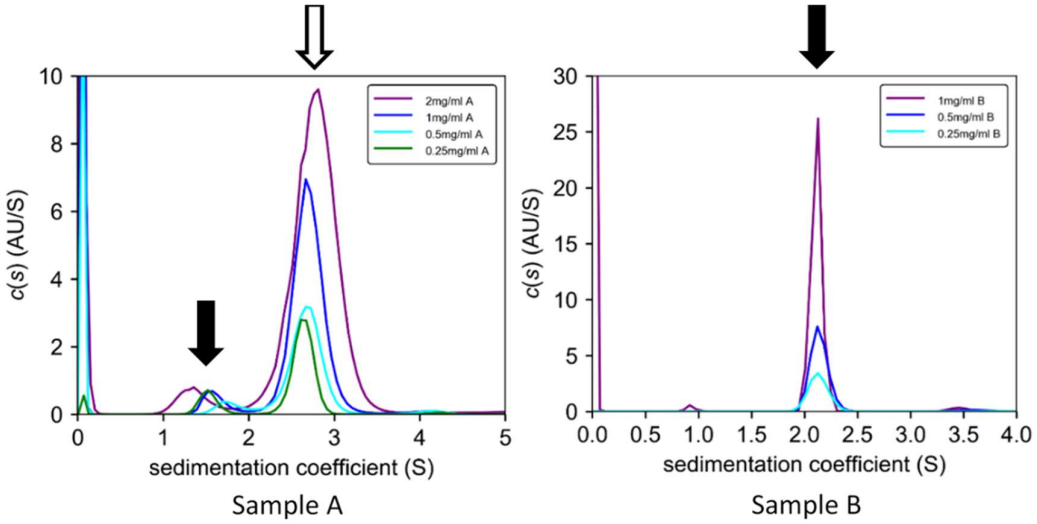
Figure 4.24: Mass spectra of Sample B reduced by β-mercaptoethanol.

Mw peak of suspected AquaCT monomer is indicated by solid arrow, clear arrow indicates suspected dimer. The mass to charge ratio (m/z) is shown in the left-hand panel, and that has been converted to molecular mass in the right-hand panel.

Mass Spectrometry analysis (Figure 4.21-4.24) shows that Sample B is most likely the desired AquaCT product, as it has a molecular weight peak at 14 and 28 kDa, which is the same or double the expected molecular weight of the protein, respectively. There is a small peak in Sample A at 14 kDa when β-mercaptoethanol is added, suggesting that there is a small amount of AquaCT monomer.

4.3.3.8 Analytical Ultra Centrifugation

Partial specific volume (Vbar) used was 0.735, buffer density 0.010859 g/cm^3 and buffer viscosity 0.0301 mPa. Results are shown in Figure 4.25. This shows that Sample A is a fairly heterogeneous sample, with multiple peaks, while Sample B shows the correct approximate weight expected for AquaCT.



		Peak 1		Peak 2		
Sample	Concentration (mg/ml)	Sedimentation coefficient (S)	Protein mass (kDa)	Sedimentation coefficient (S)	Protein mass (kDa)	Frictional ratio (F_0)
A	2	1.2	17.2	2.8	49.6	1.65
	1	1.5	21.6	2.7	46.2	1.62
	0.5	1.7	24.9	2.7	45.3	1.62
	0.25	1.5	15	2.7	33.3	1.33
B	1	2.1	28.4	-	-	1.49
	0.5	2.1	28.6	-	-	1.48
	0.25	2.1	27.6	-	-	1.45

Figure 4.25: AUC results.

Values for the peaks shown in the graphs are in the table below. The solid arrow indicates peak 1 in each sample and the clear arrow indicates peak 2 in Sample A.

4.3.3.9 Size Exclusion Chromatography Multi Angle Light Scattering (SEC-MALS)

The data in Figure 4.26 shows that Sample A is mainly eluted at 8.75 ml, and a smaller secondary peak was observed at approximately 10.4 ml, which indicates a small amount of AquaCT (indicated on figure). The molar mass points calculated across the peak for Sample A are not uniform, which is indicative of heterogeneity in the sample. Sample B is shown by the blue trace, and the green trace shows Sample B after the addition of β-mercaptoethanol. Both peaks show that Sample B elutes off the column at approximately 10.4 ml.

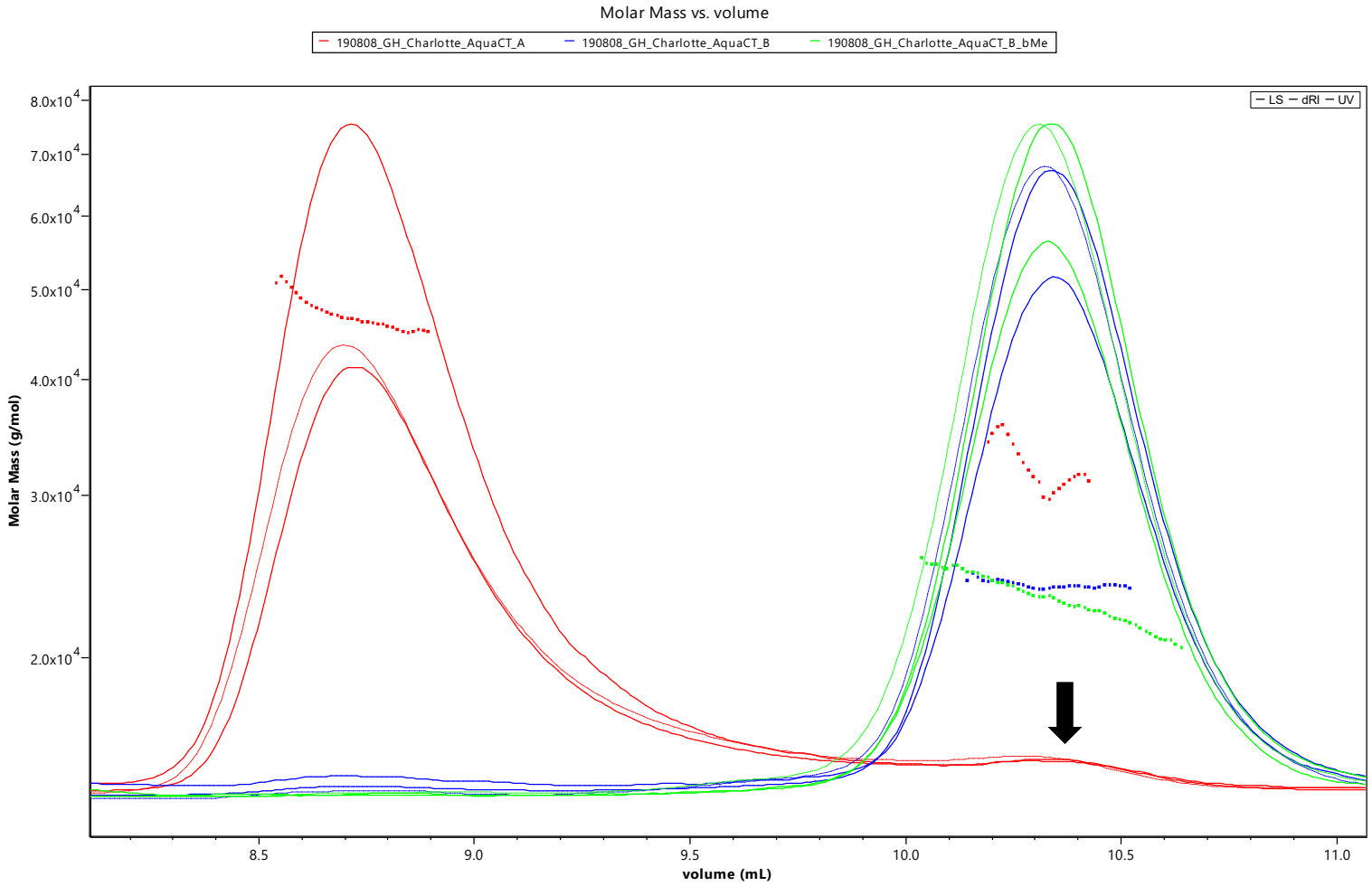


Figure 4.26: Size Exclusion Chromatography trace. *Sample A is shown in red and Sample B in blue. Sample B in a reduced state is shown in green. Peaks show where the proteins are eluted from the column, and the dots represent the molecular weight of the sample. Solid arrow indicates small secondary peak in Sample A. (The molecular weight predictions are anticipated to be imprecise because as generic correction factors were used out of necessity to calculate these, however they are a useful guide).*

4.4 Discussion

Using the principles of synthetic silk production used to make 4RepCT, a novel synthetic aquatic silk, AquaRepCT was successfully expressed and purified. By changing the expression cassette to a more powerful T7 expression system and the strain of *E. coli* employed, it was possible to obtain purified protein at scales of up to 75 mg/L using shake flasks. Subsequent studies using batch fermentation have produced this protein at up to 150 mg/L (data not shown), which demonstrates a great improvement in recombinant spidroin production compared to published values. The high expression levels seen with the AquaRepCT construct could have an impact on the reduction of industrial production costs of recombinant spidroins, as economic modelling has previously shown that increasing yield is the most valuable area in which advancements can be made (Edlund *et al.*, 2018).

AquaRepCT stays soluble at concentrations of over 40 mg/mL with no observable precipitation. This could be beneficial in terms of downstream processing, as the need for re-solubilising proteins can often affect protein folding integrity and increase processing costs and time (Sørensen and Mortensen, 2005). It has been reported that the solubility of recombinant spidroins is generally very low (2–4 mg/mL), and this increase in solubility may be due to the properties of the repetitive region of this protein (Debabov and Bogush, 2020). AquaRepCT does not form fibres upon thrombin-mediated removal of thioredoxin but fibre formation was expected based on past experience with 4RepCT and similarities in sequence (as seen in Figure 4.9). Precipitation observed when non-cleaved AquaRepCT was introduced to different buffers may be either due to the properties of the buffer or the change in pH, which could be investigated further with different concentrations and at different temperatures.

It was hypothesised that the inability of AquaRepCT to form fibres upon thrombin cleavage was due to the inability of the CTD to form dimers. To investigate this behaviour, the C-terminal domain of the AquaRepCT construct (AquaCT) was expressed without the repetitive region. Characterisations of this protein suggest that the CTD is capable of forming dimers using a disulphide bond, which is broken upon the addition of a reducing agent. The purification of AquaCT resulted in two

samples; A and B. Mass spectrometry analysis and AUC data of Sample A suggest that it is a fairly heterogeneous sample, with multiple protein species (Figures 4.20-22 and Figure 4.25). There is a small peak in the mass spectrometry traces for sample A at 14 kDa when β-mercaptoethanol is added, suggesting that there is a small amount of AquaCT monomer (Figures 4.21 and 4.22). The data in Figure 4.26 shows that Sample A is mainly eluted at 8.75 ml, and a smaller secondary peak was observed at approximately 10.3 ml. The molar mass points calculated across the peak for Sample A are not uniform, which is indicative of heterogeneity in the sample. This, coupled with the observed molecular weight, suggests that Sample A is a loosely associated tetramer in equilibrium. This peak is electrostatic, which suggests that the tetramer formation is concentration dependent. The molecular weight masses at around 10.4 mL suggest that this protein is also dimer or trimerizing as it moves in solution, or that impurities are present from sample preparation.

Figure 4.20 shows Sample B which contains dimer bands (≈25 kDa) reduced to monomer sizes with the addition of β-mercaptoethanol, indicating the presence of AquaCT. This data suggests that AquaCT dimers are held together by a disulphide bridge between two cysteines in the CTD, which has been both implicated and shown to be important for spider silk dimerisation in previous studies (Rat et al., 2018; Strickland et al., 2018). The AUC data of Sample B shows a strong peak at a sedimentation coefficient of 2.1 for all protein concentrations and again indicates that this contains AquaCT (Figure 4.25). This also gives the correct approximate molecular weight expected for AquaCT, showing that the CTD of AquaRepCT in isolation dimerises in solution in a non-concentration dependent association. SEC-MALS analysis suggests that AquaCT (Sample B), when associated with a column resin, is probably a concentration dependent dimer held together by disulphides and other intermolecular forces. When reduced by β-mercaptoethanol, AquaCT still seems to be a dimer but less strongly associated, as demonstrated by the reduction in molecular weight points. This shows that the hydrophobic residues are enough to keep the dimers together loosely even after the cysteine bonds are reduced (Figure 4.26). These data suggest that the CTD of AquaRepCT is able to covalently link to form a dimer and then a loosely associated tetramer in solution, and even with the

addition of a reducing agent it will still loosely associate. Taken together, this suggests that AquaRepCT should dimerise and then form fibres, and points to the reason for not doing so as being due to the properties of the repetitive regions. Both samples also halve their molecular weight when the highly reducing agent β-mercaptoethanol is added, further suggesting that these are dimers in solvent, and the reducing agent is reducing the disulphide bonds (formed by the cysteine) thus breaking the dimers apart, as shown in Figures 4.21-4.24.

Another reason for the lack of fibre formation may be because of the addition of an extra cysteine at the end of the CTD, which is not present in 4RepCT or indeed any of the CTDs surveyed during the course of this thesis. While this domain has been shown to dimerise, this residue may prevent the whole AquaRepCT protein from aligning as expected to enable fibre formation. It is also possible that by including the last 16 residues of the C-terminal domain in this construct (not usually designated as C-term), extra secondary structure is introduced that hinders dimerisation upon the addition of the repetitive region by constraining the rotation of monomers when binding. The above data demonstrate that this is not a hinderance to binding of the CTD when expressed on its own.

It has been demonstrated that the stiffness of 4RepCT fibres is largely independent of covalent interaction (including disulphide bonds) between the CTDs, and relies mainly on intermolecular interactions between poly-alanine blocks (Grip, Johansson and Hedhammar, 2009). This furthers the idea that it is the addition of other residues in the repetitive region that destabilises these possible interactions between repetitive domains. The differences between 4RepCT and AquaRepCT are not pronounced, however the poly(A) blocks are disrupted by charged, polar amino acids (arginine, glutamic acid and aspartic acid) in AquaRepCT. Hydropathy plots show that the main differences between the two proteins are in the repetitive regions, which is evident from by-eye analysis (Figure 4.8). Garnier secondary structure prediction (Garnier, Osguthorpe and Robson, 1978) predicts that there are more β-strands in AquaRepCT. This is agreed upon when using Phyre2 (Kelley *et al.*, 2015), which predicts more order in AquaRepCT than 4RepCT, which may occlude or prevent the formation of bonds between the Poly(A) blocks that would otherwise help to form

fibres (Figure 4.10 and Figure 4.11). In comparison to the miniaturised spidroin 4RepCT, AquaRepCT contains a higher proportion of aromatic and non-polar amino acids (Table 4.E). This could add to silk fibre stability in water by repelling water molecules and proving favourable for protein association.

Some studies have shown that the addition of the N-terminal domain into a synthetic silk construct can aid in both expression, solubility and fibre formation (Hedhammar *et al.*, 2008; Askarieh *et al.*, 2010; Kronqvist *et al.*, 2017). To this end a full-length TRX-NT-AquaRep-CT construct was designed, cloned, expressed and purified (data not shown). However, cleavage of the solubility tags using thrombin was unsuccessful, which was probably due to the lack of linker region between the thioredoxin and the start of the protein. As it has been reported that the NTD may act as a solubility tag (Kronqvist *et al.*, 2017), removing the need for thioredoxin, another construct was designed and created that lacked the thioredoxin tag. Early attempts to express this version were unsuccessful during this investigation. A protein based on the NTD of an *Argyroneta aquatica* spidroin, expressed in isolation from the rest of the protein, would be beneficial to study in the future to investigate dimerisation properties and behaviour in different pH ranges; this construct was also designed and created as part of this investigation (Appendix A.4.5).

Protein crystallography of the CTD could go a long way to investigating the bonds between these dimers. In this study, initial attempts to form crystals using AquaCT were unsuccessful, despite attempts using several screen types. This may be because there exists a tract of residues (34 amino acids) before the start of the protein, which could add significant disorder and instability; unfavourable conditions for crystal formation. In addition, thrombin cleavage of recombinant proteins is often incomplete and this enzyme can lack specificity. Progress was made to address this by cloning a TEV cleavage site into pET32b-AquaCT directly before the start of the protein (Appendix A.4.4.) TEV is an efficient cleavage enzyme, which should allow cleavage of the solubility tags and trimming of AquaCT to make it more amenable to crystallisation. This approach, coupled with mutational studies, will be the subject of further work to see if this construct is more amenable to crystallisation.

4.4.1 Future Work

Tags which increase protein solubility in bacterial expression systems and also facilitate subsequent purification could eliminate the need for the current thioredoxin and S-tag while increasing production efficiency. This would make the study of this protein more time and energy efficient and could prove beneficial for future production in larger quantities for downstream applications. One alternative includes the approach identified earlier, which requires the addition of a spidroin NTD to the miniature spidroin, as demonstrated by the construct NT2Rep (Kronqvist *et al.*, 2017). Other alternatives could include Glutathione S-transferase (GST) or maltose binding protein (MBP), which have been shown to increase solubility and purification efficiency of other proteins (Esposito and Chatterjee, 2006; Kimple, Brill and Pasker, 2013). The expression cassette itself could be improved by employing a kanamycin resistance cassette rather than ampicillin, as this allows for longer gene expression under selection pressure and kanamycin is both more stable and more reliable in culture (Hinman, Jones and Lewis, 2000; Teulé *et al.*, 2009).

Recombinant silks can often more reliably recapitulate the mechanical properties of their natural counterparts with larger protein sizes. However, merely increasing the size of the gene sequence in the construct could lead to aggregation and translational errors. It was shown with 4RepCT that by engineering in covalent links between poly-alanine blocks, multimerization can be increased, increasing the size of the protein blocks that make up the fibre and increasing strength and stiffness (Grip, Johansson and Hedhammar, 2009). This could be employed with AquaRepCT and future bio-inspired recombinant silk constructs.

Future studies are needed to focus on the characterisation and behaviour of AquaRepCT, especially in regard to its 3D structure. By obtaining a crystal structure, accurate molecular models could be produced to help guide further research into the way that synthetic silk folds and behaves. However, creating a crystal structure may prove intractable using current methods due to its repetitive nature. It would also help to explore the dimerisation properties of AquaCT by mutating the last cysteine residue to an alternative, ablating its ability to form disulphide bonds. This was

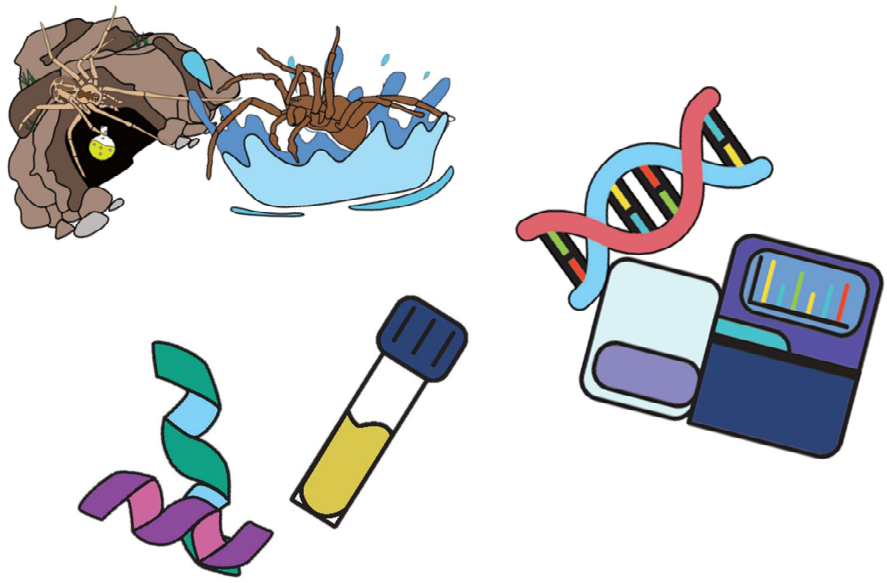
attempted during the present study but was not successful due to the repetitive nature of the DNA template.

Recombinant proteins, like spidroins, can be assembled into various morphologies such as films, capsules and hydrogels which makes them more useful for practical applications and provides a promising avenue for future research (Schacht and Scheibel, 2011; Agostini, Winter and Engert, 2015; Debabov and Bogush, 2020; Guo *et al.*, 2020). As a soluble protein, AquaRepCT may lend itself more readily to post processing manipulation, allowing the easy formation of 3D structures or bioprinting. It has been reported that the solubility of recombinant spidroins in water is very low and does not exceed 2–4 mg/mL (Debabov and Bogush, 2020). AquaRepCT is more soluble than this and could be used as a basis to dramatically simplify the wet spinning process of transforming protein dope into fibres. This approach has been used successfully in the past, lending hope to its future success (Andersson *et al.*, 2017). Additionally, as AquaRepCT is based on an aquatic spidroin, it could show more promise and versatility in applications that are based in humid/wet environments. These proteins (and possible fibres) could hold their structure and shape, retaining strength and flexibility more readily than other recombinant spidroins in humid environments. More research into the behaviour of these proteins is needed and some possible ways are expanded upon in Chapter 5.

4.4.2 Conclusions

This chapter has demonstrated that it is possible to produce and purify recombinant AquaRepCT using standard laboratory techniques based on existing and modified protocols for 4RepCT expression and purification in *E. coli*. This study has demonstrated the versatility of the 4RepCT expression system at creating a unique type of spidroin that behaves differently when expressed in *E. coli*. Also, a recombinant silk sequence has been created that does not polymerise in the same way as 4RepCT despite striking sequence similarity. This could possibly be used to find the appropriate solubility point for other materials by being used as a titratable element, controlling the material polymerisation properties dependent upon final application. This study has also designed and created several constructs which could be used in order to understand this recombinant aquatic spidroin more thoroughly.

There exists a wide range of silk proteins in nature, tailored to environments that may not have been discovered or fully characterised yet. Herein, it has been shown that it is possible to reproduce a miniaturised natural spidroin in the laboratory based on an unusual natural silk. Using this information, a possible pipeline for creating other unusual silks has been created, for example when trying to re-create acid or humidity-resistant silk from the spiders *Nesticus eremita, Lessertia barbara* or *Pirata piraticus*.



Chapter Five

GENERAL DISCUSSION

Final summaries and recommendations for future work

5.1 Summary

This thesis has examined the general theme of spider silk adapted to extreme environments, showing what approaches can be used to examine their genetic sequence, and then create these as recombinant materials. It has also started to examine the intron-exon structure of silk genes and speculated as to the ways that these genetic elements could be harnessed for use in eukaryotic heterologous expression systems.

Chapter 1 summarises the pertinent silk research landscape today and presents some new research focussing on the ecology and physical morphology of *Argyroneta aquatica* and its silk. Chapter 2 presents the first sequenced linyphiid silk genes from the species *Lessertia barbara* and *Nesticus eremita.* These findings help in the understanding of how this family of proteins has diversified while maintaining important key properties in order to fulfil all of the necessary functional requirements to persist even under extreme environmental condition of low pH. It was found that linyphiid silks assemble in the expected way and a potential novel silk type was highlighted. It was also found that the use of proline is not as expected in silks used in humid and resource scarce conditions. This study highlighted both the versatility and also the potential limits to short read RNA sequencing technology, especially when dealing with samples of lower quality.

Chapter 3 uses SMRT long read technology to present the first full length genes of *A. aquatica* and *C. citricola.* This study compared the results of different sequencing technologies for silk identification and classification and suggests appropriate approaches for future studies. Contrasting amino acid usage was found in the long-lasting silks used in either humid or dry environments, but the retention of specific motifs and modular spidroin construction was observed. These spidroin sequences have the potential for incorporation into future recombinant spidroin designs.

Chapter 4 uses the principles of recombinant spidroin expression using a modular system to create the first aquatic mini-spidroin in *E. coli*. This protein, based on sequences retrieved from an earlier *Argyroneta aquatica* transcriptome, creates a soluble silk-like protein even at high concentrations. Protein characterisation of the

C-terminal domain reveals a greater role of the repetitive region in occlusion polymerization than previously thought. This study shows the flexibility of this approach for possible future constructs based on silks used in extreme environments, and allows the design of a panel of constructs for further investigation.

In in the following section, experiments and routes of research are summarised which could be completed to carry on this research, as well attention drawn to areas which should be expanded upon in the pursuit of understanding the way from natural to synthetic spider silk.

5.2 Future Study of Silks in Extreme Environments

To further this research, it is recommended that future studies of spider silks used in extreme environments should increase the link between structure, property and function by characterising the material properties of the natural silk in tandem with in-depth genetic analyses. This will allow a clearer link to be drawn between primary-secondary and tertiary structure, which can then be used to the researchers' advantage when designing recombinant spidroins. Chapters 2 and 3 have shown that ideally, the sequencing approach of spider silk genes should involve a combination of preliminary study for terminal region identification and short repetitive region sequencing, with subsequent long read technology employed to understand the whole protein, especially in areas of specific interest.

In Chapter 3, evidence of frameshifts were found in the silk transcripts of *A. aquatica*, which could potentially change the properties of the silken fibre. By using conventional PCR and Sanger DNA sequencing technology, it could be determined whether these frameshifts are a genetic mechanism to increase diversity in silk genes. It could also open up a new avenue to explore previously discarded transcripts that are not considered to be coding sequences by conventional software packages, and the effects these amino acid changes have on the silk fibres as produced by a range of spiders. Protein sequencing using mass spectrometry or Edman degradation can be used to investigate silk peptide sequences and reveal whether this genetic phenomenon is represented in the resulting protein, and whether the mechanism of frameshifting is a natural way of generating more silk diversity.

It is also recommended that future studies examine the expression profiles of different silk types when the spider is exposed to different environments, through transcript quantification or qPCR. This, coupled with MS/MS analysis of the produced silks will give a higher level of confidence as to when and where each specific silk is used, which could be beneficial, for example to determine whether the different silk types produced by *Lessertia barbara* and *Nesticus eremita* are in fact in response to external acidic conditions. Behavioural studies too would be important in determining whether amino acid differences in the silk of *L. barbara* and *N. eremita* are due to the differences in successful resource capture. This combined approach could also be used to examine the silks of other spiders that live in extreme environments, for example the intertidal spider *Desis marina* which creates a silken retreat to trap air in times of high tide, *Thomisus nepenthiphilus*, the crab spider that inhabits carnivorous pitcher plants and uses silk to ensure its release from the plant, or the fishing spider *Dolomedes plantarius,* which creates silken dragline supports on the surface of water, the genome of which is expected to be published shortly. This would also show whether the expression of certain silk types is localised to extreme conditions or the sex or specific life stages of the spider.

During the course of this investigation, it became clear that there are more proteins produced by *Argyroneta aquatica* that could prove to be of considerable interest in biotechnology or drug research. This includes the broad class of venom toxins, which act against both invertebrates and vertebrates, and adhesive proteins, of which transcriptomic analysis showed high homology to molluscan mucins. The long-read data presented here could be used to further investigate these proteins, and assembly of these transcripts would allow the sequence recovery of all full length genes, including those over the 10 kb sequencing limit of SMRT Sequencing.

5.3 Future Recombinant Mini-Spidroin Expression

As shown in Chapter 4, a modular recombinant mini-spidroin expression system can be employed to create different silk types, with higher expression levels than seen before. In bacteria, constructs should be designed in a T7 vector rather than T5. They should also utilize a TEV cleavage system to allow efficient cleavage of tags. Incorporating a native N-terminal domain will remove the need for large or bulky

solubilization tags like thioredoxin and allow the production of 'NTD - repetitive region - CTD' containing recombinant spidroins with properties amenable to polymerisation. It is suggested here that suggest that kanamycin is used as a selection marker, as it is more stringent and reliable than ampicillin, and that an *E. coli* strain engineered for high protein production, such as Rosetta, is also used.

As discussed in Chapter 4, the future of recombinant spidroin production in *E. coli* is limited and eukaryotic hosts pose are an attractive alternative when producing larger and more complex proteins, even at industrial scales. Both *Pichia pastoris* and *Saccharomyces cerevisiae* are well-characterised eukaryotic expression hosts, suited to recombinant protein expression, which preliminary research suggests are able to produce miniaturised spidroins (unpublished data). It is also known that post translational modifications are important to retain strength and stability in some MaSp fibres, a process which these organisms can carry out.

It is important to further characterise the physical properties of AquaRepCT, especially to investigate its fibre forming potential. This could include electrospinning, full characterisation in a range of buffers, 3D printing and Förster resonance energy transfer (FRET) using Total Internal Reflection Fluorescent (TIRF) microscopy. Sequences for inclusion in the design of several novel MaSp spidroins have been suggested and included in the Appendices, which represent the first steps to creating a 'toolbox' of recombinant spidroin resources for future applications which require specific properties using the AquaRepCT expression system. Our study of silks used in extreme environments has suggested that they assemble in the expected way, making them interesting candidates for recombinant expression.

5.4 Implications of results

This thesis has laid down the framework for the development of a robust pipeline that could be used to identify, characterise and recreate other silk genes and proteins in the laboratory. With further development and a focus on the physical characterisation of these fibres, it will become easier to create recombinant spidroins that more accurately mimic the properties of their natural counterparts. It has been shown that a range of bioinformatic analysis can be carried out on the genetic

material of spiders and that silks can be identified with relative ease. This, coupled with a knowledge of the ecology and lifestyle of each spider can help in the creation of recombinant silks with properties that enable their use in extreme environmental conditions.

List of References

Adams, R. J. (2019) *Field Guide to the Spiders of California and the Pacific Coast States*, *Field Guide to the Spiders of California and the Pacific Coast States*. University of California Press. doi: 10.1525/9780520957046.

Adrianos, S. L. *et al.* (2013) '*Nephila clavipes* flagelliform silk-like GGX motifs contribute to extensibility and spacer motifs contribute to strength in synthetic spider silk fibers', *Biomacromolecules*, 14(6), pp. 1751–1760. doi: 10.1021/bm400125w.

Afgan, E. *et al.* (2016) 'The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update', *Nucleic acids research*, 44(W1), pp. W3–W10. doi: 10.1093/nar/gkw343.

Agnarsson, I., Boutry, C. and Blackledge, T. A. (2008) 'Spider silk aging: Initial improvement in a high performance material followed by slow degradation', *Journal of Experimental Zoology Part A: Ecological Genetics and Physiology*, 309(8), pp. 494–504. doi: 10.1002/jez.480.

Agnarsson, I., Kuntner, M. and Blackledge, T. A. (2010) 'Bioprospecting finds the toughest biological material: Extraordinary silk from a giant riverine orb spider', *PLoS ONE*, 5(9), pp. 1–8. doi: 10.1371/journal.pone.0011234.

Agostini, E., Winter, G. and Engert, J. (2015) 'Water-based preparation of spider silk films as drug delivery matrices', *Journal of Controlled Release*, 213, pp. 134–141. doi: 10.1016/j.jconrel.2015.06.025.

Alam, P. *et al.* (2016) 'The toughest recorded spider egg case silks are woven into composites with tear-resistant architectures', *Materials Science and Engineering*, 69, pp. 195–199. doi: 10.1016/j.msec.2016.06.063.

Albertson, A. E. *et al.* (2014) 'Effects of different post-spin stretching conditions on the mechanical properties of synthetic spider silk fibers', *Journal of the Mechanical Behavior of Biomedical Materials*, 29, pp. 225–234. doi: 10.1016/j.jmbbm.2013.09.002.

Alfaro, C. *et al.* (2013) 'Effect of thermal acclimation on preferred temperatures in two mygalomorph spiders inhabiting contrasting habitats', *Physiological Entomology*, 38(1), pp. 20–25. doi: 10.1111/j.1365-3032.2012.00853.x.

Altschul, S. F. *et al.* (1990) 'Basic Local Alignment Search Tool', *Journal of Molecular Biology*, 215(3), pp. 403–410. doi: 10.1016/S0022-2836(05)80360-2.

Álvarez-Padilla, F. and Hormiga, G. (2011) 'Morphological and phylogenetic atlas of the orb-weaving spider family Tetragnathidae (Araneae: Araneoidea)', *Zoological Journal of the Linnean Society*, 162(4), pp. 713–879. doi: 10.1111/j.1096-3642.2011.00692.x.

Amarasinghe, S. L. *et al.* (2020) 'Opportunities and challenges in long-read sequencing data analysis', *Genome Biology*, 21(1), pp. 1–16. doi: 10.1186/s13059-020-1935-5.

Ambardar, S. *et al.* (2016) 'High Throughput Sequencing: An Overview of Sequencing Chemistry', *Indian Journal of Microbiology*, 56(4), pp. 394–404. doi: 10.1007/s12088-016-0606-4.

Anderson, J. F. (1974) 'Responses to Starvation in the Spiders *Lycosa Lenta* (Hentz) and *Filistata Hibernalis* (Hentz)', *Ecology*, 55(3), pp. 576–585. doi: 10.2307/1935148.

Andersson, M. *et al.* (2013) 'Morphology and composition of the spider major ampullate gland and dragline silk', *Biomacromolecules*, 14(8), pp. 2945–2952. doi: 10.1021/bm400898t.

Andersson, M. *et al.* (2014) 'Carbonic anhydrase generates CO_2 and H^+ that drive spider silk formation via opposite effects on the terminal domains'. *PLoS Biology.* 12(8): e1001921.

Andersson, M. *et al.* (2017) 'Biomimetic spinning of artificial spider silk from a chimeric minispidroin', *Nature Chemical Biology*, 13(3), pp. 262–264. doi: 10.1038/nchembio.2269.

Andersson, M., Johansson, J. and Rising, A. (2016) 'Silk spinning in silkworms and spiders', *International Journal of Molecular Sciences*, 17(8). doi: 10.3390/ijms17081290.

Andrews, S. (2010) 'Babraham Bioinformatics - FastQC A quality control tool for high throughput sequence data', *Soil*, 5(1), pp. 47–81. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Accessed: 20 October 2020).

Anton, B. P. and Raleigh, E. A. (2016) 'Complete genome sequence of NEB 5-alpha, a derivative of *Escherichia coli* K-12 DH5α', *Genome Announcements*, 4(6), pp. 1245–1261. doi: 10.1128/genomeA.01245-16.

Archana, T. *et al.* (2018) 'Fracture toughness and fatigue behavior of spider silk and S-glass epoxy composites: An FEM approach', *Materials Today: Proceedings*, 5(1), pp. 2627–2634. doi: 10.1016/j.matpr.2018.01.042.

Arnedo, M. A. *et al.* (2004) 'From a comb to a tree: Phylogenetic relationships of the comb-footed spiders (Araneae, Theridiidae) inferred from nuclear and mitochondrial genes', *Molecular Phylogenetics and Evolution*, 31(1), pp. 225–245. doi: 10.1016/S1055-7903(03)00261-6.

Arthur, L. L. *et al.* (2015) 'Translational control by lysine-encoding A-rich sequences', *Science Advances*, 1(6). doi: 10.1126/sciadv.1500154.

Asakura, T. *et al.* (2004) 'Structural role of tyrosine in *Bombyx mori* silk fibroin, studied by solid-state NMR and molecular mechanics on a model peptide prepared as silk I and II', in *Magnetic Resonance in Chemistry*, pp. 258–266. doi: 10.1002/mrc.1337.

Asakura, T., Okushita, K. and Williamson, M. P. (2015) 'Analysis of the structure of *Bombyx mori* silk fibroin by NMR', *Macromolecules*, 48(8), pp. 2345–2357. doi: 10.1021/acs.macromol.5b00160.

Ashton, N. N. and Stewart, R. J. (2015) 'Self-recovering caddisfly silk: Energy dissipating, Ca^{2+}-dependent, double dynamic network fibers', *Soft Matter*, 11(9), pp. 1667–1676. doi: 10.1039/c4sm02435d.

Askarieh, G. *et al.* (2010) 'Self-assembly of spider silk proteins is controlled by a pH-sensitive relay', *Nature*, 465(7295), pp. 236–238. doi: 10.1038/nature08962.

Ayoub, N. A. *et al.* (2007) 'Blueprint for a high-performance biomaterial: Full-length spider dragline silk genes', *PLoS ONE*, 2(6), p. e514. doi: 10.1371/journal.pone.0000514.

Ayoub, N. A. *et al.* (2013) 'Ancient properties of spider silks revealed by the complete gene sequence of the prey-wrapping silk protein (AcSp1)', *Molecular Biology and Evolution*, 30(3), pp. 589–601. doi: 10.1093/molbev/mss254.

Babb, P. L. *et al.* (2017) 'The *Nephila clavipes* genome highlights the diversity of spider silk genes and their complex expression', *Nature Genetics*, 49(6), pp. 895–903. doi: 10.1038/ng.3852.

Baker-Austin, C. and Dopson, M. (2007) 'Life in acid: pH homeostasis in acidophiles', *Trends in Microbiology*, 15(4), pp. 165–171. doi: 10.1016/j.tim.2007.02.005.

Balakrishnan, P. *et al.* (2016) 'Natural fibre and polymer matrix composites and their applications in aerospace engineering', *Advanced Composite Materials for Aerospace Engineering*, pp. 365–383. doi: 10.1016/b978-0-08-100037-3.00012-2.

Balzer, S. *et al.* (2010). 'Characteristics of 454 pyrosequencing data - enabling realistic simulation with flowsim.' *Bioinformatics*, 26(18), pp. i420–i425. https://doi.org/10.1093/bioinformatics/btq365

Barrantes, G. and Eberhard, W. G. (2007) 'The evolution of prey-wrapping behaviour in spiders', *Journal of Natural History*, 41(25–28), pp. 1631–1658. doi: 10.1080/00222930701464364.

Bartonek, L., Braun, D. and Zagrovic, B. (2020) 'Frameshifting preserves key physicochemical properties of proteins', *Proceedings of the National Academy of Sciences of the United States of America*, 117(11), pp. 5907–5912. doi: 10.1073/pnas.1911203117.

Beattie, A. J. *et al.* (2011) 'Ecology and bioprospecting', *Austral Ecology*, 36(3), pp. 341–356. doi: 10.1111/j.1442-9993.2010.02170.x.

Beckwitt, R. and Arcidiacono, S. (1994) 'Sequence conservation in the C-terminal region of spider silk proteins (Spidroin) from *Nephila clavipes* (tetragnathidae) and *Araneus bicentenarius* (Araneidae)', *Journal of Biological Chemistry*, 269(9), pp. 6661–6663. doi: 10.1016/s0021-9258(17)37425-2.

Benjamin, S. P. and Zschokke, S. (2004) 'Homology, behaviour and spider webs: Web construction behaviour of *Linyphia hortensis* and *L. triangularis* (Araneae: Linyphiidae) and its evolutionary significance', *Journal of Evolutionary Biology*. John Wiley & Sons, Ltd, pp. 120–130. doi: 10.1046/j.1420-9101.2004.00667.x.

Blackledge, T. A. and Hayashi, C. Y. (2006) 'Unraveling the mechanical properties of composite silk threads spun by cribellate orb-weaving spiders', *Journal of Experimental Biology*, 209(16), pp. 3131–3140. doi: 10.1242/jeb.02327.

Blackledge, T. A. *et al.* (2009) 'Reconstructing web evolution and spider diversification in the molecular era', *Proceedings of the National Academy of Sciences of the United States of America*, 106(13), pp. 5229–5234. doi: 10.1073/pnas.0901377106.

Blamires, S. J. *et al.* (2016) 'Spider web and silk performance landscapes across nutrient space', *Scientific Reports*, 6(1), pp. 1–9. doi: 10.1038/srep26383.

Blamires, S. J., Chao, I. C. and Tso, I. M. (2010) 'Prey type, vibrations and handling interactively influence spider silk expression', *Journal of Experimental Biology*, 213(22), pp. 3906–3910. doi: 10.1242/jeb.046730.

Blamires, S. J., Wu, C.-L. L. and Tso, I.-M. M. (2012) 'Variation in protein intake induces variation in spider silk expression', *PLoS ONE*. Edited by T. Bilde, 7(2), p. e31626. doi: 10.1371/journal.pone.0031626.

Blanke, R. (1972) 'Untersuchungen zur Okophysiologie and Okethologie von Cyrtophora citricola Forskil (Araneae, Araneidae) in Andalusien.', *Forma et Functio*, 5, pp. 126–206.

Blasingame, E. *et al.* (2009) 'Pyriform spidroin 1, a novel member of the silk gene family that anchors dragline silk fibers in attachment discs of the black widow spider, *Latrodectus hesperus*', *Journal of Biological Chemistry*, 284(42), pp. 29097–29108. doi: 10.1074/jbc.M109.021378.

Bleckmann, H. and Barth, F. G. (1984) 'Sensory ecology of a semi-aquatic spider *(Dolomedes triton*) - II. The release of predatory behavior by water surface waves', *Behavioral Ecology and Sociobiology*, 14(4), pp. 303–312. doi: 10.1007/BF00299502.

Bogush, V. G. *et al.* (2011) 'Recombinant analogue of spidroin 2 for biomedical materials', *Doklady Biochemistry and Biophysics*, 441(1), pp. 276–279. doi: 10.1134/S1607672911060093.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: A flexible trimmer for Illumina sequence data', *Bioinformatics*, 30(15), pp. 2114–2120. doi: 10.1093/bioinformatics/btu170.

Bolt Threads – Technology (2018) [online]. Available at: https://boltthreads.com/ (Accessed: 11 May 2021).

Bosmans, R. (2006) 'Contribution to the knowledge of the Linyphiidae of the Maghreb.', *Belgian Journal of Zoology*, 136, pp. 173–191.

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 Genetics and Physiology*, 309(9), pp. 542–552. doi: 10.1002/jez.487.

Bowen, C. H. *et al.* (2018) 'Recombinant spidroins fully replicate primary mechanical properties of natural spider silk', *Biomacromolecules*, 19(9), pp. 3853–3860. doi: 10.1021/acs.biomac.8b00980.

Brandman, O. *et al.* (2012) 'A Ribosome-Bound Quality Control Complex Triggers Degradation of Nascent Peptides and Signals Translation Stress', *Cell*, 151(5), pp. 1042–1054. doi: 10.1016/j.cell.2012.10.044.

Brignoli, P. M. (1971) 'Note su ragni cavernicoli italiani (Araneae)', *Fragmenta Entomologica*, 7, pp. 121–229.

Bromhall, C. (1988) '*Argyroneta aquatica* (Clerck) over-wintering behaviour and super-cooling point.', *Bulletin of the British Arachnological Society*. doi: 10.2476/asjaa.6.117.

Brown, C. P. *et al.* (2014) 'With great structure comes great functionality: Understanding and emulating spider silk', *Journal of Materials Research*. Cambridge University Press, pp. 108–120. doi: 10.1557/jmr.2014.365.

Bushmanova, E. *et al.* (2019) 'RNASPAdes: A *de novo* transcriptome assembler and its application to RNA-Seq data', *GigaScience*, 8(9), pp. 1–13. doi: 10.1093/gigascience/giz100.

Canals, M., Veloso, C. and Solís, R. (2015) 'Adaptation of the spiders to the environment: The case of some Chilean species', *Frontiers in Physiology*, 6(Aug), p. 220. doi: 10.3389/fphys.2015.00220.

Candelas, G. C. *et al.* (1990) 'Spider silk glands contain a tissue-specific alanine tRNA that accumulates in vitro in response to the stimulus for silk protein synthesis', *Developmental Biology*, 140(1), pp. 215–220. doi: 10.1016/0012-1606(90)90069-U.

Carlson, D. E. and Hedin, M. (2017) 'Comparative transcriptomics of Entelegyne spiders (Araneae, Entelegynae), with emphasis on molecular evolution of orphan genes', *PLoS ONE*, 12(4). doi: 10.1371/journal.pone.0174102.

Castillo-Davis, C. I. *et al.* (2002) 'Selection for short introns in highly expressed genes', *Nature Genetics*, 31(4), pp. 415–418. doi: 10.1038/ng940.

Challis, R. J., Goodacre, S. L. and Hewitt, G. M. (2006) 'Evolution of spider silks: Conservation and diversification of the C-terminus', *Insect Molecular Biology*, 15(1), pp. 45–56. doi: 10.1111/j.1365-2583.2005.00606.x.

Chaw, R. C. *et al.* (2014) 'Intragenic homogenization and multiple copies of prey-wrapping silk genes in *Argiope* garden spiders', *BMC Evolutionary Biology*, 14(1), pp. 1–12. doi: 10.1186/1471-2148-14-31.

Chaw, R. C. *et al.* (2018) 'Egg case silk gene sequences from *Argiope* Spiders: Evidence for multiple loci and a loss of function between paralogs', *G3: Genes, Genomes, Genetics*, 8(1), pp. 231–238. doi: 10.1534/g3.117.300283.

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 Biochemistry and Molecular Biology*, 81, pp. 80–90. doi: 10.1016/j.ibmb.2017.01.002.

Chen, F., Porter, D. and Vollrath, F. (2012) 'Structure and physical properties of silkworm cocoons', *Journal of the Royal Society Interface*, 9(74), pp. 2299–2308. doi: 10.1098/rsif.2011.0887.

Chen, G. *et al.* (2012) 'Full-Length Minor Ampullate Spidroin Gene Sequence', *PLoS ONE*, 7(12). doi: 10.1371/journal.pone.0052293.

Chen, R. (2012) 'Bacterial expression systems for recombinant protein production: *E. coli* and beyond', *Biotechnology Advances*, 30(5), pp. 1102–1107. doi: 10.1016/j.biotechadv.2011.09.013.

Cho, M. *et al.* (2018) 'An observational study of ballooning in large spiders: Nanoscale multifibers enable large spiders' soaring flight', *PLoS Biology*, 16(6), p. e2004405. doi: 10.1371/journal.pbio.2004405.

Clark, D. P., Pazdernik, N. J. and McGehee, M. R. (2019) 'DNA Sequencing', in *Molecular Biology*. Elsevier, pp. 240–269. doi: 10.1016/B978-0-12-813288-3.00008-2.

Coddington, J. A. and Levi, H. W. (1991) 'Systematics and evolution of spiders (Araneae)', *Annual Review of Ecology and Systematics*, 22(1), pp. 565–592. doi: 10.1146/annurev.es.22.110191.003025.

Coddington, J. A. *et al.* (2002) 'The unique ribbon morphology of the major ampullate silk of spiders from the genus *Loxosceles* (recluse spiders)', *Biomacromolecules*, 3(1), pp. 5–8. doi: 10.1021/bm010108m.

Coddington, J. A. (1989) 'Spinneret Silk Spigot Morphology : Evidence for the Monophyly of Orbweaving Spiders, Cyrtophorinae (Araneidae), and the Group Theridiidae Plus Nesticidae', *The Journal of Arachnology*, 17(1), pp. 71–95.

Cokendolpher, J. C. and Reddell, J. R. (2001) 'New and rare nesticid spiders from Texas caves (Araneae: Nesticidae)', *Texas Memorial Museum, Speleological Monographs*, 5, pp. 25–34.

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 Science*, 7(3), pp. 667–672. doi: 10.1002/pro.5560070315.

Collin, M. A. *et al.* (2016) 'Evidence from multiple species that spider silk glue component ASG2 is a spidroin', *Scientific Reports*, 6, pp. 1–12. doi: 10.1038/srep21589.

Collin, M. A. *et al.* (2018) 'Genomic perspectives of spider silk genes through target capture sequencing: Conservation of stabilization mechanisms and homology-based structural models of spidroin terminal regions', *International Journal of Biological Macromolecules*, 113, pp. 829–840. doi: 10.1016/j.ijbiomac.2018.02.032.

Conesa, A. *et al.* (2016) 'A survey of best practices for RNA-seq data analysis', *Genome Biology*, 17(1), p. 13. doi: 10.1186/s13059-016-0881-8.

Copeland, C. G. *et al.* (2015) 'Development of a Process for the Spinning of Synthetic Spider Silk', *ACS Biomaterials Science and Engineering*, 1(7), pp. 577–584. doi: 10.1021/acsbiomaterials.5b00092.

Correa-Garhwal, S. *et al.* (2018) 'Silk genes and silk gene expression in the spider *Tengella perfuga* (Zoropsidae), including a potential cribellar spidroin (CrSp)', *PLoS ONE*. Edited by P. Heneberg, 13(9), p. e0203563. doi: 10.1371/journal.pone.0203563.

Correa-Garhwal, S. *et al.* (2019) 'Spidroins and Silk Fibers of Aquatic Spiders', *Scientific Reports*, 9(1), pp. 1–12. doi: 10.1038/s41598-019-49587-y.

Correa-Garhwal, S. *et al.* (2019) 'Semi-aquatic spider silks: transcripts, proteins, and silk fibres of the fishing spider, Dolomedes triton (Pisauridae)', *Insect Molecular Biology*, 28(1), pp. 35–51. doi: 10.1111/imb.12527.

Craig, C. L. (1997) 'Evolution of arthropod silks', *Annual Review of Entomology.* 42, pp. 231-267. doi: 10.1146/annurev.ento.42.1.231.

Craig, C. L. (2003). Spiderwebs and silk : tracing evolution from molecules to genes to phenotypes. Oxford University Press, New York.

Craig, C. L. *et al.* (1999) 'A comparison of the composition of silk proteins produced by spiders and insects', in *International Journal of Biological Macromolecules*. Int J Biol Macromol, pp. 109–118. doi: 10.1016/S0141-8130(99)00006-9.

Cruz Da Silva, E. L. and Sierwald, P. (2014) 'Description of the males of Euprosthenops australis (Simon, 1898) and Euprosthenopsis pulchella (Pocock, 1902) (Araneae: Pisauridae)', *Zootaxa*, 3857(1), pp. 137–150. doi: 10.11646/zootaxa.3857.1.8.

De Bakker, D. *et al.* (2006) 'Description of the structure of different silk threads produced by the water spider *Argyroneta aquatica* (Clerck, 1757) (Araneae: Cybaeidae)', *Belgian Journal of Zoology*, 136(2), pp. 137–143.

De Maio, N. *et al.* (2019) 'Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes', *Microbial Genomics*, 5(9). doi: 10.1099/mgen.0.000294.

Debabov, V. G. and Bogush, V. G. (2020) 'Recombinant spidroins as the basis for new materials', *ACS Biomaterials Science and Engineering*, 6(7), pp. 3745–3761. doi: 10.1021/acsbiomaterials.0c00109.

dos Santos-Pinto, J. R. A. *et al.* (2014) 'Structure and post-translational modifications of the web silk protein spidroin-1 from *Nephila* spiders', *Journal of Proteomics*. Elsevier, pp. 174–185. doi: 10.1016/j.jprot.2014.01.002.

dos Santos-Pinto, J. R. A. *et al.* (2018) 'Spider silk proteome provides insight into the structural characterization of *Nephila clavipes* flagelliform spidroin', *Scientific Reports*, 8(1). doi: 10.1038/s41598-018-33068-9.

Duffey, E. (1956) 'Aerial dispersal in a known spider population', *The Journal of Animal Ecology*, 25(1), p. 85. doi: 10.2307/1852.

Duffey, E. (1998) 'Aerial dispersal in spiders', *Proceedings of the 17th European Colloquium of Arachnology*, (1937), pp. 187–191. Available at: http://www.european-arachnology.org/proceedings/17th/23Duffey.pdf (Accessed: 27 January 2021).

Dufton, M. J. (1997) 'Genetic code synonym quotas and amino acid complexity: cutting the cost of proteins?'. *Journal of Theoretical Biology.* 187 (2). P. 165-173.

Eberhard, W. G. (2010) 'Possible functional significance of spigot placement on the spinnerets of spiders', *Journal of Arachnology*, 38(3), pp. 407–414. doi: 10.1636/B09-97.1.

Edlund, A. M. *et al.* (2018) 'Economic feasibility and environmental impact of synthetic spider silk production from *Escherichia coli*', *New Biotechnology*, 42, pp. 12–18. doi: 10.1016/j.nbt.2017.12.006.

Eid, J. *et al.* (2009) 'Real-time DNA sequencing from single polymerase molecules', *Science*, 323(5910), pp. 133–138. doi: 10.1126/science.1162986.

Elena, C. *et al.* (2014) 'Expression of codon optimized genes in microbial systems: Current industrial applications and perspectives', *Frontiers in Microbiology*. Frontiers Research Foundation, p. 21. doi: 10.3389/fmicb.2014.00021.

Esposito, D. and Chatterjee, D. K. (2006) 'Enhancement of soluble protein expression through the use of fusion tags', *Current Opinion in Biotechnology*, 17(4), pp. 353–358. doi: 10.1016/j.copbio.2006.06.003.

Fahnestock, S. R. and Bedzyk, L. A. (1997) 'Production of synthetic spider dragline silk protein in *Pichia pastoris*', *Applied Microbiology and Biotechnology*, 47(1), pp. 33–39. doi: 10.1007/s002530050884.

Fathi-Roudsari, M., Akhavian-Tehrani, A. and Maghsoudi, N. (2016) 'Comparison of three *Escherichia coli* strains in recombinant production of reteplase', *Avicenna Journal of Medical Biotechnology*, 8(1), pp. 16–22. Available at: /pmc/articles/PMC4717461/ (Accessed: 2 March 2021).

Finnigan, W. *et al.* (2020) 'The effect of terminal globular domains on the response of recombinant mini-spidroins to fiber spinning triggers', *Scientific Reports*, 10(1). doi: 10.1038/s41598-020-67703-1.

Flynn, M. R. and Bush, J. W. M. (2008) 'Underwater breathing: The mechanics of plastron respiration', *Journal of Fluid Mechanics*, 608, pp. 275–296. doi: 10.1017/S0022112008002048.

Foelix, R. F. (1983) 'Biology of Spiders', *Insect Systematics and Evolution*, 14(1), pp. 16–16. doi: 10.1163/187631283X00371.

Fox, E. J. and Reid-Bayliss, K. S. (2014) 'Accuracy of Next Generation Sequencing Platforms', *Journal of Next Generation Sequencing & Applications*, 01(01). doi: 10.4172/2469-9853.1000106.

Frandsen, P. B. *et al.* (2019) 'Exploring the underwater silken architectures of caddisworms: Comparative silkomics across two caddisfly suborders', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1784). doi: 10.1098/rstb.2019.0206.

Fu, L. *et al.* (2012) 'CD-HIT: Accelerated for clustering the next-generation sequencing data', *Bioinformatics*, 28(23), pp. 3150–3152. doi: 10.1093/bioinformatics/bts565.

Gaines IV, W. A. and Marcotte, W. R. (2008) 'Identification and characterization of multiple Spidroin 1 genes encoding major ampullate silk proteins in *Nephila clavipes*', *Insect Molecular Biology*, 17(5), pp. 465–474. doi: 10.1111/j.1365-2583.2008.00828.x.

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(32), pp. 11379–11384. doi: 10.1073/pnas.0502473102.

Garb, J. E. *et al.* (2007) 'Expansion and intragenic homogenization of spider silk genes since the triassic: evidence from mygalomorphae (tarantulas and their kin) spidroins', *Molecular Biology and Evolution*, 24(11), pp. 2454–2464. doi: 10.1093/molbev/msm179.

Garb, J. E. *et al.* (2019) 'The transcriptome of Darwin's bark spider silk glands predicts proteins contributing to dragline silk toughness', *Communications Biology*, 2(1), pp. 1–8. doi: 10.1038/s42003-019-0496-1.

Garb, J. E., Ayoub, N. A. and Hayashi, C. Y. (2010) 'Untangling spider silk evolution with spidroin terminal domains', *BMC Evolutionary Biology*, 10(1), p. 243. doi: 10.1186/1471-2148-10-243.

Garnier, J., Osguthorpe, D. J. and Robson, B. (1978) 'Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins', *Journal of Molecular Biology*, 120(1), pp. 97–120. doi: 10.1016/0022-2836(78)90297-8.

Garrison, N. L. *et al.* (2016) 'Spider phylogenomics: Untangling the spider tree of life', *PeerJ*, 2016(2), p. e1719. doi: 10.7717/peerj.1719.

Gasteiger, E. *et al.* (2003) 'ExPASy: The proteomics server for in-depth protein knowledge and analysis', *Nucleic Acids Research*, 31(13), pp. 3784–3788. doi: 10.1093/nar/gkg563.

Gavish-Regev, E. and Bosmans, R. (2012) 'A new synonymy in a linyphiid spider from Egypt (Araneae : Linyphiidae)', *SERKET The Arachnological Bulletin of the Middle East and North Africa*, 13(1–2), pp. 1–5.

Gavrielatos, M. *et al.* (2021) 'Benchmarking of next and third generation sequencing technologies and their associated algorithms for de novo genome assembly', *Molecular Medicine Reports*, 23(4), pp. 1–1. doi: 10.3892/mmr.2021.11890.

Geckil, H. *et al.* (2010) 'Engineering hydrogels as extracellular matrix mimics', *Nanomedicine*, 5(3), pp. 469–484. doi: 10.2217/nnm.10.12.

Geerlof, A. (2010) 'M9 mineral medium', *Helmholtz Center Munich*, pp. 14–15. Available at: https://www.helmholtz-muenchen.de/fileadmin/PEPF/Protocols/M9-medium_150510.pdf (Accessed: 22 July 2021).

Geneious (2018) 'Bioinformatics software for sequence data analysis'. Available at: https://www.geneious.com/ (Accessed: 20 October 2020).

Gertsch, W. J. (1958) 'The spider genus Loxosceles in North America, Central America and the West Indies', *American Museum Novitates*, 1907(1907), pp. 1–46. Available at: https://repositories.lib.utexas.edu/handle/2152/29952 (Accessed: 6 July 2021).

Giesa, T. *et al.* (2011) 'Nanoconfinement of spider silk fibrils begets superior strength, extensibility, and toughness', *Nano Letters*, 11(11), pp. 5038–5046. doi: 10.1021/nl203108t.

Glime, J. and Lissner, J. (2013) 'Volume 2, Chapter 7-3 Arthropods: Arachinda - Spider Habitats', *Bryophyte Ecology Subchapters*. Available at: https://digitalcommons.mtu.edu/bryo-ecol-subchapters/93 (Accessed: 6 May 2021).

Goodwin, S., McPherson, J. D. and McCombie, W. R. (2016) 'Coming of age: Ten years of next-generation sequencing technologies', *Nature Reviews Genetics*, 17(6), pp. 333–351. doi: 10.1038/nrg.2016.49.

Gosline, J. *et al.* (2002) 'Elastic proteins: Biological roles and mechanical properties', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 357(1418), pp. 121–132. doi: 10.1098/rstb.2001.1022.

Gosline, J. M. *et al.* (1999) 'The mechanical design of spider silks: From fibroin sequence to mechanical function', *Journal of Experimental Biology*, 202(23), pp. 3295–3303. Available at: https://jeb.biologists.org/content/202/23/3295 (Accessed: 20 October 2020).

Gosline, J. M., Denny, M. W. and Demont, M. E. (1984) 'Spider silk as rubber', *Nature*, 309(5968), pp. 551–552. doi: 10.1038/309551a0.

Grabherr, M. G. *et al.* (2011) 'Full-length transcriptome assembly from RNA-Seq data without a reference genome', *Nature Biotechnology*, 29(7), pp. 644–652. doi: 10.1038/nbt.1883.

Gregoriž, M. *et al.* (2011) 'Darwin's bark spider: Giant prey in giant orb webs', *Journal of Arachnology*, 39(2), pp. 287–295. doi: 10.1636/CB10-95.1.

Grip, S. *et al.* (2006) 'Transient expression of a major ampullate spidroin 1 gene fragment from *Euprosthenops sp.* in mammalian cells', *Cancer Genomics and Proteomics*, 3(2), pp. 83–88.

Grip, S., Johansson, J. and Hedhammar, M. (2009) 'Engineered disulfides improve mechanical properties of recombinant spider silk', *Protein Science*, 18(5), pp. 1012–1022. doi: 10.1002/pro.111.

Guehrs, K. H. *et al.* (2008) 'Environmental conditions impinge on dragline silk protein composition', *Insect Molecular Biology*, 17(5), pp. 553–564. doi: 10.1111/j.1365-2583.2008.00826.x.

Guerette, P. A. *et al.* (1996) 'Silk properties determined by gland-specific expression of a spider fibroin gene family', *Science*, 272(5258), pp. 112–115. doi: 10.1126/science.272.5258.112.

Guo, C. *et al.* (2020) 'Engineering silk materials: From natural spinning to artificial processing', *Applied Physics Reviews*. American Institute of Physics Inc., p. 11313. doi: 10.1063/1.5091442.

Gurevich, A. *et al.* (2013) 'QUAST: Quality assessment tool for genome assemblies', *Bioinformatics*, 29(8), pp. 1072–1075. doi: 10.1093/bioinformatics/btt086.

Gustafsson, L. *et al.* (2020) 'Recombinant spider silk forms tough and elastic nanomembranes that are protein-permeable and support cell attachment and growth', *Advanced Functional Materials*, 30(40), p. 2002982. doi: 10.1002/adfm.202002982.

Gutiérrez-Preciado, A., Romero, H. and Peimbert, M. (2010) 'An evolutionary perspective on amino acids', *Nature Education*, 3(9).

Haas, B. J. *et al.* (2013) 'De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis', *Nature Protocols*, 8(8), pp. 1494–1512. doi: 10.1038/nprot.2013.084.

Hadley, N. F. and Quinlan, M. C. (1989) 'Cuticular permeability of the black widow spider *Latrodectus hesperus*', *Journal of Comparative Physiology*, 159(3), pp. 243–248. doi: 10.1007/BF00691500.

Han, L. *et al.* (2013) 'Analysis of a new type of major ampullate spider silk gene, MaSp1s', *International Journal of Biological Macromolecules*, 56, pp. 156–161. doi: 10.1016/j.ijbiomac.2013.01.034.

Harvey, D. *et al.* (2017) 'Antibiotic spider silk: site-specific functionalization of recombinant spider silk using "click" chemistry', *Advanced Materials*, 29(10), p. 1604245. doi: 10.1002/adma.201604245.

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', *Biological Journal of the Linnean Society*, 77(1), pp. 1–8. doi: 10.1046/j.1095-8312.2002.00099.x.

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', *Journal of Molecular Biology*, 275(5), pp. 773–784. doi: 10.1006/jmbi.1997.1478.

Hayashi, C. Y. and Lewis, R. V. (2001) 'Spider flagelliform silk: Lessons in protein design, gene structure, and molecular evolution', *BioEssays*, 23(8), pp. 750–756. doi: 10.1002/bies.1105.

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', *Molecular Biology and Evolution*, 21(10), pp. 1950–1959. doi: 10.1093/molbev/msh204.

Hayashi, C. Y., Shipley, N. H. and Lewis, R. V. (1999) 'Hypotheses that correlate the sequence, structure, and mechanical properties of spider silk proteins', in *International Journal of Biological Macromolecules*. Elsevier, pp. 271–275. doi: 10.1016/S0141-8130(98)00089-0.

Haynl, C. *et al.* (2020) 'Free-standing spider silk webs of the thomisid *Saccodomus formivorus* are made of composites comprising micro- and submicron fibers', *Scientific Reports*, 10(1), pp. 1–10. doi: 10.1038/s41598-020-74469-z.

Heather, J.M and Chain, B. (2016) 'The sequence of sequencers: The history of sequencing DNA.' *Genomics,* 107(1), pp. 1-8. doi: https://doi.org/10.1016/j.ygeno.2015.11.003.

Hedhammar, M. *et al.* (2008) 'Structural properties of recombinant nonrepetitive and repetitive parts of major ampullate spidroin 1 from *Euprosthenops australis*: Implications for fiber formation', *Biochemistry*, 47(11), pp. 3407–3417. doi: 10.1021/bi702432y.

Hedin, M. C. (1997a) 'Molecular phylogenetics at the population/species interface in cave spiders of the southern Appalachians (Araneae: Nesticidae: Nesticus', *Molecular Biology and Evolution*, 14(3), pp. 309–324. doi: 10.1093/oxfordjournals.molbev.a025766.

Hedin, M. C. (1997b) 'Speciational history in a diverse clade of habitat-specialized spiders (araneae: nesticidae: *Nesticus*): Inferences from geographic-based sampling', *Evolution*, 51(6), pp. 1929–1945. doi: 10.1111/j.1558-5646.1997.tb05115.x.

Hedin, M.C. and Maddison, W. P. (2001) A combined molecular approach to phylogeny of the jumping spider subfamily Dendryphantinae (Araneae: Salticidae). *Molecular Phylogenetics and Evolution.* 18(3), pp. 386-403. doi: 10.1006/mpev.2000.0883.

Heger, A. and Holm, L. (2000) 'Rapid automatic detection and alignment of repeats in protein sequences', *Proteins: Structure, Function and Genetics*, 41(2), pp. 224–237. doi: 10.1002/1097-0134(20001101)

Heidebrecht, A. *et al.* (2015) 'Biomimetic fibers made of recombinant spidroins with the same toughness as natural spider silk', *Advanced Materials*, 27(13), pp. 2189–2194. doi: 10.1002/adma.201404234.

Heim, M., Keerl, D. and Scheibel, T. (2009) 'Spider silk: From soluble protein to extraordinary fiber', *Angewandte Chemie - International Edition*, pp. 3584–3596. doi: 10.1002/anie.200803341.

Hénaut, Y. *et al.* (2018) 'An arboreal spider protects its offspring by diving into the water of tank bromeliads', *Comptes Rendus - Biologies*, 341(3), pp. 196–199. doi: 10.1016/j.crvi.2018.02.002.

Hendrickson, W. A., Horton, J. R. and LeMaster, D. M. (1990) 'Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): A vehicle for direct determination of three dimensional structure', *EMBO Journal*, 9(5), pp. 1665–1672. doi: 10.1002/j.1460-2075.1990.tb08287.x.

Henrikson, B. -I (1993) '*Sphagnum* mosses as a microhabitat for invertebrates in acidified lakes and the colour adaptation and substrate preference in *Leucorrhinia dubia* (Odonata, Anisoptera)', *Ecography*, 16(2), pp. 143–153. doi: 10.1111/j.1600-0587.1993.tb00066.x.

Hernández Duran, L. *et al.* (2021) 'Beyond spider personality: The relationships between behavioral, physiological, and environmental factors', *Ecology and Evolution*. John Wiley & Sons, Ltd, pp. 2974–2989. doi: 10.1002/ece3.7243.

Hinman, M. B., Jones, J. A. and Lewis, R. V. (2000) 'Synthetic spider silk: A modular fiber', *Trends in Biotechnology*. Elsevier Ltd, pp. 374–379. doi: 10.1016/S0167-7799(00)01481-5.

Holland, C. *et al.* (2012) 'Silk and synthetic polymers: Reconciling 100 degrees of separation', *Advanced Materials*, 24(1), pp. 105–109. doi: 10.1002/adma.201103664.

Holland, G. P. *et al.* (2008) 'Determining secondary structure in spider dragline silk by carbon-carbon correlation solid-state NMR spectroscopy', *Journal of the American Chemical Society*, 130(30), pp. 9871–9877. doi: 10.1021/ja8021208.

Hu, X. *et al.* (2005) 'Egg Case Protein-1', *Journal of Biological Chemistry*, 280(22), pp. 21220–21230. doi: 10.1074/jbc.M412316200.

Huang, X. *et al.* (2021) 'Frame-shifted proteins of a given gene retain the same function', *Nucleic Acids Research*, 48(8), pp. 4396–4404. doi: 10.1093/NAR/GKAA169.

Huelsenbeck, J. P. and Ronquist, F. (2001) 'MRBAYES: Bayesian inference of phylogenetic trees', *Bioinformatics*, 17(8), pp. 754–755. doi: 10.1093/bioinformatics/17.8.754.

Huemmerich, D. *et al.* (2004) 'Primary structure elements of spider dragline silks and their contribution to protein solubility', *Biochemistry*, 43(42), pp. 13604–13612. doi: 10.1021/bi048983q.

Huemmerich, D., Helsen, C. W., *et al.* (2004) 'Primary structure elements of spider dragline silks and their contribution to protein solubility', *Biochemistry*, 43(42), pp. 13604–13612. doi: 10.1021/bi048983q.

Huemmerich, D., Scheibel, T., *et al.* (2004) 'Novel assembly properties of recombinant spider dragline silk proteins', *Current Biology*, 14(22), pp. 2070–2074. doi: 10.1016/j.cub.2004.11.005.

Huemmerich, D., Slotta, U. and Scheibel, T. (2006) 'Processing and modification of films made from recombinant spider silk proteins', *Applied Physics A: Materials Science and Processing*, 82(2), pp. 219–222. doi: 10.1007/s00339-005-3428-5.

Humenik, M., Smith, A. M. and Scheibel, T. (2011) 'Recombinant Spider Silks—Biopolymers with Potential for Future Applications', *Polymers*, 3(1), pp. 640–661. doi: 10.3390/polym3010640.

Illumina (2021) *Illumina Sequencing Platforms*. Available at: https://emea.illumina.com/systems.html.

Ittah, S. *et al.* (2006) 'An essential role for the C-terminal domain of a dragline spider silk protein in directing fiber formation', *Biomacromolecules*, 7(6), pp. 1790–1795. doi: 10.1021/bm060120k.

Ittah, S. *et al.* (2007) 'A model for the structure of the C-terminal domain of dragline spider silk and the role of its conserved cysteine', *Biomacromolecules*, 8(9), pp. 2768–2773. doi: 10.1021/bm7004559.

Jansson, R. *et al.* (2015) 'Rational Design of Spider Silk Materials Genetically Fused with an Enzyme', *Advanced Functional Materials*, 25(33), pp. 5343–5352. doi: 10.1002/adfm.201501833.

Jeong, H., Kim, H. J. and Lee, S. J. (2016) 'Complete genome sequence of Escherichia coli strain BL21', *Genome Announcements*, 3(2), pp. 134–149. doi: 10.1128/genomeA.00134-15.

Jin, H. J. and Kaplan, D. L. (2003) 'Mechanism of silk processing in insects and spiders', *Nature*, 424(6952), pp. 1057–1061. doi: 10.1038/nature01809.

Joel, A. C. *et al.* (2015) 'Cribellate thread production in spiders: Complex processing of nano-fibres into a functional capture thread', *Arthropod Structure and Development*, 44(6), pp. 568–573. doi: 10.1016/j.asd.2015.07.003.

Johannesson, H. *et al.* (2005) 'Concerted evolution in the repeats of an immunomodulating cell surface protein, SOWgp, of the human pathogenic fungi Coccidioides immitis and C. posadasii', *Genetics*, 171(1), pp. 109–117. doi: 10.1534/genetics.105.040923.

Jones, J. A. *et al.* (2015) 'More than just fibers: An aqueous method for the production of innovative recombinant spider silk protein materials', *Biomacromolecules*, 16(4), pp. 1418–1425. doi: 10.1021/acs.biomac.5b00226.

Karatzas, C., Turner, J. and Lazaris-Karatzas, A. (2007) 'Production of biofilaments in transgenic animals'. Available at: https://patents.google.com/patent/US7157615B2/en (Accessed: 11 May 2021).

Kelley, L. A. *et al.* (2015) 'The Phyre2 web portal for protein modeling, prediction and analysis', *Nature Protocols*, 10(6), pp. 845–858. doi: 10.1038/nprot.2015.053.

Keten, S. and Buehler, M. J. (2010) 'Nanostructure and molecular mechanics of spider dragline silk protein assemblies', *Journal of the Royal Society Interface*, 7(53), pp. 1709–1721. doi: 10.1098/rsif.2010.0149.

Kim, H., Yoo, S. J. and Kang, H. A. (2015) 'Yeast synthetic biology for the production of recombinant therapeutic proteins', *FEMS Yeast Research*, 15(1), pp. 1–16. doi: 10.1111/1567-1364.12195.

Kim, H.-J., Sun, Y. and Moon, M.-J. (2020) 'Fine structure of the silk spinning system in the caddisworm, *Hydatophylax nigrovittatus* (Trichoptera: Limnephilidae)', *Applied Microscopy*, 50(1), pp. 1–11. doi: 10.1186/s42649-020-00036-5.

Kimple, M. E., Brill, A. L. and Pasker, R. L. (2013) 'Overview of affinity tags for protein purification', *Current Protocols in Protein Science*, 73. doi: 10.1002/0471140864.ps0909s73.

Kiseleva, A. P., Krivoshapkin, P. V. and Krivoshapkina, E. F. (2020) 'Recent advances in development of functional spider silk0based hybrid materials'. *Frontiers in Chemistry.* 8. doi: 10.3389/fchem.2020.00554

Koren, S. *et al.* (2017) 'Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation', *Genome Research*, 27(5), pp. 722–736. doi: 10.1101/gr.215087.116.

Koutmou, K. S. *et al.* (2015) 'Ribosomes slide on lysine-encoding homopolymeric A stretches', *eLife*, 2015(4). doi: 10.7554/eLife.05534.

Krafft, B., Cookson, L. J. (2012) 'The role of silk in the behaviour and sociality of spiders'*, Journal of Entomology,* 2012(529564). doi: 10.1155/2012/529564.

Kronqvist, N. *et al.* (2017) 'Efficient protein production inspired by how spiders make silk', *Nature Communications*, 8. doi: 10.1038/ncomms15504.

Kuhbier, J. W. *et al.* (2010) 'Interactions between spider silk and cells - NIH/3T3 fibroblasts seeded on miniature weaving frames', *PLoS ONE*, 5(8), p. e12032. doi: 10.1371/journal.pone.0012032.

Lamboni, L. *et al.* (2015) 'Silk sericin: A versatile material for tissue engineering and drug delivery', *Biotechnology Advances*, 33(8), pp. 1855–1867. doi: 10.1016/j.biotechadv.2015.10.014.

Lammel, A. *et al.* (2011) 'Recombinant spider silk particles as drug delivery vehicles', *Biomaterials*, 32(8), pp. 2233–2240. doi: 10.1016/j.biomaterials.2010.11.060.

Lang, G., Jokisch, S. and Scheibel, T. (2013) 'Air filter devices including nonwoven meshes of electrospun recombinant spider silk proteins.', *Journal of visualized experiments*, (75), p. 50492. doi: 10.3791/50492.

Lazaris, A. *et al.* (2002) 'Spider silk fibers spun from soluble recombinant silk produced in mammalian cells', *Science*, 295(5554), pp. 472–476. doi: 10.1126/science.1065780.

Lehtinen, P. T. and Saaristo, M. I. (1980) 'Nesticidae', in *Spiders of the Oriental-Australian region*. 1st edn. Finnish Zoological and Botanical Publishing Board, pp. 47–66. Available at: https://www.jstor.org/stable/23733723?seq=2#metadata_info_tab_contents (Accessed: 5 July 2021).

Levi, H. W. (1967) 'Adapations of respiratory systems of spiders', *Evolution*, 21(3), p. 571. doi: 10.2307/2406617.

Lewis, R.V. (2006) 'Spider Silk: Ancient ideas for new biomaterials', *Chemical Reviews*. 106 (9) 3762-3774.

Li, X. *et al.* (2017) 'The correlation between the length of repetitive domain and mechanical properties of the recombinant flagelliform spidroin', *Biology Open*, 6(3), pp. 333–339. doi: 10.1242/bio.022665.

Liao, X. *et al.* (2019) 'Current challenges and solutions of *de novo* assembly', *Quantitative Biology*, 7(2), pp. 90–109. doi: 10.1007/s40484-019-0166-9.

Liebmann, B. *et al.* (2008) 'Formulation of poorly water-soluble substances using self-assembling spider silk protein', *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, 331(1–2), pp. 126–132. doi: 10.1016/j.colsurfa.2008.04.005.

Lin, Y., Ballarin, F. and Shuqiang, L. (2016) 'A survey of the spider family Nesticidae (Arachnida, Araneae) in Asia and Madagascar, with the description of forty-three new species', *ZooKeys*, 627, pp. 1–168.

Liu, Y. *et al.* (2008) 'Proline and processing of spider silks', *Biomacromolecules*, 9(1), pp. 116–121. doi: 10.1021/bm700877g.

Liu, Y. *et al.* (2021) 'Progress in the application of spider silk protein in medicine', *Journal of Biomaterials Applications*. doi: 10.1177/08853282211003850.

Liu, Y., Shao, Z. and Vollrath, F. (2008) 'Elasticity of spider silks', *Biomacromolecules*, 9(7), pp. 1782–1786. doi: 10.1021/bm7014174.

Locket, G. and Millidge, A. (1951) *British Spiders*. London: Ray Society.

Logsdon, G. A. *et al.* (2021) 'The structure, function and evolution of a complete human chromosome 8', *Nature*, 593(7857), pp. 101–107. doi: 10.1038/s41586-021-03420-7.

Lovett S. (2004) 'Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences'. *Molecular Microbiology*, 52(5), pp. 1243-1253. doi:10.1111/j.1365-2958.2004.04076.x

Lubin, Y. D. (1974) 'Adaptive advantages and the evolution of colony formation in *Cyrtophora* (Araneae: Araneidae)', *Zoological Journal of the Linnean Society*, 54(4), pp. 321–339. doi: 10.1111/j.1096-3642.1974.tb00806.x.

Madeira, F. *et al.* (2019) 'The EMBL-EBI search and sequence analysis tools APIs in 2019', *Nucleic Acids Research*, 47(W1), pp. W636–W641. doi: 10.1093/nar/gkz268.

Madsen, B., Shao, Z. Z. and Vollrath, F. (1999) 'Variability in the mechanical properties of spider silks on three levels: Interspecific, intraspecific and intraindividual', in *International Journal of Biological Macromolecules*, pp. 301–306. doi: 10.1016/S0141-8130(98)00094-4.

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(8), pp. 1–16. doi: 10.1371/journal.pone.0183397.

Mammola, S. *et al.* (2018) 'A synthesis on cave-dwelling spiders in Europe', *Journal of Zoological Systematics and Evolutionary Research*, 56(3), pp. 301–316. doi: 10.1111/jzs.12201.

Mammola, S., Cavalcante, R. and Isaia, M. (2016) 'Ecological preference of the diving bell spider *Argyroneta aquatica* in a resurgence of the Po plain (Northern Italy) (Araneae: Cybaeidae)', *Fragmenta Entomologica*, 48(1), pp. 9–16. doi: 10.4081/fe.2016.158.

Marais, G. *et al.* (2005) 'Intron size and exon evolution in *Drosophila*', *Genetics*, 170(1), pp. 481–485. doi: 10.1534/genetics.104.037333.

Margulies, M. *et al.* (2005) 'Genome sequencing in microfabricated high-density picolitre reactors', *Nature*, 437(7057), pp. 376–380. doi: 10.1038/nature03959.

Massimiliano, M. (2010) Nesticus eremita, *Wikimedia Commons*. Available at: https://commons.wikimedia.org/wiki/File:Nesticus_eremita_2.jpg (Accessed: 16 July 2021).

Mattina C, L. *et al.* (2008) 'Spider Minor Ampullate Silk Proteins Are Constituents of Prey Wrapping Silk in the Cob Weaver *Latrodectus hesperus*', *Biochemistry,* 47, pp. 4692-4700. doi: 10.1021/bi800140q.

Mclay, C. L. and Hayward, T. L. (1987) 'Reproductive biology of the intertidal spider *Desis marina* (Araneae: Desidae) on a New Zealand rocky shore', *Journal of Zoology*, 211(2), pp. 357–372. doi: 10.1111/j.1469-7998.1987.tb01539.x.

McGrath K., Kaplan, D. (1996) Protein-based Materials: Silk. Chapter 4, pp. 103-131. doi: 10.1007/978-1-4612-4094-5_4.

McQueen, D. J., Pannell, L. K. and McLay, C. L. (1983) 'Respiration rates for the intertidal spider *Desis marina* (Hector)', *New Zealand Journal of Zoology*, 10(4), pp. 393–399. doi: 10.1080/03014223.1983.10423934.

Menassa, R. *et al.* (2004) 'Spider dragline silk proteins in transgenic tobacco leaves', *Plant Biotechnology Journal*, 2(5), pp. 431–438. doi: 10.1111/j.1467-7652.2004.00087.x.

Messina, M. *et al.* (2015) 'Sulfidic spring in the gypsum karst system of Monte Conca (Italy): chemistry and microbiological evidences', *International Journal of Speleology*, 44(2), pp. 125–139. doi: 10.5038/1827-806X.44.2.3.

Mestre, L. and Lubin, Y. (2011) 'Settling where the food is: Prey abundance promotes colony formation and increases group size in a web-building spider', *Animal Behaviour*, 81(4), pp. 741–748. doi: 10.1016/j.anbehav.2011.01.002.

Miller, J. A. (2005) 'A redescription of *Porrhomma cavernicola* (Keyserling) (Araneae, Linyphiidae) with notes on Appalachian troglobites', *Journal of Arachnology*, 33(2), pp. 426–438. doi: 10.1636/04-51.1.

Miller, J. R., Koren, S. and Sutton, G. (2010) 'Assembly algorithms for next-generation sequencing data', *Genomics*. Genomics, pp. 315–327. doi: 10.1016/j.ygeno.2010.03.001.

Miller, J. S. (2015) 'The realized benefits from bioprospecting in the wake of the convention on biological diversity', *Washington University Journal of Law & Policy*, 47(1).

Morris, J., *et al.* (2013) 'Chapter 4, Translation and Protein Structure' in *Biology: How Life Works.* 3rd edition. New York: Macmillan, pp 4-2, 4-3.

Motriuk-Smith, D. *et al.* (2005) 'Analysis of the conserved N-terminal domains in major ampullate spider silk proteins', *Biomacromolecules*, 6(6), pp. 3152–3159. doi: 10.1021/bm050472b.

Müller-Herrmann, S. and Scheibel, T. (2015) 'Enzymatic Degradation of Films, Particles, and Nonwoven Meshes Made of a Recombinant Spider Silk Protein', *ACS Biomaterials Science and Engineering*, 1(4), pp. 247–259. doi: 10.1021/ab500147u.

Munro, T. *et al.* (2017) 'Investigation of Synthetic Spider Silk Crystallinity and Alignment via Electrothermal, Pyroelectric, Literature XRD, and Tensile Techniques', *Macromolecular Materials and Engineering*, 302(4), p. 1600480. doi: 10.1002/mame.201600480.

Murigneux, V. *et al.* (2021) 'Comparison of long-read methods for sequencing and assembly of a plant genome', *GigaScience*, 9(12), pp. 1–11. doi: 10.1093/gigascience/giaa146.

Nakano, K. *et al.* (2017) 'Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area', *Human Cell*. Springer Tokyo, pp. 149–161. doi: 10.1007/s13577-017-0168-8.

Neumann, D. and Kureck, A. (2013) 'Composite structure of silken threads and a proteinaceous hydrogel which form the diving bell wall of the water spider *Agyroneta aquatica*', *SpringerPlus*, 2(1), pp. 1–5. doi: 10.1186/2193-1801-2-223.

Neumann, D. and Woermann, D. (2013) 'Stability of the volume of air trapped on the abdomen of the water spider *Argyroneta aquatica*', *SpringerPlus*, 2(1), pp. 1–7. doi: 10.1186/2193-1801-2-694.

Nentwig W., Heimer S. (1987) Ecological Aspects of Spider Webs. In: Nentwig W. (eds) Ecophysiology of Spiders. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-71552-5_15

NMBE (2021) *World Spider Catalog*. Available at: https://wsc.nmbe.ch/reference/10410 (Accessed: 27 January 2021).

Nørgaard, E. (1951) 'On the Ecology of Two Lycosid Spiders (*Pirata piraticus* and *Lycosa pullata*) from a Danish Sphagnum bog', *Oikos*, 3(1), p. 1. doi: 10.2307/3565128.

Nyffeler, M. and Pusey, B. J. (2014) 'Fish predation by semi-aquatic spiders: A global pattern', *PLoS ONE*, 9(6), p. 99459. doi: 10.1371/journal.pone.0099459.

Oger, P. (2019) '*Lessertia barbara*: Les araignées de Belgique et de France', *Les araignées de Belgique et de France*. Available at: https://arachno.piwigo.com/picture?/41996/category/2739-lessertia_barbara (Accessed: 16 July 2021).

PacBio (2019) *SMRT Sequencing - How it works*. [online] Available at: https://www.pacb.com/smrt-science/attachment/infographic_smrt-sequencing-how-it-works/ (Accessed: 9 March 2021).

PacBio (2020) *HIFI Reads for highly accurate long read sequencing*. [online] Available at: https://www.pacb.com/smrt-science/smrt-sequencing/hifi-reads-for-highly-accurate-long-read-sequencing/ (Accessed: 9 March 2021).

PacBio (2020) *Sequel System*. [online] Available at: https://www.pacb.com/products-and-services/sequel-system/previous-system-releases/.

Pantini, P. and Isaia, M. (2019) 'Araneae.it: the online Catalog of Italian spiders, with addenda on other Arachnid Orders occurring in Italy (Arachnida: Araneae, Opiliones, Palpigradi, Pseudoscorpionida, Scorpiones, Solifugae)', *Fragmenta entomologica*, 51(2), pp. 127–152. Available at: www.araneae.it, (Accessed: 27 January 2021).

Papadopoulos, P. *et al.* (2009) 'Similarities in the structural organization of major and minor ampullate spider silk', *Macromolecular Rapid Communications*, 30(9–10), pp. 851–857. doi: 10.1002/marc.200900018.

Partlow, B. P. *et al.* (2016) 'Tyrosine Templating in the Self-Assembly and Crystallization of Silk Fibroin', *Biomacromolecules*, 17(11), pp. 3570–3579. doi: 10.1021/acs.biomac.6b01086.

Patro, R. *et al.* (2017) 'Salmon provides fast and bias-aware quantification of transcript expression', *Nature Methods*, 14(4), pp. 417–419. doi: 10.1038/nmeth.4197.

Peng, C. A. *et al.* (2017) 'Polyelectrolyte Fiber Assembly of Plant-Derived Spider Silk-like Proteins', *Biomacromolecules*, 18(3), pp. 740–746. doi: 10.1021/acs.biomac.6b01552.

Pérez-Rigueiro, J. *et al.* (2019) 'Emergence of supercontraction in regenerated silkworm (*Bombyx mori*) silk fibers', *Scientific Reports*, 9(1), pp. 1–14. doi: 10.1038/s41598-019-38712-6.

Perry, D. J. *et al.* (2010) 'Piriform spider silk sequences reveal unique repetitive elements', *Biomacromolecules*, 11(11), pp. 3000–3006. doi: 10.1021/bm1007585.

Persi, E., Wolf, Y. I. and Koonin, E. V (2016) 'Positive and strongly relaxed purifying selection drive the evolution of repeats in proteins', *Nature Communications*, 7(1), pp. 1–11. doi: 10.1038/ncomms13570.

Peters, H. M. (1993) 'Functional organization of the spinning apparatus of *Cyrtophora citricola* with regard to the evolution of the web (Araneae, Araneidae)', *Zoomorphology*, 113(3), pp. 153–163. doi: 10.1007/BF00394856.

Peters, H. M. and Kovoor, J. (1991) 'The silk-producing system of *Linyphia triangularis* (Araneae, Linyphiidae) and some comparisons with Araneidae - Structure, histochemistry and function', *Zoomorphology*, 111(1), pp. 1–17. doi: 10.1007/BF01632706.

Pollard, M. O. *et al.* (2018) 'Long reads: Their purpose and place', *Human Molecular Genetics*. NLM (Medline), pp. R234–R241. doi: 10.1093/hmg/ddy177.

Posnien, N. *et al.* (2014) 'A comprehensive reference transcriptome resource for the common house spider *Parasteatoda tepidariorum*', *PLoS ONE*. Edited by D. J. Jackson, 9(8), p. e104885. doi: 10.1371/journal.pone.0104885.

Pouchkina-Stantcheva, N. N. and McQueen-Mason, S. J. (2004) 'Molecular studies of a novel dragline silk from a nursery web spider, *Euprosthenops* sp. (Pisauridae)', *Comparative biochemistry and physiology. Part B, Biochemistry and molecular biology*, 138(4), p. 371—376. doi: 10.1016/j.cbpc.2004.04.020.

Pugh, P. J. A. (2004) 'Biogeography of spiders (Araneae: Arachnida) on the islands of the Southern Ocean', *Journal of Natural History*, 38(12), pp. 1461–1487. doi: 10.1080/0022293031000155403.

Rabotyagova, O. S., Cebe, P. and Kaplan, D. L. (2009) 'Self-assembly of genetically engineered spider silk block copolymers', *Biomacromolecules*, 10(2), pp. 229–236. doi: 10.1021/bm800930x.

Rachwał, M., Rybak, J. and Rogula-Kozłowska, W. (2018) 'Magnetic susceptibility of spider webs as a proxy of airborne metal pollution', *Environmental Pollution*, 234, pp. 543–551. doi: 10.1016/j.envpol.2017.11.088.

Rammensee, S. *et al.* (2008) 'Assembly mechanism of recombinant spider silk proteins', *Proceedings of the National Academy of Sciences of the United States of America*, 105(18), pp. 6590–6595. doi: 10.1073/pnas.0709246105.

Rat, C. *et al.* (2018) 'Two-step self-assembly of a spider silk molecular clamp', *Nature Communications*, 9(1), pp. 1–11. doi: 10.1038/s41467-018-07227-5.

Reed, C. J. *et al.* (2013) 'Protein adaptations in archaeal extremophiles', *Archaea*, 2013. doi: 10.1155/2013/373275.

Rhoads, A. and Au, K. F. (2015) 'PacBio Sequencing and Its Applications', *Genomics, Proteomics and Bioinformatics*. Beijing Genomics Institute, pp. 278–289. doi: 10.1016/j.gpb.2015.08.002.

Ripley, L. S. (1990) 'Frameshift mutation: determinants of specificity'. *Annual Review of Genetics*, 24, pp. 189-213.

Rising, A. and Johansson, J. (2015) 'Toward spinning artificial spider silk', *Nature Chemical Biology*. Nature Publishing Group, pp. 309–315. doi: 10.1038/nchembio.1789.

Rising, A. *et al.* (2005) 'Spider silk proteins - Mechanical property and gene sequence', *Zoological Science*, 22(3), pp. 273–281. doi: 10.2108/zsj.22.273.

Robert, X., Gouet, P. (2014) 'Deciphering key features in protein structures with the new ENDscript server'. *Nucleic Acids Research,* 42(W1), pp.W320-324. doi: 10.1093/nar/gku316.

Römer, L. and Scheibel, T. (2008) 'The elaborate structure of spider silk: structure and function of a natural high performance fiber.', *Prion*, 2(4), pp. 154–161. doi: 10.4161/pri.2.4.7490.

Rosano, G. L. and Ceccarelli, E. A. (2014) 'Recombinant protein expression in *Escherichia coli*: Advances and challenges', *Frontiers in Microbiology*, 5(APR), p. 172. doi: 10.3389/fmicb.2014.00172.

Rousseau, M. E., Lefèvre, T. and Pézolet, M. (2009) 'Conformation and orientation of proteins in various types of silk fibers produced by *Nephila clavipes* spiders', *Biomacromolecules*, 10(10), pp. 2945–2953. doi: 10.1021/bm9007919.

Růžička, V., Vaněk, J. and Šmilauer, P. (2012) 'Spiders in mountain habitats of the Giant Mountains', *Russian Journal of Ecology*, 43(4), pp. 341–347. doi: 10.1134/S1067413612040133.

Salehi, S., Koeck, K. and Scheibel, T. (2020) 'Spider silk for tissue engineering applications', *Molecules*, 25(3). doi: 10.3390/molecules25030737.

Santos-Pinto, J. R. A. dos *et al.* (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(10), pp. 1444–1454. doi: 10.1016/j.bbapap.2016.05.007.

Savage, K. N. and Gosline, J. M. (2008) 'The effect of proline on the network structure of major ampullate silks as inferred from their mechanical and optical properties', *Journal of Experimental Biology*, 211(12), pp. 1937–1947. doi: 10.1242/jeb.014217.

Schacht, K. and Scheibel, T. (2011) 'Controlled hydrogel formation of a recombinant spider silk protein', *Biomacromolecules*, 12(7), pp. 2488–2495. doi: 10.1021/bm200154k.

Scheibel, T. (2004) 'Spider silks: Recombinant synthesis, assembly, spinning, and engineering of synthetic proteins', *Microbial Cell Factories*. BioMed Central, p. 14. doi: 10.1186/1475-2859-3-14.

Scheller, J. *et al.* (2001) 'Production of spider silk proteins in tobacco and potato', *Nature Biotechnology*, 19(6), pp. 573–577. doi: 10.1038/89335.

Schuster, S. C. (2008) 'Next-generation sequencing transforms today's biology', *Nature Methods*, 5(1), pp. 16–18. doi: 10.1038/nmeth1156.

Schütz, D. and Taborsky, M. (2003) 'Adaptations to an aquatic life may be responsible for the reversed sexual size dimorphism in the water spider, Argyroneta aquatica', *Evolutionary Ecology Research*, 5, pp. 105–117.

Schütz, D. and Taborsky, M. (2005) 'The influence of sexual selection and ecological constraints on an extreme sexual size dimorphism in a cichlid', *Animal Behaviour*, 70(3), pp. 539–549.

Seemann, T. and Gladman, S. (2012) 'Fasta Statistics: Display summary statistics for a fasta file'. Available at: https://github.com/galaxyproject/tools-iuc.

Selden, P. A. (2002) 'Missing links between *Argyroneta* and *Cybaeidae* revealed by fossil spiders', *Journal of Arachnology*, 30(2), pp. 189–200. doi: 10.1636/0161-8202(2002)030[0189:MLBAAC]2.0.CO;2.

Seymour, R. S. and Hetz, S. K. (2011) 'The diving bell and the spider: The physical gill of *Argyroneta aquatica*', *Journal of Experimental Biology*, 214(13), pp. 2175–2181. doi: 10.1242/jeb.056093.

Seyyar, O. and Demir, H. (2009) 'Distribution and habitats of the water spider *Argyroneta aquatica* (Clerck, 1757) (Araneae, Cybaeidae) in Turkey', *Archives of Biological Sciences*, 61(4), pp. 773–776. doi: 10.2298/ABS0904773S.

Shah, D. U., Porter, D. and Vollrath, F. (2014) 'Can silk become an effective reinforcing fibre? A property comparison with flax and glass reinforced composites', *Composites Science and Technology*, 101, pp. 173–183. doi: 10.1016/j.compscitech.2014.07.015.

Shao, Z. and Vollrath, F. (1999) 'The effect of solvents on the contraction and mechanical properties of spider silk', *Polymer*, 40(7), pp. 1799–1806. doi: 10.1016/S0032-3861(98)00266-3.

Sheldon, K. S. *et al.* (2017) 'Revisiting the physics of spider ballooning', in *Association for Women in Mathematics Series*, pp. 163–178. doi: 10.1007/978-3-319-60304-9_9.

Shortreed, C. G. *et al.* (2020) 'Characterization of 100 extended major histocompatibility complex haplotypes in Indonesian cynomolgus macaques', *Immunogenetics*, 72(4), pp. 225–239. doi: 10.1007/s00251-020-01159-5.

Simão, F. A. *et al.* (2015) 'BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs', *Bioinformatics*, 31(19), pp. 3210–3212. doi: 10.1093/bioinformatics/btv351.

Slatko, B.E. *et al.* (2018) 'Overview of Next-Generation Sequencing technologies.' *Current Protocols in Molecular Biology*, 122(1):e59. doi: 10.1002/cpmb.59. PMID: 29851291.

Slotta, U. K. *et al.* (2008) 'An engineered spider silk protein forms microspheres', *Angewandte Chemie - International Edition*, 47(24), pp. 4592–4594. doi: 10.1002/anie.200800683.

Sørensen, H. P. and Mortensen, K. K. (2005) 'Soluble expression of recombinant proteins in the cytoplasm of Escherichia coli', *Microbial Cell Factories*, 4(1), pp. 1–8. doi: 10.1186/1475-2859-4-1.

Sparkes, J. and Holland, C. (2018) 'The rheological properties of native sericin', *Acta Biomaterialia*, 69, pp. 234–242. doi: 10.1016/j.actbio.2018.01.021.

Spiess, K., Lammel, A. and Scheibel, T. (2010) 'Recombinant Spider Silk Proteins for Applications in Biomaterials'. *Macromolecular Bioscience*. 10, pp. 998-1007. https://doi.org/10.1002/mabi.201000071

Sponner, A. *et al.* (2004) 'Conserved C-termini of spidroins are secreted by the major ampullate glands and retained in the silk thread', *Biomacromolecules*, 5(3), pp. 840–845. doi: 10.1021/bm034378b.

Sponner, A. *et al.* (2005) 'The conserved C-termini contribute to the properties of spider silk fibroins', *Biochemical and Biophysical Research Communications*, 338(2), pp. 897–902. doi: 10.1016/j.bbrc.2005.10.048.

Stark, M. *et al.* (2007) 'Macroscopic fibers self-assembled from recombinant miniature spider silk proteins', *Biomacromolecules*, 8(5), pp. 1695–1701. doi: 10.1021/bm070049y.

Stauffer, S. L., Coguill, S. L. and Lewis, R. V (1994) 'Comparison of Physical Properties of Three Silks from *Nephila clavipes* and *Araneus Gemmoides*', *Journal of Arachnology*, 22(1), pp. 5–11. Available at: https://www.jstor.org/stable/3705704?seq=1 (Accessed: 20 October 2020).

Stellwagen, S. D. and Renberg, R. L. (2018) 'Towards spider glue: long-read scaffolding for extreme length and repetitious silk family genes AgSp1 and AgSp2 with insights into functional adaptation', *bioRxiv*, p. 492025. doi: 10.1101/492025.

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, Genetics*, 9(6), pp. 1909–1919. doi: 10.1534/g3.119.400065.

Stewart, R. J. and Wang, C. S. (2010) 'Adaptation of caddisfly larval silks to aquatic habitats by phosphorylation of h-fibroin serines', *Biomacromolecules*, 11(4), pp. 969–974. doi: 10.1021/bm901426d.

Strickland, M. *et al.* (2018) 'Conservation of a pH-sensitive structure in the C-terminal region of spider silk extends across the entire silk gene family', *Heredity*, 120(6), pp. 574–580. doi: 10.1038/s41437-018-0050-9.

Sutherland, T, D. *et al.,* (2010) 'Insect Silk: One Name, Many Materials', *Annual Review of Entomology*. 55:1, 171-188

Sutherland, T, D. *et al.,* (2011) 'The coiled coil silk of bees, ants and hornets'. *Biopolymers*. 97(6), pp. 446-454. doi: 10.1002/bip.21702.

Suryawanshi, R. *et al.* (2020) 'Sericin: A Versatile Protein Biopolymer with Therapeutic Significance', *Current Pharmaceutical Design*, 26(42), pp. 5414–5429. doi: 10.2174/1381612826666200612165253.

Swanson, B. O., Blackledge, T. A. and Hayashi, C. Y. (2007) 'Spider capture silk: Performance implications of variation in an exceptional biomaterial', *Journal of Experimental Zoology Part A: Ecological Genetics and Physiology*, 307(11), pp. 654–666. doi: 10.1002/jez.420.

Tasiopoulos, C. P., Widhe, M. and Hedhammar, M. (2018) 'Recombinant Spider Silk Functionalized with a Motif from Fibronectin Mediates Cell Adhesion and Growth on Polymeric Substrates by Entrapping Cells during Self-Assembly', *ACS Applied Materials and Interfaces*, 10(17), pp. 14531–14539. doi: 10.1021/acsami.8b02647.

Teulé, F. *et al.* (2007) 'Modifications of spider silk sequences in an attempt to control the mechanical properties of the synthetic fibers', *Journal of Materials Science*, 42(21), pp. 8974–8985. doi: 10.1007/s10853-007-1642-6.

Teulé, F. *et al.* (2009) 'A protocol for the production of recombinant spider silk-like proteins for artificial fiber spinning', *Nature Protocols*, 4(3), pp. 341–355. doi: 10.1038/nprot.2008.250.

Teulé, F., Addison, B., *et al.* (2012) 'Combining flagelliform and dragline spider silk motifs to produce tunable synthetic biopolymer fibers', *Biopolymers*, 97(6), pp. 418–431. doi: 10.1002/bip.21724.

Teulé, F., Miao, Y. G., *et al.* (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(3), pp. 923–928. doi: 10.1073/pnas.1109420109.

Thomas, C. F. G., Brain, P. and Jepson, P. C. (2003) 'Aerial activity of linyphiid spiders: Modelling dispersal distances from meteorology and behaviour', *Journal of Applied Ecology*, 40(5), pp. 912–927. doi: 10.1046/j.1365-2664.2003.00844.x.

Thorpe, W. H. (1950) 'Plastron respiration in aquatic insects', *Biological Reviews*, 25(3), pp. 344–390. doi: 10.1111/j.1469-185X.1950.tb01590.x.

Tian, M. and Lewis, R. V. (2006) 'Tubuliform silk protein: A protein with unique molecular characteristics and mechanical properties in the spider silk fibroin family', *Applied Physics A: Materials Science and Processing*, 82(2), pp. 265–273. doi: 10.1007/s00339-005-3433-8.

Tokareva, O. *et al.* (2014) 'Structure-function-property-design interplay in biopolymers: Spider silk', *Acta Biomaterialia*, 10(4), pp. 1612–1626. doi: 10.1016/j.actbio.2013.08.020.

Treangen, T. J. and Salzberg, S. L. (2012) 'Repetitive DNA and next-generation sequencing: Computational challenges and solutions', *Nature Reviews Genetics*. NIH Public Access, pp. 36–46. doi: 10.1038/nrg3117.

Tsuboi, T. *et al.* (2012) 'Dom34: Hbs1 plays a general role in quality-control systems by dissociation of a stalled ribosome at the 3' end of aberrant mRNA', *Molecular Cell*, 46(4), pp. 518–529. doi: 10.1016/j.molcel.2012.03.013.

Tszydel, M. *et al.* (2015) 'Research on possible medical use of silk produced by caddisfly larvae of *Hydropsyche angustipennis* (Trichoptera, Insecta)', *Journal of the Mechanical Behavior of Biomedical Materials*, 45, pp. 142–153. doi: 10.1016/j.jmbbm.2015.02.003.

Urry, D. W., Luan, C. H. and Peng, S. Q. (1995) 'Molecular biophysics of elastin structure, function and pathology.', *Ciba Foundation symposium*. doi: 10.1002/9780470514771.ch2.

Uzenbaev, S. D. and Lyabzina, S. N. (2009) 'An experimental study of the effects of spider venom on animals', *Entomological Review*, 89(4), pp. 479–486. doi: 10.1134/S0013873809040125.

Van Beek, J. D. *et al.* (2002) 'The molecular structure of spider dragline silk: Folding and orientation of the protein backbone', *Proceedings of the National Academy of Sciences of the United States of America*, 99(16), pp. 10266–10271. doi: 10.1073/pnas.152162299.

Venkatesan, H., Chen, J. and Hu, J. (2019) 'Fibers made of recombinant spidroins—a brief review', *AATCC Journal of Research*, 6(1), pp. 37–40. doi: 10.14504/ajr.6.S1.8.

Vicherová, E. *et al.* (2017) '*Sphagnum* establishment in alkaline fens: Importance of weather and water chemistry', *Science of the Total Environment*, 580, pp. 1429–1438. doi: 10.1016/j.scitotenv.2016.12.109.

Vienneau-Hathaway, J. M. *et al.* (2017) 'Duplication and concerted evolution of MiSp-encoding genes underlie the material properties of minor ampullate silks of cobweb weaving spiders', *BMC Evolutionary Biology*, 17(1). doi: 10.1186/s12862-017-0927-x.

Vink, C. J. and Dupérré, N. (2011) '*Nesticus eremita* (Araneae: Nesticidae): redescription of a potentially invasive European spider found in New Zealand', *Journal of Arachnology*, 39(3), pp. 511–514. doi: 10.1636/A11-57.1.

Vollrath, F. (1999) 'Biology of spider silk', in *International Journal of Biological Macromolecules*. Int J Biol Macromol, pp. 81–88. doi: 10.1016/S0141-8130(98)00076-2.

Vollrath, F. (2000) 'Strength and structure of spiders' silks,' *Reviews in Molecular Biotechnology*, 74 (2) pp 67-83, doi: https://doi.org/10.1016/S1389-0352(00)00006-4.

Vollrath, F. and Knight, D. P. (2001) 'Liquid crystalline spinning of spider silk', *Nature*, 410(6828), pp. 541–548. doi: 10.1038/35069000.

Vollrath, F. *et al.* (1990) 'Compounds in the droplets of the orb spider's viscid spiral', *Nature*, 345(6275), pp. 526–528. doi: 10.1038/345526a0.

von Byern, J. *et al.* (2016) 'Characterization of the fishing lines in titiwai (*Arachnocampa luminosa* Skuse, 1890) from New Zealand and Australia', *PLoS ONE*, 11(12), p. 162687. doi: 10.1371/journal.pone.0162687.

von Byern, J. *et al.* (2019) 'Biomechanical properties of fishing lines of the glowworm *Arachnocampa luminosa* (Diptera; Keroplatidae)', *Scientific Reports*, 9(1), pp. 1–14. doi: 10.1038/s41598-019-39098-1.

Walker, A. A. *et al.* (2015) 'The other prey-capture silk: Fibres made by glow-worms (Diptera: Keroplatidae) comprise cross-β-sheet crystallites in an abundant amorphous fraction', *Comparative Biochemistry and Physiology Part - B: Biochemistry and Molecular Biology*, 187, pp. 78–84. doi: 10.1016/j.cbpb.2015.05.008.

Wasowska, S. (1977) 'Studies on the spinning apparatus in spiders.', *Zoologica Poloniae*, 26, pp. 355–407.

Weichert, N. *et al.* (2014) 'Transglutamination allows production and characterization of native-sized ELPylated spider silk proteins from transgenic plants', *Plant Biotechnology Journal*, 12(2), pp. 265–275. doi: 10.1111/pbi.12135.

Weirather, J. L. *et al.* (2015) 'Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing', *Nucleic Acids Research*, 43(18), p. e116. doi: 10.1093/nar/gkv562.

Wen, R., Liu, X. and Meng, Q. (2017) 'Characterization of full-length tubuliform spidroin gene from *Araneus ventricosus*', *International Journal of Biological Macromolecules*, 105(Pt 1), pp. 702–710. doi: 10.1016/j.ijbiomac.2017.07.086.

Westbrook, C. J. *et al.* (2015) 'No assembly required: Full-length MHC class I allele discovery by PacBio circular consensus sequencing', *Human Immunology*, 76(12), pp. 891–896. doi: 10.1016/j.humimm.2015.03.022.

Wheeler, W. C. *et al.* (2017) 'The spider tree of life: phylogeny of Araneae based on target-gene analyses from an extensive taxon sampling', *Cladistics*, 33(6), pp. 574–616. doi: 10.1111/cla.12182.

Whittall, D. R. *et al.* (2020) 'Host systems for the production of recombinant spider silk', *Trends in Biotechnology*, 0(0). doi: 10.1016/j.tibtech.2020.09.007.

Widmaier, D. M. and Voigt, C. A. (2010) 'Quantification of the physiochemical constraints on the export of spider silk proteins by *Salmonella* type III secretion', *Microbial Cell Factories*, 9(1), pp. 1–12. doi: 10.1186/1475-2859-9-78.

Widmaier, D. M. *et al.* (2009) 'Engineering the *Salmonella* type III secretion system to export spider silk monomers', *Molecular Systems Biology*, 5. doi: 10.1038/msb.2009.62.

Winkler, S., Wilson, D. and Kaplan, D. L. (2000) 'Controlling β-sheet assembly in genetically engineered silk by enzymatic phosphorylation/dephosphorylation', *Biochemistry*, 39(41), pp. 12739–12746. doi: 10.1021/bi001335w.

Wolff, J. O. *et al.* (2015) 'Spider's super-glue: Thread anchors are composite adhesives with synergistic hierarchical organization', *Soft Matter*, 11(12), pp. 2394–2403. doi: 10.1039/c4sm02130d.

Woolley, C. *et al.* (2016) 'Aerial dispersal activity of spiders sampled from farmland in southern England', *Journal of Arachnology*, 44(3), pp. 347–358. doi: 10.1636/P15-56.1.

Work, R. W. (1985) 'Viscoelastic behaviour and wet supercontraction of major ampullate silk fibres of certain orb-web-building spiders (Araneae)', *Journal of Experimental Biology*, VOL. 118(1), pp. 379–404.

World Spider Catalog (2020) *World Spider Catalog (2020)*, *Natural History Museum Bern*. Available at: https://wsc.nmbe.ch/ (Accessed: 20 October 2020).

Wray, L. S. *et al.* (2011) 'Effect of processing on silk-based biomaterials: Reproducibility and biocompatibility', *Journal of Biomedical Materials Research - Part B Applied Biomaterials*, 99 B(1), pp. 89–101. doi: 10.1002/jbm.b.31875.

Wunderlich, J. (1995) 'On the taxonomy of European genera of dwarf spiders (Arachnida: Araneae: Linyphiidae: Erigoninae)', *Contributions to Araneologie*, 4, p. 643/654.

Xia, X. X. *et al.* (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(32), pp. 14059–14063. doi: 10.1073/pnas.1003366107.

Xu, L. *et al.* (2012) 'Recombinant Minimalist Spider Wrapping Silk Proteins Capable of Native-Like Fiber Formation', *PLoS ONE*. Edited by A. Mitraki, 7(11), p. e50227. doi: 10.1371/journal.pone.0050227.

Xu, L. *et al.* (2017) 'Structural and mechanical roles for the C-Terminal nonrepetitive domain become apparent in recombinant spider aciniform silk', *Biomacromolecules*, 18(11), pp. 3678–3686. doi: 10.1021/acs.biomac.7b01057.

Yang, J. *et al.* (2005) 'High yield recombinant silk-like protein production in transgenic plants through protein targeting', *Transgenic Research*, 14(3), pp. 313–324. doi: 10.1007/s11248-005-0272-5.

Yang, Z. *et al.* (2000) 'Supercontraction and backbone dynamics in spider silk: 13C and 2H NMR studies', *Journal of the American Chemical Society*, 122(37), pp. 9019–9025. doi: 10.1021/ja0017099.

Yip, E. C. *et al.* (2016) 'Coy males and seductive females in the sexually cannibalistic colonial spider, *Cyrtophora citricola*', *PLoS ONE*, 11(6). doi: 10.1371/journal.pone.0155433.

Yoshida, M. (1999) 'Spinnerets Silk Spigots Morphology of Araneidae, Tetragnathidae, Theridiidae and Linyphiidae (Araneae: Araneoidea)', *Acta arachnologica*, 48(1), pp. 1–22.

Zhang, H. *et al.* (2015) 'Microbial production of amino acid-modified spider dragline silk protein with intensively improved mechanical properties', *Preparative Biochemistry and Biotechnology*, 46(6), pp. 552–558. doi: 10.1080/10826068.2015.1084637.

Zhao, L. *et al.* (2017) 'Studies on the use of recombinant spider silk protein/polyvinyl alcohol electrospinning membrane as wound dressing', *International Journal of Nanomedicine*, 12, pp. 8103–8114. doi: 10.2147/IJN.S47256.

Zhao, S. *et al.* (2021) 'Recombinant silk proteins with additional polyalanine have excellent mechanical properties', *International Journal of Molecular Sciences*, 22(4), pp. 1–14. doi: 10.3390/ijms22041513.

Zimmermann, M. and Spenee, J. R. (1989) 'Prey use of the fishing spider *Dolomedes triton* (Pisauridae, Araneae): an important predator of the neuston community', *Oecologia*, 80, pp. 187–194.

Zurn, J. D. *et al.* (2020) 'A rosaceae family-Level approach to identify loci influencing soluble solids content in blackberry for DNA informed breeding', *G3: Genes, Genomes, Genetics*, 10(10), pp. 3729–3740. doi: 10.1534/g3.120.401449.

PIP Reflective 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 original placement was due to take place between May and July of 2020, but unfortunately this was cancelled because of the COVID-19 outbreak and subsequent lockdown precautions. Because of this, I applied to complete a three-month placement working for Severn Trent water, who had just issued an urgent request for help in their microbiology labs, as they were experiencing staffing shortages because of the effects of COVID-19. To this end, I started work as a microbiology analyst in the Church Wilne labs of Severn Trent Water in the second week of the UK-wide lockdown, analysing samples from across the water network. It was quite strange from the outset, as I was suddenly designated a 'key worker' and was allowed to be outside when others were confined to their houses! Coupled with this, I had to navigate a social distancing approach in the laboratories, as I was trained in all aspects of the sample testing and result analysis workflow.

It was very exciting getting to learn about Severn Trent, the standards for water quality and process management in a professional testing lab, as well as how samples were tracked and managed; all things I had to learn very quickly so I was able to be a help from day one. As part of Severn Trent's commitment to ensuring that all results are reported correctly and in a timely manner, the laboratory was undertaking a new

method validation. I was able to take ownership of this project, alongside my daily testing duties, carrying out the tests and interpreting the results. I also undertook a critical appraisal of the method that was being used, as well as a cost analysis compared to current methodology. I then produced a report for the regulatory authority, UKAS, which was used as part of submitted evidence to apply for accreditation of this new method, as well as formed part of an internal review to implement a new testing process.

During my placement I found that I could quite easily apply skills I had learned from my PhD and previous work experience to a new laboratory setting, which gave me a head start to settle in as part of the team. It was very rewarding to be able to follow a method validation process from start to finish, and produce the final report, as well as understand what was needed from the new testing regime by the lab and the accreditation authority. I also gained a real insight into how water is supplied across the country, and how many samples are taken every day to ensure that high quality standards are met. I really learnt the importance of a reliable sample management software, as well as the value of cooperating teams from across the business. I was also able to see the implementation of a new method for monitoring water quality using the flow cytometer. This was a new technique to me, and it was interesting seeing how new methods are adopted into the water testing process and what regulation is needed.

There were added pressures that came about from working during the early stages of a pandemic, such as adapting to regular checks to make sure we were social distancing, a very strange socially distanced lunch, and long shifts which minimised the number of people needed each day. However, I was grateful that I was able to help during such a strange time and use my time wisely to minimise the initial impact that not being in the lab had on my PhD. Some days I longed to be at home (especially when it was such lovely weather and it felt like the whole world was ‘lockdown sunbathing’), but I enjoyed being productive in a laboratory environment. While this placement was not what I had planned for, I left with a range of new skills, knowledge, and a new view of a highly regulated working environment.



Figure P.1: Pictures of the Severn Trent laboratory. (a) filtration membrane incubated on MLGA for 16 hrs at 37°C after filtering 1 ml, 10 mL or 100 mL of untreated water. Green colonies are presumptive E. coli *and yellow colonies are coliforms.*

The Appendices

Appendix A.2.1

Silk sequence designations, origins and number of times the sequence appears per million transcripts (Transcripts Per Million, TPM) in each transcriptome (from Chapter 2), calculated using Salmon (Patro et al., 2017). Column 3 shows whether the sequence contains the N-terminal region (NT), repetitive region (R) or the C-terminal region (CT).

Silk Type	Name	NT, CT or Repetitive region	Species	Transcripts Per Million (TPM)	Found in both acid/non?
MaSp	LBN_DN2841_c0_g1_i1_MaSp1_CT	R, CT	*Lessertia* (Non-acid)	2223.1	Y
	LBA_DN39510_c0_g1_i1_MaSp1_NT	NT, R	*Lessertia* (Acid)	26.9	N
	NES_DN24404_c1_g2_i1_MaSp CT	R, CT	*Nesticus*	45.0	
	NES_DN24404_c1_g2_i2_MaSp CT	R, CT	*Nesticus*	528.5	
	NES_DN27466_c0_g1_i1_MaSp CT	CT	*Nesticus*	22.3	
MiSp	LESA_DN34551_c0_g1_i1_MiSp_CT	CT, R	*Lessertia (*Acid)	6.49	N
	NES_DN27100_c0_g1_i1_MiSp_CT	CT	*Nesticus*	125.6	
PySp	LESA_DN39694_c0_g1_i1_Acid_PySp	R	*Lessertia (*Acid)	60.20	N
	LESN_DN3248_c0_g2_i1_PySp_CT	CT	*Lessertia (*non-acid)	578.81	Y
TuSp	LBN_DN4358_c0_g1_i1_TuSp_NT	NT	*Lessertia* (non-acid)	55.09	Y
	LBN_DN1824_c0_g1_i1	R	*Lessertia* (non-acid)	29.25	N
	LBA_DN38972_c0_g1_i1 _TuSp_NT	NT	*Lessertia* (Acid)	67.12	Y
	NES_DN29553_c0_g1_i1_TuSp CT	CT	*Nesticus*	87.97	
	NES_DN27103_c0_g1_i1	R	*Nesticus*	18.96	
AgSp	LESA_37082_c2_g2_i1_PySp	R	*Lessertia* (Acid)	54.1	Y
	LESA_37082_c2_g1_i5_PySp	R	*Lessertia* (Acid)	188.3	Y
	LESA_31683_c0_g2_i1_PySp	R	*Lessertia* (Acid)	73.1	Y
	LESA_22521_c0_g1_i2_PySp_CT	CT	*Lessertia* (Acid)	11.3	N
	LESN_5109_c0_g2_i1_PySp	R	*Lessertia* (non-acid)	351.1	Y
	LESN_4768_c0_g1_i5_PySp	R	*Lessertia* (non-acid)	7.2	N
	LESN_4768_c0_g1_i2_PySp	R	*Lessertia* (non-acid)	11.3	N
	NES_DN32135_c0_g1_i1_AgSp2 NT?	R	*Nesticus*	6.12	
	NES_DN31477_c0_g1_i1_AgSp2	R	*Nesticus*	267.17	
	NES_DN30928_c0_g1_i1_AgSp2	R	*Nesticus*	10.41	
	NES_DN29978_c0_g5_i1_AgSp2	R, CT	*Nesticus*	145.80	
	NES_DN29978_c0_g2_i2_AgSp2 NT?	R	*Nesticus*	98.66	
	NES_DN29978_c0_g2_i1_AgSp2 NTR?	R	*Nesticus*	367.57	
	NES_DN29955_c0_g2_i1_AgSp2	R	*Nesticus*	3.85	
	NES_DN29955_c0_g1_i1_AgSp2	R	*Nesticus*	131.84	
	NES_DN29917_c0_g1_i5_AgSp2	R	*Nesticus*	18.43	
	NES_DN29610_c1_g7_i1_AgSp2 NT?	R	*Nesticus*	14.48	

Table continued on next page

	NES_DN29610_c1_g6_i1_AgSp2	R	*Nesticus*	26.97	
	NES_DN29610_c1_g5_i1_AgSp2	NT? R	*Nesticus*	0.002	
	NES_DN29472_c0_g2_i1_AgSp2 NT?	NT, R	*Nesticus*	13.78	
	NES_DN28860_c0_g1_i1_AgSp2	R	*Nesticus*	9.47	
	NES_DN28593_c1_g1_i1_AgSp2	R	*Nesticus*	18.82	
	NES_DN28593_c0_g1_i1_AgSp2	R	*Nesticus*	7.33	
	NES_DN27911_c0_g1_i1_AgSp2	NT? R	*Nesticus*	25.48	
	NES_DN27651_c0_g1_i1_AgSp2 NTR?	R	*Nesticus*	180.12	
	NES_DN27587_c1_g1_i1_AgSp2 NT?	R	*Nesticus*	7.266183	
	NES_DN27484_c0_g1_i1_AgSp2	R	*Nesticus*	32.37	
	NES_DN26217_c0_g2_i1_AgSp2	R	*Nesticus*	10.12	
	NES_DN25081_c0_g1_i1_AgSp2	R	*Nesticus*	9.45	
	NES_DN24609_c0_g1_i1_AgSp2 NT?	R	*Nesticus*	18.69	
	NES_DN23934_c0_g1_i1_AgSp2	R	*Nesticus*	21.86	
	NES_DN20217_c0_g2_i1_AgSp2 NT?	R	*Nesticus*	19.50	
	NES_DN11781_c0_g2_i1_AgSp2	R	*Nesticus*	2.47	
	NES_DN7033_c0_g1_i1_AgSp2 NT?	R	*Nesticus*	11.69	

Appendix A.2.2

Amino acid composition of the repetitive regions of major ampullate silk types used for analysis in Chapter 2. Included are the MaSp sequences from this study and a selection of published sequences.

Accession number/ name	DN39510_c0_g1_i1	DN2841_c0_g1_i1	DN24404	DN24404_c1_g2_i2	DN27466_c0_g1_i1	CAM32272	AWK58758	AAZ15322	ABD66603	ABD61594	ABD61593	AAZ15371	AF350278	AWK58730	AWK58638	AWK58729	ABD66602	CAM32262	AF350266_1	AAY28935
Species	*L. barbara*	*L. barbara*	*N. eremita*	*N. eremita*	*N. eremita*	*E. australis*	*T. clavipes*	*T. madagascariensis*	*L. hesperus*	*D. spinosa*	*D. spinosa*	*A. trifasciata*	*T. madagascariensis*	*L. hesperus*	*L. hesperus*	*A. argentata*	*L. hesperus*	*E. australis*	*A. trifasciata*	*L. hesperus*
MaSp type	1	1	2	2	2	2	2	2	2	2	2	2	2	3	3	3	1	1	1	1
Ala (A)	53.30%	52.20%	25.60%	25.60%	28.10%	40.60%	29.30%	27.70%	32.80%	23.60%	19.40%	20.50%	22.60%	11.30%	11.40%	11.70%	34.90%	42.80%	29.60%	35.60%
Arg (R)	2.20%	1.30%	1.80%	1.80%	9.40%	1.80%	2.00%	2.40%	1.40%	1.10%	1.20%	1.40%	0.60%	2.40%	4.10%	2.90%	1.40%	1.40%	0.90%	1.60%
Asn (N)	0.00%	3.10%	0.00%	0.00%	3.10%	1.80%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	4.10%	3.30%	3.50%	0.00%	1.40%	0.00%	0.00%
Asp (D)	2.20%	1.90%	0.00%	0.00%	3.10%	0.20%	1.00%	0.00%	0.10%	0.10%	0.40%	0.00%	0.20%	2.80%	2.90%	1.20%	0.00%	0.00%	0.40%	0.00%
Cys (C)	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Gln (Q)	0.00%	0.60%	0.60%	0.60%	3.10%	12.50%	7.10%	9.30%	6.80%	18.80%	18.60%	14.60%	12.60%	11.10%	11.00%	4.70%	11.10%	10.90%	11.30%	10.80%
Glu (E)	0.70%	0.00%	4.20%	4.20%	0.00%	0.00%	0.00%	0.10%	0.00%	0.30%	0.20%	0.20%	0.00%	3.20%	3.70%	4.70%	0.00%	0.00%	0.50%	0.00%
Gly (G)	27.70%	27.00%	34.50%	34.50%	18.80%	32.50%	26.30%	38.00%	35.50%	32.90%	36.20%	35.60%	34.20%	17.30%	13.10%	26.30%	44.00%	31.90%	43.80%	43.30%
His (H)	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.10%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Ile (I)	0.00%	0.00%	0.00%	0.00%	6.20%	0.00%	1.00%	0.00%	0.00%	0.10%	0.10%	0.20%	0.00%	4.10%	2.40%	4.10%	0.10%	0.00%	0.00%	0.20%
Leu (L)	0.00%	0.60%	1.20%	1.20%	3.10%	0.20%	1.00%	0.10%	0.10%	0.10%	0.20%	0.00%	0.40%	1.70%	2.40%	2.30%	0.30%	1.10%	3.10%	0.50%
Lys (K)	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.10%	0.00%	0.00%	0.00%	1.10%	1.60%	2.30%	0.00%	0.00%	0.00%	0.00%

Table continued on next page

Met (M)	0.00%	0.00%	0.00 %	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.20%	0.00%	0.00%	0.00 %	0.60%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Phe (F)	0.00%	0.00%	0.60 %	0.60%	0.00%	4.60%	0.00%	0.00%	0.40%	0.10%	0.00%	0.00%	0.00 %	1.70%	1.20%	0.60%	0.10%	0.00%	0.20%	0.00%
Pro (P)	0.00%	0.00%	3.00 %	3.00%	3.10%	2.00%	13.10 %	14.50 %	9.90%	11.80 %	11.20 %	13.80 %	16.90 %	4.30%	4.90%	4.70%	0.40%	0.00%	0.00%	0.20%
Ser (S)	9.50%	8.80%	17.90 %	17.90%	18.80%	3.10%	13.10 %	2.80%	6.60%	4.60%	4.60%	7.90%	7.30 %	13.90 %	14.70 %	20.50 %	2.10%	6.00%	5.30%	1.90%
Thr (T)	0.00%	0.00%	1.80 %	1.80%	0.00%	0.20%	0.00%	0.40%	0.80%	1.30%	2.20%	0.80%	0.00 %	7.50%	7.30%	2.30%	0.10%	0.20%	0.40%	0.50%
Trp (W)	0.00%	0.00%	0.00 %	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00 %	0.20%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Tyr (Y)	4.40%	4.40%	7.10 %	7.10%	0.00%	0.00%	5.10%	4.40%	5.20%	3.60%	3.80%	4.30%	4.90 %	6.00%	6.50%	4.10%	5.30%	4.00%	4.00%	5.20%
Val (V)	0.00%	0.00%	1.80 %	1.80%	3.10%	0.00%	1.00%	0.20%	0.40%	1.20%	1.70%	0.80%	0.20 %	6.80%	9.40%	4.10%	0.10%	0.40%	0.50%	0.20%
Negatively charged	2.92%	1.89%	4.17 %	4.17%	3.13%	0.22%	1.00%	0.10%	0.09%	0.43%	0.62%	0.20%	0.20 %	5.98%	6.53%	5.85%	0.00%	0.00%	0.91%	0.00%
Positively charged	2.19%	1.26%	1.79 %	1.79%	9.38%	1.75%	2.00%	2.40%	1.44%	1.20%	1.23%	1.38%	0.61 %	3.42%	5.71%	5.26%	1.64%	1.41%	0.91%	1.64%
Length	137	159	168	168	32	456	100	1916	1110	920	810	508	491	468	245	171	856	568	550	427

Appendix A.2.3

A list of spidroins identified in the *Lessertia barbara* acid (LBA), non-acid (LBN) or *Nesticus eremita* transcriptomes. Each name identified the transcriptome of origin, unique identified, silk type and, if applicable, terminal region covered.

>LBA_DN38972_NT

ALEIHHDLACILLVAALAAPAAHSQFAQSQVPSVFGSASLASGFLQCVQLGIGSSPIFPVQE
QEDLDAISSVILSAVGSSGGSQRKAQALSNAFASSFVNLVYAESGASFDTQVEALSSILTDC
FVQTTGSDNSGFVNSITQQLNALASASAGASAGAQAQADSQVQASAQAQAQAAAQ

>LBA_DN39510_MaSp1_NT

AIMTWSIRLSLAFLVVLCTQSLVAFAQSPWANPQTADAFMNNFLSSVTSNSGGIFTGDQLSD
MSDIGFSIMNQLERLNRGGHSSARKMQMMDMAFASTMAEVALAEADNPQAALGAISNALSQS
FMQTTGSSNQASINSIMSLISMFAQNAGENVIPSGPVISGPSRGPSFNIQTYKQPAQVVQTQ
TSTGQNVGYSVTSYGGDGRGSAASSASSASAAAAGGAGGGYGGDGGAAAAAAAAAAAAAAGA
AEGYGRGAGSGSSAASAAAAGAGAAGAGGYADGGAAAAAAAAASAAAAAAAGGAGSGYGARY
GGYGAGGAGAASAAAAAAAAAAAAGA

>LBN_DN2841_MaSp1_CT

AAGAADGYGRGAGAGSNAASAAAAAAGAGAGSGGAGYGDGGAAAAAAAAAAAAAAAGSAGSG
YGAGYGGYGAGGAGAASAAAAAAAAAAAAASAGDGYGAGGAGSGASAASAAAAAGSAGAGGA
GGYGNGGAAAAAAAASAAAQASAAAARLNSANAGNRVSSAVNTLVTSGPTSVNSLSNVIGNA
VAQVRASSPGLSDCDVLVQALLEVTSAMVHILGSASLGAVNAAAASNSASIVSQSVQNAIG

>LBN_DN4358_TuSp_NT

QHLRFIMIWLATLLVAALAAPAAHSQFAQSQVPSVFGSASLASGFLQCVQLGIGSSPIFPVQ
EQEDLDAISSVILSAVGSSGGSQRKAQALSNAFASSFVNLVYAESGASFDTQVEALSSILTD
CFVQTTGSDNSGFVNSITQQLNALASASAGASAGAQAQADSQVQASAQAQAQAAAQAQAQAA
AQAQAQA

>LESA_DN22521_AgSp2_CT

SNYNLPPTDVLSEALEIEENSLGQHAGNEGTEENAEEVVYRRSAPSFNLEDFENIFVLTLKD
GSERIKTLTKALLKTKGHHSHNSLKLEHFVKKLSEVFGGMRYNNEYLTYSELYVQLLTETLI
AALEIIRETDSSCLTAPSTPSELPQYVNALYNILI

>LESA_DN31882_PySp_CT

SQVQQSALSQSSVSEVQQSDSSQIEDSDVSQIQQTTDVEVQAQQGYDNYDENSSFATSKNDL
SQDYSETSDSYFEQQSEFAASESSSSSEISSDIAASQSISASYSSNSALVSSAQKLMSPAAK
NRIAALSRSLASVVSGGQVNYSALSSSLSSVASQISGSSHGLSKSEVVVEALLETMAALLGT
MSSSSSGSNDFANCVLQAFA

>LESA_DN34551_MiSp_CT

QGSGYVSGAGAASGAGTAASAGSSGAGGYGQGAGYRTGAGVGAASGAVVAAGQSATVTASGV
SQVGQTLASTTARLSSPEASYRMSNIASALSSGGVLNLDALPSIISNQFAQVSSSSAGASSG
EVLVQMLLEIVSSLIHISNSSSVGQVDMAGIGNSAAAVSQSMTAVLG

>LESA_DN39694_PySp

NVQQSSVSQVQQSAVSQQASVQVQQQSYFNSQIIQSSLYSSLQSAQFIRAISTGQTVATYSE
IASTIAASLYSEFSLSQSVVQSSIYQRLSSLGSGASSQSIAQAISSAVSSCIQQSSIARSGQ

EQAISIRLYSSISASLTSLIQRRPRPAPRPVVVQVQQAAVSQVQQSALSQVQQSSVSQVQQS
AVSQVQQSSVSNVQQSSVSQVQQSAVSQQASVQV

>LESA_DN31683_AgSp2

GPNVPSGSSPSGPNVPGGSSPSGPNVPRGSSPVGPNVPGGASPSGPNVPGGSSLSGPTKPGA
TTPSSIPGPNGKLVPVELAGPGTTPGTVTGKDKKPVKYIVPKGAYSTPGTIPGPDGQEQIPV
GPAGPGTTPGTETGKNGKVVKIILPSTPSSPNVPG

>LESA_DN31683_AgSp

GPNVPSGSSPSGPNVPGGSSPSGPNVPRGSSPVGPNVPGGASPSGPNVPGGSSLSGPTKPGA
TTPSSIPGPNGKLVPVELAGPGTTPGTVTGKDKKPVKYIVPKGAYSTPGTIPGPDGQEQKAS
KIH

>LESA_DN37082_AgSp2

ETGKNGKVVKIILPSTPSSPSGGSSPSGPNVPSGSSPSGPKVPGGSSPSGPIVPGGSSPSGP
NVPGGSSPSGPTQPKATTPSSIPGPNGKQVPVETAGPGTTPGTVTGKDKQPVKYIVPKGAFS
TPGTIPGPNGKEQLPVGPAGPGTTPGVETGKHGKVVKIILPSTPSSPSGGSSPSGPKVSGGS
SPSGPIVPGGSSPSGPKVSGGS

>LESA_DN37082_AgSp2

PGTTPGTVTGKDKQPVKYIVPKGAFSTPGTIPGPDGKEQLPVGPAGPGTTPGTETGKNGKVI
KIILPSTPSSPNVPGGSTPSVPNTPKGPDSGENPEVPKGPDGKPCSPECLRACKKNIPFIFP
GDPDPIFKGGFIGFNKLPDTLNKLLKGTFTYPDLIDSLREFFPGKLDIDAVPITDLIGANNQ
FVIPGFVGSYFKLRLAAPVNTSNPYRGLFYIPEFVHLVPYLPVGALSGKPDNPAFPMQFPQS
YGEPVFICDHLDSHESSQGSQ

>LESN_DN3248_PySp_CT

STIAASLYSEFSLSQSVVQSSIYQRLSSLGSGASSQSIAQAISSAVSSCIQQSSIARSGQEQ
AISIRLYSSISASLTSLIQRRPRPAPRPVVVQVQQAAVSQVQQSALSQVQQSSVSQVQQSAV
SQVQQSAVSQSSVSEVQQSDSSQIEDSDVSQIQQTTDVEVQAQQGYDNYDENSSFATSKNDL
SQDYSETSDSYFEQQSEFAASESSSSSEISSDIAASQSISASYSSNSALVSSAQKLMSPAAK
NRIAALSRSLASVVSGGQVNYSALSSSLSSVASQISGSSHGLSKSEVVVEALLETMAALLGT
MSSSSSGSNDFANCVLQAFA

>LESN_DN4768_AgSp2

YESESGSYPSKPSKGHKPQKGSSPSGPNAPPKGTTPSVSGPGGKNIPVVQAGPGTTPGTITG
KDGKPKKFVVPKGAFSTPGTIPGPDGKPLPVGPAGPGTTPGTHTGKDGKIDKVVLPTTPKPT
SSSSEYSTYEYSSSEYSSSEYSTYESESGSYSTYESESGSYPSKPSKGHKHHKGSSPSGPNA
PPKATTPSSIPGPNGK

>LESN_DN4768_1_AgSp2

YESESGSYPSKPSKGHKHPKRSSPSGPNAPPKGTTPSVSGPGGKKIPVVQAGPGTTPGTITG
KDGKPTKFVVPKGAFSTPGTIPGPDGKPLPVGPAGPGTTPGTHTGKDSKIDKVVLPTTPKPT
SSSSEYSTYEYSSSEYSSSEYSTYESESGSYSTYESESGSYPSKPSKGHKPQKGSSPSGPNA
PPKGTTPSVSGPGGKNIPVVQAGPGTT

>LESN_DN5109_AgSp

SPSGPIVPGGSSPSGPSKPEATTPSSIPGPNGKLVPVETAGPGTTPGTVTGKDKQPVKYIVP
KGAFSTPGTIPGPNGKEQLPVGPAGPGTTPGVETGKNGKVVKIILPSTPSSPSGGSSPSGPN
VPRGSSPVGPNVPGGASPSGPNVPGGSSLSGPTKPGATTPSSIPGPNGKLVPVELAGPGTTP

GTVTGKDKKPVKYIVPKGAYSTPGTIPGPDGQEQIPVGPAGPGTTPGTETGKNGKVVKIILP
STPSSPNVPGGSTPSVPNTPKGPDSGENPEVPKGPDGKPCSPECLRACKKNIPFIFPGDPDP
IFKGGFIGFNKLPDTLNKLLKGTFTYPDLIDSLREFFPGKLDIDAVPITDLIGANNQFVIPG
FVGSYFKLRLAAPVNTSNPYRGLFYIPEFVHLVPYLPVGALSGKPDNPAFPMQFPQSYGEPV
FICDHLDSHESSQGSQ

>NES_DN24404_MaSp_CT

AAAAAAAAASGGEGGYGGGYSSGGYGGSSSSASAASSASGVGQGGYGGGTGGASGASAGASG
GSGGYGPSSFEVTEYGPGGRGPGGVGGYGPGGLGGYGRGGEGYGTGEGAASAAAAAAAAAAA
EAAGGEGSYGGYGAGGYGGGSSAASSASAASAAASRLASPASSSRISNAASSLVSSGVSNAG
ALSSVISNLSSQVSASNPGLSGCDVLVQVLMEVVSALVHIASSSNVGQVNYGSSGQYARLVS
QSVNQAFG

>NES_DN24404_1_MaSp_CT

AAAAAAAAASGGEGGYGGGYSSGGYGGSSSSASAASSASGVGQGGYGGGTGGASGASAGASG
GSGGYGPSSFEVTEYGPGGRGPGGVGGYGPGGLGGYGRGGEGYGTGEGAASAAAAAAAAAAA
EAAGGEGSYGGYGAGGYGGGSSAASSASAASAAASRLASPASSSRISNAASSLVSSGVSNAG
ALSSVISNLSSQVSASNPGLSGCDILVQVLMEVVSALVHIVSSSNVGQVNYGSSGQTAQLVS
QSVNQAFV

>NES_DN27100_MiSp_CT

QGYGTGAAVGAGSAASAGAGAAAGAGAAAGGYGYGRGAAQGYGSSAGAGAGAGAGAGAGAGA
AGSAGYGAGAASGAGSAVGNAFAQSMASNLLSSNDFVSMISTTTTTDQAVNVATNVAQNVAS
QLGLDTNALNNLIGTVQGYVSSLGAAADAAAYANAISFALGNVLANTGTISESSASTTASSV
ASSVTTTLSTYGPSVFVRDQGYGLGAAAGAAAGAAASGGSGYGRGSGQGYGAGAGAGAAAGA
GAGAVAGGAGYGRGAGQGYGAGSASSAGAGGYGSGQGYAAGSAAAAGASSGSGYGRGVGQGF
GAGAGAAAGATAGAAGFGGGGYGAAVGGGDVGATVSSTTSRLNSASATSRISSAASSLVAGG
ALNTAALPSVISDLYSQIAASSAGVSAGEVLVQVLLEIVSSLIHILSSSNVGQIDFSSVGSS
ASAVGQSLQAVMG

>NES_DN27466_MaSp_CT

GIDGGQGGARRGNASSAASIAAARLVSPASSARVSNAASNLVSSGVSNVASLSNVIGNLSSQ
VSSSNPELNGCDLLIQILLEVVSALIHILGSSVVGKVNYEAASQYTQLVSQSMNQAFG

>NES_DN7033_AgSp2

KKPGSVNGPYGPIKIVPETPGTTPGVVTGSNGKINKVIVPHGPPLPGSGSYTLKGPNNEVIE
LIPSGLTPGYITGEKGNPIRVVPAGPGTTPGAVTRSDGKLLKIAVPQGTTTGVKTVKGPDGK
VIMLLPGGPTPGAVTDAGGSILRFVL

>NES_DN11781_AgSp2

GPDGEPIPVEPAGPGSTPGAQTGRDGRIIKILLPTTPKSGAQQGTQQPEKPSTTPKSIRGPG
GREIQIVPAGPGTTPGTITGRDGRPVKFIVPLGAFSTPGEIPGPDGE

>NES_DN20217_AgSp2

LPAQTEQEVHTIEISVAGPGTTPGTVTDSSGHPSQIYVPLDAFYRPTSIPGPNGNPIPVGPN
GPGTTPGEVTDSDGNIIQIYLPAQPSTKPPDSDHAIDIVVAGPGTTPGTVTNSGGYPVKFVV
PLDAFYKPASFPGPHGEPILV

>NES_DN23934_AgSp2

PIAVVPAGPGTTPGAKRGPDGHINKIILPAMPEDLGSPGGLPAGPNGPGRQRPGIPGPGNSH
IFIVPAMPGVTPGTKLGPNGKPVMFILPQGVFRQPGTIPGPNGKPLIVRPVPKGSTPGAVTD
NSGNIVKVFLPTKAPKQPAIKPTRVEGPGGKKIEIVPARKGATPAIVA

>NES_DN24609_AgSp2

GNVVKVIIPDGSGTTPGVLSDRSGKFLQIVVPVGVDFNPGVVTGSDGKAYLIEPVSSGQTPG
VVTGPDGNIVKILVPSAVTPGPREPITIPVPNGRSIIIIPGGPTPGAVTGA

>NES_DN25081_AgSp2

GHPVKFVVPVGAYSHPGSIPGPNGKPLSVEPAEPGTTPGVVTDPNRSIVKIYVPKIAGGGGG
RQPSHVPKNPVDIKGPDDQQIMIVPAQGGSTPGTVTDSDGHPVKFVVP

>NES_DN26217_c0_g2_i1_AgSp2

SDGQPIEFVVGGPGNKLEEVRGPDGNLVFIIPKGGQSTPADAQDINGNKVRLVPERPGTTPG
VVTGKDGKPSKIVFPTGGPKGGIPQGHKREPIQILPRS

>NES_DN27484_c0_g1_i1_AgSp2

GTVTDSQGNPVQFIVPVDAFRAGGSIPGSDGEQIPVVIAGPGTTPGAEIGNDGSIIKIVLPS
TPSPPTQPAIPPAKRRRGSSSSIEIVPAGPGDTPGTVTNSDGDPIKFIVPVAAFSTPGYIPG
PDGNPLRVDTAGPGTTPGIETADDGTVVKIVVPKV

>NES _DN27587_c1_g1_i1_AgSp2

ADPGTITGPDGKPLIFLVPIGSFDNQDTIPGPDGGSIILEPSSPVTTPGAEIGPFGDIVKIH
LPPYPDQTGYPPTSTIKGPNGRIILIVPASEGATPGTVTDPDGYPIKFIIPLGAYTSQGSIP
GPNGNTIIVQ

>NES_DN27651_c0_g1_i1_AgSp2

ETPGAVTGPDGMITQILLPGYAGQYINSGETTVTGPDGMVIIITLAGPGRTPGTTVGPNGNI
RYIVPLDSLTAPSSITGPNGQTIIVRPAAPGDTPGAVSGSDGSITILIIPAYLSQYISPLTT
YIPSPYRGQIIIVPARPGTSPGTVTGPDGRPVEFIVPVGAFFAPSSLPGPNGQTIIVQPAGP
GETPKAITNSDGNIITIIIRSSMQVYPPQIRRQSGGQILIVPAGPGTTPGTVTDSDGNPTKF
IVPLSAFSSPASIPGPYGYQIPVVTAGPGSTPGAVTDSNGNIIQIVLPSTPRNPYAQQTVAT
RPPRKYNYGTRPIQIAVAGPGTTPGTVTDSDGSPVQFIVPVSAFSIGGYIPGPNGNMISVEI
AGPGTTPGAETGDDGSIVRLVLPSTPSPPAQPATPAPRPRRRGSSNAIEIVPAGEGTTPGTI
TDSDGNPVKFIVPLGAFSTPGYIPGPDGNPLEVDTAGPGTTPGVETASDGTIIKIIVPKV

>NES_DN27911_c0_g1_i1_AgSp2

GTVTSSDGKTFKYIVPISAYSSPGSIPGPNGNQIPVGPAGPGTTPGAVTQSDGTVLRIYLPK
MQISGGNSGGSGNVIQGQGRGSKNIQIVPAGPGESPGTVQSPDGKSLKFIVPISAFSSPSSI
PGPDGNQIPVGPAGPGTTPGAEYGGNGNVIRIYLPAMPGQSGNIPGGTQVLSGRSSKQIQIV
PAGPGESPGTVQSPDGKSLKFIVPVSAFSSPSSIPGPDGNQIP

>NES_DN28593_c0_g1_i1_AgSp2

GNIIKIVLPTTTSGPRLPEHKPERPIHVVGPTEQKILIVPAGPGSTPGTVTGSDGNPLMFVV
PLGAYSTPGSIPGLKGKPLPVETAGPGTTPGAVTGSDGSIVKLILPKSSGGGRQPPHVPEKP
VNLVGPGDRQIIIVPAEGGSTPGTVTDSDGKPVKFVVPVAAYSHPGSFPGPNGKPVYVEPAG
PGTT

>NES_DN28593_c1_g1_i1_AgSp2

NPYQPGTPGTTPATIHDPSGRTIIIVPAGPGDTPGTVTGPDGYPVKFIVPLGAFRTPGSIPG
PNGYPISVVTEGPGVTPGAETGPDGNIIRIILPTTPSFPVTNERIYDPNGRQIIIVPAGPGD
TPGTVTGPDGYPVKFIVPLGAFSTPGFI

>NES_DN28860_c0_g1_i1_AgSp2

PGTVTDSNGYVIKFIVPLSAFSTPGTIPGPHGNPVFVLTAGPGTTPAVETGPDGSIIRIILP
STPYRPVYPTTPKRKHNYNKTPIKIVPEGPGSTPGTVTDSDGNPTIFIVPLSAFSTPGSIPG
PDGYPIAVETAGPGTTPGAETGPDGSIIRIVLPTTPRSYSTKKPRRIQDPNGNQIIIVPAGP
GDTPGTVTGPDGYPTKFIVPLGAFSTPGSIPAPDGNPLPVEIAGPGTTPGVETGPDGNNQNI
FTNNPQ

>NES_DN29472_c0_g2_i1_AgSp2

PGPNGYQIPVEIAGPDTTPGTITRADGSIVKIILPTTPSGPNHPTTTAQTIRDASGRTIIIV
PAGPGTTPGTVTDSEGHVIRFIVPFGVYATPGYIPGPNGYQIPVEIAGPDTTPGTITRADGS
IVKIILPTTPSGPNQPTTTAQTIRDSSGRTIIIVPAGPGTTPGTVTDSQG

>NES_DN29553_c0_g1_i1_TuSp_CT

LGISNAQAFASSLAQSVSAVGVGASASASAYASAISSAVGQYLAGQGILTSSNAASLASSFASA
FSASAAASSRNSAFTQSAAAERSQAAAKAFSQAASQAASRASSQAASKSASNATSQSSAESK
SSSTTTSTSTSGSQASSQASSQAASQAASQSASQSASRSYSASSSFARSSAQSLASSRSFSS
AFSSATSVSAVSSIAYRLGLSAASSLGISNAQSFASSLAQSVSAVGVGASTSTYARAISSAV
GKYLSDEGVLTSTNAASLASSFASAISSSSTNGRSSGFAASNSQMALQKVFGPVLSPSGLVS
EESKSRVNSLAQSISSTVSSTGALQPTVFFNLLASNIAPQLISSAPSAEYPELVTQVLLEGV
SALLKVMNGGQNSLGNLSGFPNRGADAINVLS

>NES_DN29610_c1_g5_i1_AgSp2

SSPSSIPGPDGNQIPVQPAGPGTTPGAEYGPNGNVIRIYLPAMPGQSGNINGGTRVIQGKGT
KNIKIVPAGPGESPGTVQSPDGKSLKFIVPISAFSSPSSIPGPDGNQIPVQPAGPGTTPGAE
YGPNGNVIRIYLPAMPGQSGNINGGTRVIQGKGTKNIKIVPAGPGESPGTVQSPDGKSLKFI
VPISAFSSPSSIP

>NES_DN29610_c1_g6_i1_AgSp2

QIQIVPEGPGQSPGTVQSPDGKSLTFIVPISAFSSPTSIPGPDGNQIPVQPAGPGTTPGADF
GPNGNVIPIYLPFVQQQPGNPGVLPEFIPGNGGRQILIMPANPGTTPGTVTDADGFVVKFVV
PYDAFSSPSSIPGPDGNPIPVEPAGPGTTPGTQYGPNGKVIKIYLPYQQPSGGSN

>NES_DN29610_c1_g7_i1_AgSp2

LPYQQPSGGSNPFTVRIPDRGGRQIFIVPAAPGEQPKTVTDSDGYPVKFIVPFGVFGSPGSI
PGPDGQPIPVQPAGTGATPGVQYDPKGRIIRIYLPRFPQLPGIRGPKGRQIVIVPQGAGSTP
GTVTDSDGYPIQFIVPFGAFNTPGSIPGPDGKKIPVQPAGPGTTPGTETDSNGRIIRIYVPS
LQQPSGNQQTTPRSLIGPHGRRILIQPAGPGSTPGTITDSDGYPIQFIVPIGAFSTPGSIPG
PDGKQIPVETAGPGTTPGAETDGDGNVIRIVLPTTPQNPNQGGGTQPMPSRRIKGPHGRQIL
IQPAGPGST

>NES_DN29917_c0_g1_i5_AgSp2

PTTKRYPLPNGHQIIIVPARPGDKPGTVTGPDGYPIKFIVPLGAFSTPGFIPGLNGQQIPVE
IVGPGTTPGVQTGPDGSIIKIYLPTTPRANTPMTAPTTKRYQLPNGHQIIIVPAGPGDTPGT
VTGPGGYPIKFIVPLGAFSTPGSIPGLNGQQIPVEIVGPGTTPGVQTAPDGSIIKIYLPTT

>NES_DN29955_c0_g1_i1_AgSp2

SDGRPVKFVVPVEAYSSPGSIPGPNGKSLPVEPAGPGTTPGAVTDSNGNIVKIYVPKVSGGG
RQPGHRPERPTRIEGPGDQQIIIVPAQGGSTPGIVTGSDGRPVKFVVPISAYSSPGSIPGAN
GKNLPVGPAGPGTTPGVVTDSDGNVIHLYLPSASVSGGQPGSSYEIPGGPFTSLGYLNFSLV
PHNSNSEGPFDYSQLPGAPEIDSGLPGGFYSRNKLPQGVRSRLSNTFSYPDLVESLKPLFPG
KSINLEAIPKDNLKDTDGSIPGFKGSYNSLNLGKLSSYPSGSLYYLTILEYLIAGLPDGVFN
SASFNFPGADGTPSYICDSA

>NES_DN29955_c0_g2_i1_AgSp2

SDGRPVKFVVPVEAYSSPGSIPGPNGKSLPVEPAGPGTTPGAVTDSNGNIVKIYVPKVSGGG
RQPGHRPERPTRIEGPGDQQIIIVPAQGGSTPGIVTGSDGRPVKFVVPVEAYSSPGSIPGPN
GKSLPVEPAGPGTTPGAVTDSNGNIVKIYVPKVSGGGRQPGHRPERPTRIEGPGDQQIIIVP
AQGGSTPGIVTGSDGRPVK

>NES_DN29978_c0_g2_i1_AgSp2_NTR?

YPIQFIVPLGAFSTPGSIPGPDGNPIPVVTAGPGVTPGAETDGDGNIIRIVLPTTPQNPYQG
GGGKPTPLRRIKGPHGRQILIQPAGPGSTPGTVTDSEGYPIEFIVPLGAYSTPGFIPGPNGN
QIPVETAGPGTTPGTETDGDGNIIRIYLPPTSSNPNEQFPPGMQQQIKGPNGHQIIVVPAGP
GTTPGTVTDSDGNPTEFIVPLDAYSQPSSIPGPNGNTIIVVPAGPGQNPGVETDSNGSITKI
ILPRLPSSNGQQYLQGPNGRQILIMPAGPGTTPGTVTDSNGDPTEFIVPLDAYSRPSSIPGP
NGNTIIVVPAGPGMNPGVETDSNGYIIKIILPRMTQQGGQTLISKHKPGFSTQIQILPVAPG
TTPGTVTDSDGNPVTFYVPLSAYSSDTSIMGPDGRPIPVGPVGLGSTPGAVTNAEGNIIRIY
LPNVPGSANMGGFGGSGGSGSIQRISIVPAGPGTTPGTETDPDGTVRYIVPLSAFSSPGTVM
GPNGNPIPVGPAGPGTTPGAVFGSNGQVSRIYLPPVQYSGSGGGSQQMSGNSQGNGYQSGGS
GGSGSIQRISIVPAGPGQNAGIETEQDGTVKYIVPLSAFSSPGSIMGLDGNPIPVGPAGPGV
TPGAEFGSNGQVSRIYLPPIQYSPASGGGAGQQMYGGNSQGSGYQYSGGQNGGGQQMVGGNS
GNGYQSYGGQIEGSPQMFGGNSGNGYQTYGGQNGGVVQYEY

>NES_DN29978_c0_g2_i2_AgSp2_NT?

IKGPHGRQILIQPAGPGSTPGTVTDSEGYPIEFIVPLGAYSTPGFIPGPDGSQIPVETAGPG
TTPGAETDGDGNILRIVLPTAGNPKQGGGQPTPPRRIKGPHGRQILIQPAGPGSTPGTVTDS
EGYPIQFIVPLGAFSTPGSIP

>NES_DN29978_c0_g5_i1_AgSp2

PAGPGQNPGIETEEDGTVKYIVPLDAFSSPGSIMGPDGNPIPVGPAGPGTTPGAEFGSNGQV
SRIYLPPVQYSGSGSSSQEMNGNSQNNGFQSGSSGGTFQRISIVPAGPGQNPGIETEEDGTV
KYIVPLDAFSSPGTIMGPDGNSIPVGPAGPGTTPGAEFGSNGHVSRIYLPPVQYSPGSGGGY
SQLMNGGSQQMLDGNSGNGYQSFGEQNGGSRSDGGQYGNQMYGPDLGEGVMGTQMMPDDLVG
AAAPNGDPVPPGASSQRPSHPYNHVPQVAPVPPGGPSLSSDIFASGGSPEGIQNLLQNLMGR
SNLLNPQSNCGNDAYGTFRKAKSALSELSSSTGEQRISVLVNAIISGINASANTVDFNGFFT
KLYSLFYEIRTDSSTPDVEFLSILMEALVAALEAINAAQISGFVPVDFMSDIPVYANFLSDI
L

>NES_DN29989_c0_g1_i1_FlaG

PQGPGFGGPQGPGFGGPQGPGFGGPQGPGFGGPQGPPGYGPSGSDSSNGKFAGTAGGVPKTG
PVTVSDSVSVNVGGPPSALSGNRGDSGTRPYNSGQSSSSRLSSLVNGILNSLQPYGFNYQNF
GNVLSQYAIATGTCNSNDLGLLIDALQASLHCLSSFGASPFPSVPSPAAMNAFSQYIQRMFQ

>NES_DN30928_c0_g1_i1_AgSp2

AGPGTTPGAVTQSDGSVIRIYLPKMQIHGGNIGGHGKVIQGKAKGSKSIQIVPAGPGESPGT
VNSPDGKSLKFIVPISAFSSPGSIPGPNGNQIPVGPAGPGTTPGAEYRKN

>NES_DN31477_c0_g1_i1_AgSp2

DSDGHPVKFVVPVGAYSHPGSIPGPNGKPLSVEPAEPGTTPGVVTDPNGSIVKIYVPKVAGG
GGGRQPPHVPKNPVDIKGPGDQQIIIVPAQGGSTPGTVTDSDGHPVKFVVPVEAYSSPGSIP
GPNGKPLPVEPAGPGTTPGAETGSDGNIVKLFVPRVAGGGRQPHPRPQQPTRIVGPGDSQIM
IVPAGPGDTPGTVNGPDGKPVKFVVPISAYSSPGSIPGPNGKSIPVGPAGPGTTPGAVTGSD
GSIVQIYLPPASISHGQSGGTYEVPGGPFTSLGYLNFSLIPDNSNSQGPFDYSKLPGVLEID
SGLPGGFYPRNKLPQGVRGLLSTSFSFPNLVSALSPLFPGGSINLDAIPKSKLQGSDGSIPG
FDGSYNSLKLRRLSSFPSGGFYYLTIIEYLINVLPDGVFNGASFSFSGANGTPTYICVG

>NES_DN32135_c0_g1_i1_AgSp2_NT?

DGKSIKFIVPVSAYNTPGYIPGPDGGQIPVGPAGPGTTPGAVTKGDGSVIQIYLPFIGHTTT
RKPDSNPHEIQISVAGPGTTPGTVTDSDGYPVQFLVPLDTFYRPGFVPGPNGNPIPVGPAQP
GTTPGLETESD

Appendix A.3.1

Sequences used in phylogenetic analysis of C-terminal domains in Chapter 3 (Figures 3.13 and 3.14), including description and accession numbers.

Sequence Name	Description	Accession number	Species
Aa MaSp1_1a CT	Full-length isoform 1	-	*Argyroneta aquatica*
Aa MaSp1_1b CT	MaSp1 full-length isoform 2	-	*Argyroneta aquatica*
Aa MaSp1_2a CT	MaSp1 full-length version 2	-	*Argyroneta aquatica*
Aa MaSp aq CT	MaSpAq full-length	-	*Argyroneta aquatica*
Aa MaSp	major ampullate spidroin variant 1, partial	AVH80563	*Argyroneta aquatica*
Ad MaSp2	major ampullate spidroin 2 variant 1, partial	AWK58752	*Araneus diadematus*
Ag a MaSp	major ampullate spidroin mRNA, partial cds	AY566305.1	*Agelenopsis aperta*
Cm MaSp1	dragline silk spidroin 1, partial	AAV48951	*Cyrtophora moluccensis*
Cm MaSp1_	major ampullate spidroin 1 short	AGQ04592	*Cyrtophora moluccensis*
Ds MaSp2	MaSp2b, partial	ABD61594	*Deinopis spinosa*
Ea MaSp	major ampullate spidroin-like, partial	CAM32250	*Euprosthenops australis*
Ea MaSp1	major ampullate spidroin 1, partial	CAM32262	*Euprosthenops australis*
Lh MaSp1	major ampullate spidroin 1 variant 2, partial	ABD66602	*Latrodectus hesperus*
Lh MaSp1_	major ampullate spidroin 1, partial	ABD66602	*Latrodectus hesperus*
Lh MaSp1_2	major ampullate spidroin 1 locus 2, partial	ABD66602	*Latrodectus hesperus*
Lh MiSp	minor ampullate spidroin variant 1	ARA91152	*Latrodectus hesperus*
Lh MiSp_	minor ampullate spidroin, partial	ACB29694	*Latrodectus hesperus*
Lh MiSp_2	minor ampullate spidroin variant 2, partial	ARA91182	*Latrodectus hesperus*
Lt MiSp_1	minor ampullate spidroin variant 1, partial	2M0M_A	*Latrodectus tredecimguttatus*
Na MaSp1	major ampullate fibroin 1, partial	ABC72644	*Nephila antipodiana*
Np MaSp1	dragline silk spidroin 1, partial	AAV48921	*Nephila pilipes*
Np MaSp1_	dragline silk spidroin 1, partial	AAV48948	*Nephila pilipes*
Og MaSp1	MaSp1 mRNA for major ampullate spidroin 1, partial cds	BCI65229	*Octonoba grandiconcava*
Og MaSp1_1	MaSp1 mRNA for major ampullate spidroin 1, partial cds	BCI65273	*Octonoba grandiprojecta*
Og MaSp2	MaSp2 mRNA for major ampullate spidroin 2, partial cds	BCI65252	*Octonoba grandiprojecta*

Table continued on next page

Sequence Name	Description	Accession number	Species
Og MiSp	MiSp mRNA for minor ampullate spidroin, partial cds	BCI65253	*Octonoba grandiprojecta*
Oo MaSp1	MaSp1 mRNA for major ampullate spidroin 1, partial cds	BCI65238	*Octonoba okinawensis*
Oo MaSp2	MaSp2 mRNA for major ampullate spidroin 2, partial cds	BCI65240	*Octonoba okinawensis*
Oo MiSp	MiSp mRNA for minor ampullate spidroin, complete cds	BCI65241	*Octonoba okinawensis*
Os MaSp1	MaSp1 mRNA for major ampullate spidroin 1, partial cds	BCI65229	*Octonoba sybotides*
Os MaSp2	MaSp2 mRNA for major ampullate spidroin 2, complete cds	BCI65275	*Octonoba sybotides*
Os MiSp	MiSp mRNA for minor ampullate spidroin, complete cds	BCI65265	*Octonoba sybotides*
Ov MaSp1	MaSp1 mRNA for major ampullate spidroin 1, partial cds	BCI65229	*Octonoba varians*
Ov MaSp2	MaSp2 mRNA for major ampullate spidroin 2, partial cds	BCI65275	*Octonoba varians*
Ov MiSp	MiSp mRNA for minor ampullate spidroin, partial cds	BCI65253	*Octonoba varians*
Oy MaSp1	MaSp1 mRNA for major ampullate spidroin 1, partial cds	BCI65229	*Octonoba yesoensis*
Oy MiSp	MiSp mRNA for minor ampullate spidroin, partial cds	BCI65288	*Octonoba yesoensis*
Ps MaSp1	dragline silk spidroin 1, partial	AAV48948	*Psechrus sinensis*
Pt MaSp1	dragline silk spidroin 1, partial	LAA09541	*Parasteatoda tepidariorum*
Pt MiSp	minor ampullate spidroin c-terminal variant, partial	ARA91184	*Parasteatoda tepidariorum*
Pt MiSp_	minor ampullate spidroin, partial	ARA91184.1	*Parasteatoda tepidariorum*
Pt MiSp_1	minor ampullate spidroin 1 variant 1, partial	AWK58669.1	*Parasteatoda tepidariorum*
Sg MiSp	minor ampullate spidroin, partial	ARA91183	*Steatoda grossa*
Sg MiSp_1	minor ampullate spidroin 1 variant 1, partial	AWK58670	*Steatoda grossa*
Tc MaSp	major ampullate spidroin protein MaSp-a	PRD23950	*Trichonephila clavipes*
Tc MiSp	minor ampullate silk protein, partial	AAC14590	*Trichonephila clavipes*
Tc MiSp1	minor ampullate silk protein MiSp1, partial	AAC14589	*Trichonephila clavipes*
Tc MiSp_2	minor ampullate spidroin 1 variant 2, partial	AWK58668	*Trichonephila clavipes*
Tc MiSp_a	minor ampullate spidroin protein MiSp-a	PRD23654	*Trichonephila clavipes*
Tc MiSp_b	minor ampullate spidroin protein MiSp-b	PRD30268	*Trichonephila clavipes*
Tc MiSp_c	minor ampullate spidroin protein MiSp-c	PRD24510	*Trichonephila clavipes*
Tk MaSp1	MaSp1 mRNA for major ampullate spidroin 1, partial cds	AF350285_1	*Tetragnatha kauaiensis*
B. c fibroin 1	fibroin 1, partial (Outgroup)	ABW80565	*Bothriocyrtum californicum*

Appendix A.3.2

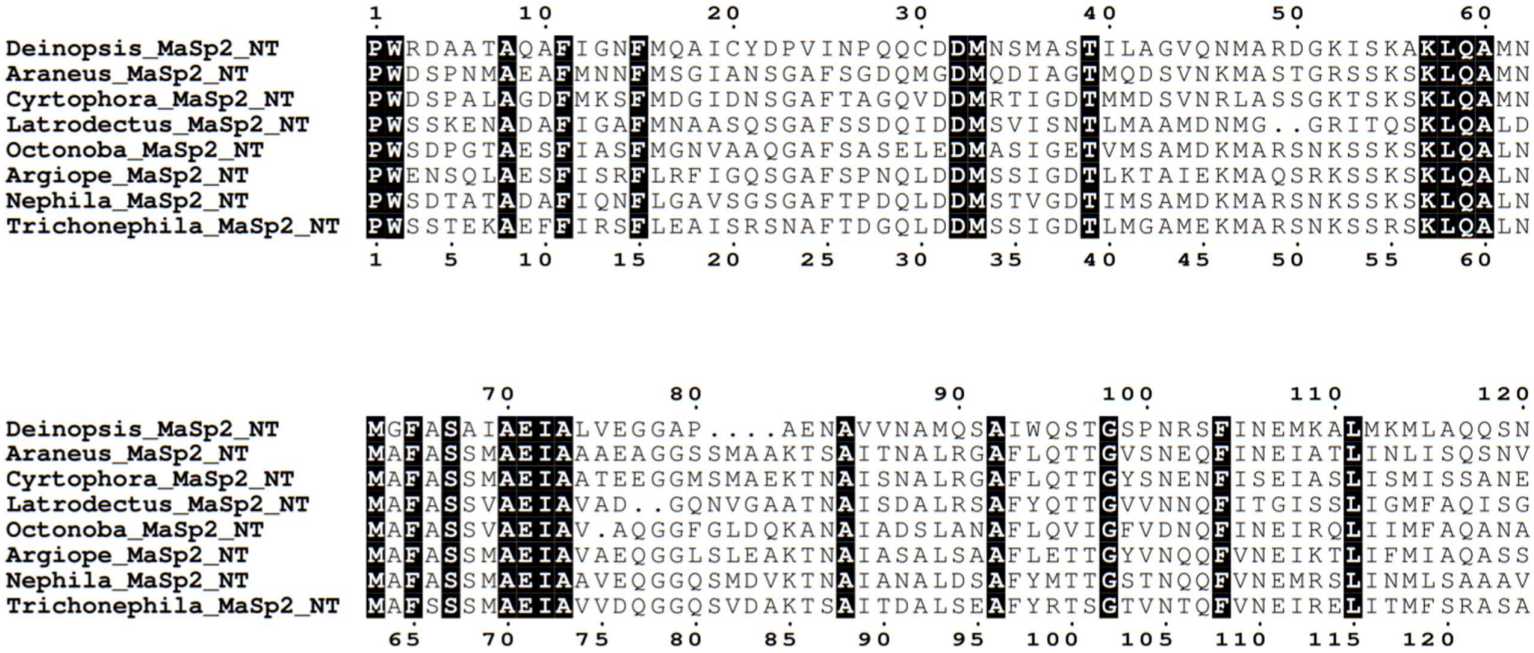


Figure A.3.2.1 Alignment of* Cyrtophora citricola *full length MaSp2 N-terminal domain with a selection of representative MaSp2 terminal domain sequences.
Sequences used can be seen in Table A.3.2.A.

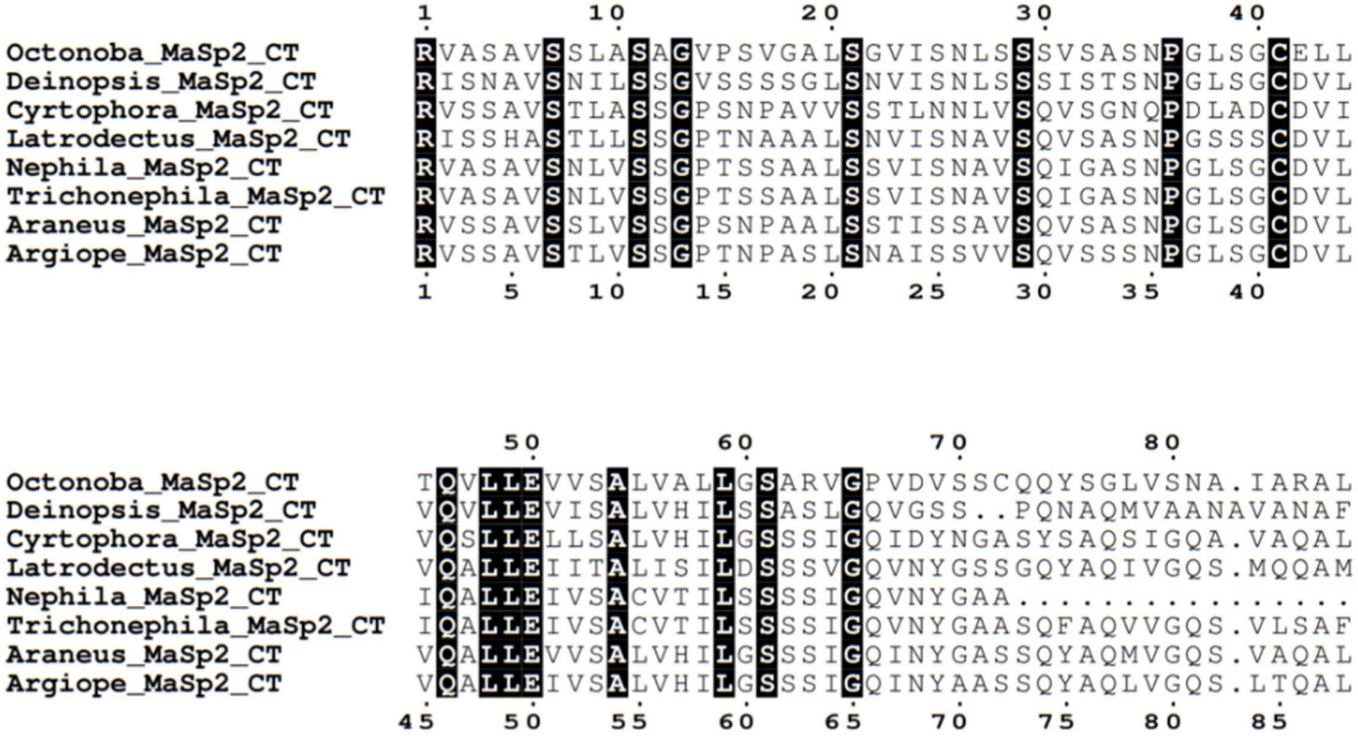


Figure A.3.2.2 Alignment of **Cyrtophora citricola** ***full length MaSp2 C-terminal domain with a selection of representative MaSp2 terminal domain sequences.***
Sequences used can be seen in Table A.3.2.A.

Table A.3.2.A Sequences used for alignments in Figures A.3.2.1 and A.3.2.2.

Sequence name	Species	Accession Number
Deinopsis_MaSp2_CT	*Deinopis spinosa*	ABD61594
Deinopsis_MaSp2_NT	*Deinopis spinosa*	ADM14319
Octonoba_MaSp2_NT	*Octonoba sybotides*	BCI65264
Octonoba_MaSp2_CT	*Octonoba grandiconcava*	BCI65231
Araneus_MaSp2_CT	*Araneus ventricosus*	GBN00527
Araneus_MaSp2_NT	*Araneus ventricosus*	GBN00527
Argiope_MaSp2_NT	*Argiope trifasciata*	AAZ15371
Argiope_MaSp2_CT	*Argiope trifasciata*	AF350267_1
Nephila_MaSp2_NT	*Trichonephila inaurata madagascariensis*	AAZ15322
Nephila_MaSp2_CT	*Trichonephila inaurata madagascariensis*	AF350278_1
Trichonephila_MaSp2_NT	*Trichonephila clavipes*	AWK58653
Trichonephila_MaSp2_CT	*Trichonephila clavipes*	AWK58758
Latrodectus_MaSp2_CT	*Latrodectus hesperus*	ABD66603
Latrodectus_MaSp2_NT	*Latrodectus hesperus*	ABD24295
Cyrtophora_MaSp2_NT	*Cyrtophora citricola*	From this study
Cyrtophora_MaSp2_CT	*Cyrtophora citricola*	From this study

Appendix A.4.1

pJE401 AquaRepCT Information

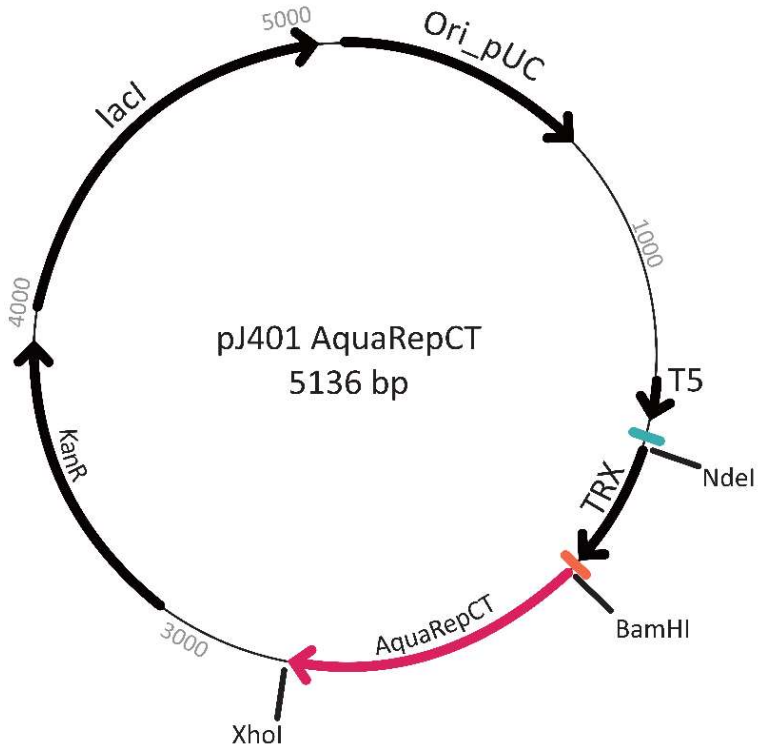


Figure A.4.1.1 Vector map of pJE401-AquaRepCT. The insert (AquaRepCT) is shown in pink. This is preceded by a thrombin cleavage sequence (orange bar) thioredoxin tag (TRX) and a HIS tag (blue bar).

ATGGCACATCACCACCACCATCACATGAGCGACAAAATCATCCACCTGACCGACGACAGCT
TCGATACCGATGTGTTAAAAGCCGACGGTGCGATTCTGGTCGACTTCTGGGCGGAGTGGT
GTGGTCCATGCAAGATGATTGCACCGATTTTGGACGAAATCGCGGACGAATATCAGGGTA
AGCTGACCGTGGCGAAGCTGAACATTGACCAGAATCCGGGTACGGCCCCTAAATACGGTA
TTCGTGGTATCCCGACCCTGCTGTTGTTTAAAAACGGTGAAGTGGCGGCGACCAAGGTCG
GTGCCCTGTCGAAGGGTCAGCTGAAAGAATTTCTGGATGCGAACTTGGCGGGTACGGACG
ATGATGATAAATTGGTCCCGCGCGGATCCGGTGGCAACGGTGGCATCGGTGGTTATGGTG
GCTATGGTGGTCTGGGCTATGGCGGCTACGGTTATGGTGCGGGTAGCTCCGCTGCCGCAG
CCGCCGCAGCAGCAGCGGCTGCTGCGGGTGGCTACGGTGGTTACGGCTACGGCGGCTAT
GGTTACGGTGGCGGCGGTTCTGCGGCGGCAGCTGCAGCCGCAGCTGCCGGTGCCGGTGC
GGGTGCGGGCGAGGGTGCCGGTGGCTATCGCCGTTACGGCGGTTACGGTGATGATGGCG
CTGCTGCAGCGGCAGCGGCGGCTGCGGCCGCGGCAGCCGCGGCGGGCGGTTATGGCGGT
TATTACTATGAAGAGGGTGGCGCTGGTGCGGGCGCGGCGGGTGCGGCCGCCGCTAGCGC
GGCAGGCGCAGCCGCTACCGGTGGCGCCCGTCGCGCGGGCGGCTACGGCTATGGCGCGG
GTGGTGCGGCGGGCGCCGCAGCGGCCGCAGGCGCGGCCGGCGGTGCTGGCACCCGTAAT
ACTGCGTACAATAGCGTCAGCCTGCTGAGCTCTAGCGCGTCAAACGCACGCGTTAGCACCG
CGGTTAGCGGCCTGGTTGGCAATGGCGTTGATGCGCTGTCTAACATTGTTTCCTCGCTGTC
CGCGTCTGTGGCAGCGACGCATCCGGGTCTGAGCGGTTGTGAAGTCCTGGTCCAGGTTCT

GATGGAAATCATCAGCGCACTCGTGCACGTGCTGAGCAGCGCGAGCATTGGCCAAGTGAA
TGTTAACGACGCGGCATACACGGCAGCACTGGTGGGTGACGTCATTAGCCAAAATTGCTA
A

Figure A.4.1.2 DNA sequence of the AquaRepCT insert (from start of histidine tag to stop codon).

HHHHHHMSDKIIHLTDDSFDTDVLKADGAILVDFWAEWCGPCKMIAPILDEIADEYQGKLTVA
KLNIDQNPGTAPKYGIRGIPTLLLFKNGEVAATKVGALSKGQLKEFLDANLAGTDDDDKLVPRG
SGGNGGIGGYGGYGGLGYGGYGYGAGSSAAAAAAAAAAAAGGYGGYGYGGYGYGGGGSA
AAAAAAAAGAGAGAGEGAGGYRRYGGYGDDGAAAAAAAAAAAAAAAGGYGGYYYEEGGA
GAGAAGAAAASAAGAAATGGARRAGGYGYGAGGAAGAAAAAGAAGGAGTRNTAYNSVSL
LSSSASNARVSTAVSGLVGNGVDALSNIVSSLSASVAATHPGLSGCEVLVQVLMEIISALVHVLS
SASIGQVNVNDAAYTAALVGDVISQNC*

Figure A.4.1.3 Amino acid sequence of the AquaRepCT insert (from start of HIS tag to stop codon). The histidine tag is in bold, the thioredoxin tag is underlined, thrombin cleavage site double underlined and the AquaRepCT sequence in purple.

Appendix A.4.2

pET32b AquaRepCT Information

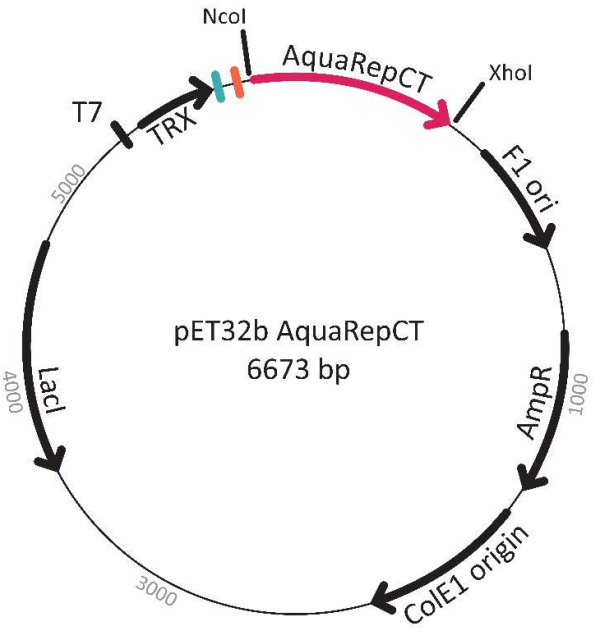


Figure A.4.2.1 Vector map of pET32b-AquaRepCT. The insert (AquaRepCT) is shown in pink. This is preceded by a thioredoxin tag (TRX), HIS tag (blue bar) and thrombin cleavage sequence (orange bar).

ATGAGCGATAAAATTATTCACCTGACTGACGACAGTTTTGACACGGATGTACTCAAAGCGG
ACGGGGCGATCCTCGTCGATTTCTGGGCAGAGTGGTGCGGTCCGTGCAAAATGATCGCCC
CGATTCTGGATGAAATCGCTGACGAATATCAGGGCAAACTGACCGTTGCAAAACTGAACAT
CGATCAAAACCCTGGCACTGCGCCGAAATATGGCATCCGTGGTATCCCGACTCTGCTGCTG
TTCAAAAACGGTGAAGTGGCGGCAACCAAAGTGGGTGCACTGTCTAAAGGTCAGTTGAAA
GAGTTCCTCGACGCTAACCTGGCCGGTTCTGGTTCTGGCCATATGCACCATCATCATCATCA
TTCTTCTGGTCTGGTGCCACGCGGTTCTGGTATGAAAGAAACCGCTGCTGCTAAATTCGAA
CGCCAGCACATGGACAGCCCAGATCTGGGTACCGACGACGACGACAAGGCCATGGCGATA
TCGGATCCAGGTGGCAACGGTGGCATCGGTGGTTATGGTGGCTATGGTGGTCTGGGCTAT
GGCGGCTACGGTTATGGTGCGGGTAGCTCCGCTGCCGCAGCCGCCGCAGCAGCAGCGGCT
GCTGCGGGTGGCTACGGTGGTTACGGCTACGGCGGCTATGGTTACGGTGGCGGCGGTTCT
GCGGCGGCAGCTGCAGCCGCAGCTGCCGGTGCCGGTGCGGGTGCGGGCGAGGGTGCCG
GTGGCTATCGCCGTTACGGCGGTTACGGTGATGATGGCGCTGCTGCAGCGGCAGCGGCG
GCTGCGGCCGCGGCAGCCGCGGCGGGCGGTTATGGCGGTTATTACTATGAAGAGGGTGG
CGCTGGTGCGGGCGCGGCGGGTGCGGCCGCCGCTAGCGCGGCAGGCGCAGCCGCTACCG
GTGGCGCCCGTCGCGCGGGCGGCTACGGCTATGGCGCGGGTGGTGCGGCGGGCGCCGCA
GCGGCCGCAGGCGCGGCCGGCGGTGCTGGCACCCGTAATACTGCGTACAATAGCGTCAGC
CTGCTGAGCTCTAGCGCGTCAAACGCACGCGTTAGCACCGCGGTTAGCGGCCTGGTTGGC
AATGGCGTTGATGCGCTGTCTAACATTGTTTCCTCGCTGTCCGCGTCTGTGGCAGCGACGC

ATCCGGGTCTGAGCGGTTGTGAAGTCCTGGTCCAGGTTCTGATGGAAATCATCAGCGCACT
CGTGCACGTGCTGAGCAGCGCGAGCATTGGCCAAGTGAATGTTAACGACGCGGCATACAC
GGCAGCACTGGTGGGTGACGTCATTAGCCAAAATTGCTAA

Figure A.4.2.2 DNA sequence of the AquaRepCT insert (from start of thioredoxin tag to stop codon).

MSDKIIHLTDDSFDTDVLKADGAILVDFWAEWCGPCKMIAPILDEIADEYQGKLTVAKLNIDQN
PGTAPKYGIRGIPTLLLFKNGEVAATKVGALSKGQLKEFLDANLAGSGSGHM**HHHHHH**SSGLV
PRGSGMKETAAAKFERQHMDSPDLGTDDDDKAMAISDPGGNGGIGGYGGYGGLGYGGYGY
GAGSSAAAAAAAAAAAAGGYGGYGYGGYGYGGGGSAAAAAAAAAGAGAGAGEGAGGYRR
YGGYGDDGAAAAAAAAAAAAAAGGYGGYYYEEGGAGAGAAGAAAASAAGAAATGGARR
AGGYGYGAGGAAGAAAAAGAAGGAGTRNTAYNSVSLLSSSASNARVSTAVSGLVGNGVDAL
SNIVSSLSASVAATHPGLSGCEVLVQVLMEIISALVHVLSSASIGQVNVNDAAYTAALVGDVISQ
NC*

Figure A.4.2.3 Amino acid sequence of the AquaRepCT insert (from start of thioredoxin tag to stop codon). The thioredoxin tag is underlined, the histidine tag in bold, thrombin cleavage site double underlined and the AquaRepCT sequence in purple.

Appendix A.4.3

pET32b AquaCT Information

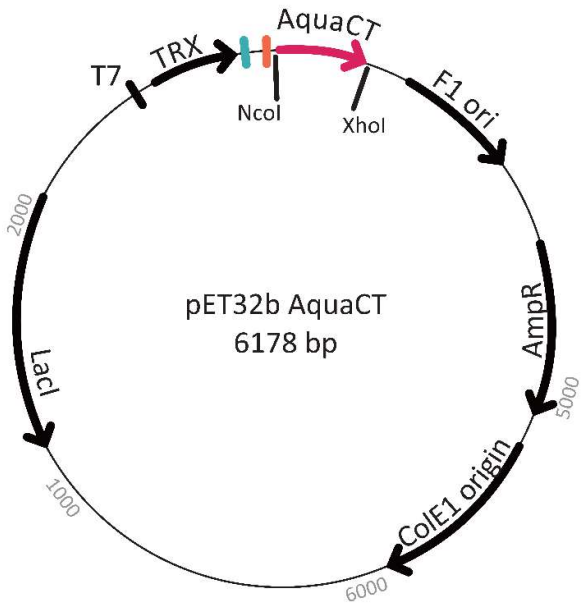


Figure A.4.3.1 Vector map of pET32b-AquaCT. The insert (AquaRepCT) is shown in pink. This is preceded by a thrombin cleavage sequence (orange bar), HIS tag (blue bar) and a thioredoxin tag (TRX).

ATGAGCGATAAAATTATTCACCTGACTGACGACAGTTTTGACACGGATGTACTCAAAGCGG
ACGGGGCGATCCTCGTCGATTTCTGGGCAGAGTGGTGCGGTCCGTGCAAAATGATCGCCC
CGATTCTGGATGAAATCGCTGACGAATATCAGGGCAAACTGACCGTTGCAAAACTGAACAT
CGATCAAAACCCTGGCACTGCGCCGAAATATGGCATCCGTGGTATCCCGACTCTGCTGCTG
TTCAAAAACGGTGAAGTGGCGGCAACCAAAGTGGGTGCACTGTCTAAAGGTCAGTTGAAA
GAGTTCCTCGACGCTAACCTGGCCGGTTCTGGTTCTGGCCATATGCACCATCATCATCATCA
TTCTTCTGGTCTGGTGCCACGCGGTTCTGGTATGAAAGAAACCGCTGCTGCTAAATTCGAA
CGCCAGCACATGGACAGCCCAGATCTGGGTACCGACGACGACGACAAGGCCATGGCGATA
TCGGATCCAACCCGTAATACTGCGTACAATAGCGTCAGCCTGCTGAGCTCTAGCGCGTCAA
ACGCACGCGTTAGCACCGCGGTTAGCGGCCTGGTTGGCAATGGCGTTGATGCGCTGTCTA
ACATTGTTTCCTCGCTGTCCGCGTCTGTGGCAGCGACGCATCCGGGTCTGAGCGGTTGTGA
AGTCCTGGTCCAGGTTCTGATGGAAATCATCAGCGCACTCGTGCACGTGCTGAGCAGCGC
GAGCATTGGCCAAGTGAATGTTAACGACGCGGCATACACGGCAGCACTGGTGGGTGACGT
CATTAGCCAAAATTGCTAA

Figure A.4.3.2 DNA sequence of the AquaCT insert (from start of thioredoxin tag to stop codon).

MSDKIIHLTDDSFDTDVLKADGAILVDFWAEWCGPCKMIAPILDEIADEYQGKLTVAKLNIDQN
PGTAPKYGIRGIPTLLLFKNGEVAATKVGALSKGQLKEFLDANLAGSGSGHM**HHHHHH**SSGLV
PRGSGMKETAAAKFERQHMDSPDLGTDDDDKAMAISDPTRNTAYNSVSLLSSSASNARVSTA
VSGLVGNGVDALSNIVSSLSASVAATHPGLSGCEVLVQVLMEIISALVHVLSSASIGQVNVNDA
AYTAALVGDVISQNC*

Figure A.4.2.3 Amino acid sequence of the AquaCT insert (from start of thioredoxin tag to stop codon). The thioredoxin tag is underlined, the histidine tag in bold, thrombin cleavage site double underlined and the AquaCT sequence in purple.

Appendix A.4.4

pET32b TEV-AquaCT

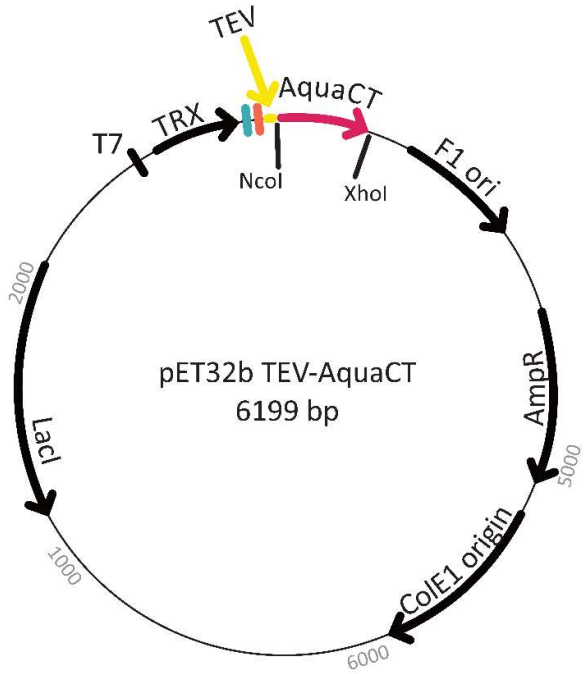


Figure A.4.4.1 Vector map of pET32b-TEV-AquaCT. The insert (AquaRepCT) is shown in pink. This is preceded by a thrombin cleavage sequence (orange bar), HIS tag (blue bar) and a thioredoxin tag (TRX). The inserted TEV site sits directly before AquaCT and is shown in yellow.

ATGAGCGATAAAATTATTCACCTGACTGACGACAGTTTTGACACGGATGTACTCAAAGCGG
ACGGGGCGATCCTCGTCGATTTCTGGGCAGAGTGGTGCGGTCCGTGCAAAATGATCGCCC
CGATTCTGGATGAAATCGCTGACGAATATCAGGGCAAACTGACCGTTGCAAAACTGAACAT
CGATCAAAACCCTGGCACTGCGCCGAAATATGGCATCCGTGGTATCCCGACTCTGCTGCTG
TTCAAAAACGGTGAAGTGGCGGCAACCAAAGTGGGTGCACTGTCTAAAGGTCAGTTGAAA
GAGTTCCTCGACGCTAACCTGGCCGGTTCTGGTTCTGGCCATATGCACCATCATCATCATCA
TTCTTCTGGTCTGGTGCCACGCGGTTCTGGTATGAAAGAAACCGCTGCTGCTAAATTCGAA
CGCCAGCACATGGACAGCCCAGATCTGGGTACCGACGACGACGACAAGGCCATGGCGATA
TCGGATCCAGAAAACCTGTATTTTCAGGGCACCCGTAATACTGCGTACAATAGCGTCAGCC
TGCTGAGCTCTAGCGCGTCAAACGCACGCGTTAGCACCGCGGTTAGCGGCCTGGTTGGCA
ATGGCGTTGATGCGCTGTCTAACATTGTTTCCTCGCTGTCCGCGTCTGTGGCAGCGACGCA
TCCGGGTCTGAGCGGTTGTGAAGTCCTGGTCCAGGTTCTGATGGAAATCATCAGCGCACTC
GTGCACGTGCTGAGCAGCGCGAGCATTGGCCAAGTGAATGTTAACGACGCGGCATACACG
GCAGCACTGGTGGGTGACGTCATTAGCCAAAATTGCTAA

Figure A.4.3.2 DNA sequence of the TEV-AquaCT insert (from start of thioredoxin tag to stop codon).

MSDKIIHLTDDSFDTDVLKADGAILVDFWAEWCGPCKMIAPILDEIADEYQGKLTVAKLNIDQN
PGTAPKYGIRGIPTLLLFKNGEVAATKVGALSKGQLKEFLDANLAGSGSGHM**HHHHHH**SSGLV
PRGSGMKETAAAKFERQHMDSPDLGTDDDDKAMAISDPENLYFQGTRNTAYNSVSLLSSSAS
NARVSTAVSGLVGNGVDALSNIVSSLSASVAATHPGLSGCEVLVQVLMEIISALVHVLSSASIGQ
VNVNDAAYTAALVGDVISQNC*

Figure A.4.4.3 Amino acid sequence of the TEV-AquaCT insert (from start of thioredoxin tag to stop codon). The thioredoxin tag is underlined, the histidine tag in bold, thrombin cleavage site double underlined, the TEV recognition site highlighted in yellow and the AquaCT sequence in purple.

Appendix A.4.5

N-terminal Sequence information used for construct design in Chapter 4. Two N-termini sequences were chosen from the *Argyroneta* transcriptome and ordered from Aruru Molecular. These were TRINITY_DN5543_ Masp1_NT and TRINITY_DN12414_AcSp_NT. The full sequences that were ordered are below and were supplied in a pET32a vector.

pET32b-NTAquaMaSp:
GGTACCGGCAGCAGCGGCGAAAACCTGTATTTTCAGGGCATGGGGTGGTCTACACAATTG
GCGTTGCCAGTAGTCCTGTTGGTGTCCTCTCTTCTGGGTAGCACGCTGGCCGCATCATCAAT
GTTCACAGACGCATCCAAGACGGAGGATTTTATGCGTTCAATGGTTTCGTCGTTGTCGCGT
AACCCTTGCTTTACAGCAGATCAAATCTCGGACATGGAAAGTCTGAGTGACACGATGACGA
TGAGCGTAGACCGTATGGTGCAGACCAATAAAGTGTCGGAGCACCGTATGAAAGCTATGT
CAACTGCGATGGCGTCTGGTATCGCTGAGATTGCTGCTTCTAGTGGGTCATGTGGGTTACA
ACAGAAGTCCAGCGCTATTGCTAATGCCATGTCCGACGCATTCCTTCAAACCAACGGGCGT
GTGGATTCAATCTTTGTTCAAGAGATGGTTTCCATGATTAATATGTTCGCCCAAGCGAGTA
GTAACGAAATCTCATATTCGACTGGGAGCACCCGCGGGTCAGCCATGGCGTGACTCGAG

pET32b-NTAquaMaSp - TEV-DN5543_MaSp ORF translation:
GTGSSGENLYFQGMGWSTQLALPVVLLVSSLLGSTLAASSMFTDASKTEDFMRSMVSSLSRNP
CFTADQISDMESLSDTMTMSVDRMVQTNKVSEHRMKAMSTAMASGIAEIAASSGSCGLQQ
KSSAIANAMSDAFLQTNGRVDSIFVQEMVSMINMFAQASSNEISYSTGSTRGSAMA*LE

pET32b-NTAquaAcSp:
GGTACCGGCAGCAGCGGCGAAAACCTGTATTTTCAGGGCATGTACTGGCTTACTGGATTTA
CGGTGTTGGTCTTACTTTTACAACAATATACATCGGTAGAGGCGGACAGCCGCTATGAGAC
ACTGTCGCCGTGGACAAATCCCACACGCGCAGAGAAATTCATTGACTGCCTTACCGATGGC
ATTGCAAAATCGCCCATCTTTCCGCAACAAGAGAAAGATGATCTGCAATACATCGTTGAAA
CGATGACCAGTGCTATCACCAGTACCTCATCGCAGCGCGGTAATCGTGCTCAACTGCAAGC
GCTTAATATGGCTTTCGCCTCATCCCTTGCAGAATTGGTGATCGCGGAAGACATTGGCAAT
ACTCAAGCCATCAGTGTGAAGACTCAAGCTCTTACGCAATCGTTAAGCAGTTGTTTTCGCTC
TGTAATTGGGACCGTCAACAAACGCTTTATTAACGAGATTAAAGACTTAATCAAAGTCTTC
GCACAGGAAGCTAATAGCGAGTTGAACGAGGTAGATGTCGAAGACTTGGGAAGTGCCAT
GGCGTGACTCGAG

pET32b-NTAquaAcSp - TEV DN12414_AcSp ORF translation:
GTGSSGENLYFQGMYWLTGFTVLVLLLQQYTSVEADSRYETLSPWTNPTRAEKFIDCLTDGIAK
SPIFPQQEKDDLQYIVETMTSAITSTSSQRGNRAQLQALNMAFASSLAELVIAEDIGNTQAISVK
TQALTQSLSSCFRSVIGTVNKRFINEIKDLIKVFAQEANSELNEVDVEDLGSAMA*LE

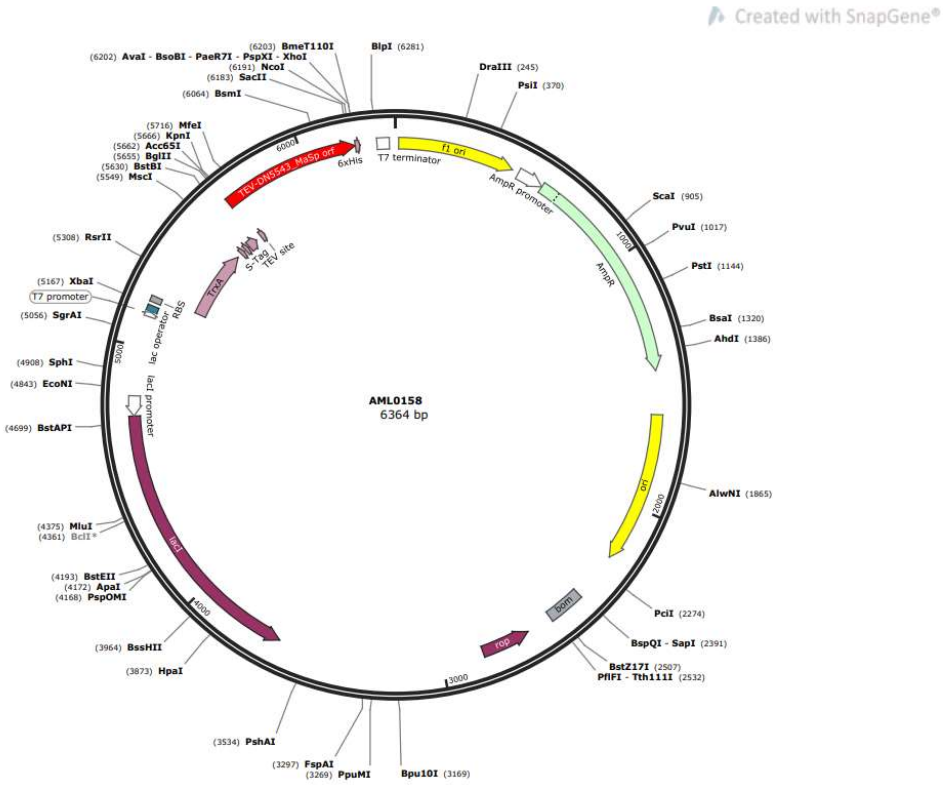


Figure A.4.5.1 Vector map of the ordered pET32b-TEV-NTAquaMaSp construct.

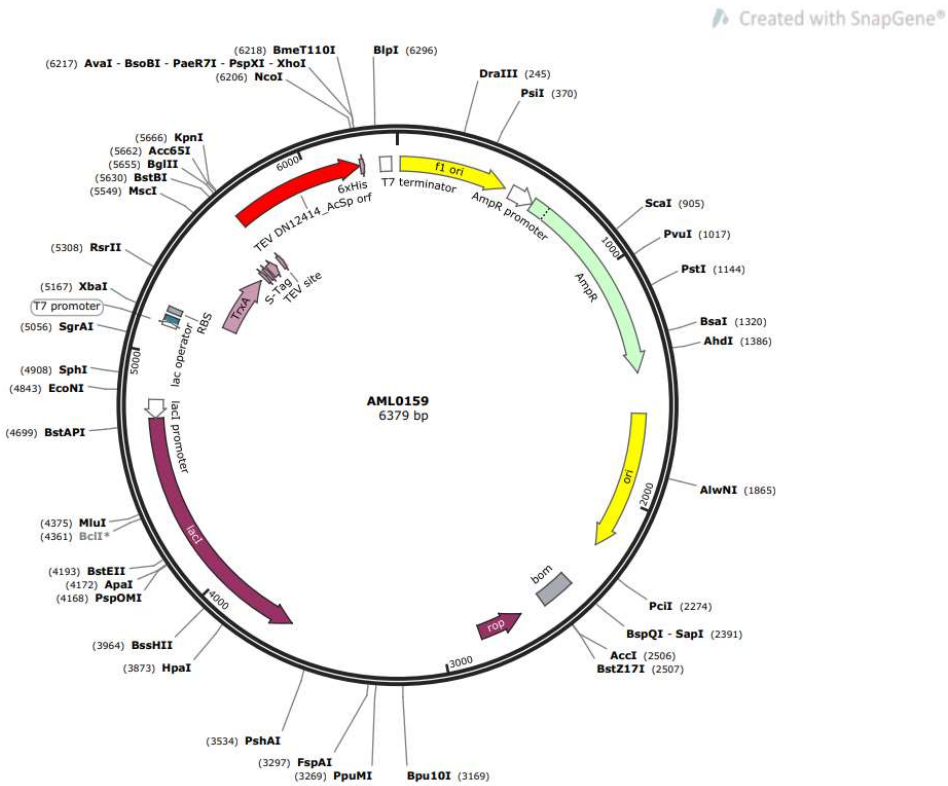


Figure A.4.5.2 Vector map of the ordered pET32b-TEV-NTAquaAcSp construct.

The sequences of the full-length NT-AquaRep-CT inserts when cloned into pET32b are below.

pET32b-NTMaSp-AquaRep-CT:
GGTACCGGCAGCAGCGGCGAAAACCTGTATTTTCAGGGCATGGGGTGGTCTACACAATTG
GCGTTGCCAGTAGTCCTGTTGGTGTCCTCTCTTCTGGGTAGCACGCTGGCCGCATCATCAAT
GTTCACAGACGCATCCAAGACGGAGGATTTTATGCGTTCAATGGTTTCGTCGTTGTCGCGT
AACCCTTGCTTTACAGCAGATCAAATCTCGGACATGGAAAGTCTGAGTGACACGATGACGA
TGAGCGTAGACCGTATGGTGCAGACCAATAAAGTGTCGGAGCACCGTATGAAAGCTATGT
CAACTGCGATGGCGTCTGGTATCGCTGAGATTGCTGCTTCTAGTGGGTCATGTGGGTTACA
ACAGAAGTCCAGCGCTATTGCTAATGCCATGTCCGACGCATTCCTTCAAACCAACGGGCGT
GTGGATTCAATCTTTGTTCAAGAGATGGTTTCCATGATTAATATGTTCGCCCAAGCGAGTA
GTAACGAAATCTCATATTCGACTGGGAGCACCCGCGGGTCAGCCATGGCGATATCGGATC
CAGGTGGCAACGGTGGCATCGGTGGTTATGGTGGCTATGGTGGTCTGGGCTATGGCGGCT
ACGGTTATGGTGCGGGTAGCTCCGCTGCCGCAGCCGCCGCAGCAGCAGCGGCTGCTGCGG
GTGGCTACGGTGGTTACGGCTACGGCGGCTATGGTTACGGTGGCGGCGGTTCTGCGGCG
GCAGCTGCAGCCGCAGCTGCCGGTGCCGGTGCGGGTGCGGGCGAGGGTGCCGGTGGCTA
TCGCCGTTACGGCGGTTACGGTGATGATGGCGCTGCTGCAGCGGCAGCGGCGGCTGCGG
CCGCGGCAGCCGCGGCGGGCGGTTATGGCGGTTATTACTATGAAGAGGGTGGCGCTGGT
GCGGGCGCGGCGGGTGCGGCCGCCGCTAGCGCGGCAGGCGCAGCCGCTACCGGTGGCG
CCCGTCGCGCGGGCGGCTACGGCTATGGCGCGGGTGGTGCGGCGGGCGCCGCAGCGGCC
GCAGGCGCGGCCGGCGGTGCTGGCACCCGTAATACTGCGTACAATAGCGTCAGCCTGCTG
AGCTCTAGCGCGTCAAACGCACGCGTTAGCACCGCGGTTAGCGGCCTGGTTGGCAATGGC
GTTGATGCGCTGTCTAACATTGTTTCCTCGCTGTCCGCGTCTGTGGCAGCGACGCATCCGG
GTCTGAGCGGTTGTGAAGTCCTGGTCCAGGTTCTGATGGAAATCATCAGCGCACTCGTGCA
CGTGCTGAGCAGCGCGAGCATTGGCCAAGTGAATGTTAACGACGCGGCATACACGGCAGC
ACTGGTGGGTGACGTCATTAGCCAAAATTGCTAACTCGAG

pET32b-NTMaSp-AquaRep-CT translation:
GTGSSGENLYFQGMGWSTQLALPVVLLVSSLLGSTLAASSMFTDASKTEDFMRSMVSSLSRNP
CFTADQISDMESLSDTMTMSVDRMVQTNKVSEHRMKAMSTAMASGIAEIAASSGSCGLQQ
KSSAIANAMSDAFLQTNGRVDSIFVQEMVSMINMFAQASSNEISYSTGSTRGSAMAISDPGG
NGGIGGYGGYGGLGYGGYGYGAGSSAAAAAAAAAAAAGGYGGYGYGGYGYGGGGSAAAA
AAAAAGAGAGAGEGAGGYRRYGGYGDDGAAAAAAAAAAAAAAAGGYGGYYYEEGGAGAG
AAGAAAASAAGAAATGGARRAGGYGYGAGGAAGAAAAAGAAGGAGTRNTAYNSVSLLSSS
ASNARVSTAVSGLVGNGVDALSNIVSSLSASVAATHPGLSGCEVLVQVLMEIISALVHVLSSASI
GQVNVNDAAYTAALVGDVISQNC*LE

pET32b-NTAcSp-AquaRep-CT:
GGTACCGGCAGCAGCGGCGAAAACCTGTATTTTCAGGGCATGTACTGGCTTACTGGATTTA
CGGTGTTGGTCTTACTTTTACAACAATATACATCGGTAGAGGCGGACAGCCGCTATGAGAC
ACTGTCGCCGTGGACAAATCCCACACGCGCAGAGAAATTCATTGACTGCCTTACCGATGGC
ATTGCAAAATCGCCCATCTTTCCGCAACAAGAGAAAGATGATCTGCAATACATCGTTGAAA
CGATGACCAGTGCTATCACCAGTACCTCATCGCAGCGCGGTAATCGTGCTCAACTGCAAGC
GCTTAATATGGCTTTCGCCTCATCCCTTGCAGAATTGGTGATCGCGGAAGACATTGGCAAT
ACTCAAGCCATCAGTGTGAAGACTCAAGCTCTTACGCAATCGTTAAGCAGTTGTTTTCGCTC
TGTAATTGGGACCGTCAACAAACGCTTTATTAACGAGATTAAAGACTTAATCAAAGTCTTC

GCACAGGAAGCTAATAGCGAGTTGAACGAGGTAGATGTCGAAGACTTGGGAAGTGCCAT
GGCGATATCGGATCCAGGTGGCAACGGTGGCATCGGTGGTTATGGTGGCTATGGTGGTCT
GGGCTATGGCGGCTACGGTTATGGTGCGGGTAGCTCCGCTGCCGCAGCCGCCGCAGCAGC
AGCGGCTGCTGCGGGTGGCTACGGTGGTTACGGCTACGGCGGCTATGGTTACGGTGGCG
GCGGTTCTGCGGCGGCAGCTGCAGCCGCAGCTGCCGGTGCCGGTGCGGGTGCGGGCGAG
GGTGCCGGTGGCTATCGCCGTTACGGCGGTTACGGTGATGATGGCGCTGCTGCAGCGGCA
GCGGCGGCTGCGGCCGCGGCAGCCGCGGCGGGCGGTTATGGCGGTTATTACTATGAAGA
GGGTGGCGCTGGTGCGGGCGCGGCGGGTGCGGCCGCCGCTAGCGCGGCAGGCGCAGCC
GCTACCGGTGGCGCCCGTCGCGCGGGCGGCTACGGCTATGGCGCGGGTGGTGCGGCGGG
CGCCGCAGCGGCCGCAGGCGCGGCCGGCGGTGCTGGCACCCGTAATACTGCGTACAATA
GCGTCAGCCTGCTGAGCTCTAGCGCGTCAAACGCACGCGTTAGCACCGCGGTTAGCGGCC
TGGTTGGCAATGGCGTTGATGCGCTGTCTAACATTGTTTCCTCGCTGTCCGCGTCTGTGGC
AGCGACGCATCCGGGTCTGAGCGGTTGTGAAGTCCTGGTCCAGGTTCTGATGGAAATCAT
CAGCGCACTCGTGCACGTGCTGAGCAGCGCGAGCATTGGCCAAGTGAATGTTAACGACGC
GGCATACACGGCAGCACTGGTGGGTGACGTCATTAGCCAAAATTGCTAACTCGAG

pET32b-NTAcSp-AquaRep-CT translation:
GTGSSGENLYFQGMYWLTGFTVLVLLLQQYTSVEADSRYETLSPWTNPTRAEKFIDCLTDGIAK
SPIFPQQEKDDLQYIVETMTSAITSTSSQRGNRAQLQALNMAFASSLAELVIAEDIGNTQAISVK
TQALTQSLSSCFRSVIGTVNKRFINEIKDLIKVFAQEANSELNEVDVEDLGSAMAISDPGGNGGI
GGYGGYGGLGYGGYGYGAGSSAAAAAAAAAAAAGGYGGYGYGGYGYGGGGSAAAAAAAA
AGAGAGAGEGAGGYRRYGGYGDDGAAAAAAAAAAAAAAGGYGGYYYEEGGAGAGAAG
AAAASAAGAAATGGARRAGGYGYGAGGAAGAAAAAGAAGGAGTRNTAYNSVSLLSSSASN
ARVSTAVSGLVGNGVDALSNIVSSLSASVAATHPGLSGCEVLVQVLMEIISALVHVLSSASIGQV
NVNDAAYTAALVGDVISQNC*LE

Appendix A.5

Future spider silk constructs set out in Chapter 5. All of the following silks have been designed to accurately represent the overall amino acid repeat content and have been looked at in terms of similarities and differences to currently used recombinant spidroins. It is expected that each will need tailored expression and purification conditions, but it is not known whether they will be expected to form fibres under laboratory conditions.

A.5.1: Recombinant silk 1 – Linyphiid silk (MaSp1)

Based on the major ampullate sequence (MaSp1) of *Lessertia barbara* (living in acid environments). It is 489 amino acids long and comprised of a native NTD, 4 repeats which are approximately 50 aa long and a native CTD. When compared to other recombinant silks (NT2RepCT, 4RepCT and AquaRepCT), this silk is more hydrophobic and has a higher alanine and aspartic acid content, with a similar glycine and lower glutamic acid. It is most similar to AquaRepCT in terms of its acidic content but is less charged. The greater hydrophobicity may contribute to enhanced solubility in buffers, but it would be interesting to see if this was able to polymerise.

Recombinant silk 1 sequence (terminal regions are underlined)

PWANPQTADAFMNNFLSSVTSNSGGIFTGDQLSDMSDIGFSIMNQLERLNRGGHSSARKMQ
MMDMAFASTMAEVALAEADNPQAALGAISNALSQSFMQTTGSSNQASINSIMSLISMFAQN
AGENGGDGRGSAASSASSASAAAAGGAGGGYGGDGGAAAAAAAAAAAAAAGAAEGYGRG
AGSGSSAASAAAAGAGAAGAGGYADGGAAAAAAAAASAAAAAAAGGAGSGYGARYGGYG
AGGAGAASAAAAAAAAAAAAAGAAAGAADGYGRGAGAGSNAASAAAAAAGAGAGSGGA
GYGDGGAAAAAAAAAAAAAAGSAGSGYGAGYGGYGAGGAGAASAAAAAAAAAAAAASA
GDGYGAGGAGSGASAASAAAAAGSAGAGGAGGYGNGGAAAAAAAASAAAQASAAAARRV
SSAVNTLVTSGPTSVNSLSNVIGNAVAQVRASSPGLSDCDVLVQALLEVTSAMVHILGSASLGA
VNAAAASNS

A.5.2: Recombinant Silk 2 – Nesticid Silk (Possible acidic specific MaSp)

This silk is composed of 5 repeats typical of the possible acid adapted MaSp identified in *Nesticus eremita.* It contains the canonical Poly(A), GG(X), GPG motifs, as well as a higher proportion of D, S and T than expected. Compared to other recombinant silks,

it contains less alanine, but more glycine, E and acidic residues. It would be most interesting to see if this silk forms fibres when expressed, and how well it reacts to acidic environments.

Recombinant silk 2 sequence (terminal regions are underlined)

AAAAAAAAASGGEGGYGGGYSSGGYGGSSSSASAASSASGVGQGGYGGGTGGASGASAGA
SGGSGGYGPSSFEVTEYGPGGRGPGGVGGYGPGGLGGYGRGGEGYGTGEGAASAAAAAAA
AAAAEAAGGEGSYGGYGAGGYGGGSSAASSASAASAAASRLASPASSSAAAAAAAAAAEAA
GGEGSYGGYGAGGYGGGSSAASSASAASAAASGVGQGGYGGGTGGASGASAGASGGSGGY
GPSSFEVTEYGPGGRGPGGVGGYGPGGLGGYGRGGEGYGTGEAAAAAAAAAAEAAGGEGS
YGGYGAGGYGGGSSAASSASAASAAASGVGQGGYGGGTGGASGASAGASGGSGGYGPSSF
EVTEYGPGGRGPGGVGGYGPGGLGGYGRGGEGYGTGEAAAAAAAAAASGGEGGYGGGYSSG
GYGGSSSSASAASSASGVGQGGYGGGTGGASGASAGASGGSGGYGPSSFEVTEYGPGGRGP
GGVGGYGPGGLGGYGRGGEGYGTGEGAASAAAAAAAAAAEAAGGEGSYGGYGAGGYGG
GSSAASSASAASAAASRLASPASSSRISNAASSLVSSGVSNAGALSSVISNLSSQVSASNPGLSGC
DILVQVLMEVVSALVHIVSSSNVGQVNYGSSGQTAQ

A.5.3: Recombinant Silk 3 - Aquatic silk (TuSp)

This recombinant silk is based on the tubuliform silk of *A. aquatica*. It contains a native NTD for solubility, 5 repeat blocks and a native CTD. As this repeat region is so long, a titratable approach to expression is suggested, starting with one repeat block and moving up to 5+, to see how large the protein can be when expressed in *E. coli.*

Recombinant silk 3 sequence (terminal regions are underlined)

VFTDTRLSGAFLQCLTQGIGASSAFPSQEQADIDSIATAILSAASTSTARARALSTALATSLADLVI
TDSVAGDYSTQITALNSILSNCFLQTTGTSNTGFIRNIKNLVGILAGGALTATSNAISGAAASGTS
NVGAQAATQTSSNSYASASASAFSQASAASLASSSAFSSAFASASSAAAAGSIAYNLALQTANA
LGISNAVGIASAVSQAVSAVGAGASSFAYATAVSNAAAQFLAAQGILSQANASALASSFASAFA
SASASASASAASSASAQSAAAARSQAAASAFSQAASQASSRAASQTASQAASQAASQAASQA
GSYSTTTTTSGSQTASQAAARAAAQAASQASSSSYASASASAFSQASAASLASSSAFSRAFASAS
SAAAAGSIAYNLALQTANALGISNAVGIASAVSQAVSAVGAGASSFAYATAVSNAAAQFLAAQ
GILSQANASALASSFASAFSSAAASASASAASSAYAQSATAARSQAAASAFSQAASQASSRSAS
QAASQAASQAASQAGSYSTTTTTSGSQAASQAAARAAAQAASQASSSSYASASASAFSQASA

ASLASSSAFSSAFASASSAAAAGSIAYNLALQTANTLGISNAVGIASAVAQAVSAVGARASSFAY
ATAVSNAAAQFLAAQGILSQANASALASSFASAFASAAASASASAASSASAQSAAAARSQSAA
SAFSQAASQASSRAASQTASQAASQAASQAASQAGSYSTTTTTSGSQAASQAAARAAAQAAS
QASSSSYASASASAFSQASAASLASSSAFSSAFASASSAAAAGSIAYNLALQTANTLGISNAVGIA
SAVAQAVSAVGARASSFAYATAVSNAAAQFLAAQGILSQANASALASSFASAFASAAASASAS
AASSASAQSAAAARSQSAASAFSQAASQASSRAASQTASQAASQAASQAGSYSTTTTTSGSQ
AASQAAARAAAQAASQASSSSYASASASAFSQASAASLASSSAFSSAFASASSAAAAGSIAYNL
ALQTANTLGISNAVGIASAVAQAVSAVGARASSFAYATAVSNAAAQFLAAQGILSQANASALA
SSFASAFASAAASASASAASSASAQSAAAARSQSAASAFSQAASQASSRAASQTASQAASQAA
SQAASQAGSYSTTTTTSGSQAASQAAARAAAQAASQASSSSYASASASAFSQASAASLASSSAF
SSAFASASSAAAAGSIAYNLALQTANTLGISNAVGIASAVAQAVSAVGARASSFAYATAVSNAA
AQFLAAQGILSQANASALASSFASAFASAAASASASAASSASAQSAAAARSQAAASAFSQAAS
QASSRAASQTASQAASQAASQAASQAGSYSTTTTTSGSQAASQAAARAAAQAASQASSSSYA
SASASAFSQASAASLASSSAFSSAFASASSAAAAGSIAYNLALQTANALGISNAVGIASAVAQAV
GAVGAGASSFAYASAVSNAAAQFLAAQGILSQANASALASSFASAFASAAASASASAASSASA
QSAAAARSQAAASAFSQAASQASSRSASQAAGQAGAAAGATAVSSLLRPVAVNLPSLVRSLFP
GFSSSSGLSSAAASARVDSLASSLGTLLSSGQQLNVQNFAYTLLSQASQGQNSNVGISSNEVIV
QALLEGLSALIQVINSSQISSVSSRSG

A.5.4: Recombinant Silk 4 - Aquatic Silk (AcSp)

This recombinant silk is based on an aciniform sequence from *A. aquatica* and is again composed of a native NTD, 3 repeat blocks and a CTD.

Recombinant silk 4 sequence (terminal regions are underlined)

PWTNPTRAEKFIDCLTDGIAKSPIFPQQEKDDLQYIVETMTSAITSTSSQRGNRAQLQALNMAF
ASSLAELVIAEDIGNTQAISVKTQALTQSLSSCFRSVIGTVNKRFINEIKDLIKVFAQEANSELNEV
DVEDLGSDVAGNIYGTGFYGAASSGIFTACAEAIASVVVDGLANAGILNTSNGYAFASNFVSSF
VQTASSAAASAAASSSASSSQTSSSTSTAASSASAAAAASAAGASYAASSAFGASGFSTLIGNVA
SGLLDSGAFQSLFGAGIAVNVRLSIIQSSLQAAFRTLNVSYAVSSSISSAVSNAISSTSSSSACAEAI
ASVVVDGLANAGILNTSNGYAFASNFVSSFVQTASSAAASAAASSSASSSQTSSSTSTAASSASA
AAAASAAGASYAASSSFGASGFSTLIGNVASGLLGSGAFQSLFGAGIAVNVRLSIIQSSLQAAFR
TLNVSYAVSSSISRAVSSAISSTSSSSACAEAIASVVVDGLANAGILNTSNGYAFASNFVSSFVQT
ASSAAASAAASSSASSSQTSSSTSNEVAGLGDGLANAGILNTSNGYAFASNFVSSFVQTASSAA

ASAAASSSASSSQTSSSTSTSASSASGAGAASAAGNGAGASSALSSGLLYDRDLFDGVGIVDNV
PYGSGSLQWASADGLTSPQARARIEQLTSELAQSFGSGKADLSAVSRAIVSSFSQAKSYGLSEV
DAKVEALFETITALIQILSSSQVGQVDTSLTESISSFLVDSLEDKFCLGI

‘Oh no...I don't understand it. But for that matter I don't understand how a spider learned to spin a web in the first place. When the words appeared, everyone said they were a miracle. But nobody pointed out that the web itself is a miracle.’

-E.B. White, ‘Charlotte’s Web’