Characterisation of next generation affinity reagents

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy

By

Rebecca Ann Pattison

September 2017

Acknowledgements

Firstly, I would like to thank my supervisor, Professor Rob Beynon for his support, guidance and knowledge provided throughout my PhD. I must also acknowledge Avacta Life Sciences for funding and collaborating on this project and my supervisors Dr Paul Ko Ferrigno and Matt Johnson. I would also like to thank Professor Enitan Carrol for collaborating on the sepsis work and for providing the samples and to the BioScreening Technology Group at the University of Leeds for providing Adhirons. Additionally, I would like to acknowledge Kit-Yee Tan, Vincent Puard and Kimberley Burrow for collaborating on the naïve Affimer work.

To all members of CPR, I am extremely grateful for your help and guidance during the past four years, particularly Philip for your mass spectrometry help and training. To Gemma, Grace, Richard and Sam, thank you for making my time in CPR memorable.

I would like to give the biggest thanks to my family and friends for your continued encouragement and support. I must say a special thank you to my Mum. For everything!! Finally, I would like to thank Rich for being unbelievably supportive throughout my entire PhD and for always believing I can do anything! And thank you for being extremely patient and for putting up with me, particularly over the last few months!

Abstract

The role of affinity reagents in biological research is essential allowing for the identification, characterisation and functional assignment of proteins. To date, antibodies are the most widely used and studied affinity molecules. The ability for antibodies to bind, with high specificity and high affinity, to the targets they were generated against are features that have been exploited in biological research. However, despite the accomplishments that have been achieved with antibodies, they do possess limitations which have driven the need to develop non-antibody based scaffold proteins that have comparable binding affinity and specificity to antibodies. In addition, exploration of highly complex proteomes, by affinity reagents will enhance our understanding of areas where analytical capabilities are currently limiting findings due to large dynamic range and number of proteins within a given proteome. Affimers, developed as antibody-alternative affinity reagents, are engineered combinatorial proteins possessing three variable interaction sites. This thesis describes the use of mass spectrometry in the characterisation of Affimers and their development for affinity purification mass spectrometry (APMS) workflows. Chapter 3 presents immobilisation strategies for Affimer APMS workflows for the identification of unknown protein binders of naïve Affimers and highlights the associated challenges of non-specific background binding of proteins and achieving sufficient target enrichment. Four immobilisation methods were assessed and despite the development of an effective method for the enrichment of IgG using Affimers via cysteine-mediated immobilisation, the method was not successful for naïve Affimer target identification. Chapter 4 describes the characterisation of Affimers that target human pepsinogen. Current assays to detect pepsinogen rely on antibodies however they lack any clinical use due to inherent problems with antibodies such as reproducibility and batch-to-batch variability. Five Affimers were expressed, purified and characterised by intact mass spectrometry. As a novel approach to overcome the issue of the large signal of capture reagents in APMS assays, a pepsinogen Adhiron resistant to Lys-C proteolysis was generated. Structural stability of the mutant, assessed by analysing the rate of proteolysis and collision induced unfolding, revealed the mutant exhibits comparable structural stability to the native protein. Chapter 5 presents an alternative approach to Affimer APMS methods for the identification of differentially expressed proteins in sepsis, by using discovery proteomics. Due to the large dynamic range of plasma, an antibody-based spin column depletion method was applied to samples within the study prior to LC-MS/MS analysis. Label-free quantification and bioinformatics analyses of patient cohort 1 and cohort 2 identified 40 and 107 differentially expressed proteins,

respectively. A panel of five candidate proteins were selected as potential markers of sepsis and for subsequent Affimer development: CRP, neutrophil-gelatinase associated protein, Protein S100 A8/A9, interleukin-1 receptor-like 1, cathepsin B. Chapter 6 outlines the preliminary development of the major urinary protein (MUP), darcin, as a novel protein scaffold. The disulfide bond in darcin was shown to be vital for providing the high structural stability of the protein, an important feature of protein scaffolds. In addition, preliminary findings demonstrated the development of a darcin resistant to proteolysis may be a suitable approach to overcome the intense signal of the affinity reagent in APMS assays. In summary, this study presents a novel approach to overcome the challenges of APMS with the development of non-digestible protein scaffolds and builds on the literature of common affinity purification contaminant proteins. In addition, the study provides a contribution to sepsis plasma proteome analysis and identifies five proteins implicated in sepsis that provided targets to guide Affimer production for future Affimer immunoassay-based strategies.

Table of Contents

List of Figures

List of Tables

Abbreviations

2-DE	Two dimensional gel electrophoresis
2D LC	Two dimensional liquid chromatography
ACN	Acetonitrile
AEX	Anion exchange chromatography
AUC	Area under the curve
BAL	Bronchoalveolar lavage
BPI	Base peak intensity
BSA	Bovine serum albumin
C	Carbon
CDR	Complement determining region
CEX	Cation exchange chromatography
CID	Collision induced dissociation
CIU	Collision induced unfolding
CRP	C-reactive protein
Da	Daltons
DA	Direct aspiration
DARPins	Designed ankyrin-repeat proteins
DDA	Data dependent acquisition
DIGE	Difference in gel electrophoresis
DNA	Deoxyribonucleic acid
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
ELISA	Enzyme linked immunosorbent assay
ESI	Electrospray ionisation
ESI-MS	Electrospray ionisation mass spectrometry
FASP	Filter aided sample preparation
Fc	Fragment crystallisable region
FDR	False discovery rate
HCAb	Heavy chain antibody
HCD	High energy collision dissociation
HILIC	Hydrophobic interaction liquid chromatography
HPLC	High performance liquid chromatography
IB	Inclusion body
IEX	Ion exchange chromatography
IMAC	Immobilised metal affinity chromatography
IMS-MS	Ion mobility mass spectrometry
IPTG	Isopropyl β-D-1-thiogalactopyranoside
ITC	Isothermal titration calorimetry
K	Lysine
kDa	Kilodalton
LB	Luria broth
LC	Liquid chromatography
LC-MS/MS	Liquid chromatography tandem mass spectrometry
m/z	mass to charge

MAb	Monoclonal antibody
MALDI	Matrix assisted laser desorption ionisation
mM	Millimolar
Mr	Relative molecular mass
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
MSIA	Mass spectrometric immunoassay
MudPIT	Multidimensional protein identification technology
MUP	Major urinary protein
N	Nitrogen
Ni-NTA	Nickel - nitilotriacetic acid
NTA	Nitilotriacetic acid
OD	Optical density
PAb	Polyclonal antibody
PBS	Phosphate buffered saline
PEG	Polyethylene glycol
pI	Isoelectric point
POCT	Point-of-care test
ppm	Parts per million
PrEST	Protein epitope signature tag
PSM	Peptide spectral match
PTM	Post translation modification
R	Arginine
RA	Reflux aspiration
RIA	Radioimmunoassay
ROC	Receiver operating characteristic
RP	Reverse phase
RPPA	Reversed phase protein arrays
QQQ	Triple quadrupole
SCX	Strong cation exchange
SDS	Sodium dodecyl sulfate
SDS-PAGE	Sodium dodecyl sulfate - polyacrylamide gel electrophoresis
SEC	Size-exclusion chromatography
SEC-MALS	Size-exclusion chromatography multi-angle light scattering
SELEX	Systematic evolution of ligands by exponential enrichment
SF	Soluble fraction
SISCAPA	Stable isotope standards and capture by anti-peptide antibodies
SPE	Solid phase extraction
SPR	Surface plasmon resonance
SQT	Stefin A Affimer with no loop insertions
SRM	Selective reaction monitoring
STM	Stefin A triple mutant
SWATH	Sequential window acquisition of all theoretical fragment ion mass spectra
TCA	Trichloroacetic acid
TCEP	Tris(2-carboxyethyl)phosphine
TEV	Tobacco etch virus

TFA	Trifluoroacetic acid
TIC	Total ion count
TOF	Time-of-flight
UHP	Ultra high pressure
UPLC	Ultra performance liquid chromatography

1. Introduction

1.1 Complexity of the proteome

The term 'proteome' was first used by Wasinger et al. in 1995[1] and Wilkins et al. in 1996[2] to describe the protein complement of a cell under defined environmental and physiological conditions. Proteomics is the large-scale study of the proteome, encompassing the analysis of protein abundance, localisation and function as well as protein-protein interactions to gain greater understanding of cellular dynamics and systems. It is estimated that the human genome encodes 19000 genes[3]. However, as numerous protein variants exist for a single gene due to alternative splicing, the number of human proteins is estimated to be a lot higher[4]. A recent large-scale transcriptomic study investigating alternative splicing estimated that approximately 100,000 different protein isoforms can be produced from the human genome[5]. In addition, post-translation modifications (PTMs) and genetic variations increase the size of the human proteome adding to the complexity. Variation in genomes between individuals is expected to be as high as 10^6 differences which is likely to have an effect on proteome complexity[6]. Furthermore, in specific pathophysiological states, genomic variation is vastly increased. For example, tumour heterogeneity is predominant in many cancer types, diversifying the genome[7] and thus contributes to proteome complexity. Proteomic analysis of an organism or cell is further complicated by the dynamic expression of proteins due to responses to stimuli such as different developmental, pathological and environmental conditions. In *Saccharomyces cerevisiae*, environmental stresses are known alter the dynamic expression of the transcriptome and proteome, again adding further complexity to the proteome[8].

Despite the vast number of proteins within a proteome, only a small portion of these are expressed by a particular cell or tissue at a given time. With the process of protein production beginning with transcription followed by translation, what determines protein expression is ultimately governed by the regulation of gene expression. Regulation occurs at many control points along the pathway of gene expression; transcription, RNA processing, RNA transport and localisation, translation, mRNA degradation and control of the activation of proteins by PTMs[9]. Transcriptional control plays a critical role in what governs the protein complement of a cell as it is the starting point for RNA production and is therefore considered the most important regulatory mechanism for many genes. The action of transcription factors, by recognising specific regulatory DNA sequences in promoter/enhancer regions and by complex formation, causes the recruitment of RNA

polymerase and chromatin remodelling[10]. It is through these mechanisms, that the protein complement of a cell is regulated.

Genomic, transcriptomic and proteomic analysis are complementary large-scale techniques, providing a comprehensive view of a biological system. However, correlation of mRNA expression levels and the resulting protein is usually poor[11]. As discussed, many factors influence the correlation between mRNA and protein expression levels including protein degradation[12] and an organism's regulation of translation[13]. As proteins make up the functional element of a cellular system they provide valuable information regarding the biological state of a sample, cell or organism. This highlights the importance of proteomic research as a complementary method to genomics and transcriptomics in biological exploration.

1.2 Challenges in the analysis of complex proteomes

The challenges in complex proteome analysis can be attributed to two key features, firstly the number of proteins and secondly, the dynamic range. As discussed, the number of proteins expressed in a proteome at a given time is vast and highly dynamic with PTMs increasing the proteomic space. It is estimated that a single human cell line in culture can express up to 12,000 proteins requiring huge breaths in analytical capabilities[14-16].

To add further complexity to this challenging environment, the quantitative amount of each protein is varied and continually changing. In proteomics, dynamic range is defined as the difference in concentration between the most abundant and least abundant protein. The dynamic range of proteins in a proteome can span at least seven orders of magnitude, ranging from one copy per cell to 10 million copies per cell[17]. In human plasma, the dynamic range of the proteome is much higher at 10^{12}. This poses a major analytical problem as current technology cannot achieve this dynamic range. Therefore, when a proteome is analysed with discovery based techniques only a fraction of the proteome is able to be accurately identified and an even smaller fraction accurately quantified[18] (Figure 1.1). In mass spectrometry (MS), sample loading is limited by the concentration of the most abundant peptides within the sample so that the detector does not become saturated. Therefore, low abundance proteins can therefore fall below the limit of detection of the mass spectrometer.

Human plasma is a prime example of the challenges in complex proteome analysis, with albumin representing 50% of the total plasma concentration and an additional 12 proteins representing a further 95% of the total plasma concentration[19]. A standard S-shaped distribution is observed when plotting protein rank against the log abundance (Figure 1. 1). This observation is not only seen in blood but also, proteomic analysis of human cell lines revealed a large dynamic range spanning seven orders of magnitude, with core histones and enzymes representing the most abundant proteins[20]. Furthermore, the yeast proteome can span up to five order of magnitude with proteins detected from one million copies per cells to 41 copies per cell[21,22].

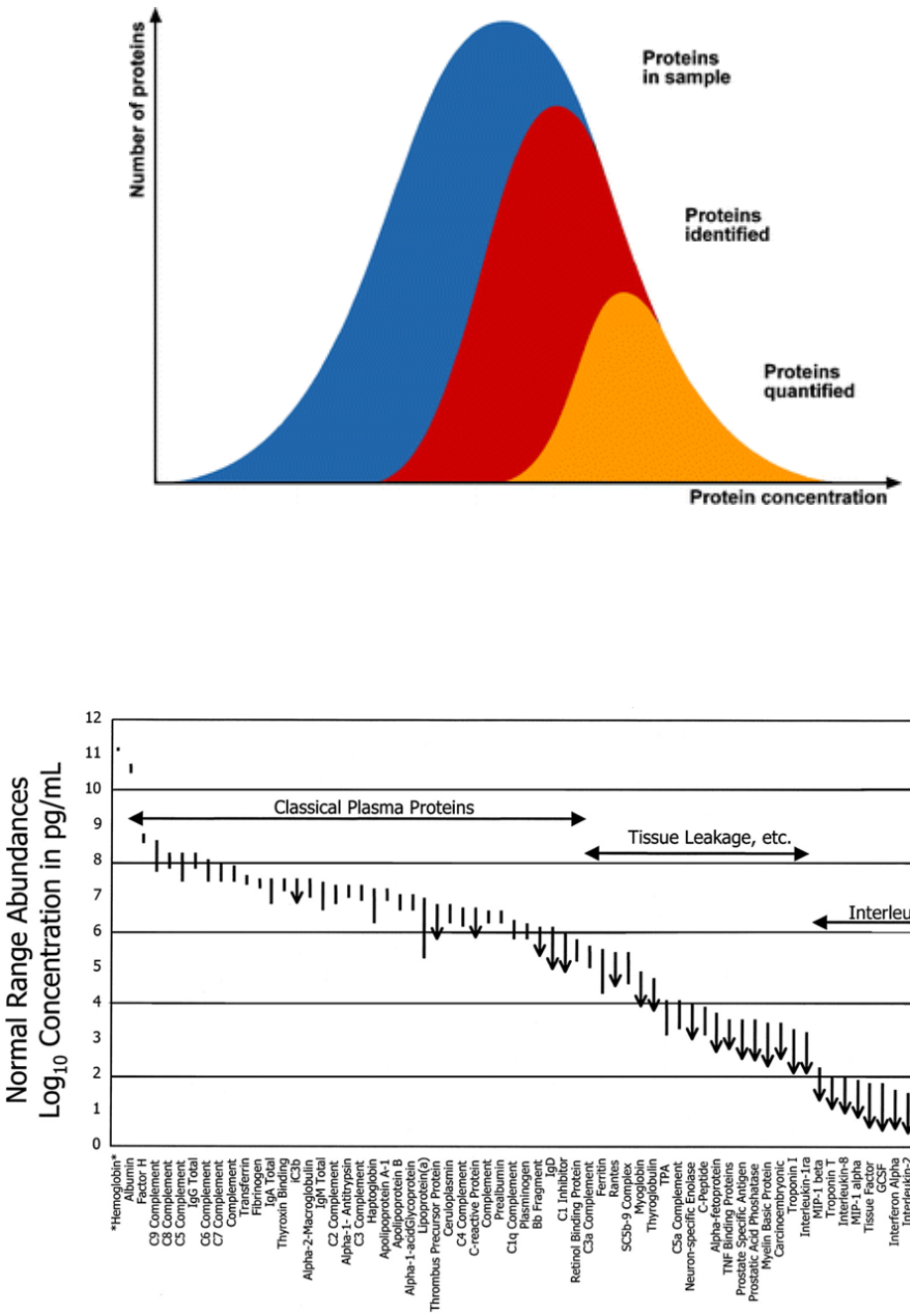


Figure 1.1 | Distribution of proteins in a proteome and dynamic range of the human plasma proteome.

Top: Expression range of a proteome and the different fractions that can be identified and quantified. Image taken from Kuster et al (2007)[23]. Bottom: The dynamic concentration range of the plasma proteome spans 12 orders of magnitude. Proteins are split into three key categories based on abundance; classical plasma proteins, tissue leakage and interleukins. Image taken from Anderson and Anderson (2002)[19].

Methods for proteome analysis at a protein level

A large range of diverse analytical methods are available for the study of complex proteomes. Traditionally, complex proteomic analysis was performed using two-dimensional polyacrylamide gel electrophoresis (2-DE) where proteins are separated based on isoelectric point (pI) and molecular weight (Mr) in the first and second dimension, respectively[24]. The proteins are then visually detected by protein staining. 2-DE has been used to resolve more than 5000 individual protein spots[25,26] and can detect less than 1 ng of protein per spot[27]. Due to the visual protein map generated, differential analysis between samples can be performed. 2-DE is an advantageous technique for complex proteome analysis because of the extensive, highly orthogonal protein fractionation and separation providing high resolution of proteins[27]. Furthermore, the technique is particularly useful for the separation of different intact proteoforms due to small difference in net charge induced with modifications[27]. An advancement of 2-DE is the technique 'difference in gel electrophoresis' (DIGE)[28]. As a limitation of 2-DE is lack of reproducibility, the DIGE approach was developed to overcome this by enabling up to three samples to be analysed on a single gel by labelling the samples with fluorescent dyes. The samples are mixed together in equal ratios and run on a single gel where the differences in samples are easily determined. However, 2-DE and DIGE do not provide sample identification. The protein spots have to be extracted for subsequent MS analysis or probed with antibodies

Protein microarrays are another analytical method to explore the proteome allowing for the large scale, high throughput analysis of samples[29]. The methodology exploits protein-interactions and the inherent nature that many proteins have of affinity for a specific ligand. The technique enables the simultaneous analysis of hundreds to thousands of samples[30]. Detection and measurements of binding is achieved through the labelling of the proteins with a fluorescent tag. Microarrays are typically available in three different formats; binder arrays that present different affinity reagents, functional arrays that present different proteins and reverse phase arrays that probe different biological samples with selected affinity reagents[30]. As reverse phase arrays analyse the proteome in a more targeted and specific nature, they will be discussed further in section 1.4. In a binder array**,** individual binding reagents (antibodies, recombinant antibodies or alternative affinity reagents) are immobilised onto a slide. Samples are then profiled by the array with the goal to identify proteins that are differentially expressed between two conditions, typically disease and control samples for biomarker studies. An early application of an antibody microarray consisted of 92 antibodies that targeted classical inflammatory proteins and was

profiled with the sera of patients with either benign pancreatic disease or pancreatic cancer[31]. A total of 142 samples were analysed and protein markers were identified that could distinguish the two conditions with high sensitivity and specificity[31].

A different type of array for proteome analysis is a functional protein array and is produced by spotting different purified recombinant proteins on the array. Large numbers of proteins can be spotted on the array and can contain over 60 % of an organism's proteome immobilised on the array. Proteome arrays are available for a number of different species including *E.coli*[32], humans[33] and *S. cerevisiae*[34,35]. Functional protein arrays have application in assessing various interactions including protein-small molecule, protein-glycan, protein-protein, protein-peptide[36] and protein-nucleic acids interactions. The identification of serological biomarkers is a common application for function protein arrays typically detecting autoantibodies[37]. Despite the progress in proteome analysis using protein arrays, technical limitations have restricted their potential one of which is that the technique relies on high quality binding reagents that target many proteins[38].

Another approach to proteome analysis is by using mass spectrometry (MS). MS is most often associated with bottom-up methods, where peptides are generated from protein digestion. However, the analysis of intact proteins (top-down proteomics) using MS is a powerful method that allows the structural features of the proteins to be preserved such as PTMs[38]. Complete protein sequencing coverage is regularly achieved allowing information on different proteoforms to be obtained[39]. Intact protein analysis is however, subject to limitations. Protein identification based on the intact mass of the protein alone requires no deviations of the observed experimental mass from that of the predicted mass from the DNA derived sequence. However, PTMs such as glycosylation, phosphorylation and acetylation, disulfide bond formation, sequence variations, truncations and proteolysis will all result in a deviation from the expected mass. Furthermore, it is likely that a single protein will carry a variation of the above modifications, vastly increasing the complexity of the mass spectrum. To overcome this, top-down approaches involve the fragmentation of the intact protein to generate sequencing information. As fragmentation methods such as electron transfer dissociation (ETD) can be used, the PTMs are retained. Top-down proteomics does have some limitations such as it is not well suited for the analysis of large proteins (greater than 70 kDa). This is due to the formation of a large number of isotopic peaks caused by the spreading of the charge resulting in a low signal to noise.

Proteome complexity also hinders top-down MS methods as protein fractionation is usually necessary in order to separate proteins. Protein fractionation methods that can be coupled with top-down mass spectrometry will be reviewed in section 1.3. But briefly, fractionation methods prior to top-down MS are very powerful and have enabled the identification of 3000 intact proteins[40]. Interestingly, using electrospray ionisation coupled with a quadrupole time-of-flight MS, Hayter *et al* were able to resolve successfully a complex mixture of proteins from chicken muscle without the need for pre-fractionation steps or protein fragmnetation[41]. The multiply charged envelopes were suitably resolved and deconvoluted to enable the correct assignment of the observed masses with the genome-predicted indicating that complex mixtures can be analysed without fractionation.

Methods for proteome analysis at a peptide level

As an alternative to intact protein analysis, bottom up proteomics is routinely employed for discovery proteomic analysis. Bottom up proteomics involves the analysis of peptides generated from the digestion of proteins using specific proteases followed by liquid chromatography tandem mass spectrometry (LC-MS/MS). The peptides generated from the protein digest are separated using liquid chromatography (LC), ionised by electrospray ionisation and analysed by MS where the peptides ions are separated by their mass-to-charge ratio (*m/z)* and detected. The peptide ions, termed precursor ions are then fragmented, routinely by collision induced dissociation (CID) to generate product ions, a technique termed tandem mass spectrometry (MS/MS)[42].

The bottom up approach has advantages over intact protein analysis relating to peptide analysis instead of intact proteins. Firstly, peptides are significantly smaller than intact proteins and therefore are easier to ionise. Additionally, peptides are generally more soluble than proteins which mean that they can be readily analysed by ESI-MS without having to be concerned about insolubility[43]. Another advantage of bottom up proteomics relate to the proteases used for protein digestion. Typically, the protease used for protein digestion is trypsin due to its highly specific and rapid cleavage at the carboxy-terminal of lysine and arginine residues (except when followed by a proline)[44]. This is an important feature as it results in a highly basic amino acid residue at the C-terminal of the peptide. The general distribution of lysine and arginine residues in proteins typically results in peptides that fall within a small, similar mass range that are in the detection limit of a mass spectrometer. Furthermore, the generated peptides are in the preferred mass range for efficient fragmentation compared to proteins, generating higher quality and complete MS/MS spectra[45].

Despite the advantages of bottom-up proteomics, numerous limitations are associated with digesting proteins into peptides. Firstly, the digestion of protein mixtures vastly increases sample complexity, a further challenge when analysing already complex proteomes. Secondly, by converting proteins into peptides lots of information can be lost. For example, only a few peptides per protein may be detected for a given protein resulting in low sequence coverage. This means PTMs, protein isoforms and truncations may not be mapped back to the associated protein[43]. Another limitation of peptide analysis is that peptides may not be unique to a single protein leading to ambiguity in protein identification.

Various MS modes are available for peptide analysis and method choice is dependent on the experimental set-up. In data-dependent acquisition (DDA), a full survey scan is performed on the precursor peptide ions (MS1) from which specific precursor ions are selected for fragmentation (MS/MS)[42]. The fragmentation spectra allows for the deduction of peptide sequencing information typically performed by database searching for complex samples. A top N approach is implemented, where N is a defined number of precursor ions with the highest intensity, are selected for fragmentation. The usual MS method for discovery proteomics is DDA. However, this approach has limitations which have led to the development of other MS methods. Firstly, the analysis and profiling of peptides is dependent on the scanning speed of the mass spectrometer[46]. Instruments with fast scanning speeds can sample more ions at a given time, resulting in a greater number of MS/MS fragmentation spectra. A study of a complex cell lysate digest analysed the number of peptide ions fragmented in a DDA LC-MS/MS experiment and found extensive under-sampling with only 16 % of the precursor ions selected for fragmentation[47]. Secondly, as mentioned, DDA methods usually select for the most intense precursor ions for fragmentation and therefore there is an inherent bias towards the most abundant peptides in the sample. As a result, acquiring MS/MS fragmentation spectra for low abundance peptides is less likely. Finally, the co-elution of chimeric peptides is highly likely when analysing complex mixtures which can lead to MS/MS spectra composed of more than one chimeric peptide ion[48,49].

As an alternative to DDA, data-independent acquisition (DIA) methods were developed to improve the dynamic range and to remove the bias sampling and selection of abundant precursor ions for fragmentation[50]. DIA methods broadly operate in two ways; either by the acquisition of MS1 spectra followed by sequential MS2 acquisition of m/z window across

the mass range or by acquiring all MS1 spectra followed by fragmentation and acquisition of all the MS2 ions[51]. The most frequently cited techniques are sequential window acquisition of all theoretical mass spectra (SWATH)[52] and MSE[53] both of which have been used for complex proteome analysis[54,55]. DIA methods do however, have limitations. As the approach generates many MS/MS spectra due to fragmentation of all precursor ions, the MS/MS spectra are highly complex which can result in challenging database searching which may compromise confidence in peptide identifications[56]. However, recent developments in data processing are working to overcome this[57].

Alternatively, selected reaction monitoring (SRM) is a targeted mass spectrometric technique used predominantly for the quantitative analysis of proteins using a triple quadrupole (QQQ) mass spectrometer. The method has the largest dynamic range capabilities and the lowest limits of detections compared to shotgun proteomic methods[58]. This can be explained by the non-scanning nature of the method allowing for extended times spent recording the signal of an individual transition, known as dwell time. Briefly, the method operates as follows. The first quadrupole (Q1) acts as a mass filter whereby selected pre-defined precursor ions are allowed through to the second quadrupole (Q2). In the second quadrupole, fragmentation by collision induced dissociation (CID) occurs and then pre-defined fragment ions are passed through to the third quadrupole (Q3) which acts as a mass filter[59]. Analysis of the precursor-fragment ion pair, termed transitions is what gives the technique such high selectivity increasing the signal to noise and therefore is able to achieve low detection limits. Although the signal from highly abundant peptides does not directly cause problems with detection of the transitions, peptides from highly complex samples may result in the interference from co-eluting peptides that have similar precursor m/z values. Another limiting aspect of the SRM methodology is that pre-defined targets are required meaning that the technique is not suitable for discovery of new proteins. Despite this, SRM is still not suitable in overcoming the plasma proteome dynamic range. However, the SRM approach, when used in combination with other fractionation and enrichment strategies to overcome the large dynamic range of proteomes it may be suitable[58].

1.3 Methods for proteome fractionation and enrichment

Numerous approaches have been developed to reduce or overcome dynamic range issues to achieve greater depth of proteome coverage. Proteome fractionation entails the separation of a sample into different fractions using protein or peptide properties in order

to simplify the sample mixture. Enrichment methods on the other hand work by increasing protein concentration, increasing the likelihood of detection. The methods available can either be carried out a protein level or at a peptide level. Typically, both protein and peptide fractionation are used in combination to achieve the enhanced proteome coverage.

Protein fractionation and enrichment

As discussed in section 1.2, 2-DE is an extremely powerful tool for complex proteome fractionation enabling high resolution and separation of proteins. However for intact protein analysis post fractionation, the sample recovery can be low[60]. Therefore other methods of protein separation are usually used. More recently an approach that is an advancement of 2-DE was proposed for intact protein fractionation. Using a computational model for proteome fractionation and identification it was hypothesised that proteins could be separated by isoelectric point and then by the linear sequence volume[61]. Although the approach was not proven experimentally, proteins are considered to differ enough in isoelectric point and linear volume to provide complete fractionation using one dimensional gel electrophoresis followed by nanopore separation[61]. As the experimental values obtained can be compared to computationally generated values, protein identification can be performed without the need for protein sequencing by MS.

A widely used method for protein separation is liquid chromatography (LC). The method, in general, involves the separation of proteins based on their different physiochemical properties using a solid stationary phase and a liquid mobile phase. Various types of chromatography are available including hydrophobic interaction liquid chromatography (HILIC), size exclusion chromatography (SEC), reversed-phase chromatography (RP) and ion exchange chromatography (IEX) and have all been applied to complex proteome separation[38]. SEC separates proteins based on their size and uses columns containing a porous material as the stationary phase that small proteins can enter, while larger proteins cannot enter. Therefore, larger proteins pass through the column earlier while smaller proteins take longer to travel through the pores and thus elute later[62]. A limitation of SEC is the low resolution; however the technique has been improved with the development of ultra-high pressure (UHP) SEC[63]. Using UHP SEC, intact proteins with a mass of 6 – 669 kDa were separated with high resolution with direct coupling for top down MS.

Alternatively, IEX separates proteins based on the charge, with two different types used; either cation exchange chromatography (CEX) where a negatively charged medium is used

or anion exchange chromatography (AEX) where a positively charged medium is used. By increasing the ionic strength of the mobile phase, proteins are eluted from the stationary phase by displacement. As charge is what drives protein separation, when used in conjunction with other chromatography methods that separate based by hydrophobicity, the separation is highly orthogonal[38]. IEX usually uses mild buffers and therefore the proteins are not denatured which can make it suitable for separating proteins whilst maintaining protein complex. Numerous intact protein samples including *E.coli*[64], human leukocytes[65] and yeast have been separated using strong anion exchange.

A different approach to reduce the dynamic range of biological samples is to employ depletion strategies. Immunodepletion columns, first described in 2003[66], are an effective and frequently used approach to deplete highly abundant proteins. Usually immunodepletion columns focus on the depletion of plasma and serum proteins for the reasons discussed in section 1.2. The columns contain antibodies to the most abundant plasma proteins and thus when the columns are incubated with sample; the respective proteins are removed from the sample. Various immunodepletion columns are available typically removing the top two, top 12 or top 20 most abundant proteins[66]. However, it is important to note that the binding capacity of the antibodies immobilised in the columns determine the effectiveness of protein depletion which typically means that a small volume of sample is able to be depleted. Furthermore, the technique may result in the loss of low abundance proteins such as cytokines[67]. Also, as there are high costs associated with the columns, more cost-effective methods may be more suitable. Using chemical-based depletion methods, highly abundant proteins can be precipitated from samples using acetonitrile[68], ammonium sulphate[69] and sodium chloride with ethanol[70]. As albumin is the most abundant plasma/serum protein, these methods tend to focus on the specific depletion of albumin.

An alternative approach from depletion strategies is equalisation of the dynamic range of a proteome. This is usually performed using bead-based combinatorial peptide ligand libraries[71-73]. Complex samples with a large dynamic range are incubated with the beads and bind to the random peptides immobilised on the beads. The approach equalises protein concentrations as once the beads become saturated with their corresponding highly abundant protein, no more can bind and thus the protein is not retained for analysis. The approach also enriches for low abundance proteins, concentrating and equalising the amount of protein to that of the highly abundant proteins. However, this is assuming that

there is enough of the low abundance protein within the sample to begin with. For proteins that do not bind to the bead-conjugated peptides, those proteins are not retained for analysis and therefore the methods may result in the loss of numerous proteins. Equalisation beads have been employed for the analysis of numerous samples including skeletal muscle[74], HeLa cells[75] and *E.coli*[76] and have been particularly successful in enriching for low abundance proteins, reducing the dynamic range of samples.

Peptide fractionation and enrichment

In addition to employing protein fractionation steps, fractionation at a peptide level can also be employed to reach deeper into the proteome. High performance liquid chromatography (HPLC) and ultra-performance liquid chromatography (UPLC) are common separation techniques for peptide mixtures and are routinely couple with electrospray ionisation (ESI) and tandem mass spectrometers for bottom-up proteomics. Reverse-phase (RP) chromatography is typically employed for peptide separation based on hydrophobicity with gradients of low to high organic solvent. This method is particularly popular due to its ability to be directly coupled with mass spectrometry. Work published by Hsieh *et al* found that increasing gradient lengths resulted in a greater number of peptide identifications when analysing a complex proteomic sample[77]. However, despite deeper proteome profiling, sample throughput is dramatically reduced when long LC gradients are used.

To further increase peptide separation and resolve complex samples, two dimensional liquid chromatography (2D LC) can be used prior to MS/MS analysis. The technique involves the separation of peptides using a first dimension that is highly orthogonal to a second dimension. As peptides have various different physiochemical properties, by employing different chromatography techniques, orthogonal separation is easily achieved[78]. First dimension techniques applied include normal phase, ion exchange and size exclusion chromatography. As the separated peptides are usually analysed by MS, the second dimension is normal RP chromatography to enable direct coupling to the MS instrument. In multidimensional protein identification technology (MudPIT)[79], peptides are separated by strong cation exchange (SCX) in the first dimension followed by RP chromatography in the second dimension[80]. An early study of the yeast proteome using MudPIT identified 1483 proteins[80] which pioneered further large-scale proteomic studies. Despite improved numbers of protein identifications, long MS analysis time is required due to the large number of fractions generated. Recent advancements in MS instrumentation with higher

acquisition rates have enabled the identification of approximately 4000 yeast proteins without the need for time-consuming fractionation[54,81].

Another approach to proteome fractionation and enrichment is to focus on specific subsets of proteins or peptides, particularly to proteins with PTMs. The fields of phosphoproteomics and glycoproteomics, involving the analysis of phosphorylated and glycosylated proteins respectively, provide extensive information on biological processes and a greater understanding of cellular dynamics. Whilst it is estimated that approximately 30 % of all proteins are phosphorylated[82] and 50 % are glycosylated[83], the level of modified protein is relatively low and is highly dynamic particularly in response to external stimuli and pathophysiological conditions. The phosphoproteome and glycoproteome represent the lower dynamic range of a proteome and therefore can be difficult to detect particularly when the unmodified peptides are more abundant. Various enrichment strategies for phosphorylated peptides have been employed including immobilised metal affinity chromatography[84], titanium dioxide (TIO_2) chromatography[85] and strong cation exchange (SCX)[84]. Glycopeptides can be enriched using lectin affinity purification, HILIC and SPE[86]. Such enrichment strategies have been able to analyse deeper areas of the proteome. For example, a method where a series of enrichment strategies were employed for PTMs identified a total of 8000 proteins with over 3000, 15000, and 20000 acetylation, ubiquitination and phosphorylation sites respectively[87].

Advancements in MS methodology and fractionation technology are allowing deeper proteome analysis. Various groups have now published the identification of over 7000 proteins in a single sample type[14,88-90]. However, they typically involve fractionation steps and long LC gradients, leading to long MS run times. A recent study, aiming to detect the entire human proteome has identified over 84 % proteome coverage of protein-coding genes using mass spectrometry[91,92]. Whilst this is impressive, portions of the proteome are still hidden from detection and other strategies for complete proteome analysis are needed. Affinity proteomics fits hand-in-hand with MS, involving the application of binding reagents to detect specific protein targets. Affinity proteomics, to enrich for low abundance proteins, is considered the ideal strategy in order to fully analyse complete proteomes.

1.4 Strategies for selective enrichment of proteins and peptides

Another approach for the exploration of complex proteomes is to use enrichment steps that utilise specific interactions between a protein and a ligand, in order to increase the

abundance of the target protein in relation to the other background proteins. Many ligands have been used for specific enrichment of proteins including Cibacron Blue, Blue B, gelatin and biotin for albumin, kinases, fibronectin and streptavidin, respectively[86]. Briefly discussed in section 1.3 was the selective enrichment of glycoproteins using lectin. This has proven particularly informative for the analysis of cancer cells, due to the associated aberrant protein glycosylation in many tumours[93]. When combined with microarray technology, lectin arrays are a high throughput method to explore glycosylation states[94,95]. Another strategy for the selective enrichment of proteins is for antibody purification that exploits the interaction between the constant region of antibodies and protein A or protein G from *Staphylococcus aureus* and *Streptococcus* bacteria, respectively[86].

One of the most common reagents for selective targeting of proteins is the use of antibodies. They have been employed in numerous techniques for the enrichment of proteins from complex samples, of which mass spectrometric immunoassays (MSIA)[96], stable isotope standards and capture by anti-peptide antibodies (SISCAPA)[97] and reversed phase protein arrays (RPPA) will be discussed.

A MSIA assays involves the immobilisation of antibodies onto a solid support for targeted protein enrichment from samples[96]. Following capture of the target, the protein is eluted for analysis by top down LC-MS or by SRM. A recent study has used the MSIA method to isolate MIF from human plasma samples[98]. Subsequent MALDI TOF MS analysis of both the intact protein and peptides from a digest was able to distinguish MIF isoforms and protein glycation demonstrating the benefits for low abundance protein identification. In addition, the high level of automation makes MSIA fairly reproducible and high throughput.

Targeted enrichment can also be carried out at a peptide level. SRM approaches analyse peptides to obtain quantitative values for protein abundances when coupled with quantification strategies and therefore the approaches pair well together. A peptide enrichment method, termed stable isotope standards and capture by anti-peptide antibodies (SISCAPA) was originally used in 2004 by Anderson et al to enrich for four specific plasma proteins[97]. A large enrichment was observed of the targeted peptides compared to the non-enriched peptides.

Reversed Phase Protein Arrays (RPPA), also known as lysate arrays, are an example where the specific binding affinity of antibodies have been exploited to explore the proteome. They differ from traditional protein arrays as instead of spotting individual proteins onto the array, tissues, cells or body fluids are immobilised onto the array slides[99]. The RPPA

allow a large number of samples from different patients or cell lines to be screened at once for a single antigen target with an individual affinity binder. The assay relies on high quality binders which require extensive antibody validation steps to ensure no cross reactivity[99]. However, most antibodies that are currently available do not pass the validation steps. Despite the limitations, RPPA have demonstrated clinical relevance and have been particularly informative for the analysis of cancer cells. A study in 2014 using RPPA showed that proteins related with cell survival and growth, cell adhesion and cellular matrix remodelling were upregulated in bone metastasis samples compared to healthy control samples[100]. However, a major limitation of these selective enrichment techniques is that they are highly reliant on antibodies which can be a problem if the target of interest does not have an associated binder.

1.5 Affinity Reagents

Affinity proteomics relies on the use of binders or affinity reagents that interact with target molecules. Whilst antibodies are currently the main affinity reagent used, there has recently been a large focus in developing antibody alternative affinity reagents to greatly increase the number of affinity tools available. An overview of the affinity reagents currently available will be discussed, including antibodies, aptamers and protein scaffolds. The timeline of key events in the development of affinity proteomic tools is displayed in Figure 1.2 and highlights the rapid development in recent years of antibody alternative binding regents. To conclude this overview, the final binder technology discussed will be Affimers, which are a key focus of this thesis.

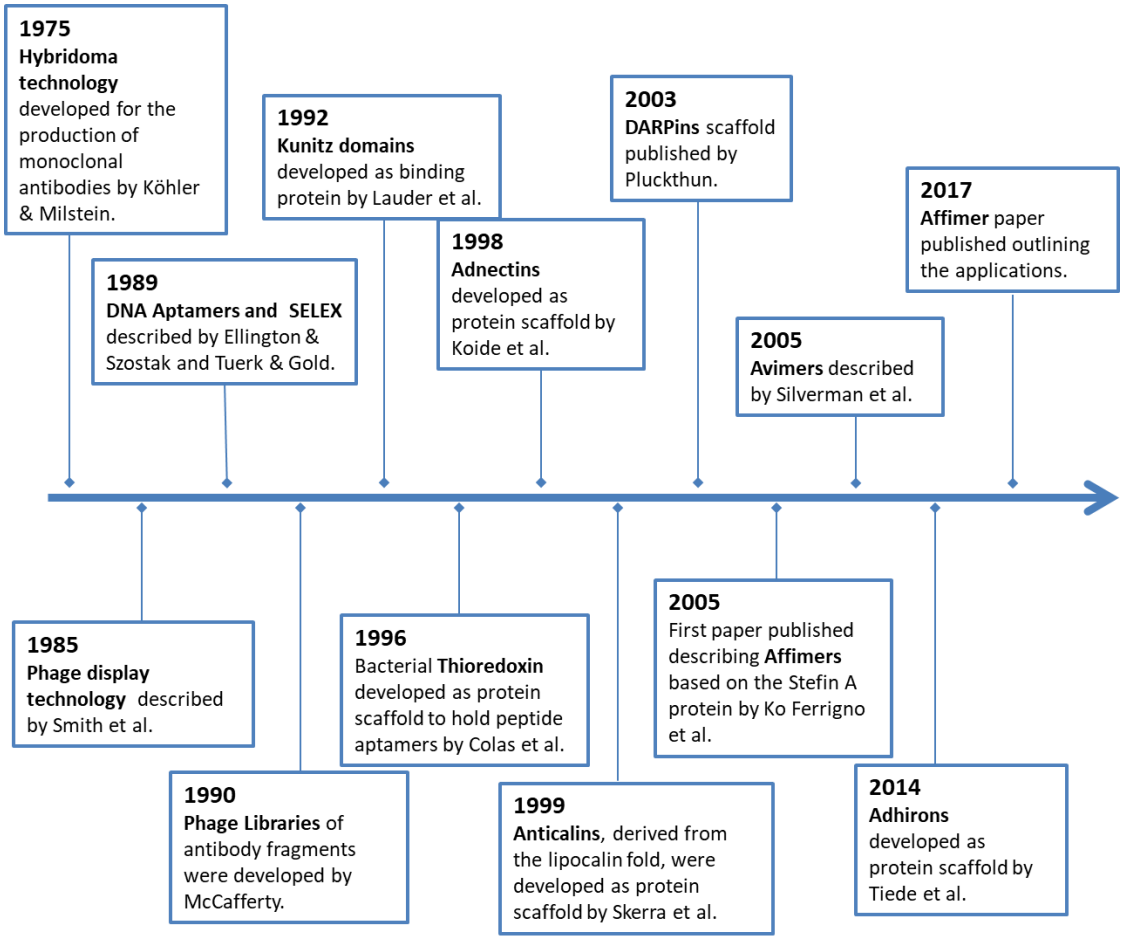


Figure 1.2 | Timeline of the development of affinity reagents.

Key events in affinity reagent development. Hybridoma technology[101], phage display technology[102], DNA aptamers[103,104], Kunitz domains, Thioredoxin[105], Adnectins[106], Anticalins[107], DARPins[108], Affimers (2005)[109], Avimers[110], Adhirons[111], Affimers (2017)[112].

1.5.1 Antibodies

Antibodies are the most widely used and studied affinity molecules. They are produced as a component of the adaptive immune system and have a vital role in defending vertebrates against pathogens and antigens that enter the body. The ability for antibodies to bind, with high specificity and high affinity, to the targets they were generated against are features that have been exploited in biological research allowing antibodies to be utilised in various immuno-affinity assays enabling the identification, characterisation and functional assignment of proteins.

Antibodies are highly conserved in mammalian species and can be divided into five major antibody classes; IgA, IgM, IgG, IgE and IgD. The distinction of antibody class is due to sequence differences in the constant region of the heavy chain of the antibody conferring different immunological responses and properties[113]. A single antibody structural unit is a 'Y-shaped' structure comprised of four polypeptide chains (Figure 1.3); two heavy chains (~ 55 kDa each) and two light chains (~ 25 kDa each), connected by disulfide bonds[114]. The 'V-shaped' portion of the antibody molecule is the area associated with antigen binding, known as the complementarity determining regions (CDRs). These regions are the most variable portion of an antibody molecule conferring diversity and specificity of binding partners. The CDRs are part of the 'fragment, antigen-binding' (Fab) region, which is made of a variable domain comprised of one heavy and one light chain and a constant domain also comprised of one heavy and one light chain. The lower portion of the antibody molecule is known as the 'fragment crystallisable' (Fc) region and is much more constant in its structure between the antibody sub-classes consisting of only constant domains[113].

Immunoglobulins have two key structural regions which can be attributed to different portions of the antibody structure; the constant region and the variable region. The constant region of the antibody is responsible for the effector function of the molecule whereas the variable region provides the antibody with the ability to bind an antigen through the Fab region. Variability in the V_H and V_L is located in three specific areas of the domains and are known as hypervariable regions. These regions are approximately located at residues 28 to 35, 49 to 59 and 92 to 103 in V_L and at similar positions in V_H[113]. The antigen-antibody binding interaction occurs through different non-covalent forces including electrostatic, hydrophobic and van der Waals forces and hydrogen bonding.

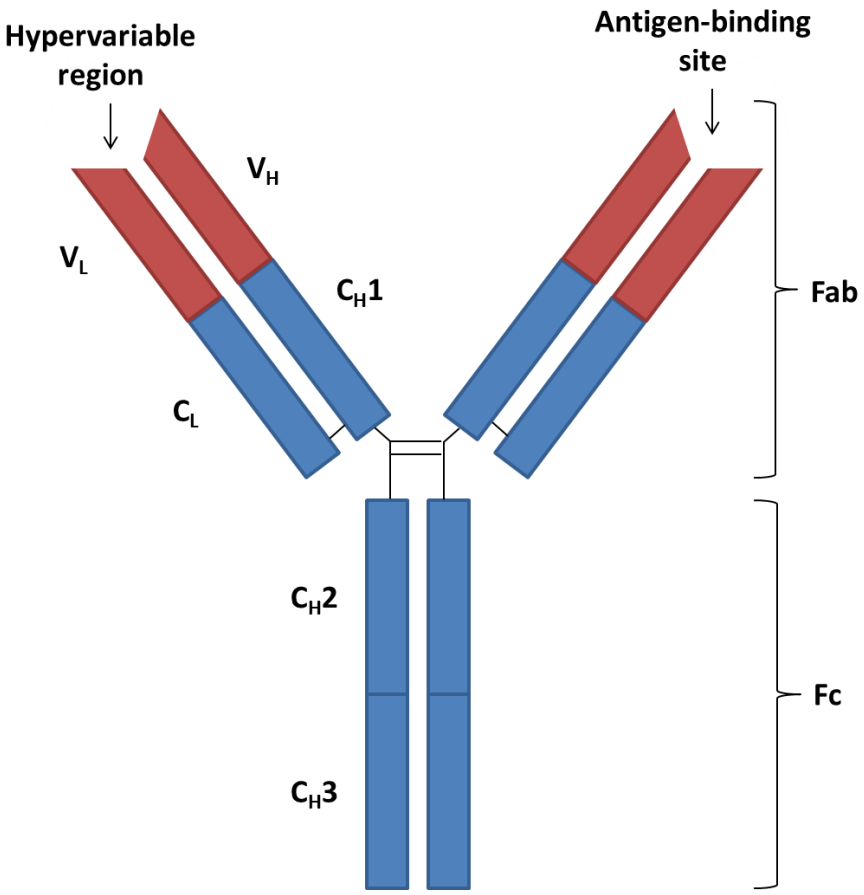


Figure 1.3 | Immunoglobulin G antibody structure
The IgG structure consists of two heavy chains and two lights chains. The antigen binding occurs in the variable regions of one heavy and one light chain. The constant domains are blue and the variable regions are red. The top portion of the antibody is the 'fragment, antigen-binding' (Fab) region and the lower portion is the 'fragment crystallisable' (Fc) region[115].

Polyclonal and Monoclonal Antibodies

There are two forms of antibodies used in biological research differing by the methods of production: monoclonal (MAbs) and polyclonal (PAbs) antibodies. In an immune response to an antigen, B-lymphocytes produce antibodies to target the antigen[116]. However, a single antigen will present numerous different epitopes due to their high complexity and as a consequence, a heterogeneous population of B-lymphocytes will produce antibodies that recognise a different epitope of the antigen. Therefore the serum will contain antibodies that recognise different epitopes of the same antigen, hence termed polyclonal antibodies[117]. To generate polyclonal antibodies for research applications, an animal is injected with an antigen which produces an immune response of polyclonal antibodies,

which are then purified from the serum. The serum may also contain antibodies against antigens previously encountered by the immunised animal. It may be advantageous to use polyclonal antibodies compared to monoclonal antibodies in certain assays, such as those that require the recognition of antigens in multiple orientations or where financial restrictions apply due to relatively low cost of production.

Conversely, monoclonal antibodies are derived from a single B-lymphocyte clone producing homogeneous antibodies with a single epitope target. Monoclonal antibody technology was first described in 1975 by Kohler and Milstein, for which they were awarded the Nobel Prize in 1984[101]. To generate monoclonal antibodies, lymphocytes are extracted from the spleen or lymph nodes of an animal immunised with a single antigen. Due to the short life span of B-lymphocytes and the inability to produce suitable quantities of antibodies through culture, the lymphocytes are fused with myeloma cells producing hybridomas. The hybridomas, an immortalised cell line, can now be grown in culture at a fast rate, producing an unlimited supply of antibodies with a single defined epitope target. As monoclonal antibodies recognise a single epitope, they exhibit exquisite specificity to their target and therefore typically demonstrate improved performance in immunoassays compared to polyclonal antibodies, generally producing less background and cross-reactivity with other proteins[118].

The first application of antibodies in life sciences as a research tool was in the 1940s with the development of a tube immunodiffusion technique used to measure the antigen-antibody interactions and later a plate double diffusion technique[119]. In 1959 the antigen-antibody interaction was exploited again with the development of the radioimmunoassay (RIA) for insulin[120]. The RIA method was the predominant technique used to measure proteins. Another advancement in the applications of antibodies was with the development of the enzyme linked immunosorbent assay (ELISA)[121,122]. The technique detects antigen-antibody binding with a colour change induced by the enzyme. ELISA is still the current and standard assay for protein detection due to its high sensitivity. Since the development of monoclonal antibody technology[101], the application and routine use of antibodies has become widespread, transforming the fields of biotechnology and life sciences. Antibody technology has enabled huge advances in many areas including basic research, understanding the biological mechanism of diseases, and diagnostic assays. A key example of the success and revolutionary impact of antibody technology is in the development of antibody-based therapeutics. To date, numerous MAb drugs have been developed and

approved for the treatment of 37 diseases, predominantly cancer, becoming a key focus for pharmaceutical drug development[123].

Whilst there is no dispute that antibodies are an invaluable tool for biological research, they do possess some limitations due to their structural properties. Firstly, high throughput antibody production can be difficult to achieve in bacterial expression systems due to extensive glycosylation and disulfide bond formation, essential to preserve the stability and function of the antibody. Secondly, the production of polyclonal antibodies has ethical implications as host animals are required. Thirdly, this process can be expensive and time consuming. Finally, in clinical research it is often beneficial to explore different protein conformations, isoforms and mutated proteins. Despite the vast numbers of antibodies used in biological research, there are still cases where antibodies are not available for this subgroup of proteins or where current antibodies are of poor quality.

Furthermore, a key challenge of antibodies is the lack of reproducibility between experiments and low specificity for their intended target due to poor validation and high variability[124]. Numerous initiatives have been suggested in order to ensure that published antibody data is both reliable and reproducible for example, publishing the full western blot gels and including all appropriate controls[125-127]. Therefore, the development of alternative affinity reagents is necessary to overcome the current limitations of antibodies, providing complementary reagents to those currently available.

Recombinant Antibodies

Recombinant antibodies were first described in 1990 as a development of phage display technology when it was discovered that the variable domains of antibodies could be expressed on the surface of bacteriophages[128]. A brief outline of phage display technology is described in section 1.5.4. The technique is very similar for the production of recombinant antibodies however, the initial step involves the creation of an antibody gene library. The variable regions of recombinant antibodies are either from genetic material from human B-cells or are synthesised *in vitro*. Human B-cell libraries are either immune libraries generated from an immunised donor species or naïve libraries, generated from a non-immunised human[129]. Diversity in the libraries is reported to be higher in the naïve libraries (10^{11} clones) compared to the immune libraries (10^{10} clones) [130]. As recombinant antibodies can be synthetically engineered, they do not require an immune response and therefore have the potential for development to most protein targets. Recombinant antibodies typically have extremely high binding affinities in the picomolar to femtomolar

range to their targets which can sometimes be greater than the affinities of conventional antibodies[131]. Due to the disulfide bonds between the heavy and light chains of antibodies, high expression in bacterial systems is difficult to achieve. Therefore, recombinant antibodies are usually based on smaller antibody fragments[132], a feature which also assists phage display technology and also increases their versatility for different applications[129].

1.5.2 Nanobodies

Another class of immunoaffinity reagents used in biological research are Nanobodies. As previously discussed, antibodies are highly conserved across mammalian species, however in addition to conventional antibodies, *Camelidae* species (camels and llamas) also have heavy chain antibodies (HCAbs) that are devoid of the light chains and the first constant domain in their sera[133]. Similarly, shark species have two typical antibodies, IgW and IgM, as well as simple antibodies known as Ig new antigen receptor (IgNAR), that lack the light chain[134]. The HCAbs have a variable region known as the V_HH (nanobody), comparable to the Fab region of antibodies, which confer the antigen binding properties of the HCAbs. The V_HH regions have been isolated and produced by recombinant technology and expressed, primarily in *E.coli* for use as affinity reagents[135].

Compared to conventional antibodies, Nanobodies are extremely small (13 – 14 kDa), a characteristic that provides them with many beneficial features. Studies on the binding mechanism of Nanobodies to their respective targets have demonstrated that Nanobodies can recognise less common epitopes such as the active site of enzymes due to their small size easily accessing 'hidden' sites[136,137]. The binding to cleft regions of proteins is typically less common with conventional antibodies as it is the antibodies that typically have cavities, clefts and flat surfaces for target interactions[138]. Additionally, research has also shown that certain nanobodies retain their recognition function after incubation at high temperatures up to 90 ^oC[139]. Despite these benefits, some of the issues relating to conventional antibodies also apply to Nanobodies, reinforcing the need for alternative non-immunoglobulin affinity reagents.

1.5.3 Aptamers

DNA or RNA Aptamers are short single stranded oligonucleotides, typically 20 – 100 bases, which fold to form tertiary structures capable of binding to target molecules with high specificity and high affinity[140]. The name 'Apatmer' is derived from the Latin word aptus, meaning 'to fit' and was first used in 1990 by Szostak *et a*l, where they discovered that RNA

molecules of random sequences can fold to form structures with binding sites for small organic dyes [104]. Simultaneously reported by another group in 1990, was the development of a selection method to identify RNA molecules that bind to bacteriophage T4 DNA polymerase, termed systematic evolution of ligands by exponential enrichment (SELEX)[103]. Aptamers have comparable binding affinities to antibodies and interact with their targets through hydrogen bonding, electrostatic interactions, salt bridges and van der Waals forces[141]. Aptamers have been developed to a wide variety of different targets including human VEGF[142]. Based on the different properties of DNA and RNA, selection for apatmer production may vary; RNA has more flexibility in its backbone compared to DNA however RNA is more susceptible to degradation compared to DNA. The process of SELEX involves incubating a library of randomised DNA sequences with an immobilised target. DNA that does not bind is washed away whilst the bound DNA is eluted and amplified by PCR. The selection process is then repeated, usually between eight and 15 rounds. The DNA is sequenced which usually reveals a pool of different DNA sequences however with a similar protein fold[143].

One key advantage of aptamers is the speed and relative ease of production compared to antibody production. Large quantities of aptamers can be quickly generated and as aptamer production is a chemical synthesis process, batch-to-batch variability is extremely low[143]. Furthermore, production does not rely on the use of animals eliminating the ethical issues of animal use in research. The small size of aptamers means that they may be able to access different epitopes or portions of their protein target, which would otherwise remain inaccessible for antibodies. Additionally, the production of an aptamer does not rely on the generation of an immune response and therefore, an aptamer can theoretically be developed for any molecular target. Aptamers can readily accommodate modifications such as the addition of functional groups and fluorescent dyes usually without altering the binding affinity of the molecules[141].

Despite the numerous benefits of Aptamers compared to antibodies as affinity reagents for use in biological research, to date, the implementation of Aptamers as research and clinical tools is extremely limited. Antibodies have been used as affinity reagents in research laboratories for numerous years and are therefore considered the gold-standard due to familiarity with the products and well-defined, standardised protocols are in place. Substituting antibodies for Aptamers in conventional immunoassays requires extensive, time-consuming method development to generate comparable and robust standardised protocols. In addition, a key reasons in the lack of widespread use of aptamer technology is

that an extremely broad patent portfolio exists, consisting of over 140 individual patents, with few companies having the appropriate licenses to develop the technology[144].

1.5.4 Peptide Aptamers

Peptide aptamers are typically 5 – 20 amino acids long, constrained within a protein scaffold and were first described by Brent et al. in 1996[145]. The work advanced on a previous study on the *E.coli* protein, thioredoxin, where the loop region of the protein was identified as a site for combinatorial peptide insertion[146]. Brent *et al* reported the development of thioredoxin as a scaffold protein to display combinatorial peptides of 20 residues, in a loop region of the protein[145]. Using the two-hybrid system, the combinatorial library of thioredoxin proteins that bound to human Cdk2 were selected and were shown to bind to different sites of the Cdk2 protein. A protein scaffold is the term used to describe the protein structure that holds the peptide fragments that enable binding to specific targets[147]. The flexible nature of peptide fragments means that when inserted into scaffold proteins they are limited in the number of conformations they can adopt and hence display greater specificity and affinity to their binding target compared to unconstrained peptide fragments[148]. Additionally, peptides constrained within a scaffold are less susceptible to degradation from cellular proteases, increasing their stability compared to free peptides[148].

Selection Techniques

To produce affinity reagents with randomised variable regions with affinities to different targets, selection of the mutated variants is necessary. Display technology enables this allowing peptide and proteins to be selected that bind to target proteins with high affinities. The most frequently used display technique is phage display and is the method used to produce the Affimers in this thesis with defined protein targets. Other display technology available includes ribosome display, mRNA display and bacterial and yeast cell-surface display. Phage display is a method used to screen libraries of proteins for binding to specific targets[102]. The process involves the use of a bacteriophage, which expresses the library of peptides or proteins on the surface of the phage. The phage selection process, termed bio-panning, consists of five main steps; the preparation of a phage peptide library; immobilisation of the target, usually onto a micotitre plate by passive adsorption; exposure of the immobilised target to the phage display library for binding; washing to remove unbound phage; and elution of the target bound phage. Following the final elution step, the recovered phages are amplified and the bio-panning is repeated between three and six times to enrich for high affinity binders[149]. Validation of the generated phages is essential to

confirm high selectivity for the target they were generated against which is typically performed by ELISA. A defining requirement of the method is that recombinant or purified endogenous protein is available for the bio-panning process.

1.5.5 Alternative Scaffolds

Considerable research has been undertaken to develop non-antibody based scaffold proteins[150,151]. The structure of antibodies, consisting of a constant region and variable interaction sites for target binding, lead to the concept of developing alternative affinity reagents based on this design[152]. Under natural conditions, the primary roles of a large number of the scaffold proteins that have been developed are to form binding interactions with other proteins. Therefore, the pre-existing binding sites of the protein scaffolds can be modified allowing peptide insertions to be introduced at these positions forming a region similar to an antibody epitope. Proteins selected for scaffold development must possess numerous beneficial features in order to be a superior affinity reagent to antibodies. The protein scaffold must be structurally tolerant to modifications such as the addition of variable epitope regions of increased lengths, purification tags and functional groups for immobilisation. The size of the scaffold should be relatively small to encourage cellular uptake and to limit non-specific binding[151]. Regardless of selection and production methods, the protein scaffold should express at high quantities and display low batch-to-batch variability. When designing the protein scaffold, the final application should be considered, as if the scaffold is intended for medical and human in vivo applications, the species of origin of the scaffold should be carefully selected so that it does not evoke an immune response.

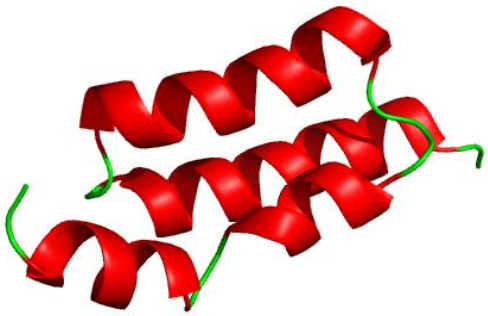
Currently, there are over 50 synthetic engineered protein scaffolds that allow specific peptide fragments to be presented as recognition sites[150,151]. Table 1.1 provides an overview of some of the scaffold proteins currently available. The protein scaffolds described are diverse in terms of the size, structure, interaction sites and protein of origin. Typically, protein scaffolds are categorised based on the structure; scaffolds with structures that mimic the structure of antibodies with a rigid frame and variable loop regions or scaffolds with rigid structures with variable regions in the flat surface of the scaffold[151]. Several protein scaffolds most frequently developed and their applications are discussed further.

Table 1.1 | Examples of protein scaffolds developed as novel affinity reagents

Abbreviations: BBP, bilin-binding protein; ApoD, apolipoprotein D; FABP, bovine heart fatty-acid binding protein; BPTI, beef pancreas trypsin inhibitor; LAC-D1, lipoprotein-associated coagulation inhibitor; ITI-D", human inter-α-trypsin inhibitor; APPI, Alzheimer's amyloid-β protein precursor inhibitor domain

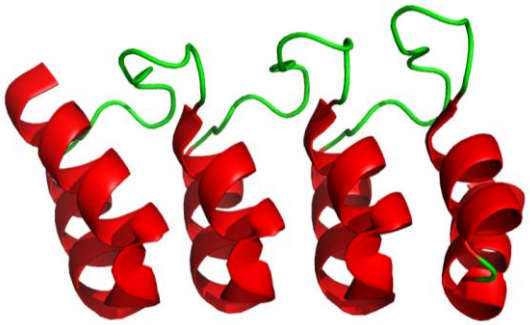
Scaffold	Protein	Molecular Weight (kDa)	Structure	No. of Residues	Randomised Elements	Company/ Reference
Affimer	Stefin A, plant cystatin consensus sequence	11 – 13 kDa	α-helix and β-sheet	98	2 loops	Avacta Life Sciences[111,153]
DARPins	Ankyrin repeats	14 -18 kDa	2 x α-helix and β-turn	1 unit – 33	3 x (α/β turn)	Molecular Partners[108]
Anticalins	Lipocalins (BBP, ApoD, FABP)	Approx. 20 kDa	β-barrel	160 - 180	4 loops	Pieris Pharmaceuticals[154]
Adnectins	Fibronectin III	Approx. 10 kDa	β-sandwich	94	2 – 3 loops	Adnexus[155]
Affibodies	Z domain of Protein A	Approx. 6 kDa	3 α-helices	58	2 x α-helix	Affibody[156,157]
Kunitz domain	BPTI, LACI-D1, ITI-D2, APPI	Approx. 7 kDa	α-helix and β-sheet	58	1 or 2 loops	DYAX[158]
Affilin	Ubiquitin/ γ-B-crystallin	Approx. 10/20 kDa	α-helix and β-sheet/ β-sheet	76 -176	β-sheet	Navigo Proteins[159]
Avimer	A domain	4 kDa	Four loops	1 unit - 43	4 loops	Amgen[110]

Affibodies



Affibodies are modified protein scaffolds based on the Z domain of the immunoglobin Fc-binding region of Protein A from *Staphylococcus aureues*[160]. The engineered proteins are made from a single polypeptide chain of 58 amino acids that folds and is highly stable despite lacking cysteine residues. Combinatorial library construction was first described in 1995 by Nord et al.[156] in which 13 residues in two of the three alpha helical regions were randomised. Affibodies are extremely small (approximately 6 kDa) and have been used in various applications including the targeting of viruses[161], affinity chromatography of human apolipoprotein A-1 and *Taq* DNA polymerase[156,162], proteome depletion[163] and imaging of tumours. A labelled Affibody designed to target epidermal growth factor receptor 2 (HER2) has been used to visualise tumours in patients with metastatic breast cancer[164,165].

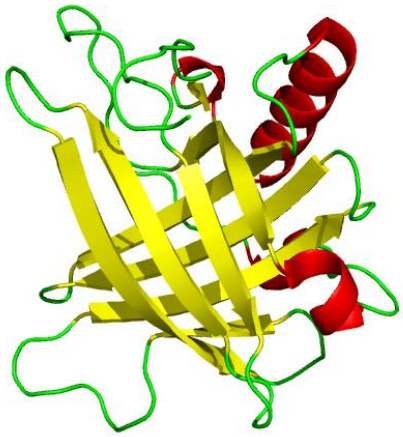
DARPins



Designed ankyrin-repeat proteins (DARPins) are based on a consensus sequence of the naturally occurring protein group, ankyrin repeat proteins. The natural biological function of ankyrin-repeat proteins in mediating protein-protein interactions[166] has been exploited for the development of engineered binding proteins. The proteins are constructed from repeating units, with a single unit typically made from 33 amino acid residues comprised of a β-turn and two anti-parallel α-helices. Typically, DARPins are made of four to six structural units but can consist of up to 29 repeat units, forming proteins with a molecular weight of 14 – 22 kDa. Early studies developed combinatorial libraries of DARPins with seven randomised positions per repeat unit[108]. Using ribosome display, DARPins for maltose binding protein were selected and had low nanomolar binding affinities[167]. Since then DARPins have been developed to various different targets for different applications including imaging[168] and as therapeutics[169]. The DARPin scaffold was developed further to generate a DARPin structure with loops to potentially increase the accessibility to binding sites on targets. After one

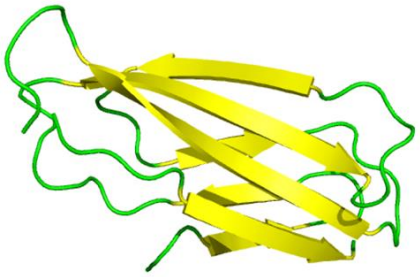
round of library screening with the Loop DARPins, binders to BCL-2 were generated with low picomolar affinities[170].

Anticalins



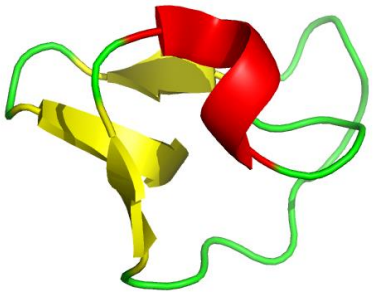
Anticalins are engineered lipocalins that have been modified to bind specific targets[171]. Lipocalins are small proteins, typically 160 -180 amino acids and are found in numerous species including plants, humans and bacteria and have a highly conserved structure across the different species despite having low sequence homology[172]. The proteins consist of a β-barrel core made up of eight antiparallel β-strands that are connected with four loops. It is this region of the lipocalin protein that is involved in interactions with their targets and thus the loop regions that are modified in development of anticalins. The typical role of lipocalins is involvement in the transportation and storage of small molecules such as steroids, lipids and odorants. The molecules bind tightly to the deep pocket of the β-barrel core of the lipocalins, a feature that can be exploited in Anticalin design enabling the binders to target small molecules for the development of antidotes[173]. Furthermore, Anticalins are highly stable withstanding temperatures of up to 70 ^{o}C, they do not contain disulfide bonds and are not glycosylated. The lipocalin initially developed as a protein scaffold was bilin-binding protein (BBP) from *Pieris brassicae*. In four of the loops, a total of 16 amino acid residues were mutated and demonstrated high binding affinities (low nanomolar range) with the plant steroid digoxigenin and the dye florescin. However, as a protein scaffold of human origin would be more preferable for medical applications, the human lipocalins apolipoprotein D, human lipocalins 1 and human lipocalins 2 (Lcn2) were identified for scaffold development[172]. Anticalins have been used in a variety of applications and have been developed against clinically relevant targets. The Lcn2 scaffold has been used to develop Anticalins targeting cytotoxic T-lympohcyte associated antigen 4[174] and vascular endothelial growth factor[175] with nanomolar binding affinities.

Adnectins



Adnectins, also termed Monobodies, are engineered protein scaffolds based on the 10th extracellular domain of human fibronectin III (^{10}Fn3) first described in 1998 by the Koide lab[106]. The protein is ideal for development as a protein scaffold as the structure of ^{10}Fn3 is similar to the variable domain of antibodies possessing three loops for target binding. ^{10}Fn3 lacks disulfide bonds but still able to maintain a highly stable structure. Studies have demonstrated that the protein is able to maintain its structure at high temperatures (up to 80 ^{o}C)[176] and express at high levels in bacterial expression systems[177]. Initial adnectin library design and selection was carried out using phage display for proteins that bound ubiquitin[106] and mRNA display to select for Adnectin binders of TNF-alpha[178]. The generated Adnectins had low micromolar[106] and low nanomolar[178] affinities to their targets respectively. More recently, various Adnectins have been generated that bind clinically relevant targets. For example, Adnectins that target epidermal growth factor receptor (EGFR) and interleukin 23 (IL-23) were produced using mRNA display[179]. EGFR and IL-23 are implicated in cancer and inflammatory diseases respectively, and the corresponding Adnectins have been shown to block intracellular signalling in cell-based assays and are potentially suitable as antibody alternatives when developing therapeutics.

Knottins



Cysteine-knot miniproteins, also termed knottins are a group of small proteins, comprised of 25 to 35 amino acids residues. The protein group is characteristically defined by having a minimum of three disulfide bonds[180]. The arrangement of the disulfide bonds, with one disulfide bond between cysteine 3 and 6 going through the disulfide bonds of cysteine 1 and 3 and cysteine 2 and 4, forming a highly stable 'knotted' structure of antiparallel β-strands[180]. This feature makes knottins ideal candidates for affinity scaffold development. The connecting loops between the antiparallel β-strands of naturally occurring knottins are highly diverse with over 2000 different sequences observed of various loop lengths indicating the loop regions are amenable to engineering and

randomisation[181,182]. Additionally, the small size and high stability of the protein makes them suitable for tumour imaging as they demonstrate rapid uptake and clearance from tissues[183]. The protein, trypsin inhibitor II from *Ecballium elaterium* (EETI-II) was initially developed as a knottin protein scaffold by determining the knottin protein tolerated mutations into the loop regions of the protein structure[184-186]. Other proteins belonging to the knottin family developed as scaffold include the C-terminal cellulose-binding domain from cellobinohydrolase I of the fungus *Trichoderma ressei*[187,188] and the human protein, Agouti-related protein[189]. Further, to generate Knottin binders that target integrins, known binding motifs were incorporated into the loops of EETI-II resulting in binders with nanomolar affinities to integrin proteins[183].

1.6 Affimers

Affimers are small engineered proteins, based on either one of two proteins; the human protein stefin A or a consensus sequence of plant phytocystatins. The development of Affimer technology was first reported in 2005 when Ko Ferrigno *et al* published work detailing the initial stages of engineering the human protein, stefin A to constrain peptide aptamers[109]. The work successfully demonstrated that the stefin A variant possessed the desired features of a scaffold protein; it was made biologically neutral by removing the natural cysteine protease inhibitor function of the protein[153], remained folded and had high thermal stability even whilst displaying a peptide insertion and that designed peptides constrained within the scaffold interacted with known targets. Additionally, it was found that the protein scaffold expressed well in yeast, mammalian and bacterial cells.

Further advancements to the protein scaffold were made and work published in 2010 described the development of five new stefin A protein scaffolds based on the STM (stefin A triple mutant) design[190]. Different modifications were made to the loop regions of the scaffold and the structure, stability, folding and expression of the proteins assessed. A variant, termed SQM, in which epitope tags were inserted into the N-terminus of the protein led to a lower yield of protein expression compared to the other variants described. However, one major limitation of the SQM scaffold was that when insertions were made into loop 1 and the N-terminus, as well as in loop 2, the stability of the scaffold was compromised; a feature that is not desirable for applications in large combinatorial libraries.

Therefore, further improvements were made to the scaffold and the final development of the Stefin A scaffold was reported in 2011[153]. The study reports the design of a new Stefin A variant called, SQT (Stefin A Quadruple Mutant – Tracy) that is highly stable and can withstand the addition of three peptide insertions, giving the scaffold enhanced diversity in combinatorial libraries. The SQM scaffold variant was re-engineered to form the SQT scaffold by changing the position of a restriction site to the beginning of loop 2 instead of the β-sheet resulting in two amino acid substitutions; E78A and L80R. Analysis of the proteins by CD spectroscopy to determine the thermal stability discovered that SQT had comparable stability to STM. The SQT scaffold is thus the Affimer scaffold (Figure 1.4).

In 2014, another group published their work describing the development of Adhirons (Figure 1.4), now commercially termed Affimers, as a protein scaffold[111]. The Adhiron protein scaffold is based on a consensus sequence of plant phytocystatins, which are small, natural cysteine protease inhibitors, of approximately 100 amino acid residues. The protein was selected for development due to inherent features desirable for an ideal protein scaffold; small, highly stable, contains no disulfide bonds or glycosylation sites. The natural interaction regions of the protein that convey the inhibitory function of the protein were modified for peptide insertion leading to the insertion of two, randomised nine amino acid loop regions. The group demonstrated that a panel of Adhirons, generated through phage display technology for the yeast SUMO protein, could be identified to bind to the yeast SUMO protein through ELISA studies. The binding affinities were measured by isothermal titration calorimetry and the KD values for three of the Adhirons were in the low nanomolar range comparable to other antibody alternative affinity reagents[111].

Despite having low sequence homology between the two Affimer protein scaffolds, structurally, they are very similar in structure (Figure 1.4). A key advantage of developing Affimers from two different species is that depending on the final application of the protein scaffold; one of the proteins may be more beneficial compared to the other. For example, a protein scaffold that is modified from a human protein would be better suited for therapeutic applications as it is less likely to induce an immune response compared to a plant protein. The Affimers based on the phytocystatin consensus sequence are ideal for use as research reagents and as diagnostic tools as they display enhanced affinity to their targets.

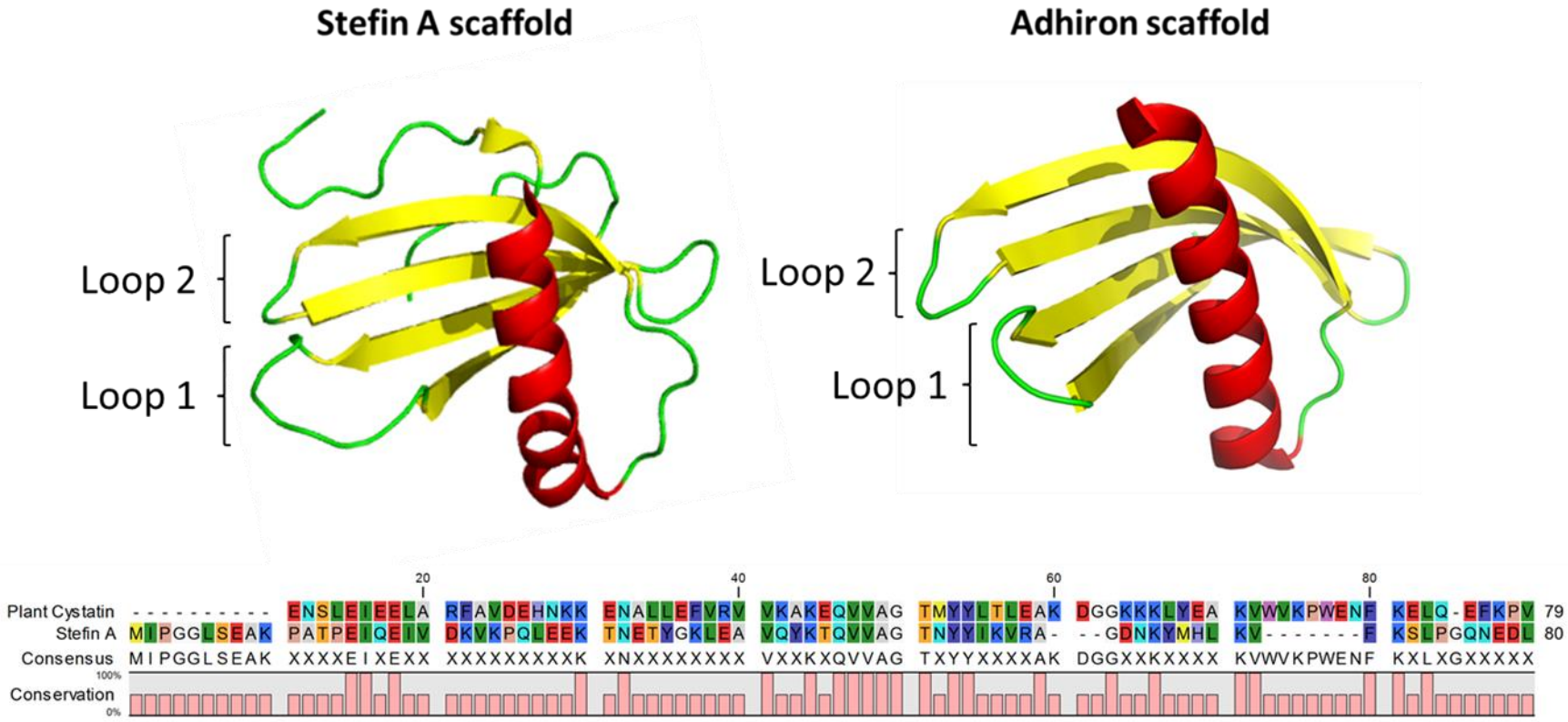


Figure 1.4 | Ribbon diagram and sequence alignment of the two different Affimer structures.

Ribbon diagrams of the two different Affimer scaffolds. The stefin A scaffold was derived from the Protein Data Bank (PDB) 4N6T and the Adhiron scaffold was derived from PDB 1NB5. Images were generated using PyMol[191].

Affimers have comparable binding affinity to antibodies, in the low nanomolar region and exhibit high stability to various pH ranges and temperatures. They also have the ability to bind to a solid surface and retain their recognition functions, proving very useful for Affimer microarray development [192]. Due to the relatively small size of Affimers (11 -13 kDa) chemical modifications with dyes and protein fusions are easily implemented. To date, Affimers have shown their versatility with various cellular and molecular applications. They have been utilised for cell imaging, affinity histochemistry, western blotting, MRI reagents and for the formation of nanoparticles[112].

1.7 Affimers in the discovery of biomarkers

Biomarkers are defined as molecules that can be detected and measured to define a specific biological state, pathology or response to therapy[193]. Biomarkers are extremely important in the clinical setting as they play a role in disease prediction, diagnosis, prognosis and management. In 1948, the very first biomarker was discovered for the detection of myeloma[194]. The light chain of immunoglobulins was identified in 70 % of patients with the disease. Since then, many other proteins have been established as biomarkers however they are not used alone; other diagnostic tests are performed. This is predominantly due to the lack of specificity of the biomarker. To be useful in the clinical setting, biomarkers must meet certain criteria. This includes involving minimally invasive sample collection methods, high specificity, low cost of testing and show significant differences between control and disease samples[193]. In addition, for a biomarker to have clinical applications, extensive validation is necessary[193]. This includes both analytical and clinical validation and assessment of the clinical utility of the biomarker in improving patient health outcomes and being economically viable.

Due to the unbiased nature of discovery LC-MS/MS, it is an ideal method for biomarker discovery. However, the method does have limitations in that it lacks the sensitivity needed for complex proteome analysis. Although biofluids such as plasma and serum are ideal samples for biomarker discovery due to their relative ease of collection and the fact they sample the pathophysiology of the whole body, they have dynamic ranges spanning 10 to 12 orders of magnitude[19]. In addition, although a large number of proteins can be identified in a single MS analysis, sample throughput is low especially if long LC gradients are employed. Examples of the other methods available for complex proteome analysis were discussed particularly, 2-DE and microarray technology.

The use of Affimer regents for biomarker discovery has great potential to enrich for specific protein targets and overcoming proteome complexity. A proof of concept study using Affimers containing randomised peptide insertions in their loop regions, termed naïve Affimers, was employed for unbiased sample analysis[192,195]. The method involved the immobilisation of Affimers onto the array followed by incubation of a fluorescently labelled sample. Affimers that bound a protein target generated a fluorescence signal and therefore was identified as displaying affinity for proteins. This strategy can be employed for the differential analysis of samples for biomarker discovery. As each array can contain up to 20000 different Affimer spots, by comparing the fluorescence signals on the Affimer array at a set position a comparison can be made between samples regarding protein capture, allowing large scale biomarker discovery screening. The Affimer are naïve and thus contain unknown peptide insertions in the loop regions, so the method is a untargeted method for biomarker analysis. In addition, as the Affimers could enrich for target, low abundance proteins may be brought into the detectable range for identification. The use of bottom-up mass spectrometry would be an ideal method for target protein identification.

1.8 Aims and Objectives

The overall aim of this thesis was to investigate and characterise next generation affinity reagents, with a focus on Affimer technology. The primary aim of the first chapter was to develop an affinity purification workflow coupled with mass spectrometry to identify unknown naïve Affimer binding targets. Naïve Affimers were selected from Affimer array data that showed differential binding between hospital control and sepsis serum and it was hypothesised that identification of proteins bound to the Affimers could reveal protein markers of sepsis[192]. Various methods for affinity purification were investigated including His-tag Affimer immobilisation, covalent Affimer immobilisation and mass spectrometric immunoassay (MSIA). The challenges and identification of non-specific background proteins in affinity purification workflows is also presented.

The focus of the second chapter was to characterise Affimers with known binding targets. Pepsin has been proposed as a protein marker to distinguish patients with reflux aspiration from direct aspiration of food and saliva, allowing clinicians to guide surgical intervention[196]. Current assays rely on antibodies as the detection reagent however they lack any clinical utility due to inherent problems with antibodies previously discussed. Therefore the aims were to generate Adhirons to target human pepsinogen and develop an affinity purification workflow for pepsinogen enrichment. To overcome the issue of the large signal of capture reagents in LC-MS analysis of affinity purifications, a pepsinogen Adhiron resistant to Lys-C proteolysis was also expressed and stability analysed.

The third study in this thesis explored an alternative approach to identify protein markers of sepsis due to the limitations in the Affimer array and affinity purification approach. Using LC-MS/MS, the aim was to undertake a comparative analysis of plasma from sepsis patients and hospital controls to identify changes in protein expression. Results from the comparative analysis will guide the selection of candidate proteins for phage display Affimer production as potential markers of sepsis.

The aims of the final chapter were to perform preliminary investigations into whether darcin was suitable for development as an alternative affinity reagents scaffold. Previously, lipocalins structures have been reported as alternative scaffolds[171,172]. Darcin, a major urinary protein (MUP), found in rodents is an example of a lipocalin and structural analysis revealed it has an extremely robust and stable structure, resisting denaturation at 8 M

urea[197]. The aims were to explore various modifications to the protein and determine the effect on protein stability.

The terms Affimer and Adhiron are used throughout this thesis. Affimer is the commercial name of both the stefin A scaffold and plant cystatin consensus sequence scaffold. The term Adhiron specifically refers to those binders that are based on the plant cystatin consensus sequence. In this thesis, various different Affimers were used of which can be grouped into three separate categories based on their intended purpose. For the Affimer array work, naïve Affimers based on the Stefin A protein scaffold were used. These Affimers consisted of random peptides within the variable loop region of the scaffold and their binding partners was unknown. The second set of Affimers had known binder partners and were used for method development and optimisation. The Affimer protein scaffold for this set was based on the plant cystatin consensus sequence. The third set of Affimers were generated to specific target to overcome a distinct biological problem. To address the need for alternative affinity reagents that target pepsinogen, a total of five pepsinogen Adhirons were generated.

Workflow

The typical approach for Affimer characterisation involved a variety of analytical techniques. Affimers were either received as pure protein from Avacta Life Sciences or as plasmid DNA from the University of Leeds BioScreening Technology Group for *E.coli* expression. Firstly, SDS-PAGE analysis was performed to confirm the Affimers were approximately the expected molecular weight and that the protein sample contained no contaminant proteins following purification.

To confirm the molecular weight was as expected, intact mass analysis of the proteins was carried out using ESI-MS on the Waters Synapt G2. Electro-spray ionisation (ESI) is the most commonly used ionisation technique in the field of proteomics and the predominant method used within this study. ESI is a soft ionisation technique and is the process of transferring charge particles from the liquid to gas phase before MS analysis. The development of ESI for protein analysis was first published by Fenn in 1984 and when coupled to a mass spectrometer, revolutionised the field of proteomics[198]. The technique was the first to create multiply charged proteins, bringing them into the mass range of the mass spectrometer for detection. The process of ESI involves three main steps; the production of nebulised droplets at the tip of the emitter, repeated evaporation of solvent to form smaller charged droplets and finally, the formation of gas phase ions[198].

Further confirmation of the Affimer sequence was achieved by peptide mapping and sequencing. Using either a single protease or a combination to achieve improved sequence coverage, the Affimers were digested to their resulting peptides. The peptides were analysed by LC-MS/MS on the Thermo Scientific QExactive Orbitrap mass spectrometer to obtain sequence information. The instrument has high resolution and high mass accuracy allowing the accurate assignment of MS/MS spectra. Following characterisation and confirmation of the correct protein product, Affimers were suitable for use in downstream applications of affinity purification method development.

Chapter 2: Materials and Methods

2.1 Protein expression

E.coli cells (BL21 (λDE3) strain) were made competent by calcium chloride and the cells heat shocked to allow the DNA to enter the cells. A total of 5 ng of plasmid, solubilised in TE buffer (10 mM Tris, 1 mM ethylenediaminetetraacetic acid (EDTA)) were incubated with competent cells and heat shocked at 42 ^{o}C. Cells were centrifuged (1600 x g), resuspended in LB and grown on agar plates containing antibiotic (dependent on plasmid used, further details in results sections). Plates were incubated overnight at 37 ^{o}C for selective growth of transformed cells only. An individual colony was selected to inoculate LB broth (10 mL) containing antibiotic and incubated overnight at 37 ^{o}C with mixing at 180 rpm. For larger-scale protein expression, the overnight broth (4 mL) was added to LB broth (200 mL) containing antibiotic and incubated at 37 ^{o}C with mixing at 180 rpm. Protein expression was induced with isopropyl β-D-1-thiogalactopyranoside (1 mM) at an absorbance reading of 0.6 at OD (600 nm). The protein was left to accumulate for either 4 hours or 16 hours, as specified in the results. The cells were harvested by centrifuged at 3500 rpm at 4 ^{o}C for 10 minutes. The LB broth removed and the protein pellets stored in the freezer until required.

2.2 Cell lysis using sonication

A cell pellet from 50 mL of culture was re-suspended in 50 mM phosphate buffer pH 8.0 plus protease inhibitors and benzonase nuclease (25mL buffer made with 1 x complete EDTA-free protease inhibitor tablet and 0.1 % (v/v) benzonase nuclease). The suspension was sonicated at 30 % amplitude for 10 seconds every minute for a total of 12 minutes. To obtain the soluble fraction, the sonicated solution was centrifuged (6000 g for 8 minutes).

2.3 Purification with Ni-NTA HisTrap column method

Manual

The pepsinogen Adhirons, present in the soluble fraction of the bacterial cell lysate, were purified using Ni-NTA affinity columns (HisTrap HP, G.E. Healthcare, Buckinghamshire, UK) by use of the His-tag following the manufacturer's guidelines. After loading the samples, columns were washed with wash buffer (0.2 M sodium phosphate, 0.5 M sodium chloride, 10 mM imidazole, 20 % glycerol, pH 7.8) and the protein was eluted from the column using elution buffer (0.2 M sodium phosphate, 0.5 M sodium chloride, 0.5 M imidazole, 20 %

glycerol, pH 7.8). Eluted fractions were analysed by SDS-PAGE with the fractions with the highest concentration of protein combined and the Adhirons stored in elution buffer.

ÄKTA™ start purification system

Darcin and the darcin variants, present in the soluble fraction of the bacterial cell lysate, were purified using Ni-NTA affinity columns (HisTrap HP, G.E. Healthcare, Amersham, UK) using the ÄKTA™ start purification system (G.E. Healthcare, Buckinghamshire, UK). After loading the samples, columns were washed with wash buffer (0.2 M sodium phosphate, 0.5 M sodium chloride, 10 mM imidazole, pH 8.0) and the protein was eluted from the column using elution buffer (0.2 M sodium phosphate, 0.5 M sodium chloride, 0.5 M imidazole, pH 8.0). Eluted fractions were analysed by SDS-PAGE with the fractions with the highest concentration of protein combined. Darcin proteins were buffer exchanged by three rounds of dialysis into 50 mM ammonium bicarbonate.

2.4 Protein assay

Protein concentration was calculated using the Coomassie Plus protein assay kit (Pierce, Rockford, USA). Bovine serum albumin (BSA) was used as a standard and diluted to create a standard curve (0 – 50 mg/mL). Samples were diluted with purified water so they were within the range of the standard curve. The samples and BSA standards were mixed with Coomassie Plus protein assay reagent (2:1, reagent: sample). The absorbance reading was measured at 620 nm using a plate reader (Thermo Scientific™ Multiskan™).

2.5 Tris((2-carboxyethyl)phosphine) (TCEP) reduction

Protein reduction was performed using Pierce™ immobilised TCEP disulfide reducing gel (Thermo Scientific) as described in the manufacturer's protocol. Briefly, equal volumes of TCEP resin and protein sample were incubated for 1 hour at room temperature on a shaking block (1300 rpm). To recover the sample, the solution was centrifuged at 1000 x g for 1 minute.

2.6 1D SDS-PAGE

SDS-PAGE was performed as described by Laemmli[199]. Samples were incubated with reducing (with dithiothreitol (DTT)) or non-reducing (No DTT added) SDS sample buffer as stated for 5 minutes at 95 ^{o}C and loaded onto 15 % polyacrylamide gels. BioRad Broad Range Molecular Markers (Watford, U.K.) were loaded onto each gel. The gels were run at a

constant voltage of 180 V until the dye front reached the bottom of the gel. Protein bands were visualised using PhastGel® Blue R Coomassie Brilliant blue stain (Sigma-Aldrich, Dorset, U.K.) and destained with destain solution (water 80 %, acetic acid 10 % and methanol 10 %).

2.7 Partial plasma depletion

Albumin was depleted from plasma samples following the protocol described by Liu et al.[200] with 10 % (v/v) TCA in HPLC grade water used.

2.8 Spin column depletion

Plasma sample depletion was performed using the Thermo Scientific™ Pierce™ Top 12 Abundant Protein Depletion Spin columns (Thermo Scientific™ Pierce™). According to the manufacturer instructions, 10 µL of plasma sample was added to the spin column and incubated at room temperature on a roller mixer for 1 hour. The columns were centrifuged at 1000 x g for 2 minutes and the flow-through collected.

2.9 Vivaspin® sample concentration

Vivaspin® 500 centrifugal filter units (Sartorius Stedim Biotech, Germany) with a 3 kDa cut-off were used to concentrate the protein depletion spin column depleted plasma sample as described by manufacturer instructions. Briefly, sample was added to the filter and centrifuged at 15000 x g for 20 minutes. The sample was removed from the filter using a gel loading tip and retained for in-solution digestion.

2.10 Strataclean™ resin concentration

For protein concentration, 16 µL of Strataclean™ resin (Agilent Technologies) was added to each sample and vortexed for 2 minutes. The samples were centrifuged for 2 minutes at 430 x g and the supernatant removed. The protein within the sample was bound to the beads and extracted by SDS-PAGE or in-solution digestion.

2.11 TCA precipitation

The standard TCA precipitation method[201] was performed for protein concentration.

2.12 His-tag affinity purification

Protein (20 µg per 10 µL bead slurry) was immobilised by the HexaHistidine-tag to Ni-NTA magnetic agarose beads (Qiagen) by incubating for 1 hour at room temperature on a roller

mixer. The beads were washed with PBS-T (Tween 20) and the liquid separated from the beads using a magnetic rack. The beads were blocked with 1 % casein in PBS-T for 1 hour at room temperature on a roller mixer. The block was removed and beads washed with PBS-T. Beads were incubated with target (as described in results) for 1 hour at room temperature on a roller mixer. The unbound sample was removed from the beads and washed with NPI-20-T (50 mM NaH_2PO_4, 300 mM NaCl, 20 mM imidazole and 0.05 % (v/v) Tween 20). PBS pH 7.4 was added to the beads until required for analysis.

2.13 Mass spectrometric immunoassay (MSIA) method

The forward MSIA streptavidin protocol was followed as manufacturer's instructions using the Thermo Scientific™ MSIA™ streptavidin D.A.R.T.'s®. The tips were loaded onto a Finnpipette Novus i Electronic 12-Channel Pipette (Thermo Scientific) and all solutions loaded into a 96-well polypropylene low bind plastic microplates (Thermo Scientific). Using the parameters in Table 2.1, the affinity purification was performed. Eluted protein was dried down in a SpeedVac concentrator until dry and re-suspended in 25 mM ammonium bicarbonate. The resin from the tips was removed and the protein extracted by SDS-PAGE or in-solution digestion.

Table 2.1| MSIA workflow Finnpipette Novus i Electronic 12-Channel Pipette settings

The speed for sample loading was set at 1.

Step	Description	Microtiter plate volume (µL)	Mixing cycle volume (µL)	No. of mixing cycle iterations
1	Pre-wash - wash buffer	200	150	10
2	Affinity ligand binding - Biotinylated Affimer solution	125	100	500
3	Wash – PBS pH 7.2 (Total of two washes)	250	150	20
4	Target binding	150	125	999
5	Wash buffer – as outlined in results (Total of three washes)	250	150	20
6	Wash - water	250	150	20
7	Elution – 33 % ACN, 0.1 % (v/v) formic acid in HPLC grade water	50	40	100

2.14 Pyridyl disulfide-activated magnetic beads affinity purification

Adhirons (10 µg per 10 µL 10 % (v/v) bead suspension) were immobilised onto the PureCube thiol-activated MagBeads (Cube Biotech, Germany) that had been re-suspended in PBS pH 7.4, for 1 hour at room temperature on a roller mixer. Using a magnetic rack, the beads were separated from the unbound Affimer and washed twice with PBS pH 7.4. Beads were incubated with target (as described in results) for 1 hour at room temperature on a roller mixer. The unbound sample was removed from the beads and washed twice (as described in results). Beads were incubated with reducing SDS sample buffer and analysed by SDS-PAGE.

2.15 SulfoLink® coupling resin affinity purification

Typically, 50 µL of 50 % (v/v) SulfoLink® coupling resin slurry (Thermo Scientific) was used for each affinity purification reaction. The resin was centrifuged at 5000 rpm for 2 minutes and the storage buffer remove. The resin was washed twice with coupling buffer (50 mM Tris, 5 mM EDTA-Na, pH 8.5) by vortexing briefly, centrifuging at 5000 rpm and removing the wash. Prior to coupling, proteins were reduced following the TCEP reduction protocol. Adhiron/Affimer coupling solution (15 µg protein in 200 µL coupling buffer per 50 µL 50 % (v/v) SulfoLink® coupling resin slurry) was incubated with the resin at room temperature for 45 minutes on a roller mixer. The Adhiron/Affimer coupling solution was removed by centrifugation at 5000 rpm for 2minutes and the resin washed twice with four resin volumes of coupling buffer. The non-specific binding sites of the resin were blocked with one resin volume of 50 mM L-cysteine·HCl in coupling buffer for 45 minutes on a roller mixer. The blocking solution was removed by centrifugation at 5000 rpm for 2minutes and the resin washed once with four resin volumes of NaCl (1 M) and twice with four resin volumes of PBS, pH 7.4. Sample containing target (total volume of four resin volumes of PBS pH 7.4) was incubated with the resin at room temperature for 45 minutes on a roller mixer. Sample was removed by centrifugation at 5000 rpm for 2minutes and the resin washed as described in the results.

2.16 In-gel proteolysis

Pieces of gel were excised from the protein bands using a glass Pasteur pipette. Gel pieces were destained with a 2:1 solution of 25mM ammonium bicarbonate: acetonitrile (ACN) for 15 minutes at 37 ^{o}C and repeated until destained. The gel pieces were incubated with DTT (10 mM) at 60 ^{o}C for 1 hour to reduce the disulfide bonds. The solution was aspirated and

the gel pieces incubated with iodoacetamide (55 mM) for 45 minutes at room temperature in the dark. This step prevents the disulfide bonds from reforming by the covalent binding of a thiol group to cysteine residues. The solution was aspirated and the gel pieces dehydrated with ACN for 15 minutes at 37 ^{o}C. Trypsin (0.25 µg) was added to the gel pieces and the sample incubated at 37 ^{o}C for 16 hours. The digestion was stopped with formic acid (final concentration 1 % v/v).

2.17 In-solution digestion

Protein samples were incubated with RapiGest SF surfactant (Waters Corporation, Milford MA) at a final concentration of 0.05 % (w/v) in 25 mM ammonium bicarbonate for 10 minutes at 80 ^{o}C. Samples were then incubated with DTT (3 mM, final concentration) for 10 minutes at 60 ^{o}C to reduce disulfide bonds followed by alkylation with iodoacetamide (9 mM, final concentration) for 30 minutes in the dark, at room temperature. The protease, trypsin (diluted in 50 mM acetic acid),) Glu-C (diluted in 50 mM ammonium bicarbonate) or Lys-C (diluted in 50 mM ammonium bicarbonate) was added to the sample and incubated at 37 ^{o}C for 16 hours. The digest was treated with trifluoroacetic acid (TFA, final concentration 0.5 % v/v) and incubated at 37 ^{o}C for 45 minutes to remove the RapiGest™ SF Surfactant. The digested samples were centrifuged at 12000 rpm for 15 minutes.

2.18 Filter-aided sample preparation (FASP) digestion

Spin filters with a 10 kDa molecular weight cut off (Vivacon® 500, Sartorius Stedim Biotech, Germany) were washed with 300 µL 1% (v/v) formic acid by centrifugation for 15 minutes at 12500 rpm to remove any contaminants. Samples were loaded onto the filters and washed twice by centrifugation at 12500 rpm for 15 minutes using 500 µL of wash solution (0.05 % (v/v) RapiGest™(Waters, Manchester) in 25 mM ammonium bicarbonate). The samples were concentrated to approximately 50 µL and incubated for 15 minutes at 80 ^{o}C. DTT was then added to each sample (final concentration of 5 mM) and incubated at 60 ^{o}C for 15 minutes followed by the addition of iodoacetamide (final concentration of 5.5 mM) and incubation at room temperature in the dark for 30 minutes. The samples were then centrifuged at 12500 g rpm for 5 minutes to remove the above solutions and the filtrate discarded. Samples were digested overnight (approx. 16 h) at 37 ^{o}C using 1 µg sequencing grade trypsin (Sigma, Poole, UK, proteomics grade). Peptides were collected by centrifugation at 12500 g rpm for 10 minutes. The filters were washed twice with 20 µl wash solution by centrifugation at 12500 g rpm for 10 min. The samples were acidified with TFA and centrifuged at 12500 rpm for 10 minutes.

2.19 Native protein digestion

Darcin or darcin [$C_{78}S$, $C_{171}S$] v 2.0 and pepsinogen Adhiron A4 or pepsinogen Adhiron [A4_K_R] were incubated with trypsin at 37 ^{o}C in ammonium bicarbonate or chymotrypsin at 25 ^{o}C in 100 mM Tris-HCl, 10 mM $CaCl_2$ respectively (ratio 20:1, sample:protease). During the reaction, a volume equivalent to 5 µg of starting protein was removed and split in two for intact mass and SDS-PAGE analysis. Proteolysis was stopped by the addition of 10 % (w/v) TCA and neutralised with ammonium bicarbonate for SDS-PAGE or by 1 % (v/v) formic acid for intact mass spectrometry.

2.20 MALDI-TOF mass spectrometry using Bruker UltrafleXtreme™

Peptide mixtures were spotted onto a target plate and mixed with an equal volume of MALDI matrix (saturated solution of α-cyano-4-hydroxycinnaminic acid in 50 % (v/v) ACN/0.2 % (v/v) TFA) and left to dry. Data was acquired in reflectron mode using a laser energy of approximately 32 % of maximum, a laser frequency of 1000 Hz over 800 – 4000 *m/z* range. The data was analysed using Bruker FlexAnalysis software and a peak list generated for peptide mass fingerprinting.

2.21 Electrospray ionisation mass spectrometry (ESI-MS) of intact proteins

Before loading onto the mass spectrometer, samples containing glycerol were desalted offline using a C4 desalting trap (Waters, Manchester, UK) and diluted to 1 pmol/µL in 0.1 % (v/v) formic acid, 5 % (v/v) ACN in HPLC grade water. The samples were loaded onto a C4 desalting trap (Waters Mass PREP™ Micro desalting column, 2.1 x 5 mm, 20 µm particle size, 1000 Å pore size) (Waters, Manchester, UK) connected to a Waters nano ACQUITY Ultra Performance liquid chromatography® (UPLC®) system. The UPLC® system was coupled to a Waters SYNAPT™ G2 QTof mass spectrometer fitted with an electrospray source. Protein was eluted using solvent A and solvent B over 10 minutes using a gradient of 5 % to 95 % solvent B at 40 µL/minute (Solvent A was HPLC grade water with 0.1 % (v/v) formic acid and solvent B was HPLC grade ACN with 0.1 % (v/v) formic acid). Data was acquired over 500 – 3500 m/z range. Data processing was performing using maximum entropy deconvolution (MAX ENT 1, Mass Lynx version 4.1, Waters). Prior to sample analysis, the mass spectrometer was calibrated using horse heart myoglobin (500 fmol, Sigma).

2.22 Liquid chromatography tandem mass spectrometry (LC-MS/MS)

Q Exactive™ Orbitrap™ mass spectrometer

Samples were loaded onto a trap column (PepMap 100 C18, Thermo Scientific) for 7 minutes with loading buffer (98 % water, 2 % acetonitrile and 0.1 % TFA) at a flow rate of 9 µL/min. Peptides were separated using a reversed-phase C18 analytical column (Easy-Spray PepMap C18 column, Thermo Scientific, 75 µm internal diameter, 500 mm length, 2 µm particle size) connected to an Ultimate 3000 RSLC™ nano system (Thermo Scientific). The initial gradient conditions were 96.7% mobile phase A (0.1 % formic acid) 3.8 % mobile phase B (80 % ACN, 20 % water and 0.1 % formic acid) and over 10 minutes went to 50 % mobile phase A, 50 % mobile phase B. Modifications to the gradient time are noted in the relevant results sections. The column temperature was 25 ^{o}C and the gradient was operated at a flow rate of 0.3 µL/min. Eluted peptides were analysed using a QExactive™ mass spectrometer (Thermo Scientific) operating in data dependent acquisition (DDA) positive ESI mode. The capillary voltage was set to 2.2 eV with a capillary temperature of 250 ^{o}C. A MS full scan range of 300 to 2000 m/z was selected with a resolution of 70,000, a maximum fill time of 200 ms and an AGC value of 1e6. The top 10 most abundant peaks were selected for MS/MS fragmentation by collision induced dissociation (CID) with a normalised collision energy of 28 %. The MS/MS scans were performed at a resolution of 35,000, a maximum fill time of 100 ms, an AGC value of 1e5 and an isolation window of 3 *m/z*.

Q Exactive™ HFHybrid Quadrupole-Orbitrap™ mass spectrometer

Samples were loaded onto a trap column (PepMap 100 C18, Thermo Scientific) for 7 minutes with loading buffer (98 % water, 2 % acetonitrile and 0.1 % TFA) at a flow rate of 9 µL/min. Peptides were separated using a reversed-phase C18 analytical column (Easy-Spray PepMap C18 column, Thermo Scientific, 75 µm internal diameter, 500 mm length, 2 µm particle size) connected to an Ultimate 3000 RSLC™ nano system (Thermo Scientific). The initial gradient conditions were 96.7 % mobile phase A (0.1 % formic acid) 3.8 % mobile phase B (80 % ACN, 20 % water and 0.1 % formic acid) and over 10 minutes went to 50 % mobile phase A, 50 % mobile phase B. Modifications to the gradient time are noted in the relevant results sections. The column temperature was 25 ^{o}C and the gradient was operated at a flow rate of 0.3 µL/min. Eluted peptides were analysed using a QExactive™ mass spectrometer (Thermo Scientific) operating in data dependent acquisition (DDA)

positive ESI mode. The capillary voltage was set to 2.2 eV with a capillary temperature of 250 ^{o}C. A MS full scan range of 350 to 2000 m/z was selected with a resolution of 60,000, a maximum fill time of 100 ms and an AGC value of 3e6. The top 16 most abundant peaks were selected for MS/MS fragmentation by collision induced dissociation (CID) with a normalised collision energy of 30 %. The MS/MS scans were performed at a resolution of 30,000, a maximum fill time of 45 ms, an AGC value of 1e5 and an isolation window of 2 *m/z*.

2.23 Ion mobility-MS (IM-MS) and collision induced unfolding (CIU)

IM-MS analysis was carried out using a Waters Synapt G2-Si mass spectrometer (Waters, Manchester). Prior to analysis, proteins were buffer exchanged into 500 mM ammonium acetate using Amicon spin filter columns (3 kDa molecular cut-off) and diluted to 5 pmol/µL. Approximately 1 – 3 µl of sample was analysed. The spray voltage was set at 3 kV and the sampling cone 50 V. A single charge state was isolated in the quadrupole and was subjected to collisional activation. The activation voltage increased from 10 V to 26 V in 2 V increments. The IM wave height was 30 V and the wave velocity was 650 m/s. The data was processed using MassLynx v4.1 (Waters, Manchester) and the data exported to generate a plot using OriginPro 9.0.

2.24 Data analysis

Label-free protein quantification using Progenesis QI

Protein quantification was performed using Progenesis QI software v2.0 (Waters Corporation, Milford, MA). A merged peak list was generated and searched against a database as described in the results, using MASCOT search engine. Proteins with similar peptides were grouped into families and only unique peptides were used for quantification. The criteria used to determine differential expression is outlined in results.

Proteome Discoverer database searching

Peak lists were generated by Proteome Discoverer 2.0 (Thermo Scientific) using default parameters. The peak lists acquired were searched against a database as stated in the results using MASCOT as the search engine (Matrix science, Inc.). The following criteria were applied to the database search: carbamidomethylation of cysteine as a fixed modification, oxidation of methionine as a variable modification, precursor mass tolerance of 10 ppm, fragment ion mass tolerance of 0.01 Da and a maximum of one missed cleavages.

PEAKS database searching

For identification of Adhiron PTMs, data was searched using PEAKS software (Bioinformatics Solutions Inc, Canada) and searched against a custom naive Affimer database. The following criteria were applied to the database search: carbamidomethylation of cysteine as a fixed modification, oxidation of methionine as a variable modification, precursor mass tolerance of 10 ppm, fragment ion mass tolerance of 0.01 Da, a maximum of one missed cleavages and fragmentation type was set to higher-energy C-trap dissociation (HCD).

Chapter 3: Approaches and Challenges of Affimer Immobilisation for Affinity Purification

3.1 Introduction

The enrichment or purification of proteins is common practice in biological research with immunoaffinity chromatography the most common approach chosen. The general principles rely on an affinity reagent immobilised to a solid support (the stationary phase) and a solution containing the target molecule (the mobile phase). The solid support is necessary in order to isolate the target proteins from the remaining sample following capture. Typically, immunoaffinity purification methods rely on the use of antibodies as the capture reagent however the principles can be applied to use Affimers as the capture reagent. The general approach of an affinity purification workflow involves four main steps.

1) The immobilisation of the capture reagent onto a solid support through specific coupling chemistry.
2) Incubation of the sample with the affinity reagent for target enrichment.
3) Removal of non-specific proteins by washing with suitable buffers that still maintain the interaction between the affinity reagent and its target.
4) The elution of target proteins using buffers to disrupt the interaction between the affinity reagent and its target.

The immobilisation of affinity reagents onto solid supports can be performed using various methods and chemistries including covalent bonding through different functional groups on the protein or through non-covalent interactions. Immobilisation of proteins through the primary amine groups is a very common technique, with many solid supports available for immobilisation. However, the specific orientation of the molecule cannot be defined. Primary amines are located in the side chain of lysine residues and sometimes the N-terminus of each protein. Immobilisation through lysine residues is not the preferred method as it could prevent accessibility to the variable loop regions on the Affimer, hindering the capture of protein targets. The engineering of specific regions for coupling chemistry into the capture reagent structure is the ideal solution. For purification of Affimers following *E.coli* expression, a His-tag had been included in the protein design which facilitates immobilised-metal ion chromatography (IMAC) purification. Another functional group that is frequently used for immobilisation of affinity reagents is the sulfhydryl group. Sulfhydryl groups are located in the side chain of cysteine residues and

typical have an important role in protein structure and stability, forming disulfide bonds. Characteristically, Affimers do not contain cysteine residues in the scaffold sequence as disulfide bonds are not required for the protein to maintain its structure. However, the addition of a single cysteine residue into the protein sequence allows for the specific orientation of the Affimers. It is essential that the Affimers are in a reduced form to allow for immobilisation as when in solution, the Affimers can form disulfide linked dimers. The immobilisation techniques, based on his-tag, sulfhydryl chemistry and streptavidin-biotin interaction are explored in is chapter.

His-tag Affinity Purification

As mentioned above, the addition a His-tag is common in recombinant protein production to allow for the relatively simple purification of proteins for other cellular components. A His-tag, comprising usually of 6 histidine residues is added to the N- or C- terminus of the protein. Due to the relatively small size of the tag, no detrimental effects on protein folding or function are usually observed. The interaction between the imidazole group in the side chain of histidine residues and transition metal ions (such as Zn^{2+}, Cu^{2+} and Ni^{2+}) has been exploited to allow for the purification of His-tagged proteins IMAC (Figure 3.1). The method was first described for protein purification in 1975 using iminodiacetic acid[202]. In 1987, nitilotriacetic acid (NTA) was described as an improved method for the purification of His-tagged proteins[203], by increasing protein yield and improving purity due to the increased strength of the interaction between the acid and the metal ions.

Various solid supports are available for His-tag purification but in this study Ni-NTA magnetic agarose beads were used. The use of agarose beads for protein purification was first described in 1970[204] and is commonly used as a solid supports as they are easily functionalised for protein binding. Magnetic bead technology allows for small scale affinity purifications in tubes by using a magnet to separate the beads from the sample. Elution of protein from the beads is achieved by high concentrations of imidazole.

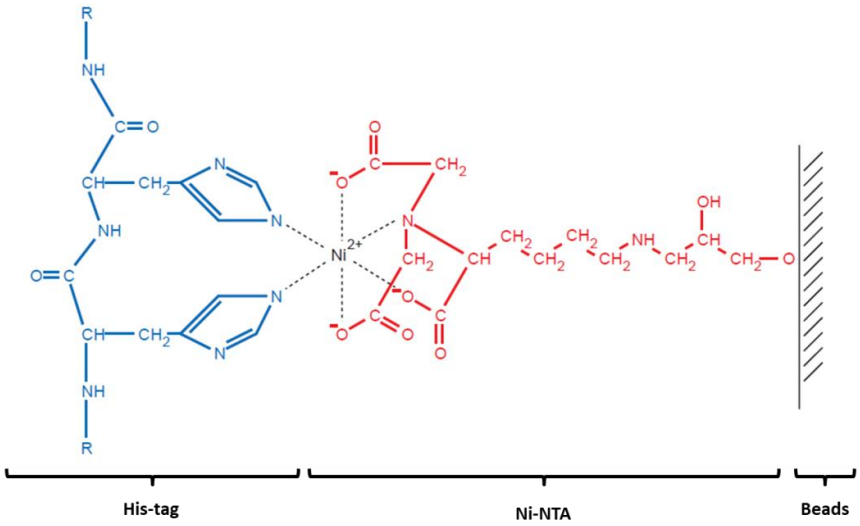


Figure 3.1 | Interaction between Ni-NTA and histidine residues of the 6XHis-tag.

A coordination complex is formed between nickel bound to NTA (red) and two adjacent histidine residues (blue) of the recombinant protein. The protein bound via the His-tag can be removed with imidazole through competitive elution. Image from Qiagen.

Streptavidin-Biotin Interaction and Mass Spectrometic Immunoassay (MSIA)

The MSIA approach was first described by Nelson et al. in 1995 and involves the immobilisation of antibodies to resin contained within a pipette tip[96]. When attached to an automated pipette, the technique enables small scale, reproducible and rapid affinity purifications to be undertaken eliminating much of the manual handling steps involved when compared to a traditional affinity purification workflow. MSIA affinity pipette tips are available from Thermo Scientific and offer various binding chemistry in the tips. The most suitable binding chemistry to immobilise the Affimers to the tips is based on streptavidin-biotin chemistry. The tips are coated with streptavidin and bind the capture protein through a biotin molecule. As the Affimers can be engineered to contain a single cysteine residue, defined orientation of immobilisation can be achieved through specific cysteine biotinylation. Proteins can be readily functionalised with biotin through a simple reaction with various functional groups on proteins. Due to the high reactivity of maleimide groups with sulfhydryls, biotin functionalised with maleimide will be used in this study (Figure 3.2). Streptavidin is a tetramer and has four binding sites for biotin[205] (Figure 3.2). The interactions between streptavidin and biotin include hydrogen bonding, van der Waals forces and hydrophobic interactions[206]. Although the streptavidin-biotin interaction is non-covalent, the affinity of streptavidin for biotin is extremely high with a $K_d = 10^{-14 \text{ to }} 10^{-15}$ M). This is an advantage over the his-tag interaction as more stringent washes can be applied whilst the capture reagent will remain immobilised on the solid support.

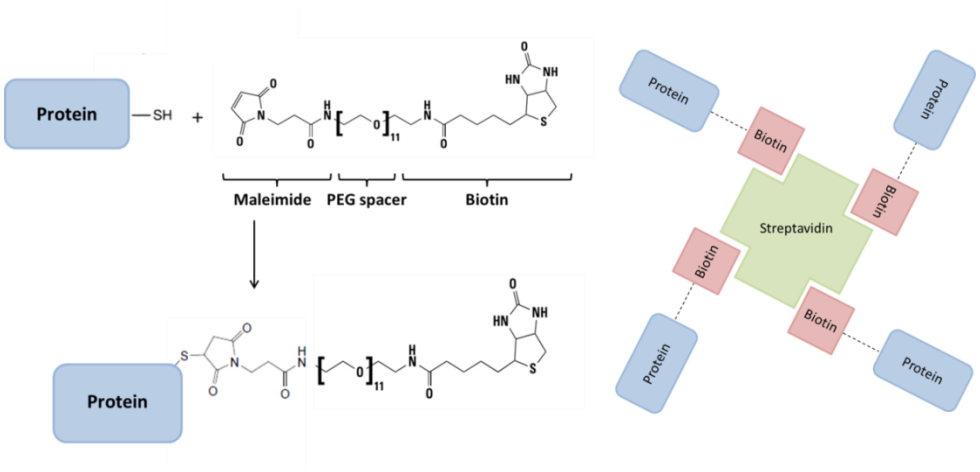


Figure 3.2 | Reaction for biotinylation of cysteine terminated proteins with maleimide-PEG11- biotin and the streptavidin-biotin interaction.

Maleimide groups react with reduced sufhydryl groups forming a thioether bond. Biotinylated proteins form strong non-covalent bonds with streptavidin. Image adapted from Thermo Scientfic.

Pyridyl disulfide-activated magnetic beads

Another immobilisation chemistry that utilises the cysteine residue of the Affimer is the interaction between pyridyl disulfide and sulfhydryl groups (Figure 3.3). Following the reaction, proteins are covalently coupled to the support forming a disulfide bond that can be cleaved upon the addition of a reducing agent such as DTT. As the proteins are covalently coupled to the beads, stringent wash buffers can be used to remove non-specific binding whilst retaining the affinity reagent.

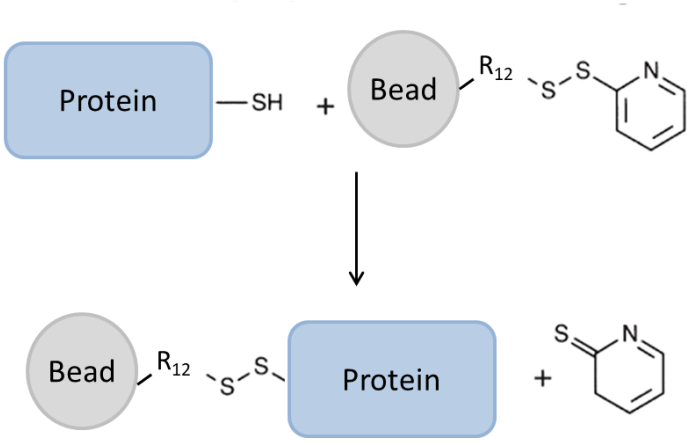


Figure 3.3 | Reaction for the immobilisation of cysteine terminated proteins to pyridyl disulfide-activated magnetic beads.

The pyridyl disulfide functional groups are immobilised onto magnetic agarose beads via a 12-atom spacer. The reduced sulfhydryl groups of cysteine residues on proteins react with the pyridyl disulfide groups forming a reversible disulfide bond and pyridine-2-thione.

Iodoacetyl-activated Resin

The final immobilisation chemistry explored within this chapter again utilise the cysteine residue for covalent Affimer immobilisation. Iodoacetyl groups react with sulfhydryl groups forming an irreversible covalent thioether bond (Figure 3.4). A commercially available product with iodoacetyl groups is the Thermo Scientific SulfoLink® coupling resin. The SulfoLink® coupling resin is typically used in a column format where antibodies are bound to the resin for affinity purification of specific protein targets. As MS can permit low amounts of protein material (low µg), small scale purification can be performed eliminating the need to use columns. The functional groups on the resin can be blocked with free cysteine which will not interfere with downstream MS analysis.

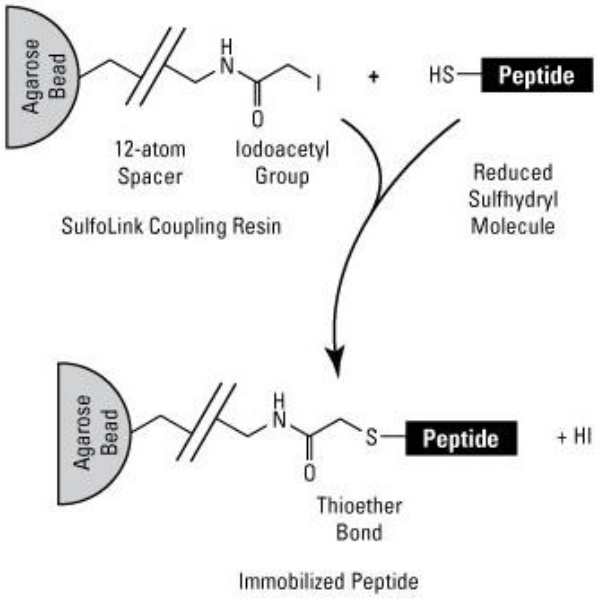


Figure 3.4 | Reaction for the immobilisation of cysteine terminated proteins to Iodoacetyl-activated resin.

The iodacetyl groups are immobilised onto agarose resin via a 12-atom spacer to help reduce steric hindrance. The reduced sulfhydryl groups of cysteine residues in peptides and proteins react with the iodoacetyl groups forming an irreversible covalent thioether bond. Image from Thermo Scientific.

Naïve Affimer Analysis

As discussed in Chapter 1, a common approach to protein biomarker discovery involves comparative mass spectrometry analysis of patient samples against healthy controls to identify differentially expressed proteins. A limitation of MS based approaches is the depth of proteome coverage achieved due to the large dynamic range of biological samples, especially plasma and serum. An approach to overcome this problem involves the use of

affinity reagents to enrich for low abundance proteins whose detection may be masked by highly abundant proteins in typical MS methods. Antibody array technology is explained further in Chapter 1 and as discussed, is typically a targeted approach to biomarker identification due to the immobilisation of antibodies with known protein targets. However, non-targeted approaches have been used with combinatorial antibodies as the capture reagent providing an unbiased approach to protein enrichment and analysis[207]. Affimer microarray technology has been described for the enrichment of cyclin dependent protein kinases[208]. A naïve Affimer discovery array is similar to antibody array however, naïve Affimers containing randomised unknown variable loop regions are immobilised onto an array[192]. Comparative analysis in binding of proteins to Affimers between control and disease samples is performed to reveal naïve Affimers that enrich for protein in a particular sample type. Identification of the naive Affimer target by mass spectrometry will reveal proteins that have an increased expression in either the control or disease samples and thus identify a protein biomarker. The naïve Affimer array work was carried out by colleagues at Avacta Life Sciences. The need for biomarkers of sepsis is discussed further in Chapter 5 and was therefore selected for naive Affimer array interrogation.

3.2 Aims and Objectives

The work in this chapter will focus on the development of an affinity purification workflow to capture, enrich and identify naive Affimer targets. The key aims were to:

- Optimise the His-tag magnetic bead affinity purification workflow and apply the protocol to the analysis of naïve Affimer.
- Develop the pyridyl disulfide-activated magnetic bead affinity purification.
- Assess the suitability of the MSIA approach for Affimer affinity purification.
- Develop the SulfoLink® resin affinity purification workflow and apply the optimised method to the analysis of naïve Affimers with subsequent target identification using mass spectrometry.

3.3 Results and Discussion

3.3.1 His-tag Experiments

The work in this chapter was performed in collaboration with Kit-Yee Tan and Vincent Puard from Avacta Life Sciences as part of the industrial experience element of CASE. The general workflow for a His-tag affinity purification is outlined in Figure 3.5 and indicates the various steps that required optimisation.

The beads selected for use in this study were the Qiagen Ni-NTA magnetic agarose beads. Other magnetic beads using nickel ion chromatography were assessed, but greater background binding of proteins was observed (data not shown). Based on the manufacturers guidelines, the binding capacity of the beads was stated as 20 µg per 10 µL bead suspension. However as steric hindrance could interfere with Affimer immobilisation by blocking the nickel sites from the His-tag, the binding capacity stated may not be accurate. Furthermore, if the full binding capacity of the His-tag beads was not achieved, then free sites of the beads may be exposed. This may cause an increase in non-specific binding. Therefore, it was vital to confirm that the beads were fully saturated with Affimer.

To assess the binding capacity of the beads, a titration experiment, increasing the amounts of Affimer incubated with the beads was performed following the His-tag immobilisation protocol described in Chapter 2.12. The binding titration was performed using 10 µL of bead slurry to ensure accuracy in pipetting. Preliminary experiments established that using less than 10 µL resulted in bead loss and inconsistencies in the amount of beads retained following the washing protocol. As the beads did not bind the Affimer with a covalent bond, incubation with sample buffer would release the Afimer from the beads. Therefore, samples were analysed by SDS-PAGE by loading the bead-Affimer solution directly into the wells of the gel (Figure 3.6). At approximately 14 kDa, the expected molecular weight of the Affimer, a band is visible across the gel decreasing in intensity from 20 µg to 1 µg of Affimer used in the titration. The amount of Affimer appears to reach a saturation point as no difference in band intensity is observed from 20 µg up to 30 µg Affimer. A band is present at approximately 6 kDa which is likely to correspond to a degradation product or contaminant retained following purification. Based on these findings 20 µg of Affimer per 10 µL bead slurry was selected as the binding capacity of the beads.

Due to the large number of optimisation experiments performed, the details of the key remaining optimisation steps are summarised in Table 3.1.

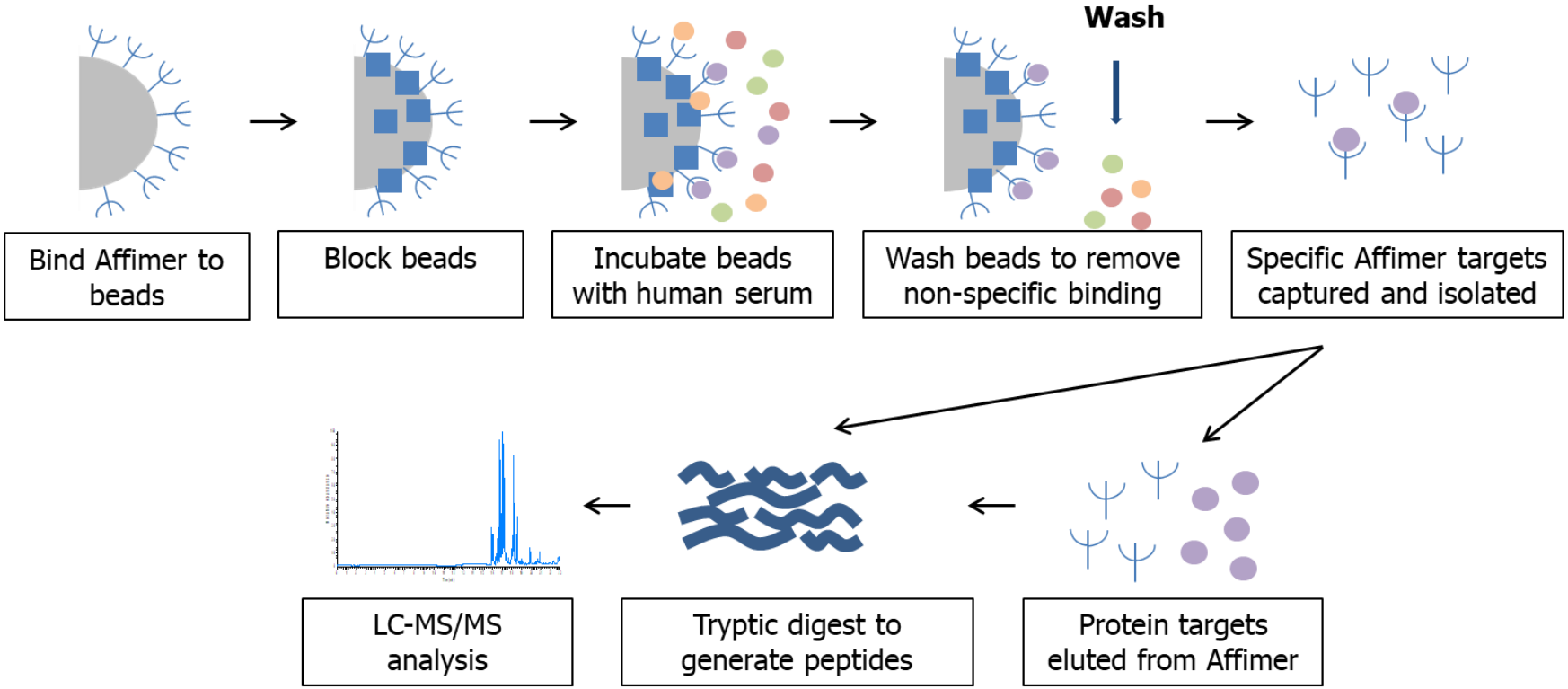


Figure 3.5 | General approach for a His-tag Affimer affinity purification workflow.

Affimers are bound to nickel coated magnetic beads through the His-tag. The beads are blocked to reduce non-specific binding and then incubated with sample to allow for the Affimers to capture target protein. The beads are then washed to remove non-specific binding. Protein targets can be eluted from the beads or the bead-Affimer-target complex digested directly for LC-MS/MS analysis.

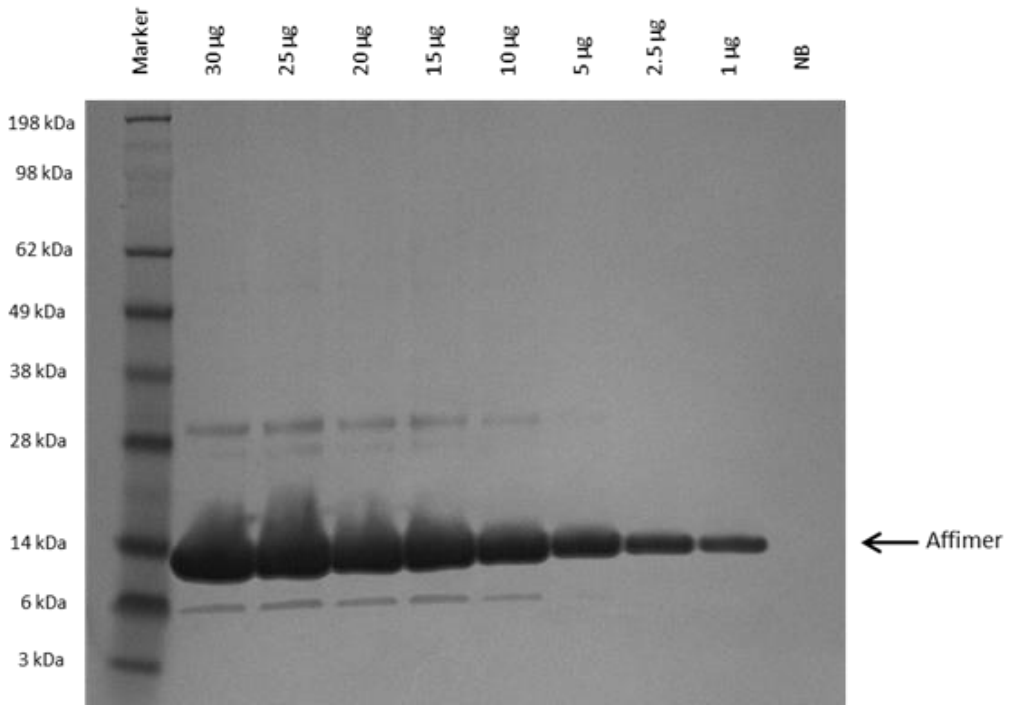


Figure 3.6 | SDS-PAGE analysis of His-tag magnetic bead binding capacity.

10 µL Ni-NTA magnetic agarose beads were incubated with different amounts of Affimer (1 µg to 30 µg) to establish the bead binding capacity. NB corresponds to beads with no Affimer bound. Beads were loaded directly onto the gel. Samples were analysed using a 4 – 12 % gradient polyacrylamide gel and the protein bands were visualised using InstantBlue™ Coomassie stain.

Table 3.1 | Summary of His-tag magnetic bead optimisation steps.

Analysis carried out in collaboration with Avacta Life Sciences.

Optimisation Step	Aim	Details	Result
Binding capacity of beads	To determine the binding capacity of 10 µL of Ni-NTA magnetic beads for Affimer.	As described above.	20 µg Affimer per 10 µL bead slurry.
Blocking buffer and time	To determine if blocking the beads with different buffers post Affimer incubation reduced the amount of non-specific binding for serum proteins.	Comparison of non-specific background from serum proteins with beads blocked with casein + 0.5 %, 1 % or 3 % BSA incubated for 1 hour or overnight.	Casein + 0.5% BSA in PBS-T for 1 hour.
Steric hindrance	To assess if steric hindrance reduced the amount of target captured by the Affimer.	Comparison between: Chicken IgY Affimer incubated with pure IgY target and then immobilised onto His-tag beads and; Chicken IgY Affimer immobilised onto His-tag beads and then incubated with pure IgY target.	SDS-PAGE analysis revealed more target captured when Affimer was immobilised onto beads and then incubated with target.
Elution	To determine a suitable elution for removing target and retaining Affimer on beads.	Elution buffer with 250 mM imidazole was used to elute the His-tag Affimer from the beads.	As the elution did not remove all the protein from the beads, an on-bead digest was carried out.
Digestion	To determine the optimal digestion method.	As tween wash added to the wash buffer to reduce non-specific binding, a method was needed to remove tween before LC-MS/MS analysis. In a FASP digest, the beads were loaded above the filter and washed extensively to with 25 mM ammonium bicarbonate to remove contaminants.	FASP on-bead digestion (Chapter 2.18).

His-tag Ni-NTA magnetic bead affinity purification parameters:

- Beads: Qiagen Ni-NTA magnetic beads.
- Binding Capacity: 10 µl bead slurry binding 20 µg his-tagged protein (2 mg/mL Affimer solution).
- Blocking solution: Casein + 0.5% BSA in PBS-T.
- Sample: 10 mg serum incubated for 45 minutes.
- Wash buffer: PBS-T, 20 mM imidazole, 50 mM NaCl.

Initially 25 naïve Affimers His-tag affinity purification samples in human serum were analysed by SDS-PAGE and LC-MS/MS. On-bead digestion of the Affimer-target complex was performed and peptides analysed on the Thermo QExactive mass spectrometer. Both SDS-PAGE and LC-MS/MS analysis confirmed the positive control Affimer, CC3, bound its target, complement C3. However, due to common background of non-specific proteins, identification of naïve Affimer targets was problematic.

Repeated analysis of the naïve Affimers was performed with each naive Affimer incubated with either control or sepsis serum so that a comparative analysis could be performed to reveal proteins that bind differentially in the different samples. The beads were divided in two and analysed by SDS-PAGE and LC-MS/MS. When comparing control sample and disease sample lanes for an individual Affimer by SDS-PAGE, subtle differences between protein bands was observed (Figure 3.7). For the positive control CC3 Affimer, bands were observed in the control and disease sample at approximately 176 kDa, the expected molecular weight of complement C3. The digested affinity purification samples were analysed on the Thermo Scientific QExactive mass spectrometer as described in Chapter 2.22. Peptides for the Affimer were highly abundant and limited the loading of more material onto the instrument (Figure 3.7). To investigate quantitative differences in the proteins captured in the naïve Affimer affinity purifications, label-free protein quantification was performed using Progenesis QI (Waters Corporation). Proteins were normalised using the 'normalise to all proteins method' and samples grouped by naïve Affimer. The data was searched using MASCOT against a human and eSQT database. Due to the high sequence homology between the naïve Affimers and eSQT, the database only contained eSQT Affimer. A total of 81 proteins were identified and quantified using non-conflicting peptides with a minimum of 2 peptides per protein. To investigate differences in protein abundances between the naïve Affimers, protein abundance values were log 10 transformed and analysed by hierarchical clustering (Figure 3.8). The portion in the centre

of the heat map represents highly abundant non-specific serum proteins that are consistently binding to the naïve Affimers. In addition, keratins were also identified as consistent background. Specific enrichment of proteins by the naïve Affimers was not observed.

Extensive optimisation of the His-tag affinity purification approach was carried out as outlined in Table 3.1, however various factors prevented the identification of naïve Affimer targets. Each naive Affimer affinity purification was only performed in a single replicate. Therefore statistical analysis of the identified proteins could not be performed reducing the confidence in identifications. Future analysis of naïve Affimers should ensure triplicate analysis.

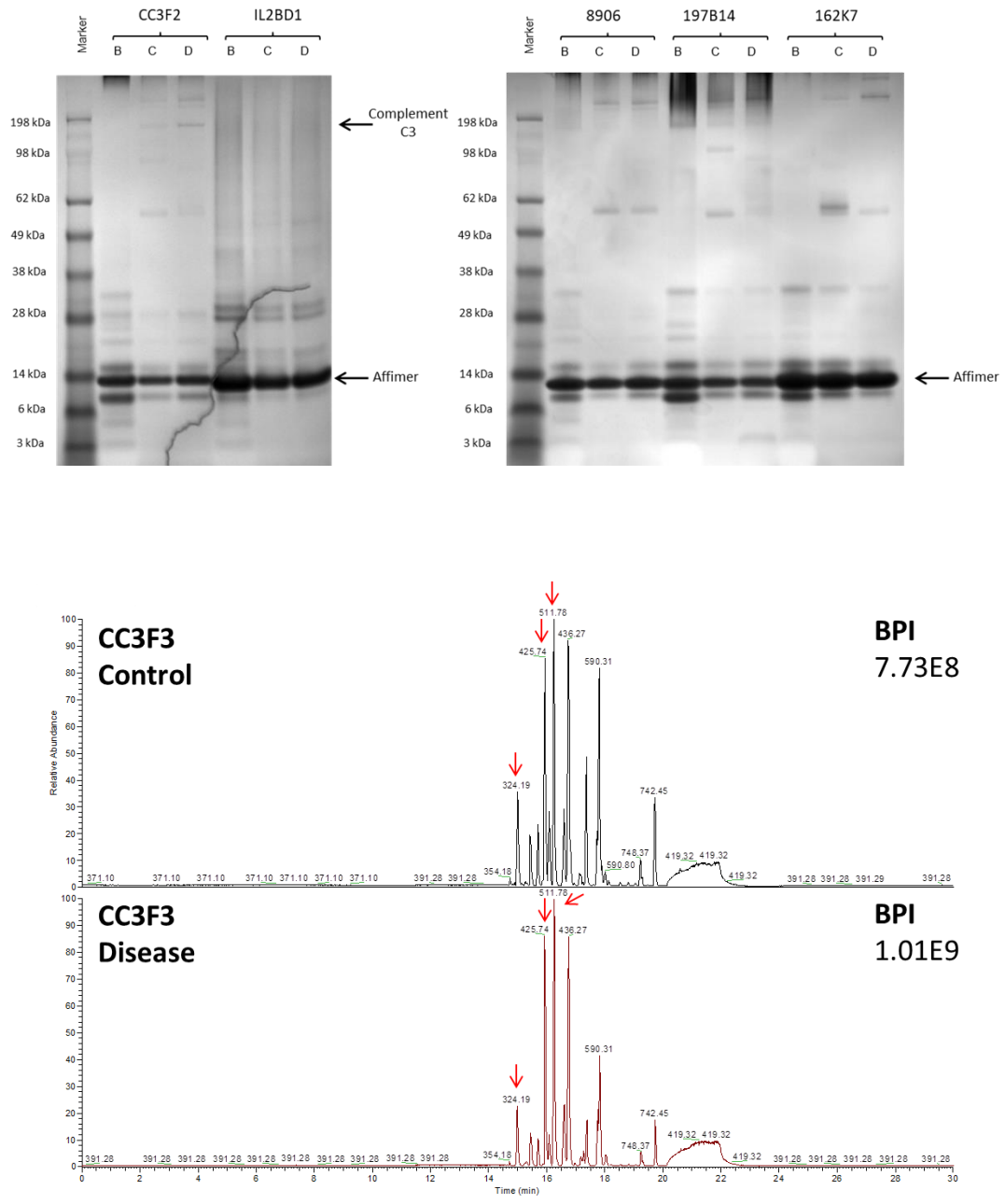


Figure 3.7 |SDS-PAGE analysis of naïve Affimer affinity purifications and representative BPI chromatogram of CC3 Affimer on-bead digest of his-tag affinity purification sample.

Top: Representative SDS-PAGE analysis of 5 of the naïve Affimer comparative affinity purifications including the CC3F2 Affimer that targets complement C3. 10 µL of the bead slurry was loaded onto a 4 – 12 % gradient polyacrylamide gel and the protein bands were visualised using InstantBlue™ Coomassie stain. His-tagged Affimers were immobilised onto Ni-NTA magnetic beads B – blocked beads and Affimer, C – blocked beads, Affimer and control serum, D – blocked beads, Affimer and sepsis serum. Bottom: Representative BPI chromatograms of on-bead digest of CC3F3 affinity purification in control and disease serum. Red arrows indicate peaks corresponding to Affimer peptides. Affinity purification and SDS-PAGE analysis undertaken by collaborators, Avacta Life Science.

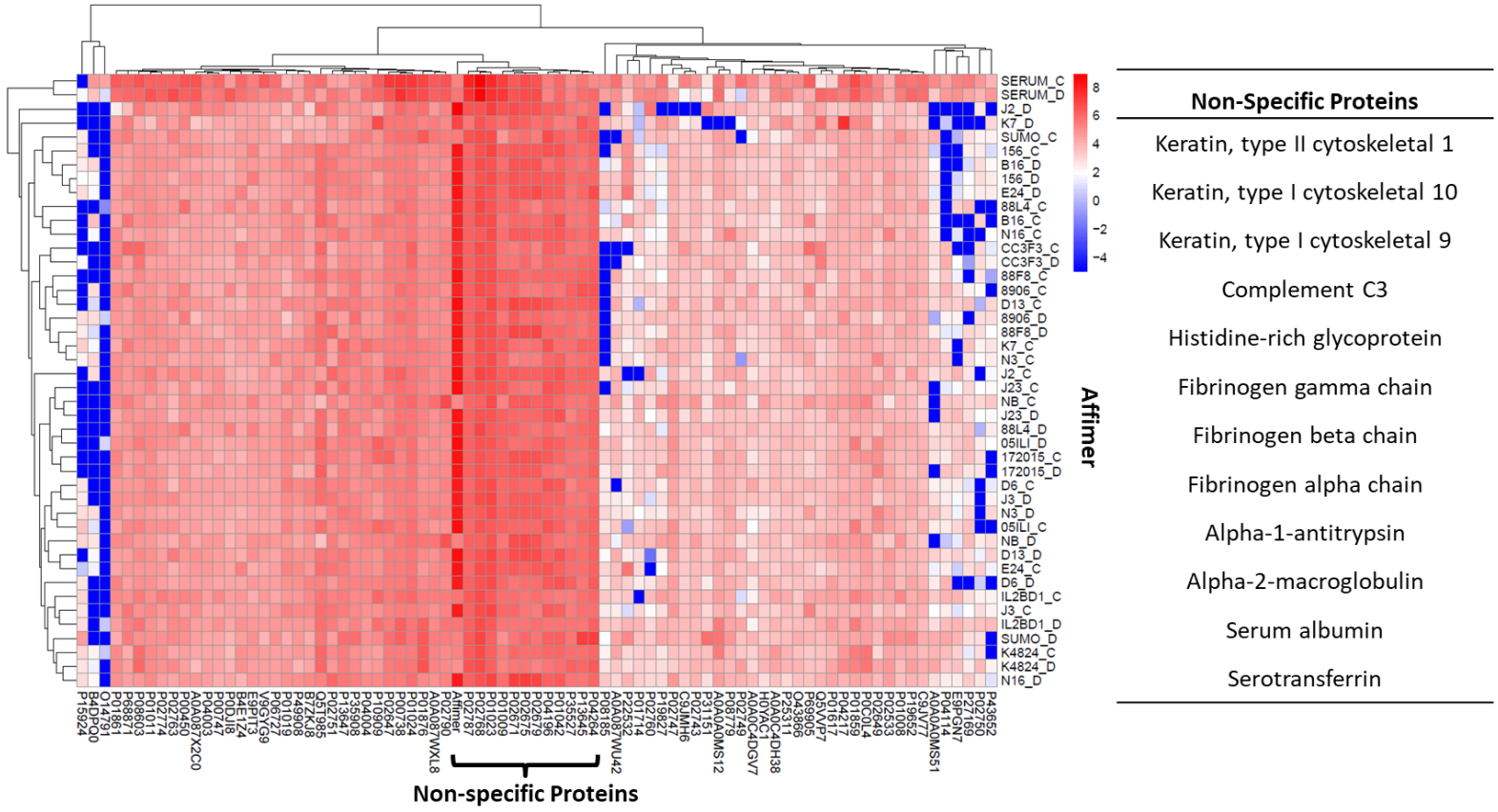


Figure 3.8 | Heatmap and hierarchical clustering of log 10-transformed protein abundance data for naïve Affimer affinity purifications.

Protein abundance values were taken from Progenesis QI and log 10—transformed. Samples clustered on protein abundance with highly abundant proteins represented in red and less abundant proteins represented in blue. The red portion in the centre of the heatmap represents proteins that are identified as highly abundant consistent non-specific background.

3.3.2 MSIA Experiments

As discussed in Section 3.1, a MSIA affinity purification assay involves using biotinylated Affimers that have a high binding affinity to streptavidin which is contained with a tip. The MSIA tips are attached to an automated pipette that automatically passes solutions through the tips, allowing the reagents to bind to the affinity reagent coupled to the resin. The general workflow employed within this study is outlined in Figure 3.9. The MSIA method has been frequently used for the enrichment of low abundance proteins captured using antibodies from complex biological samples and subsequently detected with SRM based assays[209] and top down analysis of intact proteins[210]. However, as the main application of this approach is to identify unknown binding partners of naïve Affimers, method development is required to determine if the approach can be applied in a discovery based format.

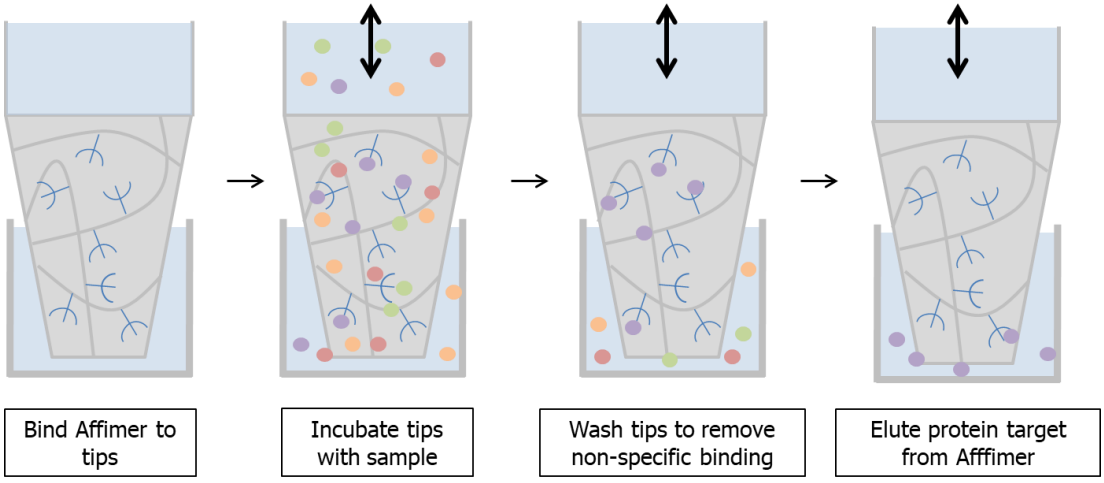


Figure 3.9 | Outline of the general MSIA affinity purification workflow.

The MSIA tips are loaded onto the automated pipette and all samples added to a 96-well microtitre plate for automated sample loading. Biotinylated Affimers are bound to the resin through streptavidin-biotin chemistry. The sample is then passed through the tip on a continuous cycle to allow for the Affimers to capture the target molecules. The tips are then washed to remove non-specific binding. The target protein in eluted from the Affimers for tryptic digestion and MS analysis.

3.3.2.1 Analysis of biotinylation efficiency

The Affimers used in this study were obtained from Avacta Life Sciences. To develop the method, Affimers with known, well characterised binding targets were used. In addition, the eSQT Affimer was analysed to act as a negative control. The eSQT Affimer is the Affimer

protein scaffold that does not contain any peptide insertions within the variable loop regions. Although biotinylated Affimers were needed for the MSIA workflow, both unbiotinylated and biotinylated Affimers were obtained for confirmation of the correct protein product prior to biotinylation. To assess the purity of the proteins, Affimers were analysed by SDS-PAGE on a reducing gel (Figure 3.10). At approximately 14 kDa, a band was observed for the unbiotinylated version of eSQT Affimer which is the expected molecular weight of the protein. For biotinylated eSQT Affimer, a band was observed at approximately 14 kDa and 16 kDa suggesting a heterogeneous population of Affimer comprised of both the biotinylated and non-biotinylated forms. Faint protein bands were observed at approximately 30 kDa in both eSQT Affimer samples corresponding to the dimer form of the Affimer. A doublet band at approximately 30 kDa and 31 kDa is visible in the biotinylated sample. Although the Affimers contain a single cysteine required for biotinylation, if the Affimers were biotinylated, the cysteine residues would not be available for disulfide linked dimerisation. Therefore, this suggests the Affimer dimers may be forming through domain swapped dimerisation[211]. This is discussed further is Chapter 4. Analysis of CC3 Affimer samples revealed similar results. A single band was observed at approximately 15 kDa for the unbiotinylated Affimer sample, corresponding to the expected molecular weight. For the biotinylated CC3 Affimer sample, two bands were observed at approximately 15 kDa and 16 kDa, again suggesting a heterogeneous population of both biotinylated and unbiotinylated versions of the Affimer. Protein bands at approximately 30 kDa were observed in both CC3 Affimer samples corresponding to the dimer form of the Affimer. SDS-PAGE analysis of the Affimer reagents revealed high purity and no contaminates (Figure 3.10).

To confirm the Affimers had the expected molecular weight and to ensure an increased mass after biotinylation, the Affimers were analysed by intact mass spectrometry following the protocol described in Chapter 2.21. Furthermore, intact mass analysis of the Affimers would determine biotinylation efficiency. After deconvolution of the multiply charged protein envelope, the average mass of the Affimer was determined (Figure 3.11 and Table 3.2). The mass adduct following biotinylation was + 922 Da. The predominant species in the mass spectrum for unbiotinylated CC3 Affimer was 13793.0 Da which corresponded to the theoretical molecular weight. An additional peak of plus 42 Da was also observed which was likely to correspond to acetylation. Two main species in the mass spectrum for biotinylated CC3 Affimer were observed at 13792.9 Da and 14715.0 Da, corresponding to unbiotinylated and biotinylated forms of the Affimer. This result supports findings from

SDS-PAGE analysis confirming that the biotinylation reaction was not 100 % efficient. The predominant species within the mass spectrum for unbiotinylated eSQT Affimer has a mass of 12000.7 Da which corresponded to the theoretical molecular weight of the Affimer. The main species in the biotinylated eSQT Affimer sample had a mass of 12922.7 Da corresponding to the expected molecular weight. A peak at 12000 Da was not observed. This result differs from SDS-PAGE findings as it suggests almost 100 % biotinylation efficiency. In addition, a peak was observed in both of the eSQT Affimer samples of 11869 Da corresponding to the loss of 103 Da, consistent with the loss of the c-terminal cysteine. Biotinylation of this Affimer fraction could not take place which explains why the peak is also observed in the biotinylated Affimer sample. Both SDS-PAGE and intact mass results confirmed the Affimers are suitable for use in the MSIA workflow; however it was noted that it was necessary to use an excess of CC3 Affimer during the immobilisation step due to inefficiencies in biotinylation.

Additional Affimers were also received for the MSIA optimisation. Intact mass analysis revealed they were unsuitable for coupling to the MSIA tips due to them lacking a C-terminal cysteine residue meaning the Affimers could not undergo the biotinylation reaction. The Affimers provided would have allowed for recombinant protein to be spiked into samples for affinity purification optimisation. Due to the cost of recombinant human complement C3 it was not possible to purchase pure protein for optimisation. Therefore, all binding studies had to be performed using human plasma or serum which contained endogenous complement C3 which would be enriched by the CC3 Affimer.

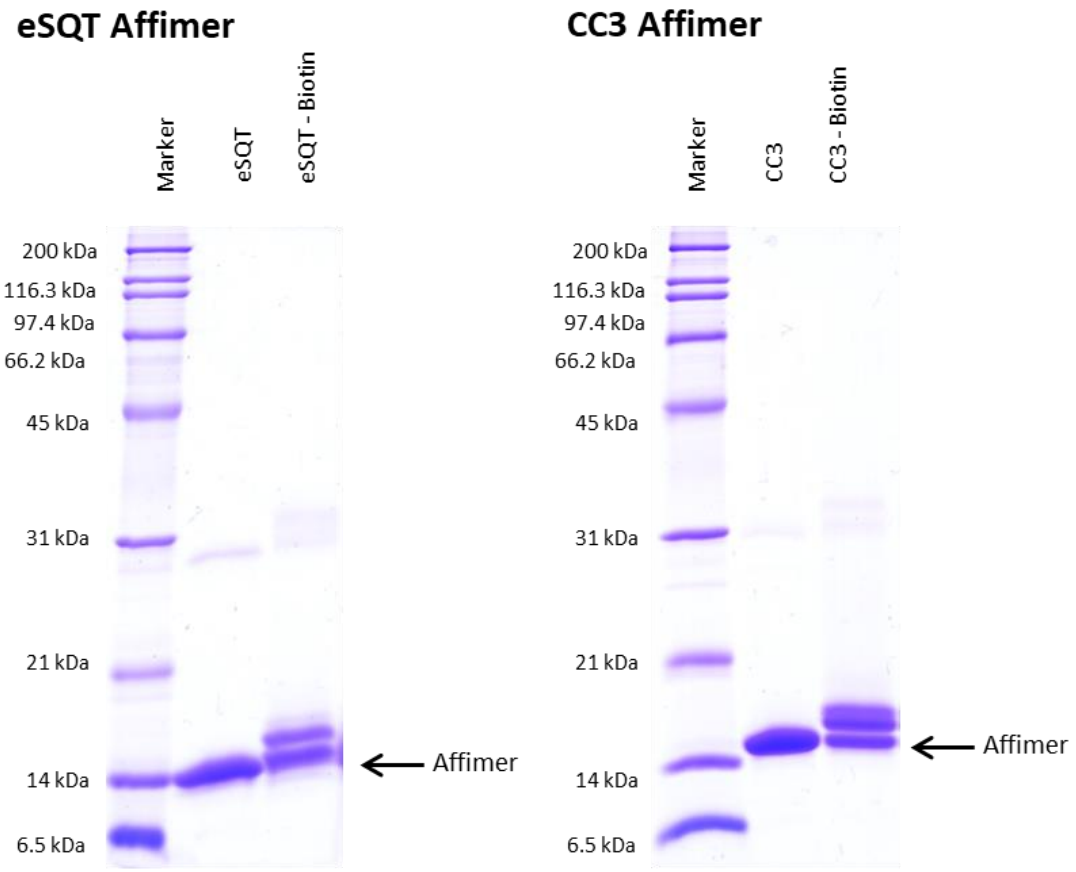


Figure 3.10 | SDS-PAGE analysis of eSQT and CC3 Affimers, pre and biotinyaltion.

eSQT and CC3 Affimers (2 µg), pre and post biotinylation were analysed by SDS-PAGE under reducing conditions on a 15 % gel and visualised with Coomassie plus stain. A shift in molecular weight was observed following biotinylation of both of the Affimers.

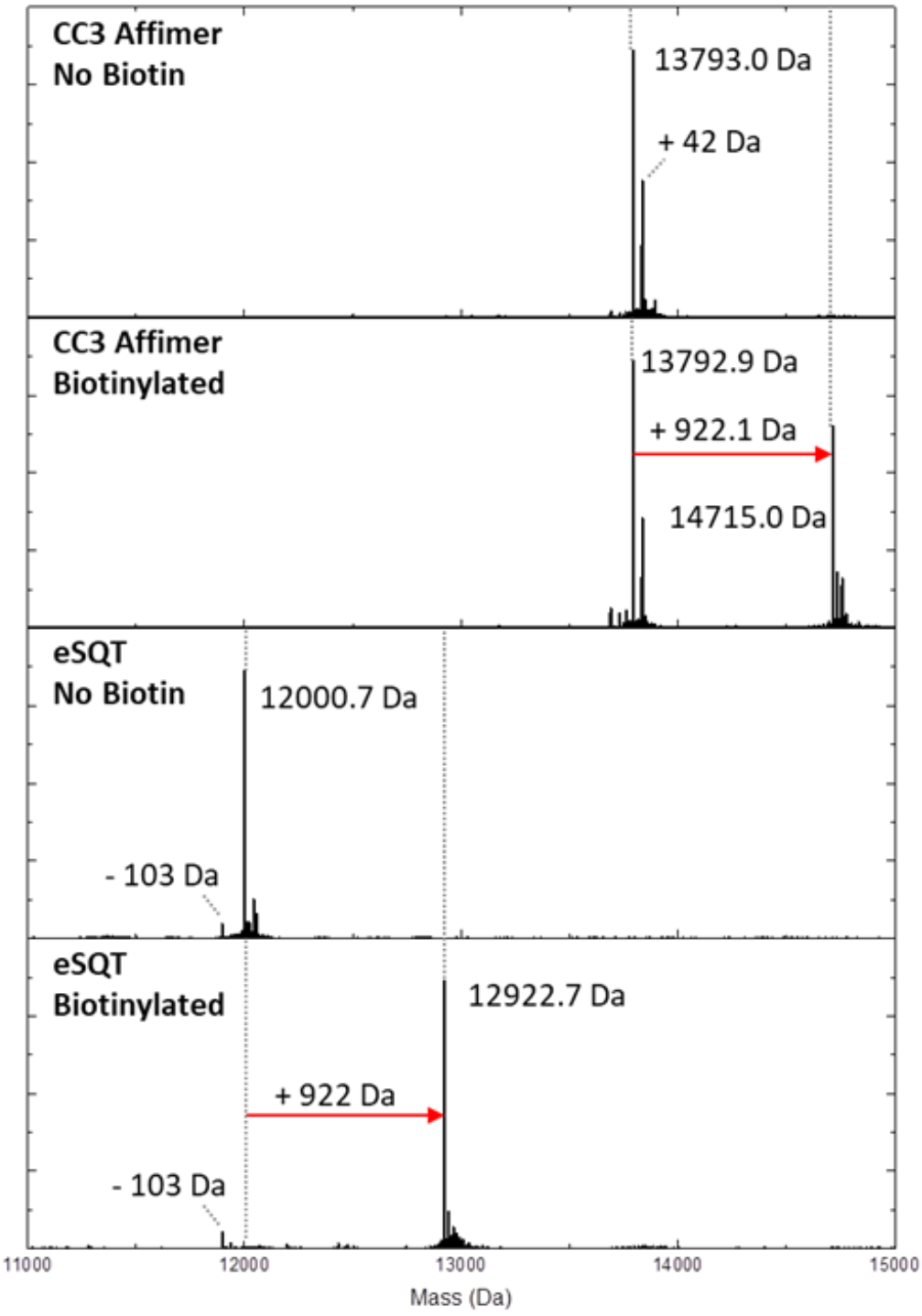


Figure 3.11 | ESI-MS analysis of eSQT and CC3 Affimer pre and post biotinylation.

Protein samples were diluted to 1 pmol/uL in 3% acetonitrile, 0.1% formic acid. A total of 500 fmol and 4 pmol for eSQT Affimers and CC3 Affimers respectively, was loaded onto a C4 desalting trapping column. The samples were analysed by ESI-MS on the waters G2 mass spectrometer. The multiply charged protein envelope was deconvoluted using Waters MAXENT 1 algorithm to determine the average mass of the intact protein.

Table 3.2 | Comparison between theoretical and observed intact mass of Affimers.

Theoretical mass of biotinylated Affimers corresponds to the addition of biotin (+ 922 Da).

Affimer	Theoretical Mass (Da)	Observed Mass (Da)	Difference (Da)	Interpretation
CC3	13792.5	13793.0	+ 0.5	Expected mass observed
Biotinylated CC3	14714.5	13792.9	- 921.6	Unbiotinylated CC3 Affimer present even after biotinylation
		14715.0	+0.5	Expected mass observed ~ 50 % biotinylation achieved
eSQT	12000.5	12000.7	+ 0.2	Expected mass observed
Biotinylated eSQT	12922.5	12922.7	+ 0.2	Expected mass observed ~ 100 % biotinylation achieved

3.3.3.2 Method development

The method development work in this section was performed in collaboration with Kimberley Burrow, an Avacta Life Sciences and University of Liverpool employee. The first step in the method development process involved determining the binding capacity of a single MSIA tip. Each tip contains a total of 4 µg streptavidin which, based on a molecular weight of 57 kDa for streptavidin, is equivalent to 70 pmol per tip. A single streptavidin molecule (57 kDa) can bind 4 biotin molecules. Therefore, a maximum of 280 pmol of Affimer can bind to each tip which is the equivalent of 3.9 µg (based on an Affimer molecular weight of 14 kDa). However, due to steric hindrance it is unlikely that the full binding capacity of 280 pmol of Affimer will be immobilised onto the tips. To assess the binding capacity of the tips, samples containing different amounts of Affimer ranging from 2 µg to 20 µg were incubated with the tips and the unbound Affimer fraction analysed by SDS-PAGE to assess for depletion of Affimer (results not shown). The resin inside the tips was also removed and analysed by SDS-PAGE. The findings from 2 µg, 3 µg and 4 µg loading are evident in Figure 3.12. A doublet band at approximately 15 kDa was observed for all samples of comparable intensity sharing a similar mobility to the control eSQT Affimer also analysed on the gel. This result suggested that a sufficient amount of Affimer was loaded onto the tips and that little difference was observed when increasing the amount to 4 µg of Affimer compared to 2 µg. However, a control tip that had no Affimer bound was also analysed and a single band at approximately 15 kDa was observed that matched the lower band in the tips where Affimer was bound. Streptavidin is a tetrameric protein and incubation with reducing sample buffer would results in the observation of the monomer by SDS-PAGE. This band was therefore likely to be streptavidin. Thus the fainter upper band in the tips where Affimer was added corresponded to Affimer. Analysis of tips with up to 20 µg Affimer was added show comparable results to the 2 µg indicating that the binding capacity of the tips had been reached. This amount of Affimer was significantly less than that used in the His-tag affinity purifications. In addition, results from his-tag affinity purifications had demonstrated that the binding of Affimer to target was not a 1:1 ratio and therefore it was suggested that the amount of Affimer immobilised onto each tip was not enough for an enrichment to be observed. Despite this, enrichment of CC3 using CC3 Affimer was assessed.

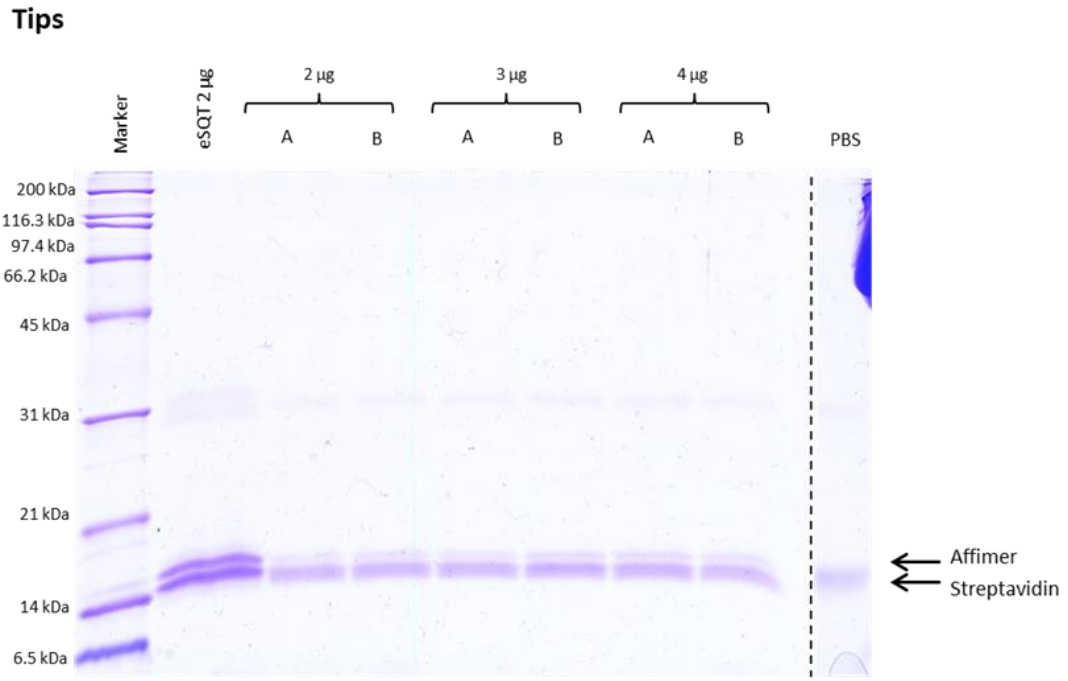


Figure 3.12 | SDS-PAGE analysis determining the binding capacity of the MSIA tips.

After Affimer immobilisation, the resin was removed from the tips and incubated with sample buffer at 95 ^{o}C for 10 minutes. Each sample was analysed in duplicate as annotated by lanes A and B. A tip containing no Affimer, incubated only in PBS, was also analysed. Samples were analysed on a 15 % gel and the protein was visualised with Coomassie plus stain. SDS-PAGE analysis undertaken in association with collaborators, Kimberley Burrow from Avacta Life Sciences.

To assess for protein enrichment with the CC3 Affimer and to determine the level of non-specific background of proteins binding to the tips, the MSIA experiment was performed as described in Chapter 2.13 with human plasma and the tips and elution fractions analysed by SDS-PAGE. To achieve enrichment of complement c3 with the CC3 Affimer using the His-tag approach, 20 mg of serum protein was used for the affinity purification. As the volume of the well was limited, 4.5 mg of plasma protein was used for the MSIA purifications. The expected concentration of complement c3 in serum was approximately 60 – 150 mg/dL[212]. Included in the analysis were tips with no Affimer bound to assess background binding of plasma proteins and tips with only Affimer bound to ensure proteins bands could not be attributed to Affimer. After the elution step, the elution fraction was retained for SDS-PAGE analysis. To analyse the remaining protein bound to the tips, the resin inside the tips was removed and incubated with sample buffer for 10 minutes at 95 ^{O}C.

SDS-PAGE analysis revealed no differences were observed in protein bands in both the elution fractions and destroyed tips between the MSIA tips with Affimer bound and those without Affimer following exposure to human plasma (Figure 3.13). This would suggest that the Affimers are not providing an enrichment of the protein target, complement C3. The low amount of Affimer immobilised onto the tips and the high level of background binding could explain this result.

To further assess protein enrichment in plasma using the Affimer MSIA method, comparative analysis of the proteins identified with eSQT or CC3 Affimer bound to the tips was performed. The MSIA experiment was performed as described in Chapter 2.13 and the elution fraction and tips digested for LC-MS/MS analysis. The resin was digested to analyse non-specifically bound proteins. Peptides were analysed on the Thermo Scientific QExactive mass spectrometer on a 30 minute gradient as described in Chapter 2.22. To investigate any enrichment of complement C3 with the CC3 Affimer compared to the eSQT Affimer, label-free protein quantification data was obtained using Progenesis QI (Waters Corporation). Retention time alignment was performed with sample eSQT Affimer Elution replicate 2 selected as the alignment file. Alignment scores achieved 90 % or greater indicating similarities in the retention time profiles between the samples. Proteins were normalised using the 'normalise to all proteins method' and scores for all but one samples were with a log factor of 0.3. The data was searched using MASCOT against a human and eSQT database. Due to the high sequence homology between eSQT and CC3, the database only contained eSQT Affimer. A total of 52 proteins were identified and quantified using non-conflicting peptides with a minimum of 2 peptides per protein. To investigate

differences in protein abundances and to asses for enrichment of complement C3 with the CC3 Affimer, protein abundance values were log 10 transformed and analysed by hierarchical clustering (Figure 3.14). Separation between the tip and elution samples was observed with the samples clustering together although within the two groups samples from the same Affimers did not cluster together. This suggests little difference in the proteins identified and their abundances. The high number of proteins identified demonstrates a consistent non-specific background binding to the tips. Abundance values for complement C3 were examined further (Figure 5.14) and little difference was observed between the CC3 Affimer samples and eSQT Affimer samples. The high levels of non-specific binding and the small amount of Affimer bound to the tips could explain this result.

Various other more stringent wash buffers were tried including PBS (range pH 6.6, 7.2 and 9.1), ammonium acetate pH4 and Tris-HCl pH 8 however the same level of non-specific background was observed. As enrichment of complement C3 was not observed with CC3 Affimer, the MSIA protocol was not developed further. The CC3 Affimer binds its target with high affinity whereas the naïve Affimers bind their target with a weak affinity. Because of this, along with the high amount of non-specific background binding it was hypothesised that identification of naïve Affimer targets would not be achievable. Although the method has numerous benefits due to the high level of automation, the MSIA method may be more suitable with capture reagents that have high binding affinities to their target where stringent washing can be applied.

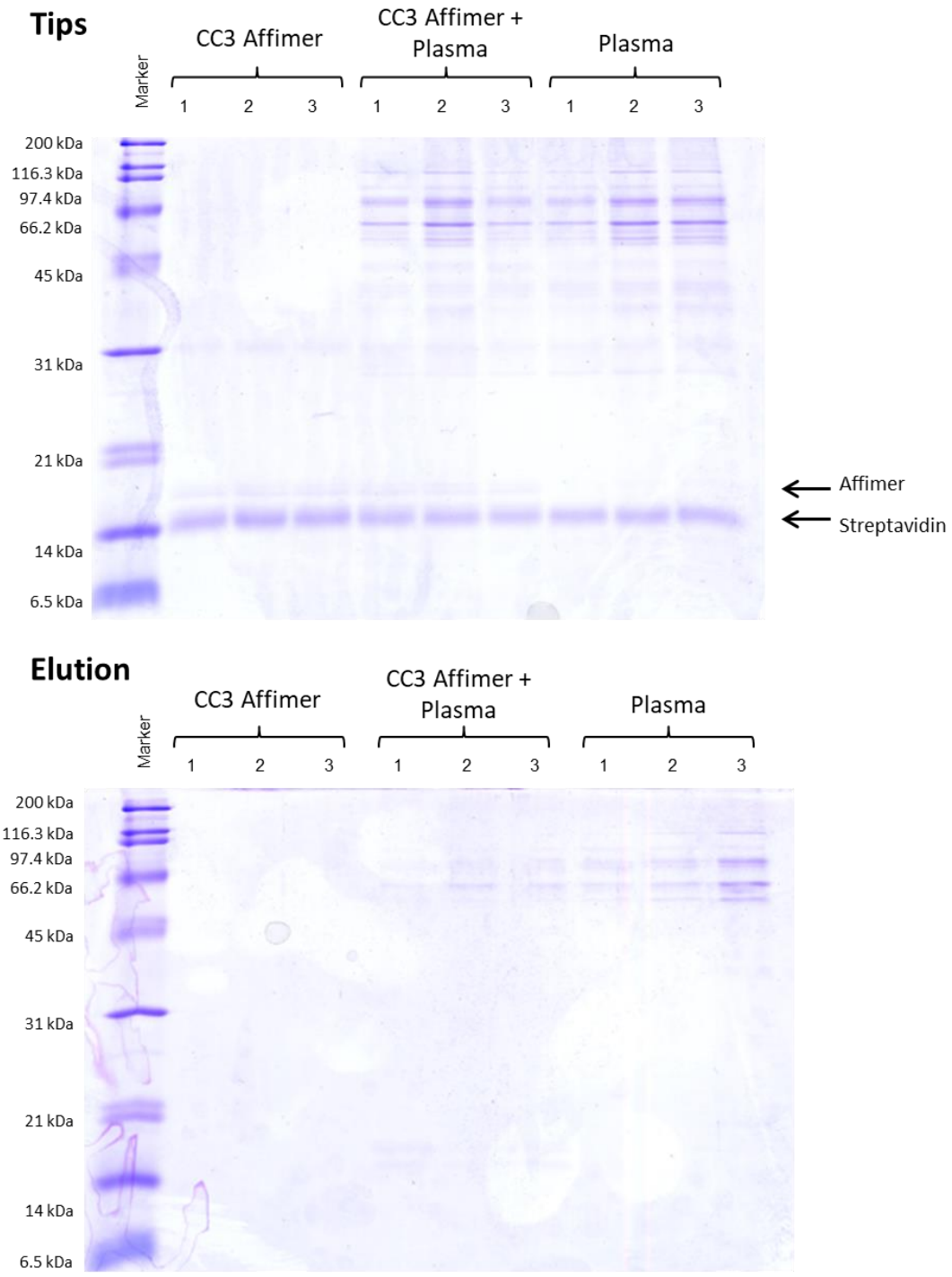


Figure 3.13 | SDS-PAGE analysis of MSIA CC3 Affimer affinity purification.

The CC3 Affimer that targets complement C3 was bound to tips and incubated with human plasma. Bottom: Protein was eluted using 33 % acetonitrile, 0.1 % formic acid. Top: The resin was removed from the tips and incubated with sample buffer at 95 ^{o}C for 10 minutes before SDS-PAGE analysis. Non-specific background from plasma proteins observed. Samples were analysed on a 15 % gel and the protein visualised with Coomassie plus stain.

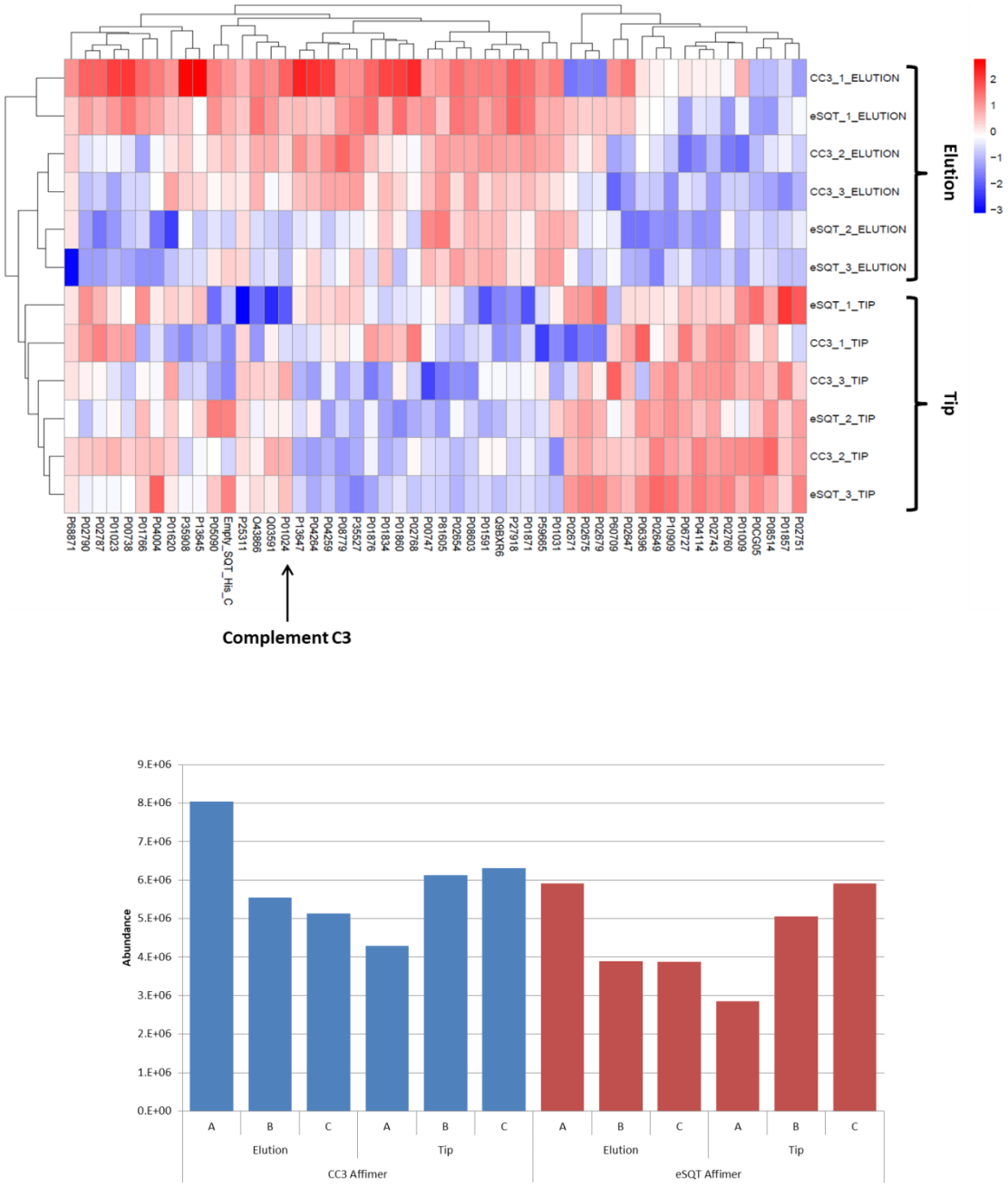


Figure 3.14 | Heatmap with hierarchical clustering of log 10-transformed protein abundance data for MSIA CC3 Affimer and eSQT affinity purification and abundance data for complement C3.

Top: Protein abundance values were taken from Progenesis QI and log 1—transformed. Samples clustered on protein abundance with highly abundant proteins represented in red and less abundant proteins represented in blue. Bottom: Normalised abundance data for complement C3 reveals little difference in enrichment with CC3 Affimer (blue) or eSQT Affimer (red).

3.3.3 Pyridyl disulfide-activated magnetic beads

The method development in this section was carried out using pepsinogen Adhirons due to the lack of availability of other Affimers containing a single cysteine residue. Further information and characterisation of these Adhirons is provided in Chapter 4. A key benefit of the pyridyl disulfide based binding chemistry is that the Affimers can be covalently immobilised onto the beads and as a disulfide bond is formed, the immobilisation is reversible. This would allow for a two-step elution procedure, first by eluting the target from the Affimer and then removing the Affimer from the beads by adding a reducing agent such as DTT to the buffer. This may be of value if sequencing information of naïve Affimers was not available. The general workflow employed for pyridyl disulfide-activated magnetic beads is very similar to the His-tag affinity purification approach outlined in Figure 3.5.

As with the other methods, the first step was to determine the binding capacity of the beads. Based on manufacturer's guidelines, 10 µL of a 10 % bead suspension binds 15 µg of protein. However, as the molecular weight of the protein the calculations were based on was not stated, then the binding capacity of the beads for the Adhirons could differ. Thus, the binding capacity of the beads was assessed with five different protein amounts ranging from 5 µg to 30 µg. Adhirons were immobilised onto the magnetic beads as described in Chapter 2.14 and incubated for 1 hour. The unbound fraction was removed from the beads and retained for analysis. The beads were washed twice with PBS pH7.4 to remove non-specific binding and retained for analysis. The beads, unbound fraction and washes were analysed by SDS-PAGE (Figure 3.15, Top). As discussed in chapter 4, the pepsinogen Adhirons ran on the gel as monomers and dimers and as higher order species. An increase in the intensity of the Adhiron bands was not observed after 10 µg suggesting that the binding capacity had been reached. However, a large portion of the Adhiron was observed in the unbound fraction for all samples regardless of the amount of Affimer used. This may have suggested incomplete binding of the Adhiron to the beads. No Adhiron was observed in the wash fractions indicating no/low levels of non-specific binding of the Affimer to the magnetic beads. Non-specific binding instead of covalent immobilisation of the Adhiron to the beads may result in Adhirons in an undesirable orientation for target capture and leaching of the Adhiron off the beads during sample incubation and thus this finding was preferred. Results for 20 µg and 30 µg Adhiron are not shown as similar results were observed to the 15 µg sample. These findings indicated that the binding capacity of the beads was 10 µg per 10 µL of a 10 % (v/v) bead slurry.

To determine the optimal incubation time for Adhiron immobilisation onto the beads, a binding time-course was carried out following the method described in Chapter 2.14 The beads were incubated with 10 µg Adhiron for either 1 hour, 2 hours, 3 hours or overnight to determine if increasing the incubation time resulted in more Affimer immobilised onto the beads. Following incubation, the beads were separated from the unbound Affimer fraction (labelled UF) using a magnet and washed twice to remove non-specifically bound Affimer. Samples were analysed by SDS-PAGE and the beads were loaded directly onto the gel (Figure 3.15, Bottom). The sample buffer contained DTT to break the disulfide bond between the Affimer and the pyridyl disulfide functional group on the magnetic beads and therefore the Adhirons were eluted off the beads and observed by SDS-PAGE. SDS-PAGE results indicate that increasing the incubation time did not result in more Adhiron immobilised onto the beads as the bands observed representing Adhiron were of similar intensity for all incubation times.

After confirming immobilisation of the Adhiron, enrichment of the Adhirons target was assessed following the initial protocol described in Chapter 2.14. The five different pepsinogen Adhirons were immobilised onto beads and incubated with 2 µg of pepsinogen in PBS pH 7.4. Beads were washed with wash buffer (0.5 M NaCl, PBS pH 7.4) to remove non-specific binding. The bead-Affimer-target complex, unbound pepsinogen fraction and washes were analysed by SDS-PAGE to assess for pepsinogen enrichment. Beads with no Affimer were also analysed to investigate non-specific binding of pepsinogen to beads. Representative SDS-PAGE analysis for pepsinogen Adhiron C2 and beads with no Affimer is shown in Figure 3.16. At approximately 14 kDa, a band was observed in the pepsinogen Adhiron bead lane confirming the Adhiron was immobilised onto the beads. A band was also observed at approximately 47 kDa, the expected molecular weight of pepsinogen suggesting capture of the target. However, a band was also observed at 47 kDa for the bead only control suggesting pepsinogen had bound non-specifically to the beads. Therefore, it could not be confirmed if the Adhirons were enriching for pepsinogen or if the pepsinogen was binding non-specifically to the beads.

Further optimisation was performed to reduce the non-specific binding of pepsinogen to the beads. Enrichment of pepsinogen was performed in a phosphate buffer containing increasing sodium chloride concentrations (250 mM & 500 mM) to assess whether increasing the ionic strength of the buffer would reduce non-specific binding to the resin. However, a decrease in non-specific binding was not observed. In addition, various wash buffers were investigated for effectiveness of reducing non-specific binding of pepsinogen.

The buffers investigated included PBS pH 6.8, 7.2 and 8.4 and PBS pH 7.4 with 0.5 M or 0.75 M sodium chloride. A reduction in non-specific binding of pepsinogen was not observed. As non-specific binding of the Affimer target could not be reduced, further optimisation of this protocol was not carried out.

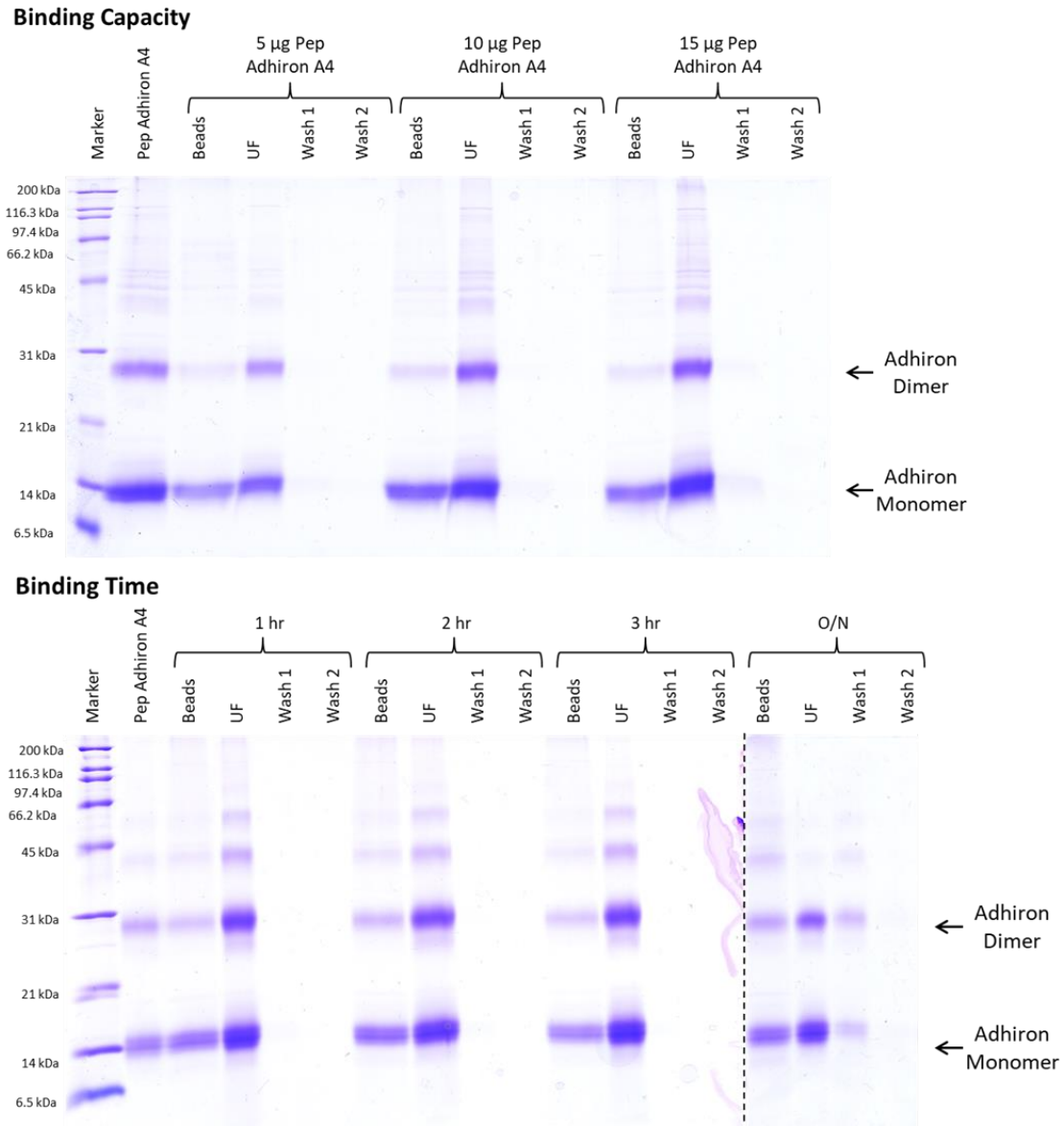


Figure 3.15 | SDS-PAGE analysis of pyridyl disulfide-activated magnetic beads binding capacity and incubation time.

Beads, unbound fraction (UF) and washes analysed by SDS-PAGE. Top: 5 µg, 10 µg or 15 µg of pepsinogen Adhiron A4 was incubated with 10 µL beads (10% suspension) and incubated for 1 hour to immobilise. Binding capacity of the beads was 10 µg Adhiron. Bottom: 10 µg of pepsinogen Adhiron A4 were immobilised onto 10 µL beads (10% suspension) and incubated for either 1 hour, 2 hours, 3 hours or overnight (O/N). Samples were analysed on a 15 % gel and the protein visualised with Coomassie plus stain.

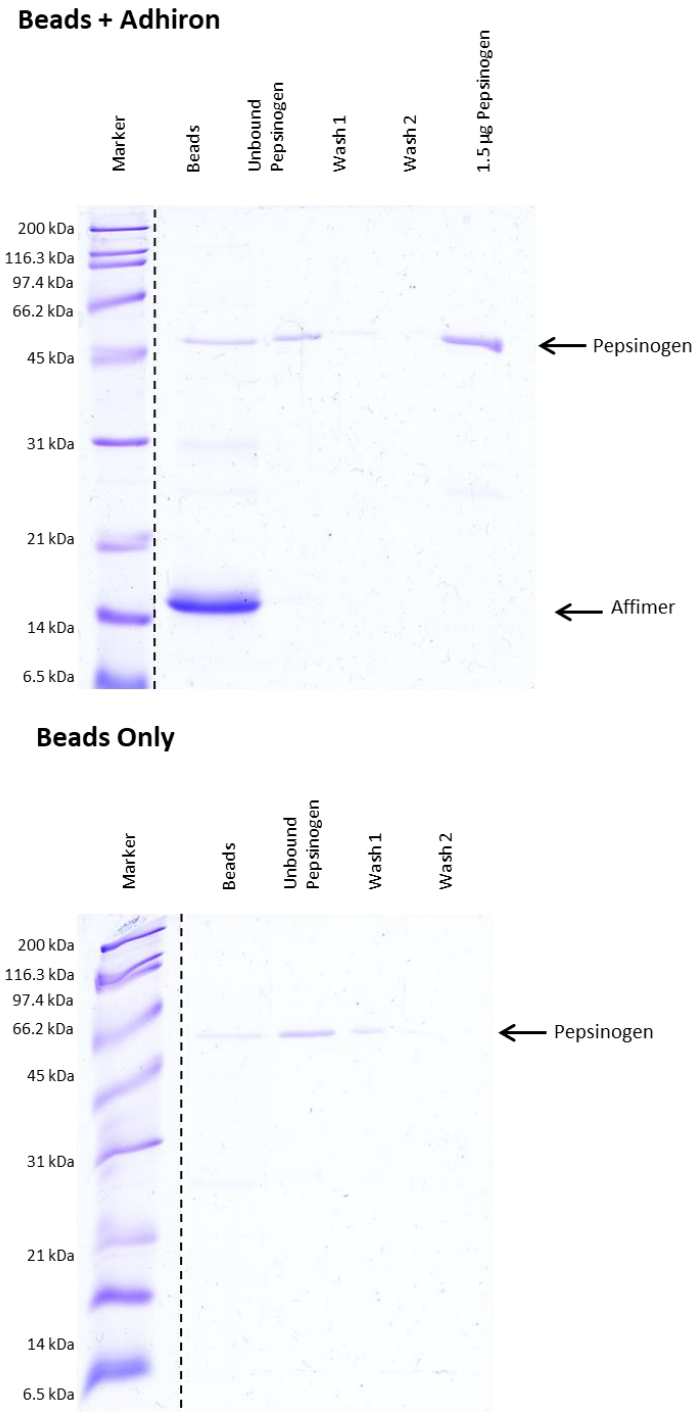


Figure 3.16 | SDS-PAGE analysis of pepsinogen Adhiron pyridyl disulfide-activated beads affinity purification of human pepsinogen.

Top: Pepsinogen Adhiron C2 (10 µg) was immobilised on 10 µL of pyridyl disulfide magnetic beads (10 % suspension) and incubated with 2 µg pepsinogen in 50 µL PBS pH 7.4. Bottom: Bead only control was also analysed to assess non-specific binding of pepsinogen to the beads. Samples were analysed on a 15 % gel and the protein visualised with Coomassie plus stain.

3.3.4 SulfoLink® resin

The final immobilisation technique assessed in this chapter is SulfoLink® resin and this work was performed in collaboration with Avacta Life Sciences by Kimberley Burrow. As discussed in section 3.3.1, the Thermo Scientific SulfoLink® resin is coated with iodoacteyl groups to enable binding of proteins through free cysteine residues. Therefore, the Affimers required a single free cysteine engineered into their protein structure. The general outline of the affinity purification workflow is the same as that described in Figure 3.5. However, as the resin is not magnetic, the particles were separated from solution by centrifugation.

3.3.4.1 Quality Control Check of Affimers

The Affimers used in this study were obtained from Avacta Life Sciences. To assess the purity of the Affimer and to confirm full length protein expression samples were analysed by SDS-PAGE and by intact mass spectrometry (Figure 3.17). SDS-PAGE results confirm high purity of the IgG Affimer with a single band observed at approximately 14 kDa, the expected molecular weight of the Affimer. After deconvolution of the multiply charged protein envelope, the average mass of the Affimer was determined (Figure 3.17). The predominant species in the mass spectrum for was 12840.6 Da which corresponded to the theoretical molecular weight minus the initiating methionine. An additional mass of + 42 Da was also observed which is likely to correspond to acetylation (investigated in Figure 3.26). Both SDS-PAGE and intact mass results confirmed the Affimers was suitable for use in the SulfoLink® workflow.

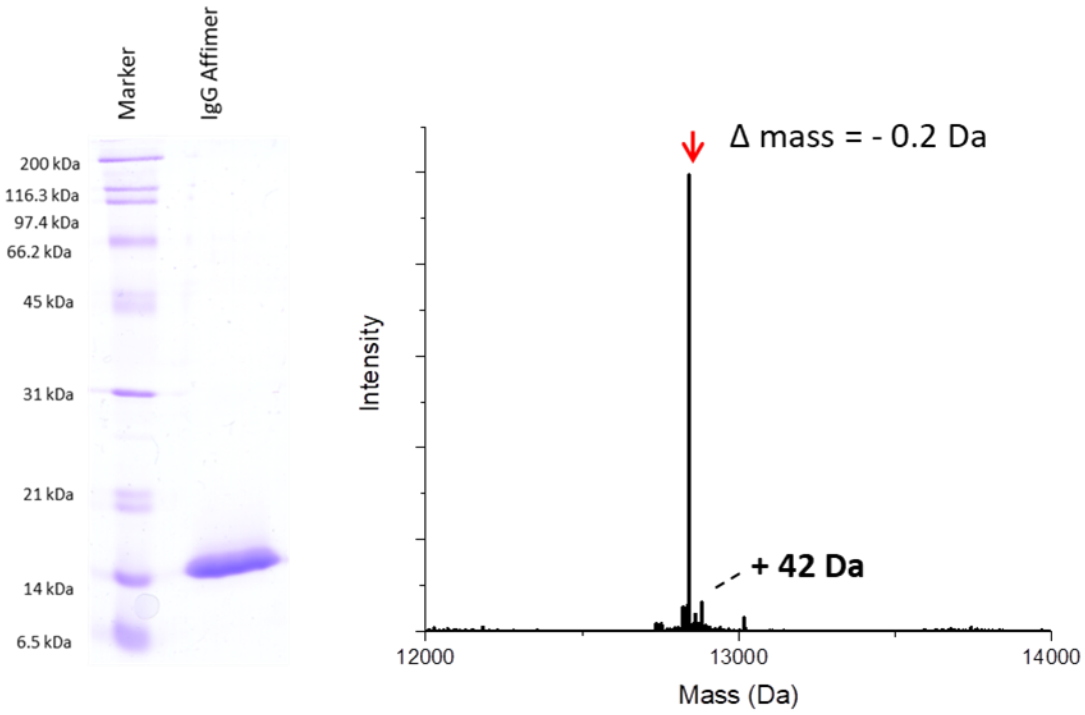


Figure 3.17 | SDS-PAGE and ESI-MS analysis of IgG Affimer.

Left: A total of 2 µg of human IgG Affimer was analysed on a 15 % gel under reducing conditions and visualised with Coomassie plus stain. Right: Protein samples were diluted to 1 pmol/uL in 3 % acetonitrile, 0.1 % formic acid. A total of 4 pmol of Affimer was loaded onto a C4 desalting trapping column. The sample was analysed by ESI-MS on the waters G2 mass spectrometer. The multiply charged protein envelope was deconvoluted using Waters MAXENT 1 algorithm to determine the average mass of the intact protein.

3.3.4.2 Method Development

The SulfoLink® resin is typically used in a column format for affinity purifications of proteins in large volumes and therefore the protocol provided by the manufacturers was adapted to support use within a 0.5 mL tube due to the smaller volumes analysed in this study. The general steps and procedures of the workflow were the same however to separate the resin from incubation solutions, the tubes were centrifuged which brought the resin to the bottom of the tube and the supernatant to the top (method described in Chapter 2.15). Careful pipetting of the supernatant using gel loading tips ensured the resin was not disturbed whilst all the liquid was removed successfully. As with the other immobilisation techniques, the first step was to determine the binding capacity of the resin. Manufacturer's guidelines state the binding capacity of 1 mL of settled resin at 5 mg for human IgG which is equivalent to approximately 33 nmol/mL. As the Affimer is a lot smaller (approximately 13 kDa) theoretically 1 mL of settled resin can bind 433 µg Affimer. To ensure for complete mixing of the resin with the sample, 50 µL of resin slurry was chosen as the volume of resin used for immobilisation which therefore contained 25 µL settled resin. Based on the above calculations, 25 µg settled resin could bind 10.8 µg Affimer.

To assess IgG Affimer immobilisation onto the resin, 15 µg Affimer in 200 µL coupling buffer was incubated with the resin for 45 minutes following the protocol described in Chapter 2.15. To determining the coupling efficiency, unbound fractions, washes and the resin was analysed by SDS-PAGE (Figure 3.18). Each fraction had the same volume and thus by loading equivalent volumes on the gel it was possible to determine if the Affimer sample had been depleted through binding to the resin. The intensity of the protein band in the unbound fraction was a lot less than the Affimer band in the starting material lane suggesting depletion of the Affimer and sufficient immobilisation of the Affimer to the resin (Figure 3.18). A faint band was observed in the first wash lane indicating the removal of Affimer non-specifically bound to the resin or to other Affimers. No Affimer band was visible in the resin lane. This result was expected as the Affimer was covalently bound to the resin and would not elute off during the SDS-PAGE process. LC-MS/MS analysis of resin digested using trypsin confirmed the presence of Affimer peptides (data not shown). Both these findings confirmed the successful immobilisation of the Affimer and the resin was therefore suitable for further affinity purification development.

The IgG Affimer was selected as the preferred Affimer for method development due to the readily and economically available human IgG protein. As this was the intended target of

the IgG Affimer, the next step in development was confirming enrichment of the target with the Affimer using the SulfoLink® resin protocol. To investigate enrichment of human IgG, resin coated with IgG Affimer and resin blocked with cysteine was incubated with 20 µg human IgG was analysed by SDS-PAGE (Figure 3.19). Sample A refers to resin with Affimer bound and sample B refers to resin with no Affimer bound but blocked with free cysteine. To determine if depletion of IgG could be observed, the unbound sample was also analysed. Following removal of the IgG solution, the resin was washed 3 times with PBS pH 7.4 to remove non-specific binding. SDS-PAGE analysis of IgG under reducing conditions reveals three main bands due to the different antibody fragments at 25 kDa, 50 kDa and 150 kDa corresponding to non-reducing IgG. Protein bands corresponding to IgG observed in the unbound fraction for the starting material and sample B were slightly more intense compared to bands visible in sample A (Figure 3.19). This suggested a small level of enrichment of IgG by the resin with Affimer bound. Analysis of resin loaded directly onto the gel supported this result with protein bands corresponding to IgG visible in sample A lane and not in sample B lanes. This finding demonstrated that the SulfoLink® affinity purification method could allow for the enrichment of protein target using the IgG Affimer and that the resin was sufficiently blocked to prevent non-specific binding of IgG.

To assess the level of non-specific binding of proteins, the resin was incubated with one of three samples; 20 µg IgG only (sample A), 20 µg IgG and 300 µg yeast lysate (sample B) or 300 µg yeast lysate only (sample C).Yeast was selected as the biological sample as no endogenous proteins would bind specifically to the Affimer. The resin analysed either had IgG Affimer bound or had no Affimer bound. The affinity purification was performed as described in Chapter 2.15. The resin for all samples was analysed by SDS-PAGE (Figure 3.20). As PBS pH 7.4 was sufficient in removing IgG bound non-specifically to the resin, a PBS pH 7.4 wash was also assessed on this sample set to determine if it was stringent enough to remove non-specific yeast proteins. SDS-PAGE analysis revealed yeast proteins remained bound to the resin after 3 washes with PBS pH 7.4 (Figure 3.20). A similar pattern was observed for resin containing Affimer and for resin with no Affimer. Enrichment of IgG was also observed for samples where IgG was added to the resin with Affimer bound. To establish whether increasing the stringency of the wash buffer would reduce non-specific binding, the experiment was repeated with two washes with a high pH wash buffer (0.1 M Tris-HCl, 0.5 M NaCl, pH 8.0). SDS-PAGE analysis of the resin revealed the non-specific background had been eliminated (Figure 3.20). In the yeast lysate only samples no protein bands were observed. In the yeast and IgG samples no protein bands were observed in the

resin with no Affimer sample but bands were observed in the resin with Affimer bound corresponding to IgG. This result indicates that the high pH wash buffer removes non-specific binding but retains the specific interaction between IgG and the IgG Affimer. The high pH wash buffer was used for future experiments.

A key challenge of the His-tag affinity purification workflow was the highly abundant signal from Affimer peptides following the on-bead digestion of the bead-Affimer-target complex. It was therefore essential to develop a suitable elution protocol that eluted off the target but retained the Affimer on the SulfoLink® resin. The experiment described above was repeated however following the final wash step, the resin was incubated with elution buffer (0.1 M ammonium acetate, 0.5 M NaCl, pH 4.0). To analyse the effectiveness of the elution buffer, both the elution fractions and remaining resin was analysed by SDS-PAGE (Figure 3.21). The elution fractions were neutralised with ammonium bicarbonate prior to SDS-PAGE analysis. Analysis of elution fractions for samples A and B with the resin containing Affimer, revealed bands corresponding to the IgG protein were present indicating the elution buffer was suitable for removing the IgG target from the Affimer. No IgG bands were observed in the elution fractions for the resin with no Affimer again demonstrating no non-specific binding of IgG to the resin. Bands representing yeast proteins were not observed in any of the elution samples indicating the wash buffer was suitable at removing non-specific binding. Analysis of the resin revealed no IgG protein bands were visible indicating that the elution buffer removed all of the captured IgG (Figure 3.21).

To assess the level of non-specific background in human plasma, the SulfoLink® affinity purification was repeated using human plasma (300 µg). As with the resin exposed to yeast, the resin was washed twice with the high pH buffer followed by a PBS pH 7.4 wash. However, SDS-PAGE analysis of the resin revealed the presence of plasma proteins binding non-specifically to the resin (data not shown). Therefore, the experiment was repeated and the resin was washed four times with the high pH wash buffer and once with PBS pH 7.4. Resin with a pepsinogen Adhiron bound was included in the analysis to confirm the IgG protein was not binding to the Affimer scaffold. The resin was analysed by SDS-PAGE (Figure 3.22). No protein bands were observed in resin samples that had the pepsinogen Adhiron bound or the resin with no Affimer bound. For resin samples that had IgG Affimer bound, IgG bands were observed in the IgG only, plasma and IgG, and the plasma only samples. No other plasma protein bands were observed indicating the wash buffer is sufficient in removing non-specific proteins.

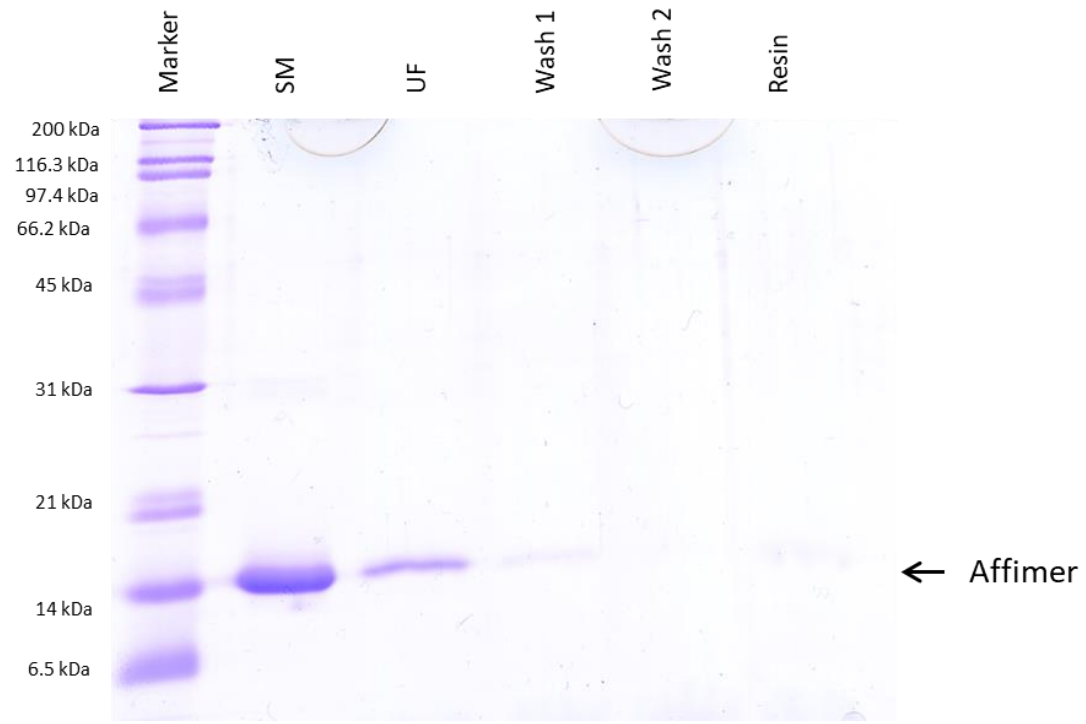


Figure 3.18 | SDS-PAGE analysis of IgG Affimer immobilisation on SulfoLink resin.

IgG Affimer (15 µg in 200 µL PBS pH 7.4) was incubated with 50 µL SulfoLink resin (50 % suspension) for 45 minutes. The resin was washed in PBS pH 7.4 (200 µL) to remove non-covalently bound Affimer. Equal volumes of starting material (SM), unbound Affimer fraction (UF) and washes were analysed by SDS-PAGE. Depletion of Affimer was observed in the unbound fraction indicating immobilisation of Affimer onto the resin. The Affimer was irreversibly covalently bound to the resin and was therefore not observed on the gel. Samples were analysed on a 15 % gel and the protein visualised with Coomassie plus stain.

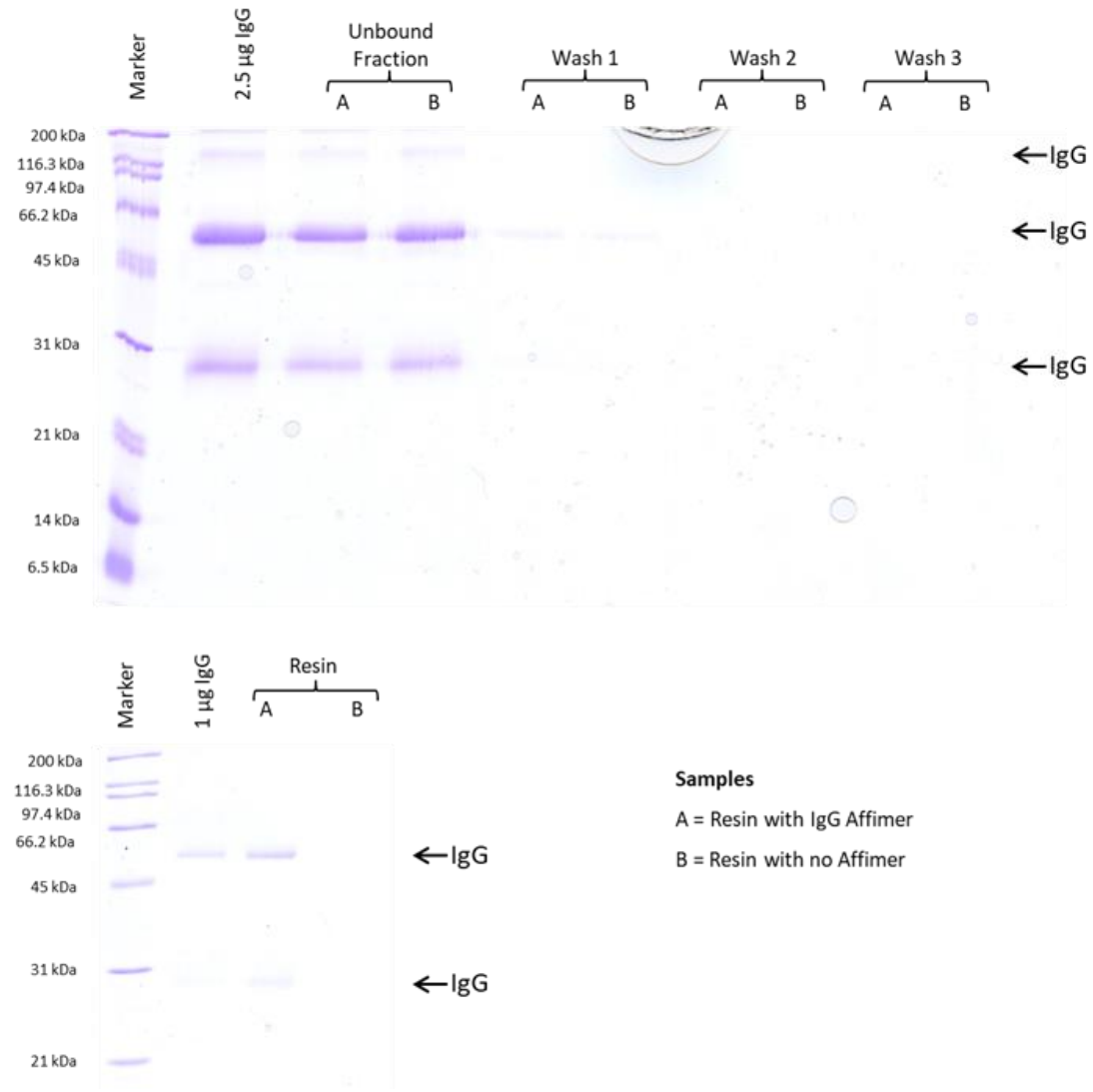


Figure 3.19 | SDS-PAGE analysis of pure IgG affinity purification with IgG Affimer.

IgG Affimer (15 µg) was immobilised onto 50 µL SulfoLink resin (50 % suspension), labelled A on SDS-PAGE. Resin-Affimer complex was incubated with pure IgG (20 µg in 200 µL in PBS pH 7.4) for 1 hour. Resin was washed three times with PBS pH 7.4. A resin control sample containing no Affimer, labelled B on SDS-PAGE, was also analysed to assess for non-specific binding of IgG to the resin. Analysis of resin confirms enrichment of IgG with IgG Affimer and no non-specific binding of IgG to the resin. Samples were analysed on a 15 % gel and the protein visualised with Coomassie plus stain. Affinity purifications and SDS-PAGE analysis undertaken in association with collaborators, Kimberley Burrow from Avacta Life Sciences.

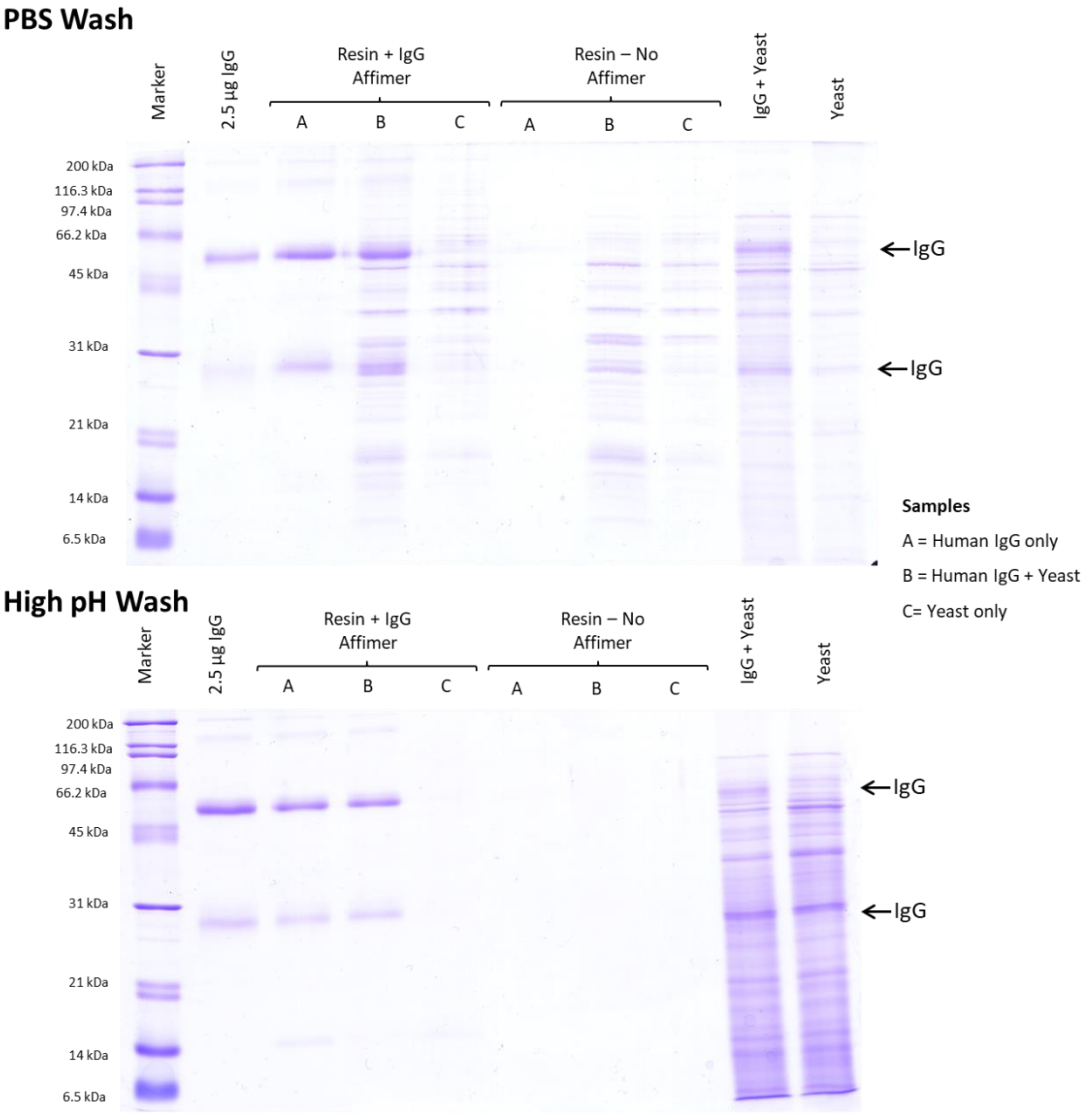


Figure 3.20 | SDS-PAGE analysis of pure IgG in yeast affinity purification with IgG Affimer washing optimisation.

IgG Affimer (15 µg) was immobilised onto 3 X 50 µL SulfoLink resin (50 % suspension), and incubated with either 20 µg IgG (labelled A), 300 µg yeast & 20 µg IgG (labelled B) or 300 µg yeast (labelled C) for 1 hour. Resin with no Affimer bound was also incubated with the above samples. Top: Resin washed with PBS pH 7.4. Bottom: Resin washed with Tris-HCl pH 8. Samples were analysed on a 15 % gel and the protein visualised with Coomassie plus stain. Affinity purifications and SDS-PAGE analysis undertaken in association with collaborators, Kimberley Burrow from Avacta Life Sciences.

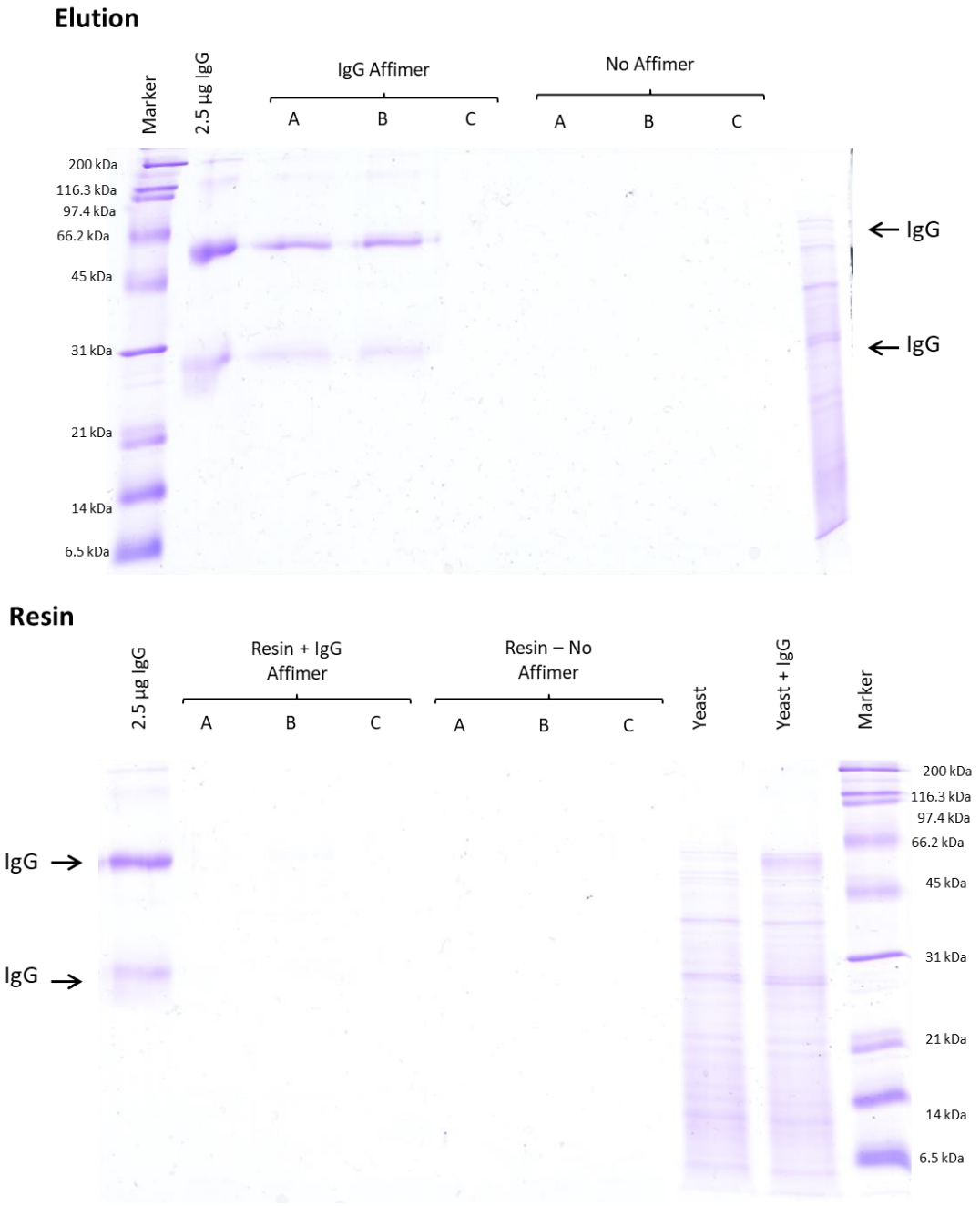


Figure 3.21 | SDS-PAGE analysis elution buffer optimisation of pure IgG in yeast affinity purification with IgG Affimer.

IgG Affimer (15 µg) was immobilised onto 3 X 50 µL SulfoLink resin (50 % suspension), and incubated with either 20 µg IgG (labelled A), 300 µg yeast & 20 µg IgG (labelled B) or 300 µg yeast (labelled C) for 1 hour. Resin with no Affimer bound was also incubated with the above samples. Resin washed with Tris-HCl pH 8. Top: Protein was eluted off the resin using ammonium acetate pH 4 and neutralised with ammonium bicarbonate for SDS-PAGE analysis. Bottom: Resin was analysed to confirm elution of IgG. Samples were analysed on a 15 % gel and the protein visualised with Coomassie plus stain. Affinity purifications and SDS-PAGE analysis undertaken in association with collaborators, Kimberley Burrow from Avacta Life Sciences.

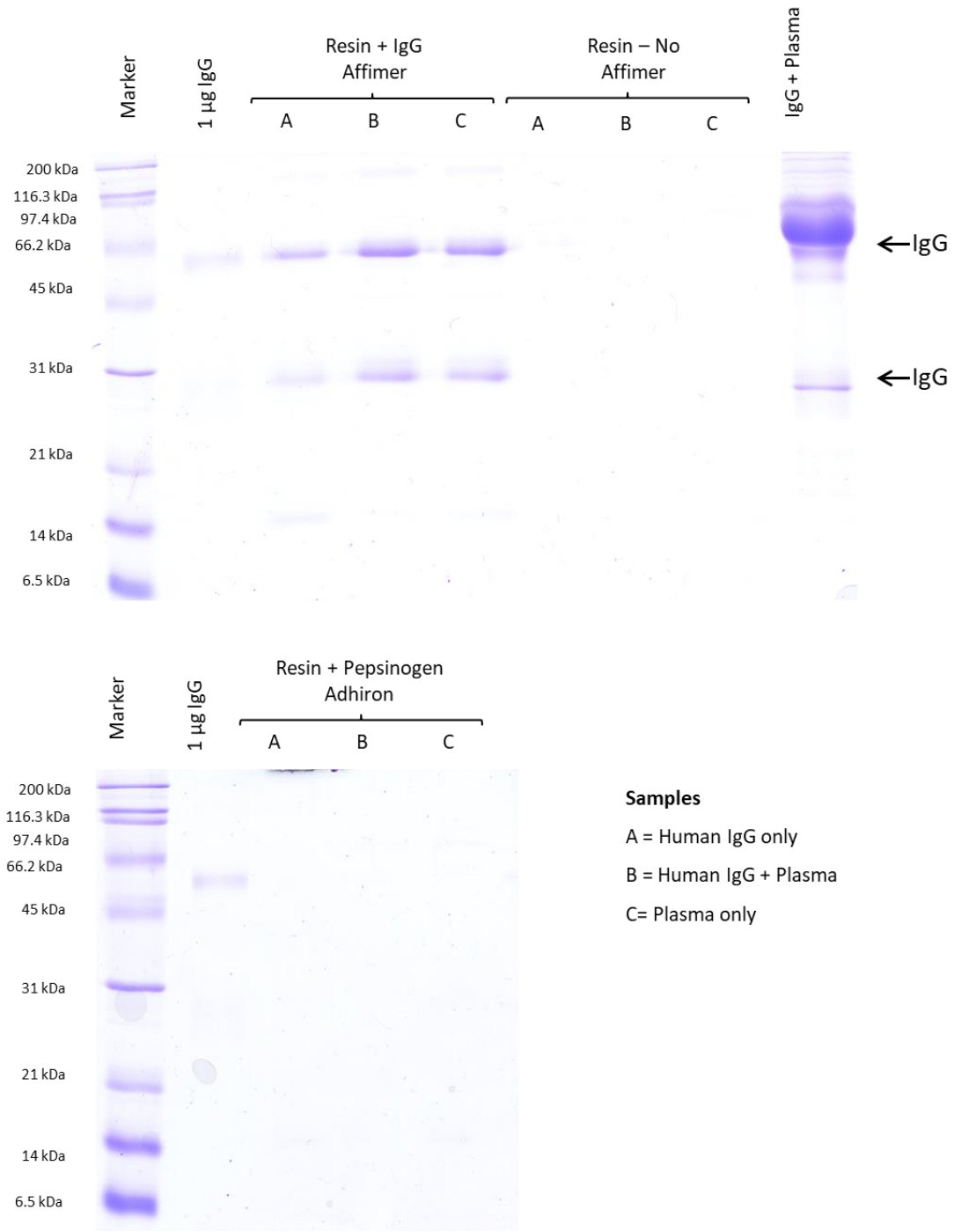


Figure 3.22 | SDS-PAGE analysis of pure and endogenous IgG in human plasma affinity purification.

IgG Affimer (15 µg) was immobilised onto 3 X 50 µL SulfoLink resin (50 % suspension), and incubated with either 20 µg IgG (labelled A), 300 µg plasma & 20 µg IgG (labelled B) or 300 µg plasma (labelled C) for 1 hour. Resin with no Affimer bound and pepsinogen Adhiron was also incubated with the above samples. Resin washed with ammonium acetate pH 4. Samples were analysed on a 15 % gel and the protein visualised with Coomassie plus stain. Affinity purifications and SDS-PAGE analysis undertaken in association with collaborators, Kimberley Burrow from Avacta Life Sciences.

3.3.4.3 Naïve Affimer characterisation and affinity purification

The final affinity purification SulfoLink® resin protocol discussed in Section 3.3.4.2 demonstrated the successful enrichment and purification of human IgG using the IgG Affimer. Therefore, the affinity purification protocol was tested on naïve Affimers with the goal of identification of binding partners using mass spectrometry. It must be noted however that the IgG Affimer was based on the cystatin consensus sequence scaffold and produced using phage display screens which yields Affimers with much greater binding affinities to their target than Affimers based on Stefin A scaffold, those produced for the array work. Therefore, the washes applied post sample incubation may be too stringent and remove the true interaction partners of the Affimers.

The Affimer array work was repeated due to inconsistences with the statistical analysis with the previous array data set obtained for the naive Affimer his-tag affinity purifications. Furthermore, a major issue of the first set of naïve Affimers selected for analysis was that they were chosen based on demonstrating the greatest differential signal between control and sepsis serum samples regardless of the direction of change. Therefore, the Affimers selected were chosen based on preferential binding in the control serum. Patients with sepsis experience a systemic immune response in which proteins belonging to the inflammatory and immune system such as inflammatory cytokines and complement proteins are released[213]. Therefore, it was thought that the most informative proteins and those indicative of sepsis would display an increased expression in the disease cohort. Therefore, the naïve Affimers selected for analysis using the SulfoLink® resin protocol had displayed increased binding in the sepsis cohort. Furthermore, for reason discussed in Chapter 5, plasma is preferred over serum for the identification of sepsis biomarkers.

A total of 13 naïve Affimers were selected for expression and purification (performed by Avacta Life Sciences). All 13 Affimers were chosen as they had the highest signal fold change in the disease cohort. Three of the Affimers failed to express and therefore a total of 10 naïve Affimers were produced for analysis. To compare the Affimer sequences, a multiple sequence alignment was performed (Figure 3. 23). The variable loop regions of the Affimers are highlighted at residue positions 50 to 55 and 80 to 91 for loops 1 and loops 2 respectively. The naïve Affimer sequences were unique. The Adhiron sequences were also analysed by generating a phylogenetic tree. Interestingly, the branch lengths are fairly similar which could indicate low divergence from one another (Figure 5.23). However, as

the Affimers share the same scaffold sequence, this could explain the low divergence observed.

To assess the purity of the naïve Affimers, samples were analysed by SDS-PAGE under reducing and non-reducing conditions (Figure 3.24). Under reducing conditions, a band was observed at approximately 14 kDa in all samples which corresponds to the expected molecular weight of the Affimers. A faint band was observed at approximately 28 kDa for 6 of the Affimers which is likely to correspond to an Affimer dimer. This could be due to insufficient reduction or domain-swapped dimerisation discussed in Chapter 4[211]. Under non-reducing conditions, a band at approximately 14 kDa and 28 kDa was observed for all Affimers representing the monomer and dimer forms. No other protein bands were observed by SDS-PAGE indicating high purity of the Affimers.

To obtain exact mass values and to confirm full length expression of the Affimers, the samples were analysed by intact mass spectrometry as described in Chapter 2.21 The multiply charged protein envelopes were deconvoluted using MAXENT 1 in Mass Lynx, to obtain the average mass of the Affimers (Figure 3.25) and summarised in Table 3.3. For all but one Affimer, the observed mass corresponded to the theoretical mass of the Affimer. An intact mass value was not obtained for naïve Affimer H8. For naïve Affimer A8, D9 and I16 a peak was observed at 103 Da less than the theoretical mass suggesting that a small portion of these Affimers had lost the c-terminal cysteine. Although the cysteine is required for immobilisation onto the resin, only a small fraction of each Affimer was missing the cysteine and therefore it should not have a detrimental effect on Affimer immobilisation.

A + 42 Da adduct was also observed for 8 of the naïve Affimers. To confirm the modification of 42 Da to the naïve Affimers, an in-solution digest was performed on the naïve Affimers using either trypsin or GluC as described in Chapter 2.17. The use of two enzymes would allow for improved sequence coverage and therefore aid in the identification of the modification. Peptides were analysed on the Thermo QExactive mass spectrometer and the data searched using PEAKS 7[214]. The data was searched against a database containing the naïve Affimer sequences using the parameters described in Chapter 2.24. Using the PEAKS PTM node, modified peptides were identified. A representative peptide map of naïve Affimer D9 is shown in Figure 3.26 and identifies the N-terminal peptides as being acetylated. The other nine naïve Affimers are also identified as having this modification which would explain the mass increased observed by ESI-MS.

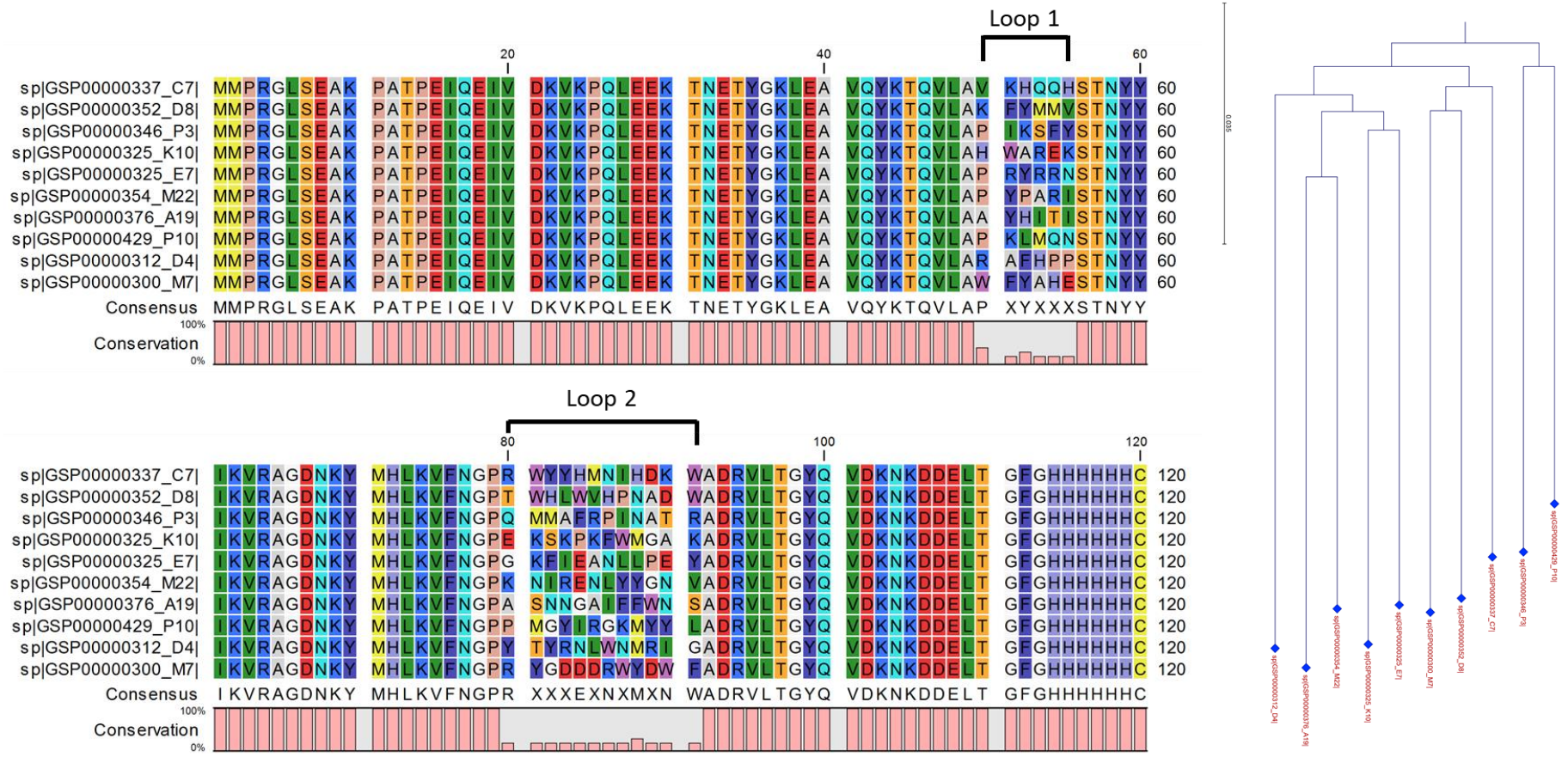


Figure 3.23 | Multiple sequence alignment and phylogenetic analysis of naïve Affimers.

Protein sequences were aligned and visualised using CLC Viewer. Amino acid residues were coloured using the Rasmol colour scheme. The loop regions of the proteins are highlighted at positions 50 to 55 and 80 to 91 for loop 1 and loop 2 respectively.

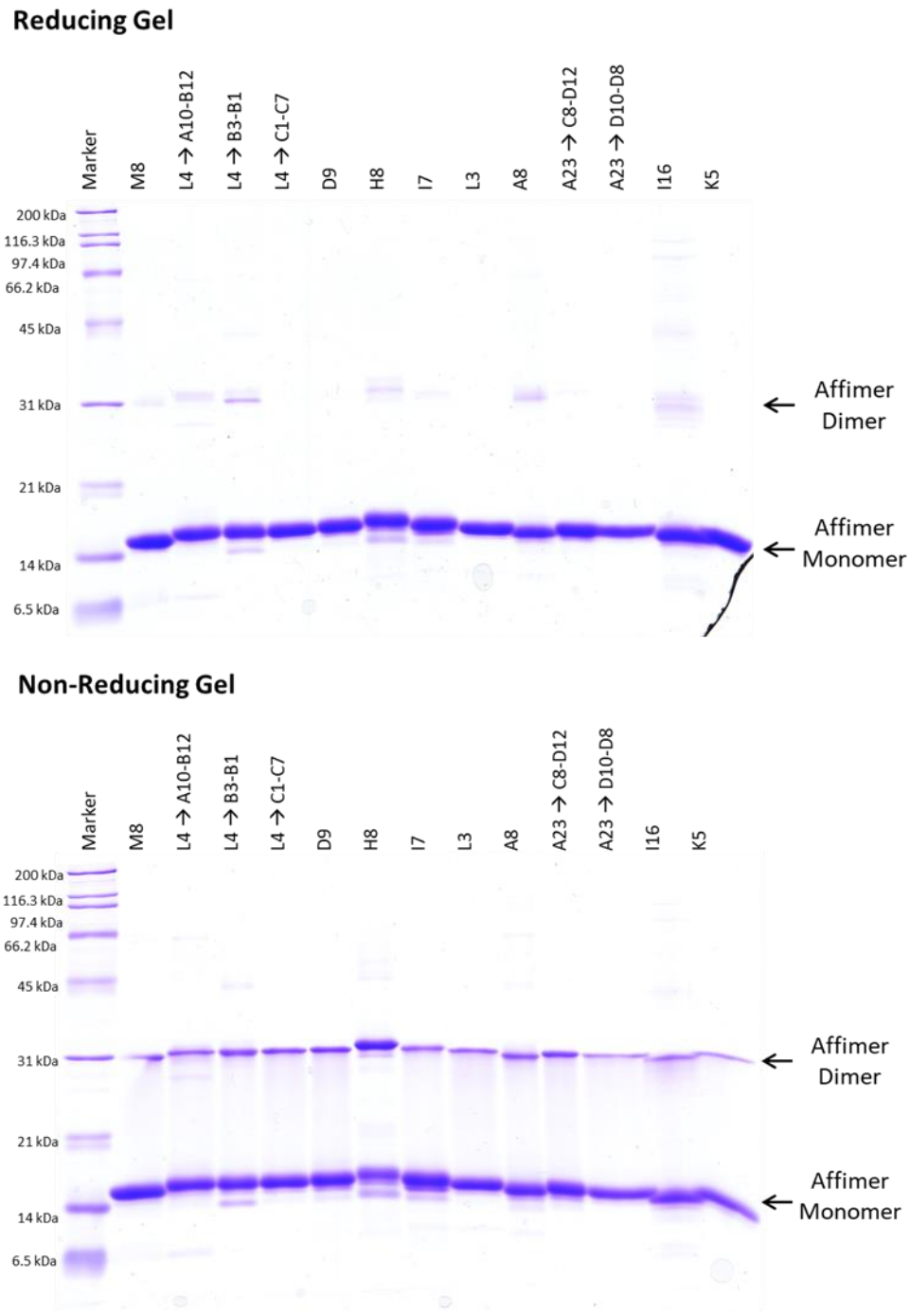


Figure 3.24 | SDS-PAGE analysis of naïve Affimers under reducing and non-reducing conditions.

A total of 2 µg of each Affimer sample was analysed by SDS-PAGE. The reducing gel samples were incubated with sample buffer containing DTT and non-reducing gel samples were incubated in sample buffer than had no DTT added. Affimers L4 and A23 were analysed in triplicate and duplicate respectively due to receiving multiple fractions following purification. The samples were analysed on a 15 % gel and the protein was visualised with Coomassie plus stain.

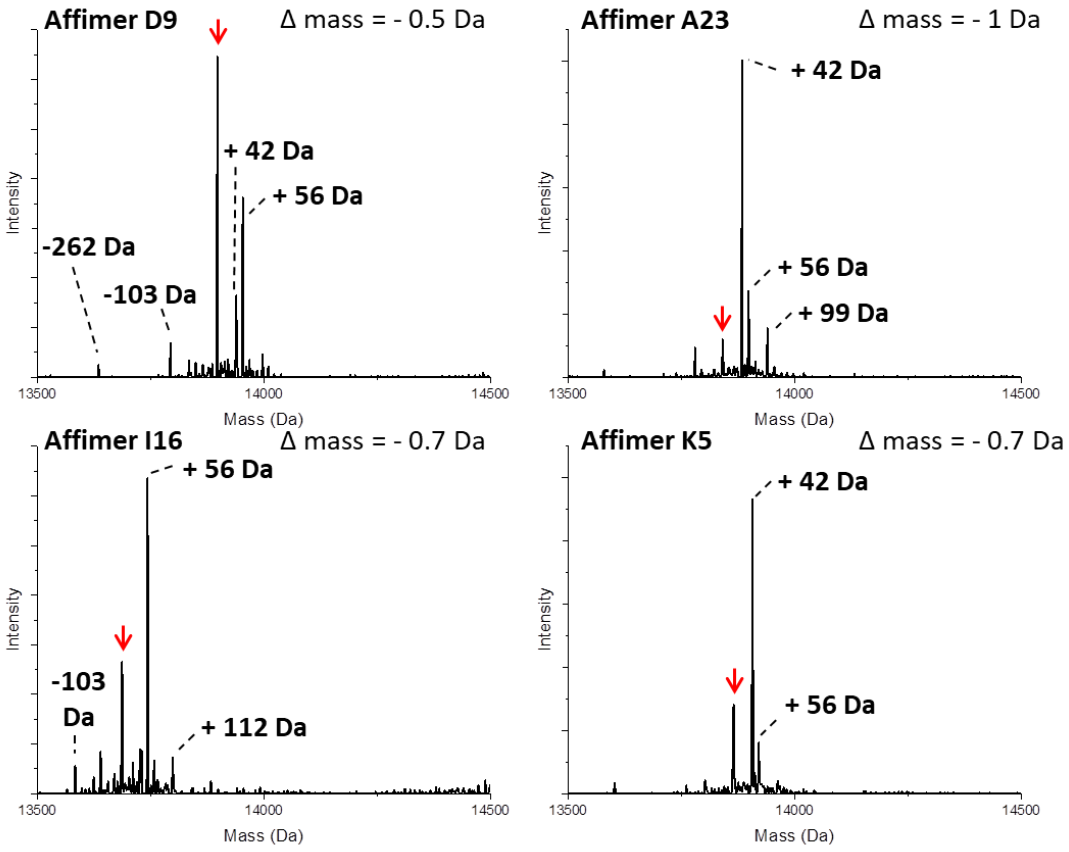


Figure 3.25 | ESI-MS analysis of naïve Affimers (Part 1).

Samples were diluted to 1 pmol/µL in 3 % acetonitrile, 0.1% formic acid. 2 µL of each sample was loaded onto a C4 desalting trapping column. Samples were analysed by ESI-MS on the Waters G2 mass spectrometer. The multiply charged protein envelope was deconvoluted using Waters MAXENT 1 algorithm to determine the average mass of the intact Affimer.

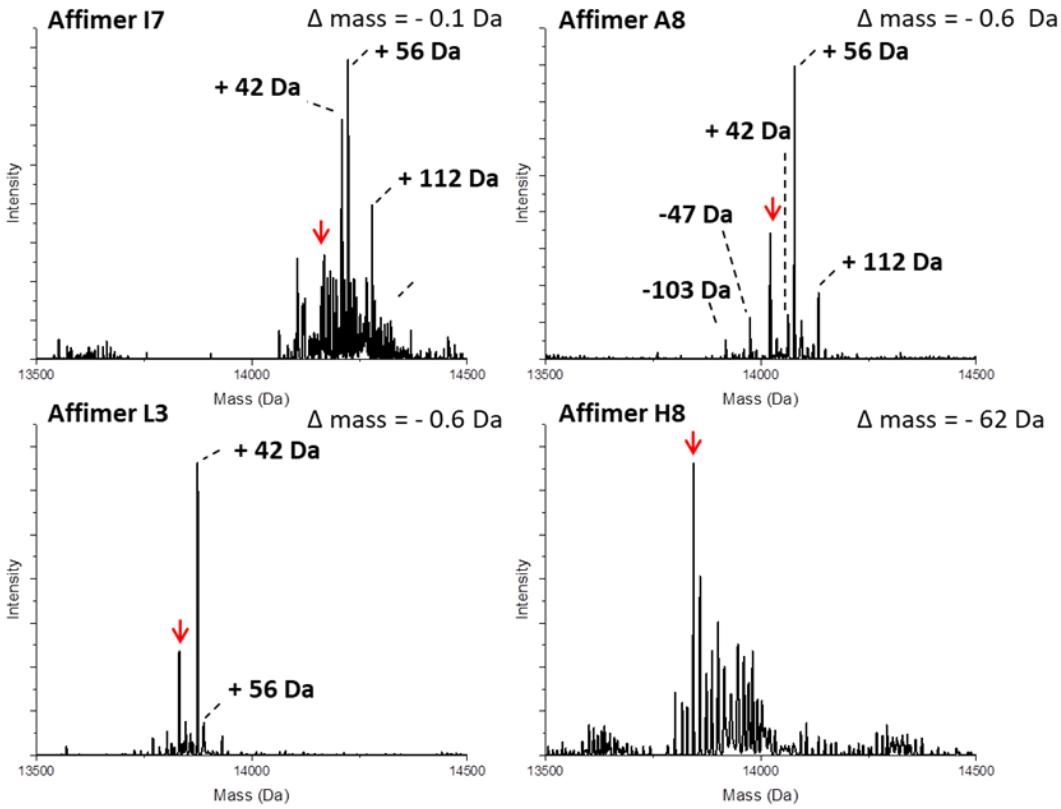


Figure 3.25| ESI-MS analysis of naïve Affimers (Part 2).

Samples were diluted to 1 pmol/µL in 3 % acetonitrile, 0.1% formic acid. 2 µL of each sample was loaded onto a C4 desalting trapping column. Samples were analysed by ESI-MS on the Waters G2 mass spectrometer. The multiply charged protein envelope was deconvoluted using Waters MAXENT 1 algorithm to determine the average mass of the intact Affimer.

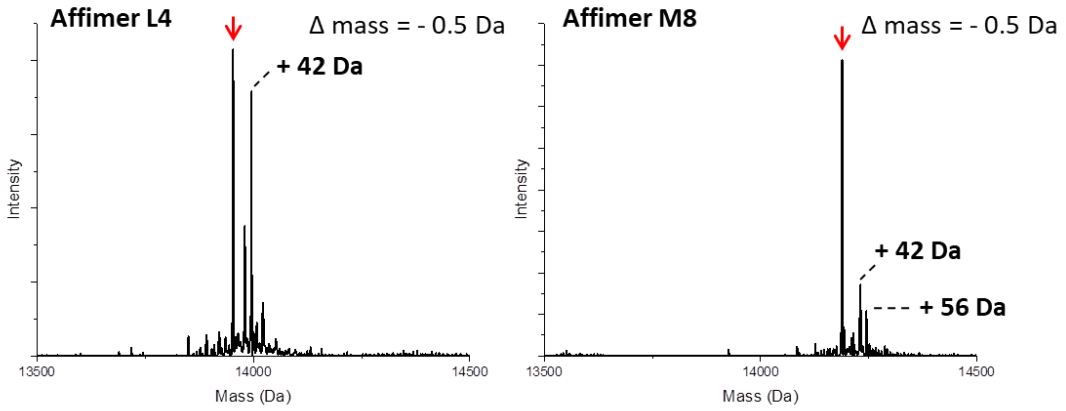


Figure 3.25 | ESI-MS analysis of naïve Affimers (Part 3).

Samples were diluted to 1 pmol/µL in 3 % acetonitrile, 0.1% formic acid. 2 µL of each sample was loaded onto a C4 desalting trapping column. Samples were analysed by ESI-MS on the Waters G2 mass spectrometer. The multiply charged protein envelope was deconvoluted using Waters MAXENT 1 algorithm to determine the average mass of the intact Affimer.

Table 3.3 | Intact mass summary.

The observed mass recorded corresponds to the unmodified intact mass.

Affimer	Theoretical Mass (Da)	Observed Mass (Da)	Difference (Da)
I7	14168.1	14168.1	- 0.1
A8	14023.0	14022.4	- 0.6
L3	13832.8	13832.2	- 0.6
H8	13905.9	13843.9	- 62
D9	13897.8	13897.3	- 0.5
A23	13841.7	13840.7	- 1.0
I16	13687.4	13686.7	- 0.7
K5	13864.9	13864.2	- 0.7
L4	13953.9	13953.4	- 0.5
M8	14188.9	14188.4	- 0.5

Figure 3.26 | Representative PEAKS PTM peptide map for naïve Affimer D9.

The naïve Affimers were separately digested with trypsin and Glu-C to obtained sequencing information. Data was searched using PEAKS against a naïve Affimer database and over 95 % sequence coverage achieved. Acetylation of the N-terminus is likely to explain the + 42 Da mass discrepancy observed via intact mass spectrometry.

The three Affimers that gave the greatest differential signal between disease and control plasma samples on the array were selected for analysis; naïve Affimer D4, E7 and M22. Each Affimer sample was analysed in triplicate following the protocol outlined in Chapter 2.15. Resin with no Affimer bound and a single eSQT sample was included in the analysis as negative controls. As the naïve Affimers were identified from array data as having the greatest differential signal in disease samples, sepsis plasma was analysed in the affinity purification. A pool of 10 patient samples was selected at random and a total of 300 µg incubated with each sample. Following the SulfoLink® resin affinity purification workflow, protein target was eluted using ammonium acetate pH 4. The elution fraction was neutralised with ammonium bicarbonate and along with the remaining resin, digested following standard protocol described in Chapter 2.17 for LC/MS-MS analysis on the Thermo QExactive mass spectrometer. To investigate quantitative difference in protein enrichment, label-free protein quantification was performed using Progenesis QI (Waters Corporation). Retention time alignment was performed with naïve Affimer replicate 2 selected as the alignment file. Alignment scores achieved 66 % or greater. The data was searched using MASCOT against a human and naïve Affimer database. A total of 36 proteins were identified and quantified using unique peptides with a minimum of two peptides per protein. To investigate if the binding partners of the three different naïve Affimers could be identified, protein abundance values were log 10 transformed and analysed by hierarchical clustering (Figure 3.27). No clustering of the replicates of each sample was observed demonstrating no quantified differences in proteins enriched by the naïve Affimers or by the resin and eSQT Affimer. The results demonstrate a consistent level of non-specific background binding in the affinity purification. The resin data was also searched using Progenesis QI but results are not shown as comparable findings were observed to elution data.

Further analysis of proteins identified from the SulfoLink® naïve Affimer affinity purification revealed highly abundant plasma proteins and skin proteins such as keratins (Table 3.4). Due to the various sample handle steps involved with the affinity purification workflow, it was inevitable that contamination from keratins will occur. An approach to overcome the non-specific binding of highly abundant proteins would be to perform depletion strategies on the plasma prior to the affinity purification. However, as the array work that identified naïve Affimers was carried out on non-depleted plasma, applying depletion strategies may result in conflicting findings. In addition, if the targets of the naïve Affimers are low

abundance and weak binders then the binding of non-specific background of highly abundant proteins may interfere with the identification of true naïve Affimer targets.

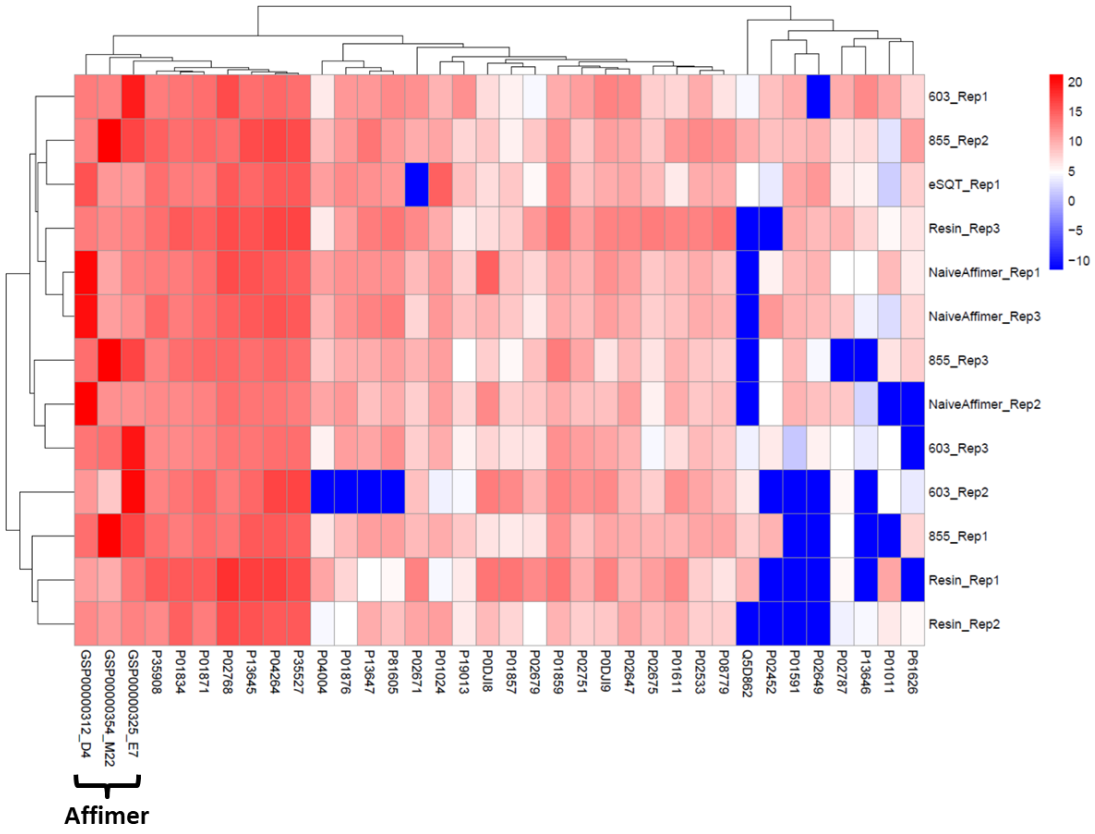


Figure 3.27 | Heatmap and hierarchical clustering of log 10-transformed protein abundance data for naïve Affimer affinity purifications.

Protein abundance values were taken from Progenesis QI and log 10—transformed. Samples clustered on protein abundance with highly abundant proteins represented in red and less abundant proteins represented in blue. The red portion to the left of the heatmap represents proteins that are identified as highly abundant consistent non-specific background. Affinity purifications undertaken by collaborators, Kimberley Burrow from Avacta Life Sciences.

Table 3.4 | Summary of proteins identified from LC-MS/MS analysis of naïve Affimer affinity purifications.

The proteins are ordered based the clustering order in the heatmap in Figure 3.27.

Accession	Protein Description	Peptides	Unique Peptides	Confidence Score
GSP00000312_D4	GSP00000312_D4	16	6	1086.74
GSP00000325_E7	GSP00000325_E7	16	6	1150.87
GSP00000354_M22	GSP00000354_M22	14	4	1033.52
P01011	Alpha-1-antichymotrypsin	3	3	112.61
P01024	Complement C3	11	11	358.15
P01591	Immunoglobulin J chain	2	2	59.43
P01611	Ig kappa chain V-I region Wes	2	2	48.34
P01834	Ig kappa chain C region	4	4	390.16
P01857	Ig gamma-1 chain C region	7	2	225.54
P01859	Ig gamma-2 chain C region	7	3	340.04
P01871	Ig mu chain C region	18	17	771.51
P01876	Ig alpha-1 chain C region	3	3	108.4
P02452	Collagen alpha-1(I) chain	2	2	54.85
P02533	Keratin, type I cytoskeletal 14	16	4	743.88
P02647	Apolipoprotein A-I	6	6	176.26
P02649	Apolipoprotein E	2	2	75.83
P02671	Fibrinogen alpha chain	6	6	173.9
P02675	Fibrinogen beta chain	5	4	175.94
P02679	Fibrinogen gamma chain	2	2	109.11
P02751	Fibronectin	2	2	100.28
P02768	Serum albumin	42	42	2303.33
P02787	Serotransferrin	2	2	43.01
P04004	Vitronectin	2	2	47.78
P04264	Keratin, type II cytoskeletal 1	43	35	2543.76
P08779	Keratin, type I cytoskeletal 16	19	7	794.6
P0DJI8	Serum amyloid A-1 protein	5	2	285.65
P0DJI9	Serum amyloid A-2 protein	5	2	293.41
P13645	Keratin, type I cytoskeletal 10	33	26	2209.02
P13646	Keratin, type I cytoskeletal 13	12	4	442.98
P13647	Keratin, type II cytoskeletal 5	20	9	901.96
P19013	Keratin, type II cytoskeletal 4	8	3	369.96
P35527	Keratin, type I cytoskeletal 9	33	32	1739.18
P35908	Keratin, type II cytoskeletal 2 epidermal	26	16	1447.11
P61626	Lysozyme C	2	2	59.88
P81605	Dermcidin	2	2	98.47
Q5D862	Filaggrin-2	2	2	60.9

3.4 Conclusions

The primary aim of this chapter was to develop an affinity purification method for the identification of naïve Affimer targets using mass spectrometry. A total of four different immobilisation chemistries were assessed for suitability in a naïve Affimer affinity purification including His-tag, streptavidin-biotin interaction, pyridyl disulfide and iodoacetyl chemistry. This work details the challenges of affinity purification and despite extensive optimisation, the methods were unsuitable for the analysis of naïve Affimers.

The SulfoLink resin immobilisation technique based on iodoacteyl chemistry was suitable for IgG affinity purification using the IgG Affimer. However, analysis of naïve Affimers with this approach failed to result in protein identifications. The optimisation step for all the affinity purification methods were performed using a catalogue Affimer which was based on the type II cystatin consensus sequence scaffold and generated through phage display. Whereas the naïve Affimers are based on the type I Affimer stefin A scaffold and have lower binding affinities to their protein target. Therefore, any optimisation steps performed using the Type II scaffold could not be directly applied to the analysis of Type I Affimers.

False positive identification of protein interactions is a common problem with affinity purifications. This is primarily due to non-specific background of proteins binding to the solid support or the capture reagent scaffold[215]. Numerous studies have identified bead proteomes with different proteins binding non-specifically to different bead types[216,217]. Databases such as the CRAPome[215] have been developed in order to establish a repository of previously identified proteins from negative controls. The goal is to help identify and distinguish true interactions from background contamination. Highly stringent wash buffers can be applied in affinity purifications to reduce non-specific interactions however it may compromise the interaction of specific targets.

Chapter 4: Development of Adhirons for Enrichment of Pepsinogen

4.1 Introduction

Novel biomarker discovery is a major goal in clinical research to advance disease diagnosis, increase the value of prognostic indicators and to provide targeted therapy[193]. The work discussed in Chapter 3 outlines the potential role of Affimer technology in the discovery of novel biomarkers. Although there is a large focus in clinical research to identify novel biomarkers, in certain cases, protein biomarkers have already been identified but the current assays and technology available prevents the accurate detection of the protein. Various diagnostic tests such as ELISAs rely on the use of antibodies which possess various limitations as discussed in Chapter 1.5. Due to the ability to generate Affimers to any protein target where the recombinant or purified form is available, using phage display technology[111], Affimers can help to overcome some of the technological issues.

An example in which a protein has been previously identified for disease diagnosis is pepsinogen. Pepsinogen has been implicated in various diseases including gastric cancers[218,219], duodenal ulcers[220], rheumatic disease[221] and gastroesophageal reflux disease[222]. More specifically, pepsinogen has been identified as a suitable marker for the diagnosis of reflux aspiration (RA) in children with severe neurodisability[196]. Patients present with recurrent respiratory tract infections, which result in regular hospital admissions and is the leading cause of premature death in these patients[196]. The cause of the respiratory problems is multifactorial such as muscle weakness, poor cough function and both direct and reflux aspiration. RA is defined as inhalation of the stomach content into the airways which differs from direct aspiration (DA) of food and saliva into the airways. Both disorders are common in this set of patients due to lack of motor control of both the oral and stomach sphincter. As patients with DA and RA present with similar symptoms, treatment and management of the disorders can be difficult. Distinguishing between RA and DA would allow clinicians to provide more specialised treatment and to select patients suitable for surgical intervention. Patients with proven RA could benefit from surgery to tighten the oesophagus, preventing aspiration of the stomach content[196].

Pepsin is an aspartic protease and is the major proteolytic enzyme of the digestive system (Figure 4.1), produced as pepsinogen by chief cells that are located in the stomach lining. Pepsinogen is a zymogen in that it is produced as an inactive enzyme with no catalytic activity until it is converted into the active form, pepsin[223]. There are two different groups

of pepsinogens: pepsinogen A (isozymogens pepsinogen A3, A4 and A5) and progastricsin (pepsinogen C) (Figure 4.2). Expression of pepsinogen A is highly specific to the stomach[224] and has not been detected in the oesophagus or lungs[225]. Therefore the detection of pepsinogen in the lungs and oesophagus would identify patients with reflux aspiration[196]. It should be noted that expression of progastricsin has been detected in the lungs.

The amino acid sequence of pepsinogen is split into three regions from residues 1 -15, 16 – 65 and 66 – 388 representing the signal peptide, the activation peptide and the active pepsin moiety, respectively[223] (Figure 4.2). During the synthesis of pepsinogen, the highly hydrophobic signal peptide (prepeptide) is removed and therefore the expressed form of the pepsinogen protein consists of the activation peptide and the pepsin moiety. The amino acid distribution within the two regions of the protein is very distinct; the activation peptide is highly basic whereas the pepsin moiety is highly acidic. Pepsinogen is the inactive precursor of pepsin and is converted into active pepsin upon exposure to acid conditions[226]. At neutral pH, the binding cleft of the pepsinogen structure is occupied by the activation peptide, held in place through hydrophobic, electrostatic and hydrogen bonds, providing stability to the protein[227]. However, at acidic pH the structure is completely different as the activation peptide is cleaved from the pepsin moiety[223]. This cleavage occurs as the carboxyl groups in the protein become protonated under acidic conditions which disrupts the interactions between the binding cleft and the activation peptide. Thus, conformational changes occur which reveal the activation site where the pepsinogen protein cleaves the activation peptide through autolysis[223].

As pepsinogen has been identified as the protein of interest, a targeted SRM approach to detection could be employed. Typically, the protease of choice for protein digestion is trypsin. However, whilst the activation peptide of pepsinogen contains many tryptic cleavage sites, the peptide fragments produced would be very small and it is likely that the activation peptide may be cleaved from the protein. Pepsin, on the other hand, contains four tryptic cleavage sites resulting in extremely large peptide fragments (Figure 4.3). The use of alternative proteases, such as endoproteinase Asp-N, could generate suitably sized peptide fragments (Figure 4.3). Interestingly, an Asp-N digest of pepsinogen would generate a peptide that spans the activation region and the pepsin moiety[228]. Developing SRM transitions that target this peptide could yield valuable information on whether pepsinogen or pepsin has been identified proposing the original source of the protein.

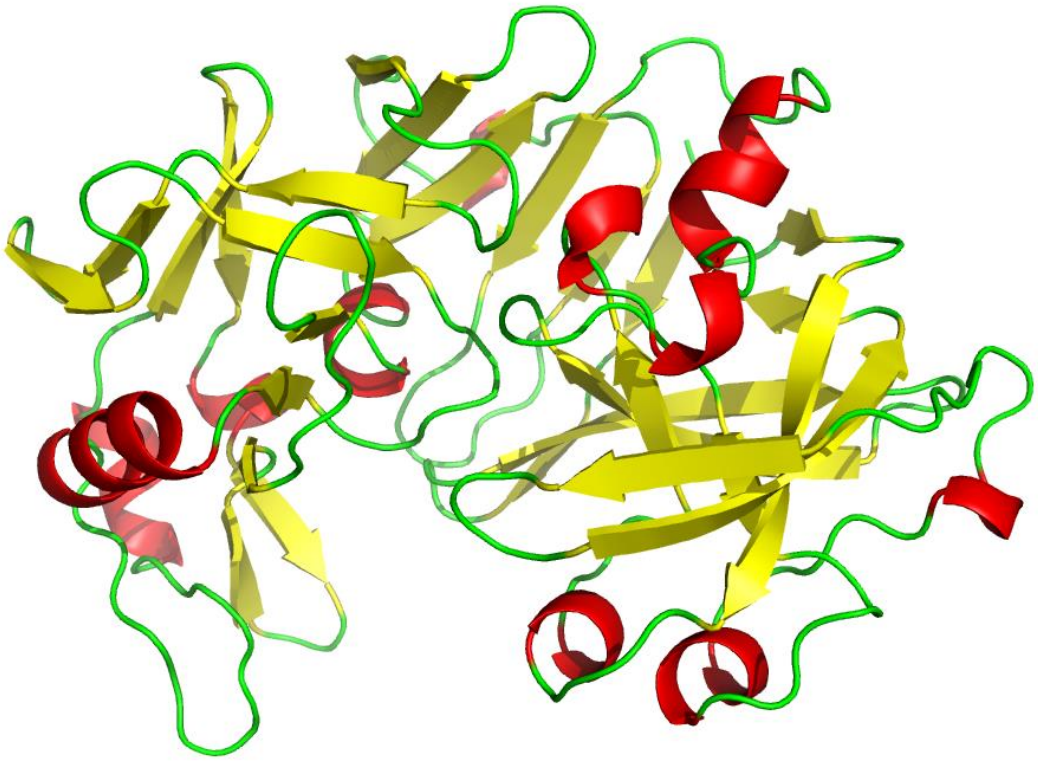


Figure 4. 1 | Tertiary structure of human pepsin.

A ribbon diagram representing the 3D structure of human pepsin 3A (PDB entry: 1PSO). The secondary domains are highlighted with the β-sheets shown in yellow and the α-helix represented in red. The image was generated and visualised using PyMol[191].

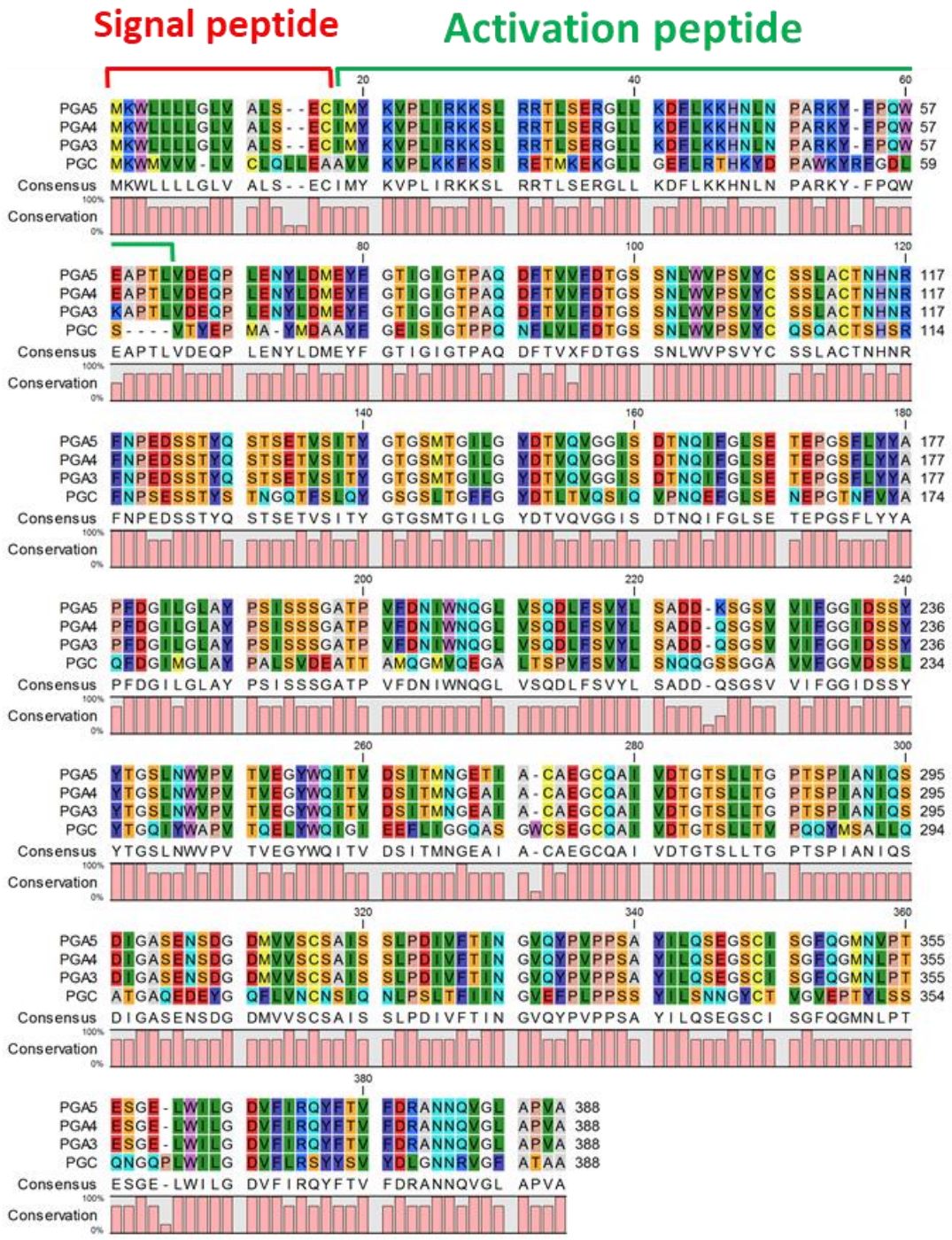


Figure 4.2 | Multiple sequence alignment of human pepsinogens.

The pepsinogen A (PGA-3, PGA-4 and PGA-5) and pepsinogen C protein sequences were aligned and visualised using CLC Viewer. Amino acid residues were coloured using the Rasmol colour scheme. The signal peptide and activation peptide and highlighted.

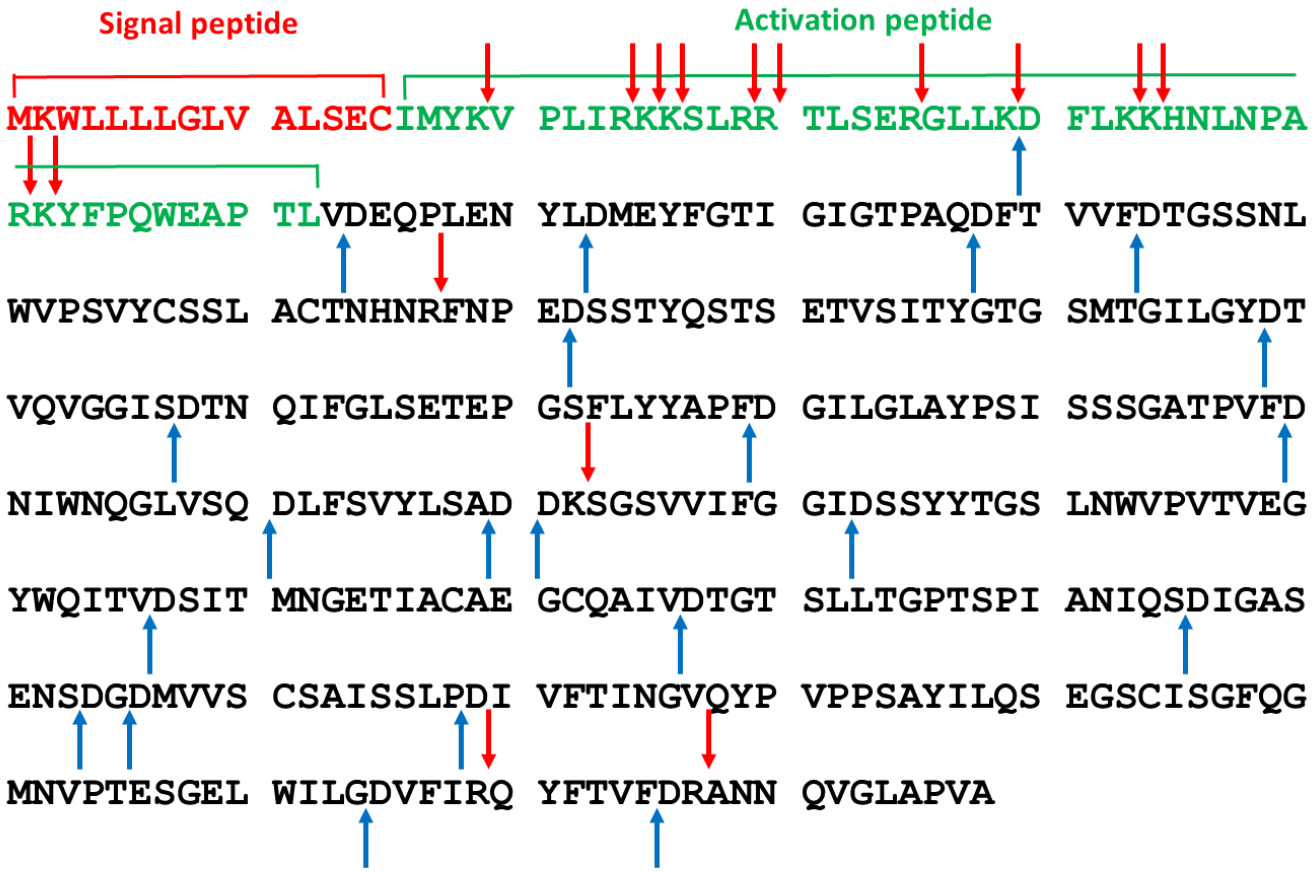


Figure 4.3 | Theoretical trypsin and endoproteinase Asp-N cleavage sites of pepsinogen A5.

The signal peptide and the activation peptide are highlighted on the pepsinogen A5 amino acid sequence. Asp-N and trypsin cleavage sites are indicated by blue and red arrows respectively.

The current assays for the detection of pepsinogen rely on the use of antibodies for ELISA or western blotting. Several commercial ELISA tests are available, including from RayBiotech, Thermo Scientific and Sigma-Aldrich, however they are not fully validated for the analysis of different biological samples. As previously discussed, antibodies possess various limitations such as lack of specificity which may be problematic for pepsinogen detection. Due to the nature of reflux aspiration, bronchoalveolar lavage (BAL) and sputum are the ideal sample types for pepsin detection[196]. Both matrixes are fairly complex which could cause further problems with antibody specificity. Another key challenge for pepsin detection is the low concentration of the protein expected in BAL. Therefore, by enriching first for pepsinogen using affinity reagents followed by targeted SRM, sensitivity and complexity of the sample would be overcome.

To overcome the issue of low pepsinogen concentration, an Affimer affinity purification strategy could be applied to enrich for the target protein. Previously, a SRM based approach was used to detect pepsin with prior enrichment using an anti-peptide antibody developed to bind an Asp-N peptide from pepsinogen[228]. However, despite the method detecting pepsinogen, by enriching for a single peptide, valuable information on the rest of the protein may be lost. Therefore, Adhirons, designed to target the complete human pepsinogen protein, would be the ideal reagent for target enrichment. A major challenge of affinity purifications and co-immunoprecipitations coupled with mass spectrometry is the identification of genuine interacting proteins[215]. Typically, the interacting proteins are of low abundance and if the elution strategy applied elutes the highly abundant capture reagent as well as the interacting proteins, then when digested for MS analysis, peptides from proteins of interest may be below the limit of detection. In addition, if an on-bead digest of the capture reagent and its interacting partners is performed then the same problem arises. Peptides from the capture reagent will limit the amount of sample loaded onto the instrument, preventing detection of the specific interacting proteins. Therefore, development of a protein scaffold that can withstand proteolysis would be advantageous. It would allow for the proteolysis of interacting partners whilst the capture reagent remains intact and thus, separation from the peptide mixture. It must be noted that although the signal of peptides from the capture reagent would be removed, peptides from non-specific proteins binding in the affinity purification would still be present. The appropriate controls would still be necessary to determine true interacting proteins from false positives.

4.2 Aims and Objectives

The work in this chapter will focus of the development and characterisation of Adhiron binders to capture human pepsinogen. They key objectives of this chapter were to:

- Express and purify Adhiron binders that target human pepsinogen.
- Design, express and characterise a pepsinogen Adhiron that is resistant to proteolysis.
- Apply the developed SulfoLink® affinity purification method to capture human pepsinogen using the Adhirons.

4.3 Results and Discussion

4.3.1 Expression and purification of pepsinogen Adhirons

A total of five unique pepsinogen Adhiron plasmids were received from Professor Mike McPherson at the University of Leeds in the BioScreening Technology Group where the phage display screening and selection process was carried out. A brief description of the phage display process can found in Chapter 1.7. The constructs contained a His-Tag sequence for purification. The constructs had been sub-cloned into plasmids that allowed the introduction of a cysteine residue in the C-terminal portion of the proteins. The addition of the cysteine residue is vital for the immobilisation of the Adhirons onto solid supports for affinity purification applications. The five Adhiron sequences are shown in Figure 4.4 and analysis of the loop regions indicate a high proportion of hydrophobic residues. It is therefore likely that the interaction between the Adhirons and pepsinogen is mediated through hydrophobic interactions[229].

The plasmids were transformed into BL21 (λDE3) *E.coli* cells following the transformation protocol (Chapter 2.1). The Adhiron plasmids contain the gene for resistance to ampicillin and thus cells containing the plasmid were selectively grown. Following overnight growth of the transformed *E.coli* cells, a single colony from the agar cultures for each of the pepsinogen Adhirons was selected and used to inoculate a small overnight LB culture. The overnight mini broth culture contained ampicillin to prevent biological contamination. To

increase the scale of protein production, the overnight cultures were used to inoculate 200 mL of LB broth, again containing ampicillin, following the protein expression protocol. The culture growth rate and protein expression post IPTG induction was monitored by removal of culture solution at time points during the growth and analysed by SDS-PAGE (Figure 4.2). After 4 hours, the culture was harvested into cell pellets by centrifugation.

For all five pepsinogen Adhirons, a band at approximately 14 kDa was visible on SDS-PAGE after IPTG induction (Figure 4.5). Based on sequence information, this was the approximate molecular weight for the Adhirons suggesting full length protein expression. For pepsinogen Adhiron A1, a band was also visible pre-induction. In addition, the post-induction band for pepsinogen Adhiron A1 was less intense compared to the other pepsinogen Adhirons.

Before protein purification, the cells were lysed by sonication as outlined in Chapter 2.2. Previous work by Tiede *et al* determined that the Adhiron proteins express in the soluble fraction of the lysate[111]. Therefore following centrifugation, the soluble fraction was taken for purification using GE healthcare Ni-NTA affinity columns using a manual elution as described in the methods. The sample was loaded onto the column and washed with 10 X column volume to remove the non-specific proteins bound to the column. The pepsinogen Adhirons were eluted through competitive elution by increasing the concentration of imidazole in the elution buffer and fractions analysed by SDS-PAGE (Figure 4.6). For all pepsinogen Adhirons, the protein elutes predominately in elution fraction 2. Pepsinogen Adhirons A3 and A4 started to form dimers upon purification due to the second band at approximate 28 kDa (Figure 4.3). Dimer formation was expected due to the unpaired cysteine residues on all of the five pepsinogen Adhirons. This may suggest that the concentration of DTT within the sample buffer is not sufficient in reducing the Adhirons fully. As the elution buffer was not compatible with LC-MS due to the high salt and glycerol concentration, the protein was dialysed into 25 mM ammonium bicarbonate. However, upon dialysis, protein precipitation was observed. Therefore, to prevent further loss of protein, the pepsinogen Adhirons were stored in elution buffer and buffer exchanged when required.

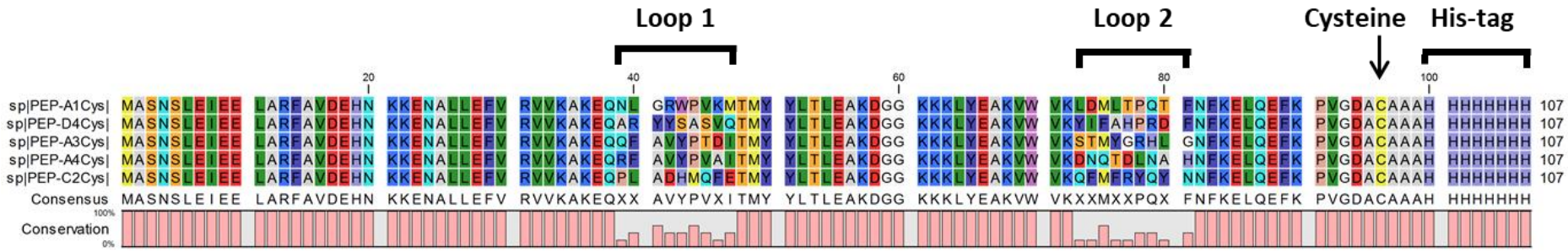


Figure 4.4 | Amino acid multiple sequence alignment of five pepsinogen Adhirons.

Protein sequences were aligned and visualised using CLC Viewer. Amino acid residues were coloured using the Rasmol colour scheme. Variable loop regions of the proteins are evident at positions 39 to 57 and 73 and 81 for loop 1 and loop 2 respectively.

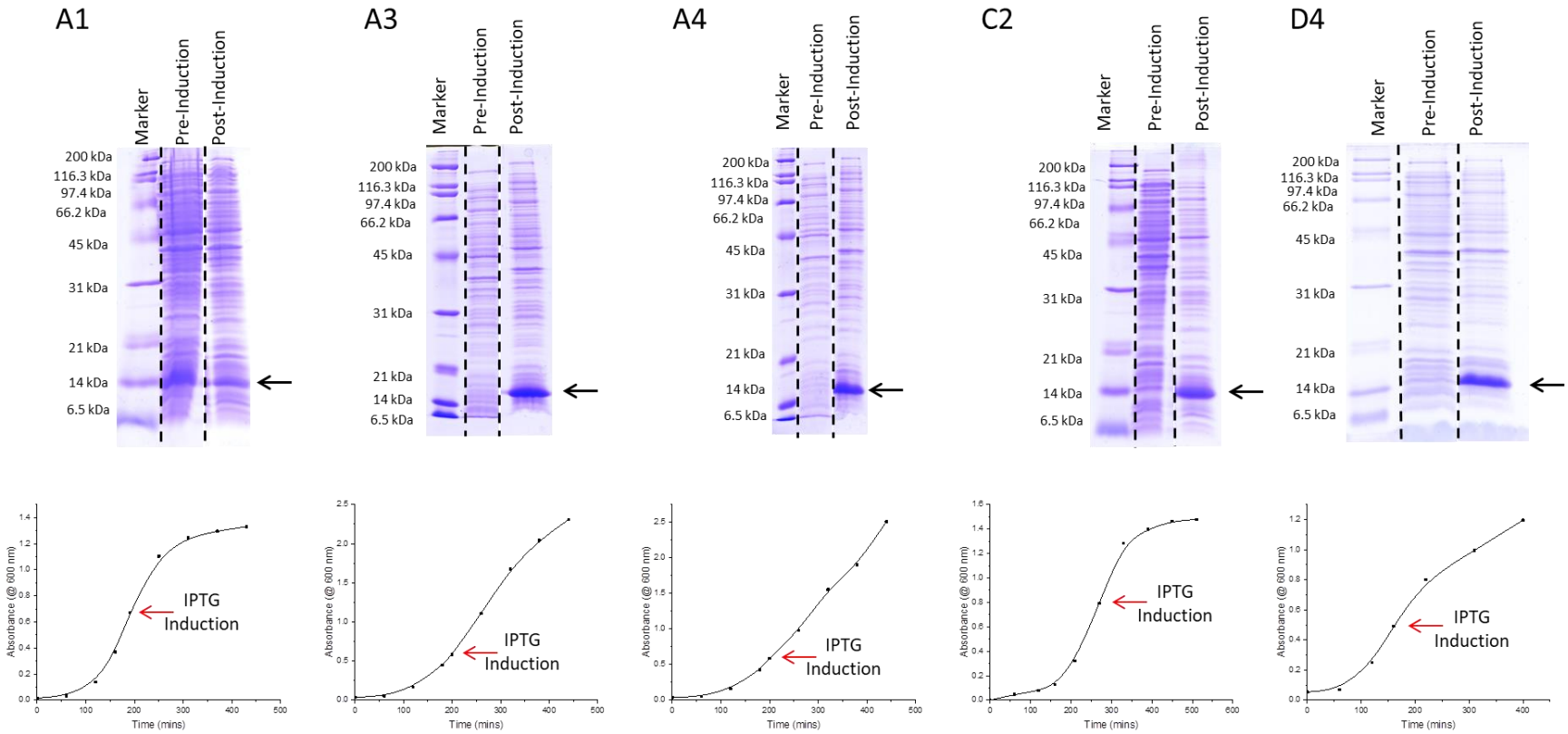


Figure 4.5 | *E.coli* cultures and expression of pepsinogen Adhirons.

SDS-PAGE analysis of pre-induction and post-induction time-points and growth curves for five pepsinogen Adhiron expression. Protein expression was induced with IPTG at an OD=600 nm and protein left to accumulate for 4 hours. All pepsinogen Adhirons expressed as evident by the band at approx. 14 kDa. The samples were ran on a 15 % gel and visualised with Coomassie plus stain.

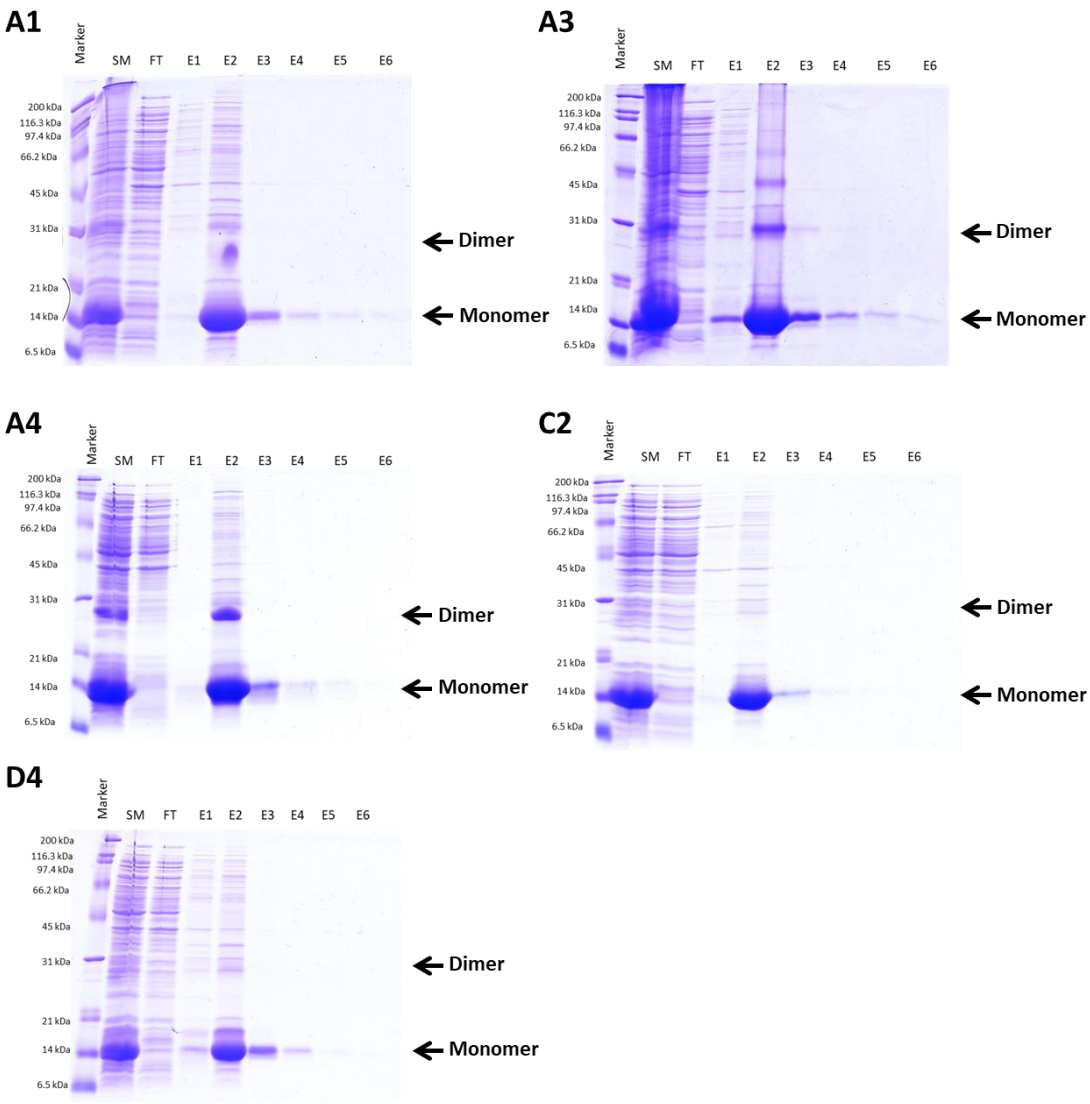


Figure 4.6 | Purification of pepsinogen Adhirons.

His-tag Adhirons were purified using GE Healthcare Ni-NTA affinity columns using a manual elution. Adhirons predominantly elute in elution fraction 2. 10 µL of each fraction was ran on a 15 % gel and visualised with Coomassie plus stain. SM - starting material, FT- flow through and E- elution.

To confirm the products of protein expression and the absence of contaminant proteins, the pepsinogen Adhirons were analysed by intact mass spectrometry as described in Chapter 2.21 (Figure 4.7). Due to the Adhirons being stored in the elution buffer, the samples were first loaded onto an offline C4 desalting trapping column and washed to removed salts and glycerol. After deconvolution of the multiply charged protein envelope, the average mass of the intact pepsinogen Adhirons was determined (Table 4.1). The predominant species within the mass spectrum for pepsinogen Adhirons A1, A4, C2 and D4 correspond to the dimer form of the Adhiron minus two methionine residues (Table 4.1). The mass resolution was not able to accurately confirm whether there was loss of an additional 2 Daltons, corresponding to disulfide bond formation. For pepsinogen Adhiron A3, only the monomer form was observed which corresponded to the theoretical mass minus the initiating methionine residue. Removal of the N-terminal methionine is expected as the residue in position two is alanine[230]. It is likely that pepsinogen Adhiron A3 was observed as a monomer only as it was analysed immediately after purification meaning there was no time for dimer formation to occur. As the pepsinogen Adhirons will be used in their reduced form, it was necessary to confirm the intact mass of the monomer and to confirm effective reduction.

To reduce the five pepsinogen Adhirons, they were incubated with DTT (final concentration 5 mM) for 30 minutes at room temperature. Although glycerol was added to the Adhiron storage buffer to increase protein stability[231], it was incompatible with the mass spectrometer. Therefore, using an offline C4 desalting trapping column, the samples were cleaned-up to remove salt and glycerol. The pepsinogen Adhirons were then analysed by intact mass spectrometry as described in Chapter 2.21. Following deconvolution of the multiply charged protein envelope, the average mass of the pepsinogen Adhiron monomers was measured (Figure 4.8). All five pepsinogen Adhirons had been fully reduced with none of the dimer form of the Adhiron detected. The predominant species within the mass spectra for all Adhirons corresponded to the theoretical mass minus the initiating methionine (Table 4.1). This result confirmed that for all pepsinogen Adhirons the expected, full length protein had expressed and that the Adhirons can be effectively reduced with DTT.

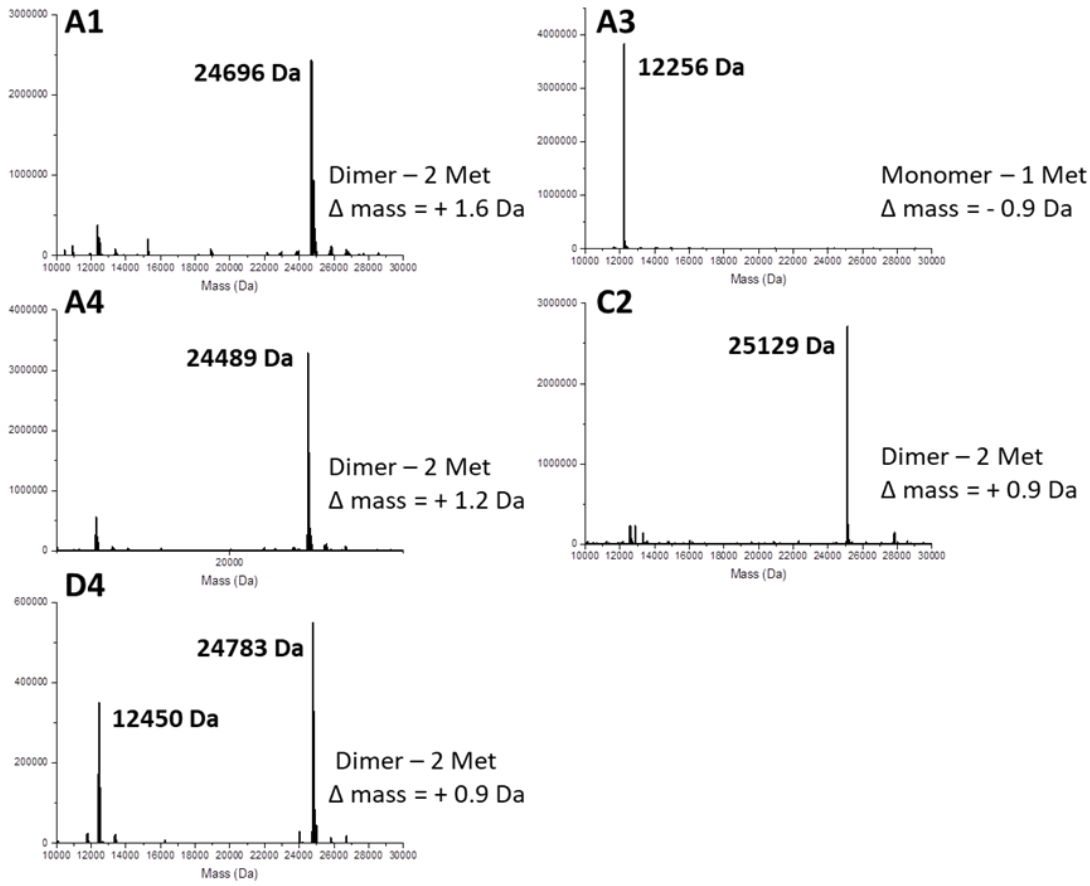


Figure 4.7 | ESI-MS intact analysis of pepsinogen Adhirons without reducing agent.

After offline C4 trap clean-up to remove glycerol and salts, protein samples were diluted to 1 pmol/µL in 3% acetonitrile, 0.1% formic acid. 2 µL of each sample was loaded onto a C4 desalting trapping column. Samples were analysed by ESI-MS on the Waters G2 mass spectrometer. Multiply charged protein envelope was deconvoluted using Waters MAXENT 1 algorithm to determine average mass of intact proteins. – 2 Met corresponds to the loss of the initiating methionine resides from the Adhirons.

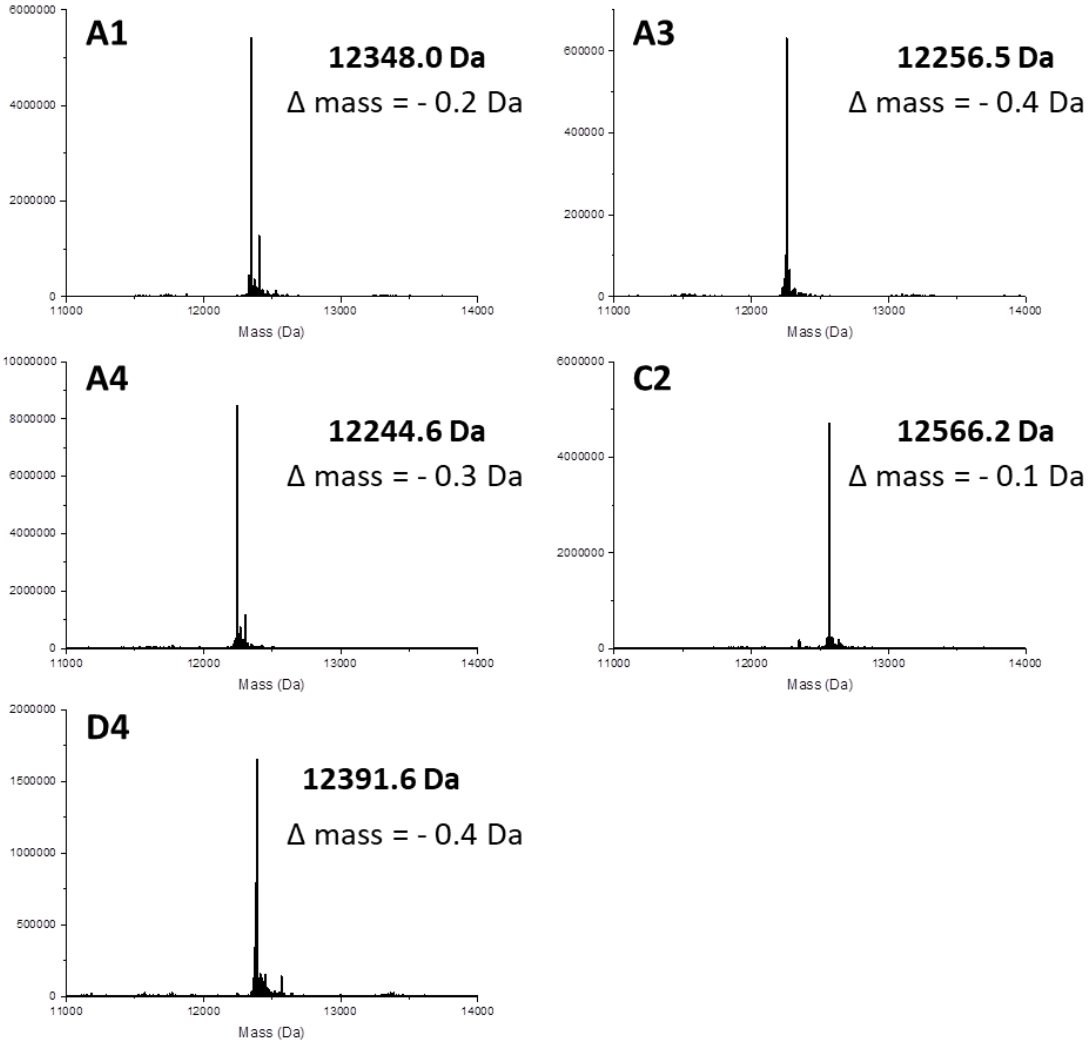


Figure 4.8 | ESI-MS analysis of pepsinogen Adhirons incubated with DTT.

Samples were incubated with DTT to reduce Adhrions. After offline C4 trap clean-up to remove glycerol and salts, protein samples were diluted to 1 pmol/µL in 3% acetonitrile, 0.1% formic acid. 2 µL of each sample was loaded onto a C4 desalting trapping column. Samples were analysed by ESI-MS on the Waters G2 mass spectrometer. Multiply charged protein envelope was deconvoluted using Waters MAXENT 1 algorithm to determine average mass of intact proteins. The observed mass of the pepsinogen Adhiron and the difference from the theoretical mass minus the initiating methionine is presented.

Table 4.1 | Comparison between theoretical and observed intact mass of pepsinogen Adhrions.

Theoretical mass (Da) of monomer calculated minus the initiating methionine. The theoretical mass (Da) of the dimer is calculated minus initiating methionine residues and -2 Da corresponding to the formation of the disulfide bond.

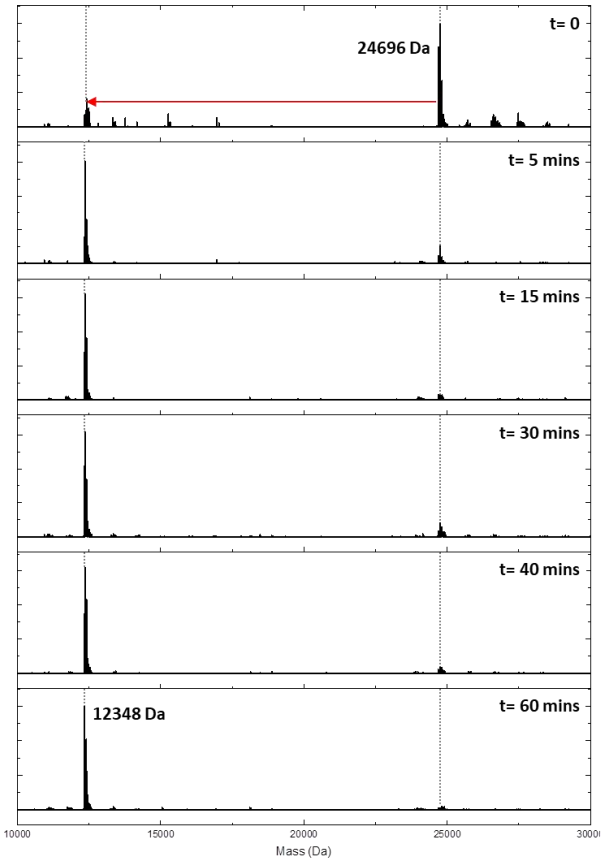
Pepsinogen Adhiron	Expected Mass of Monomer (Da)	Expected Mass of Dimer (Da)	Observed Mass of Monomer (Da)	Observed Mass of Dimer (Da)	Difference (Da) (Observed – Theoretical of Monomer)
A1	12348.2	24694.4	12348.0	24696	-0.2
A3	12256.9	24511.9	12256.5	-	-0.4
A4	12244.9	24487.7	12244.6	24489	-0.3
C2	12566.3	25130.6	12566.2	25129	-0.1
D4	12392.0	24782.1	12391.6	24783	-0.4

For downstream applications, the addition of DTT to the Adhiron buffer was unsuitable. DTT would compete with the cysteine residues on the Adhiron for binding to the sulfhydryl groups on the resin and thus reduce the coupling efficiency of the Adhirons. To overcome this, the Adhiron solution could be desalted to remove the DTT however due to losses in protein with desalting columns, this was not the preferred method. TCEP is an alternative reducing agent that efficiently reduces disulfide bonds[232]. As TCEP is thiol-free then it is not necessary to remove it from the sample before coupling the Adhirons to the resin. However, TCEP was available in an immobilised form and was therefore used so that the sample could be separated from the TCEP by centrifugation preventing the TCEP from entering the MS.

To assess whether immobilised TCEP was suitable for Adhiron reduction and to determine the time taken for complete reduction, a reduction time-course experiment was undertaken. The immobilised TCEP resin was prepared as described in Chapter 2.5. Pepsinogen Adhiron A1 was incubated with the resin and incubated for 60 minutes. During the time-course, samples were removed for intact mass analysis (Figure 4.9). After 60 minutes, the Adhiron was also analysed by SDS-PAGE and compared to unreduced Adhiron (Figure 4.6). Intact mass analysis revealed that pepsinogen Adhiron A4 underwent rapid and

nearly complete reduction after 5 minutes. After 60 minutes, complete reduction was observed with only the monomer form of the Adhiron detectable (Figure 4.9). Conversely, SDS-PAGE analysis suggests that the Adhiron has not fully reduced with monomers, dimers and higher order species as indicated by the protein bands visible on the gel (Figure 4.9). As ESI-MS results indicate complete reduction, this result could suggest that the SDS-PAGE process causes protein complexes. Irrespective of TCEP reduction, the lanes for pepsinogen Adhiron A1 without and with TCEP reduction look identical. A proposed mechanism for dimerisation is through domain swapping. This type of dimerisation does not occur through disulfide bond formation but instead by domain swapping of one protein with another, resulting in multiple species (two or more) grouping together[211]. Human cystatin C has previously been identified to be involved in domain swapping[233]. Although the pepsinogen Adhirons were based on the plant cystatin consensus sequence, their structure is highly homologous to that of the type I Affimer (cystatin A/stefin A sequence) and therefore could explain the observation of multimers. The other suggestion is that the bands observed by SDS-PAGE with a molecular weight greater than 28 kDa (Adhiron dimer) are from contaminating proteins during the purification step.

To determine the identity of the protein bands from SDS-PAGE analysis of pepsinogen Adhiron A1, an in-gel digestion was carried out on specific bands as detailed in Chapter 2.16. The samples were analysed by MALDI-TOF MS as described in Chapter 2.20. To compare TCEP unreduced and TCEP reduced samples, protein bands from both lanes were analysed (Figure 4.10 and Figure 4.11, respectively). Peptides corresponding to pepsinogen Adhiron A1 were identified in bands at approximately 14 kDa, 28 kDa, 42 kDa and 56 kDa in both samples (Figure 4.10 and Figure 4.11). As the molecular weight of the protein bands increased, the sequence coverage achieved decreased which was probably due to the decreasing intensity of the higher molecular weight bands. Despite incomplete sequence coverage of pepsinogen Adhiron A1 at the higher mass ranges, the results are sufficient to confirm that the protein bands are pepsinogen Adhiron A1.



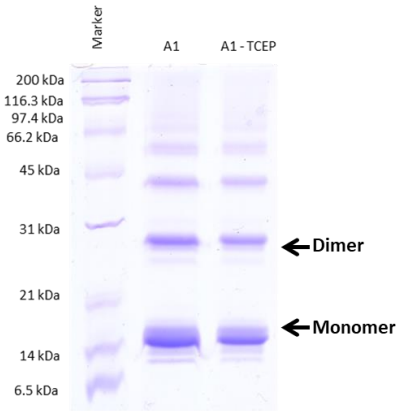


Figure 4.9 | TCEP reduction time-course of pepsinogen Adhiron A1.

Samples were taken during the time-course for ESI-MS analysis. Rapid reduction of pepsinogen Adhiron A1 occurs after the first 5 minutes. SDS-PAGE analysis of pepsinogen Adhiron A1 pre and post 1 hour TCEP reduction reveals comparable protein bands suggesting SDS-PAGE causes the Adhiron to aggregate or the presence of major contaminants.

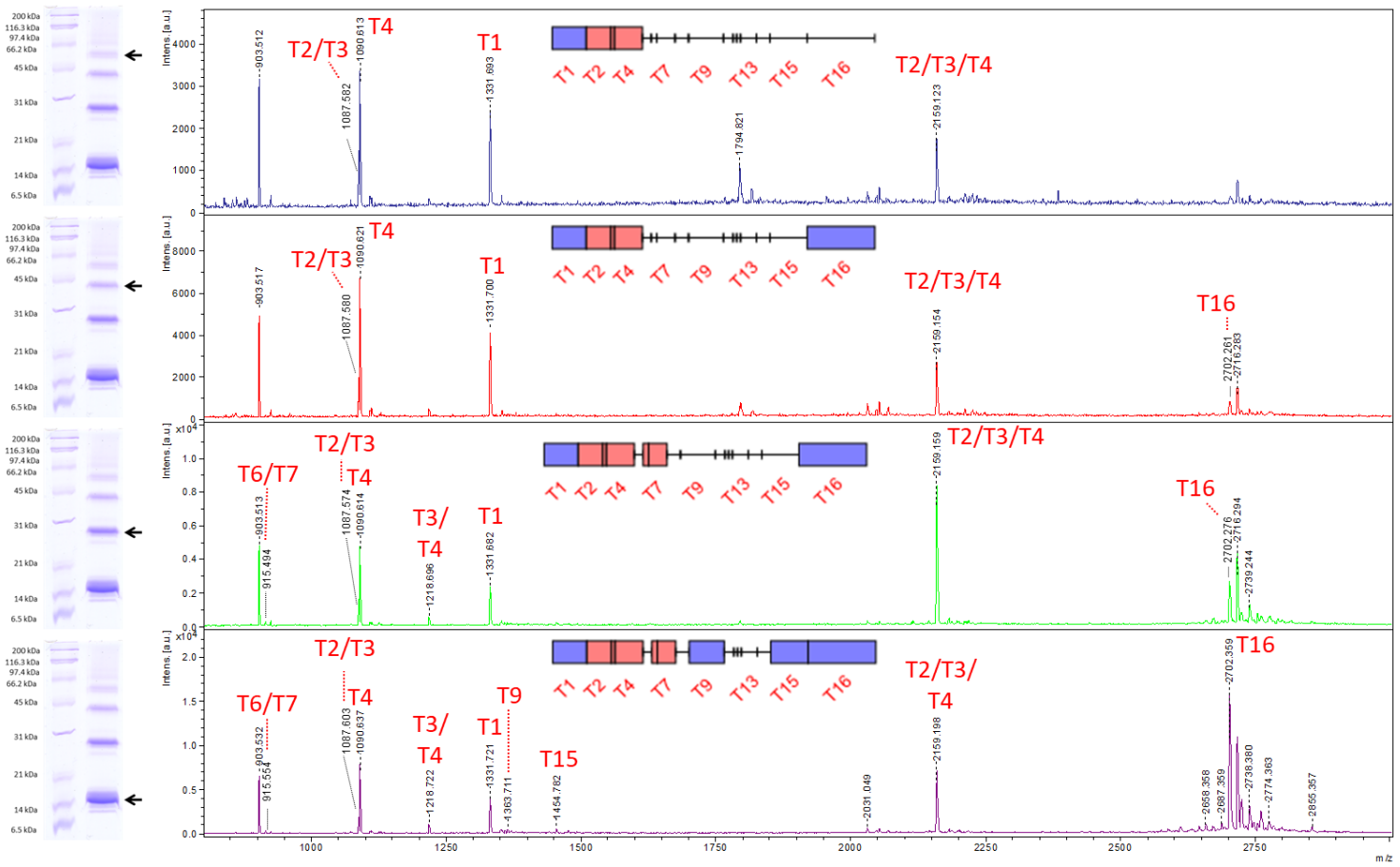


Figure 4.10 | MALDI spectra of in-gel digest of bands in non-reduced Adhiron A1 sample and peptide maps.

Main peaks labelled with corresponding peptide number. Confirmation that the Adhiron forms aggregates when analysed by SDS-PAGE. Increasing protein coverage achieved as the molecular weight of the Adhiron species decreases. Peptide maps generated using Peptide Mapper[234].

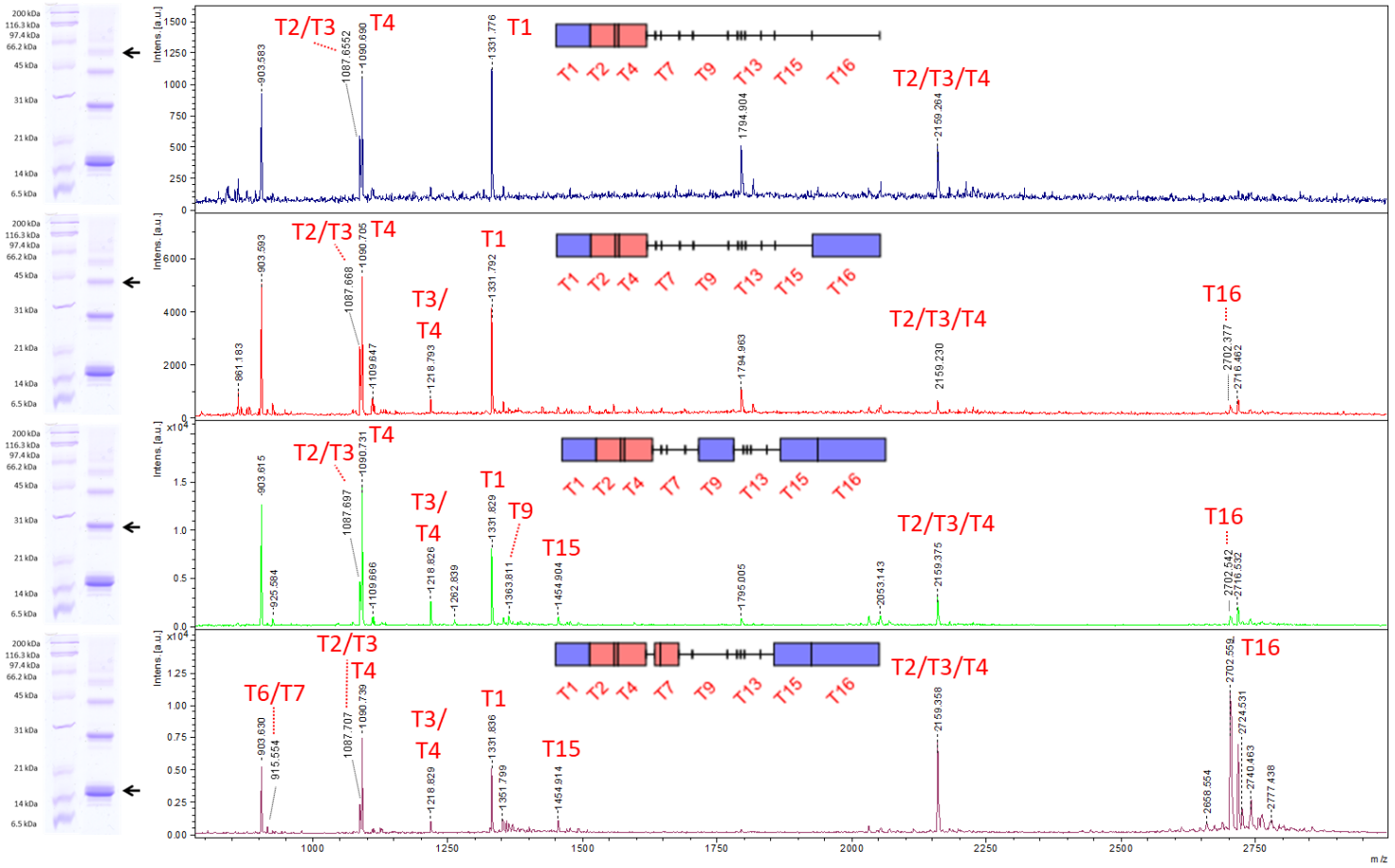


Figure 4.11 | MALDI spectra of in-gel digest of bands in TCEP reduced Adhiron A1 sample and peptide maps.

Main peaks labelled with corresponding peptide number. Confirmation that the Adhiron forms aggregates when analysed by SDS-PAGE. Increasing protein coverage achieved as the molecular weight of the Adhiron species decreases. Peptide maps generated using Peptide Mapper[234].

4.3.2 Design, expression and purification of 'non-digestible' Adhiron

To generate a protein scaffold resistant to Lys-C proteolysis, the lysine residues within the protein would have to be substituted to another residue. The most favoured amino acid residue that shares similar properties is arginine; both amino acids are polar, positively charged residues. Pepsinogen A4 was selected for modification as the protein contained no lysine residues in its variable loop regions. The non-digestible pepsinogen Adhiron A4 was termed pepsinogen Adhiron [A4_K_R] (Figure 4.12).

Pepsinogen Adhiron [A4_K_R] was expressed as described in Chapter 2.1. The gene was synthesised by GeneMill at the University of Liverpool and cloned into a pET11a plasmid and the plasmid was transformed into BL21 (λDE3) *E.coli* cells. The pET11a plasmid contains a gene for antibiotic resistance to ampicillin and therefore only transformed cells containing the plasmid were selectively grown. A single colony was selected from the overnight agar plates and used to inoculate a small overnight LB culture. For larger scale production, the overnight culture was used to inoculate a 200 mL LB broth containing ampicillin. The growth rate was monitored by removal of culture solution at time points during the expression. To investigate culture growth and protein expression, the *E.coli* growth curve was analysed and pre- and post- induction times analysed by SDS-PAGE (Figure 4.13). For both the pre- and post- lanes on SDS-PAGE, a band is present at approximately 14 kDa, although the band did become more intense following IPTG induction. This is the expected molecular weight of pepsinogen Adhiron [A4_K_R] suggesting expression of the correct protein. As with pepsinogen Adhiron A1, the band observed at 14 kDa pre-induction indicated leaky protein expression. The protein was purified as described in Chapter 2.3 using Ni-NTA affinity chromatography using a manual elution. Elution fractions were analysed by SDS-PAGE (Figure 4.13) which determined the protein eluted in elution fraction 2. The protein concentration was determined by Coomassie plus protein assay and as with the other Adhirons, the protein was stored in the elution buffer until needed for downstream analysis.

To confirm expression and purification of the pepsinogen Adhiron [A4_K_R], the protein was analysed by intact mass spectrometry as described in Chapter 2.21 (Figure 4.14). Prior to MS analysis, pepsinogen Adhiron [A4_K_R] was incubated with DTT for 30 minutes to reduce the protein into a monomer followed by sample clean-up using an offline desalting C4 trapping column. The main peak within the mass spectrum of 12581.3 Da corresponded to the theoretical mass of pepsinogen Adhiron [A4_K_R] minus the initiating N-terminal

methionine residue. This result confirmed expression of the correct protein and allowed further characterisation studies on the expressed protein to be carried out.

4.3.3 Characterisation of pepsinogen Adhiron [A4_K_R]

Although pepsinogen Adhiron [A4_K_R] had the expected molecular weight, to further validate and confirm expression of the correct protein with the right substitutions, a proteolysis study was undertaken using either trypsin or Lys-C. As pepsinogen Adhiron [A4_K_R] contained no lysine residues as they were substituted for arginine residues, the protein should be resistant to Lys-C proteolysis. Conversely, pepsinogen Adhiron A4 had 12 lysine residues and therefore the protein should be amenable to digestion with Lys-C.

The Adhirons were digested as described in in Chapter 2.17. Following overnight incubation, the digestion was stopped with the addition of TFA. As TFA can cause undigested protein to form a precipitate, a pre-TFA sample was taken to check for the presence of protein. After the addition of TFA and centrifugation, a large white protein pellet was visible in the pepsinogen Adhiron [A4_K_R] sample digested with Lys-C. This would suggest that the protein has precipitated out of solution and did not undergo proteolysis. To evaluate proteolysis, samples were analysed by SDS-PAGE (Figure 4.15). Pepsinogen Adhiron A4 undergoes complete proteolysis when digested with either trypsin or Lys-C (Figure 4.15). As the protein contains both lysine and arginine residues, this result was expected. Pepsinogen Adhiron [A4_K_R] undergoes complete proteolysis with trypsin however remains intact following incubation with Lys-C (Figure 4.15). As pepsinogen Adhiron [A4_K_R] contained no lysine residues, the cleavage site for Lys-C, the protein was resistant to Lys-C proteolysis. This result confirms the successful expression of a pepsinogen Adhiron that is resistant to Lys-C proteolysis.

To further validate the introduction of the arginine mutations and to compare the difference in peptides using different proteases, the pepsinogen Adhirons were analysed by LC-MS/MS using the Thermo QExactive mass spectrometer. As SDS-PAGE results indicate pepsinogen Adhiron [A4_K_R] did not digest with Lys-C, this sample was not analysed by LC-MS/MS (Figure 4.15). Peptide map analysis of both pepsinogen Adhiron A4 and pepsinogen Adhiron [A4_K_R] demonstrates differences in peptides generated when incubated with different proteases. A Lys-C digestion generates larger peptides due to the fewer cleavage sites. Analysis of BPI chromatograms reveals differences in peptide elution profiles when pepsinogen Adhiron A4 is digested with trypsin or Lys-C (Figure 4.16).

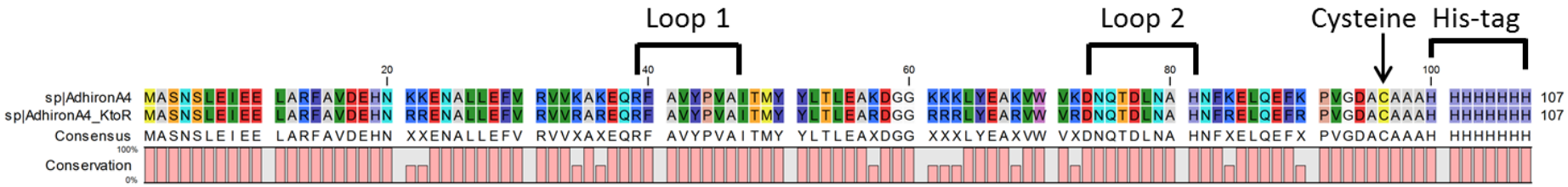


Figure 4.12 | Amino acid sequence alignment of pepsinogen Adhiron A4 and pepsinogen Adhiron [A4_K_R].

Protein sequences were aligned and visualised using CLC Viewer. Amino acid residues were coloured using the Rasmol colour scheme. Variable loop regions of the proteins are evident at positions 39 to 57 and 73 and 81 for loop 1 and loop 2 respectively. The theoretical mass of pepsinogen Adhiron A4 (minus the initiating methionine) was 12244.9 Da and pepsinogen Adhiron [A4_K_R] (minus the initiating methionine) was 12581.0 Da.

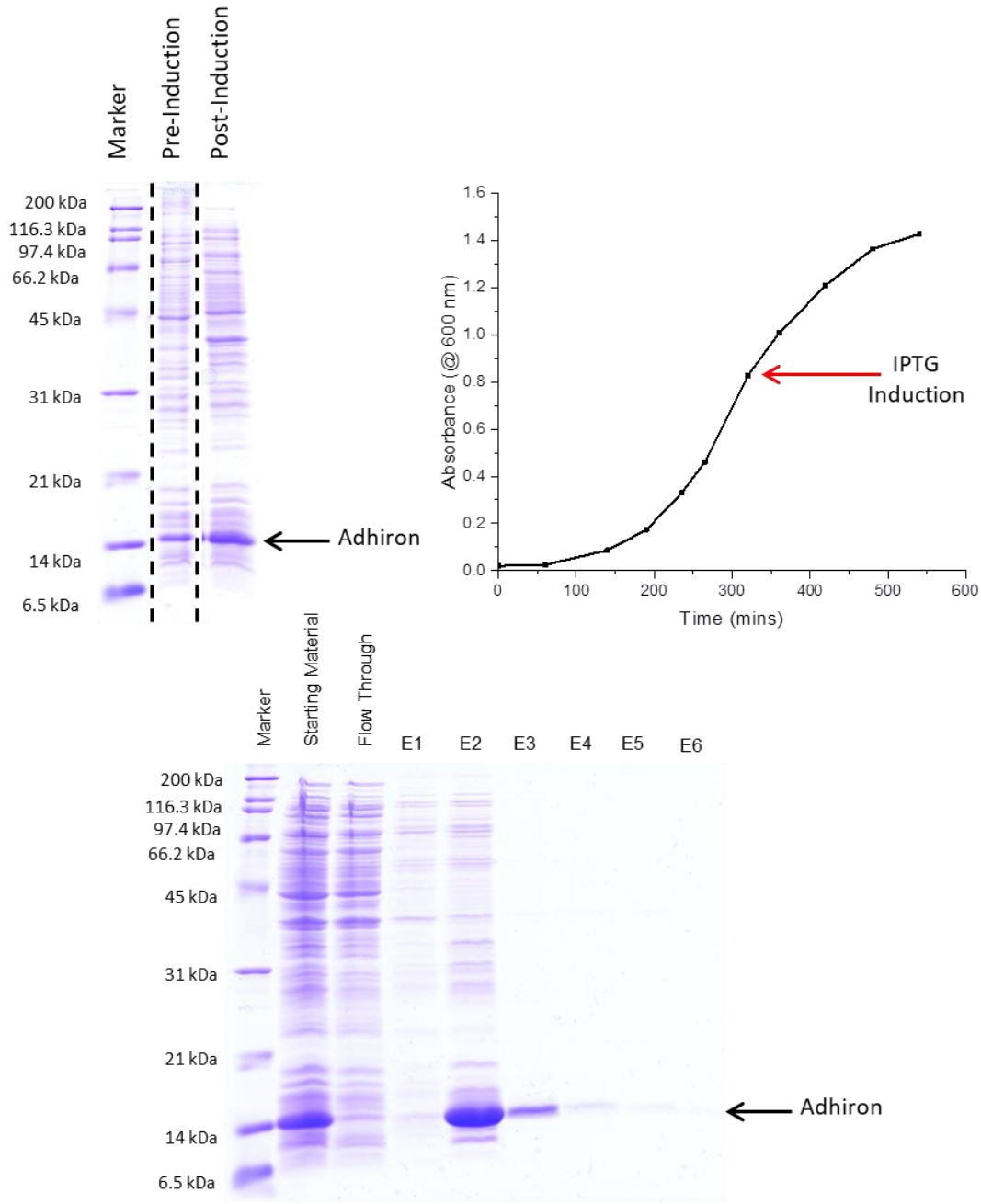


Figure 4.13 | Expression and purification of pepsinogen Adhiron [A4_K_R].

Top Left: SDS-PAGE analysis of pre-induction and post-induction time-points for pepsinogen Adhiron [A4_K_R]. Top Right: Growth curve for pepsinogen Adhiron [A4_K_R] expression. Protein expression was induced with IPTG at an OD = 600 nm and protein left to accumulate for 4 hours. Bottom: His-tag proteins purified using GE Healthcare Ni-NTA affinity columns using a manual elution. Adhirons predominantly elute in elution fraction 2. 10 µL of each fraction was ran on a 15 % gel and visualised with Coomassie plus stain. SM - starting material, FT- flow through and E- elution. Samples ran on a 15 % gel and visualised with Coomassie plus stain.

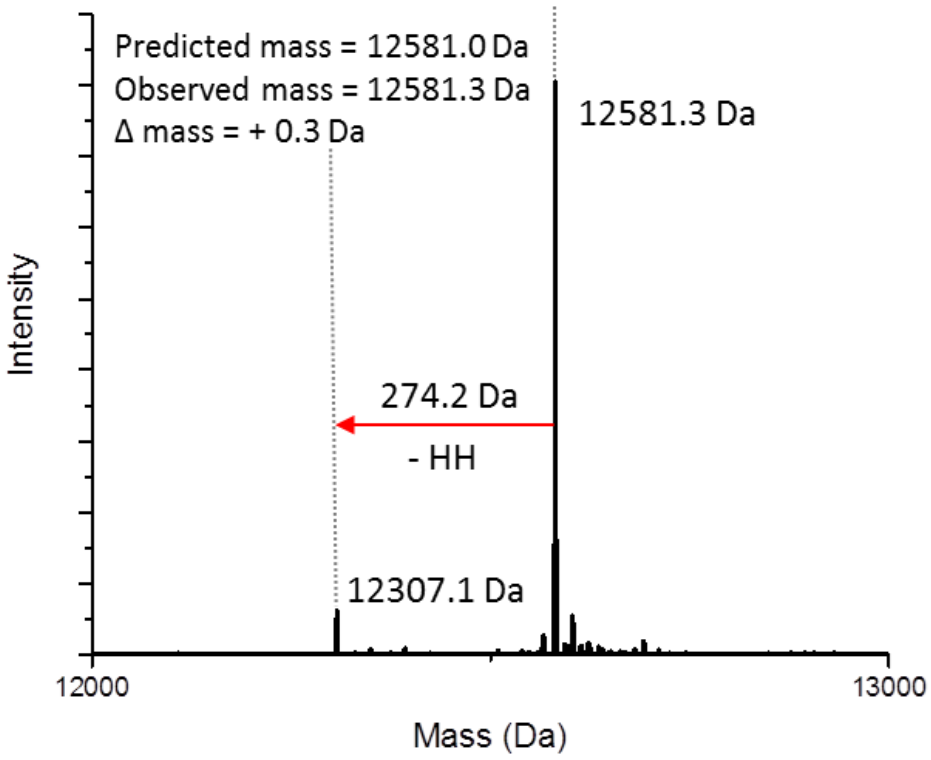


Figure 4.14 | ESI-MS analysis of pepsinogen Adhiron [A4_K_R] incubated with DTT.

The protein was incubated with DTT to reduce the Adhirons. After offline C4 trap clean-up to remove glycerol and salts, protein samples were diluted to 1 pmol/µL in 3% acetonitrile, 0.1% formic acid. 2 µL of each sample was loaded onto a C4 desalting trapping column. Samples were analysed by ESI-MS on the Waters G2 mass spectrometer. Multiply charged protein envelope was deconvoluted using Waters MAXENT 1 algorithm to determine average mass of intact proteins.

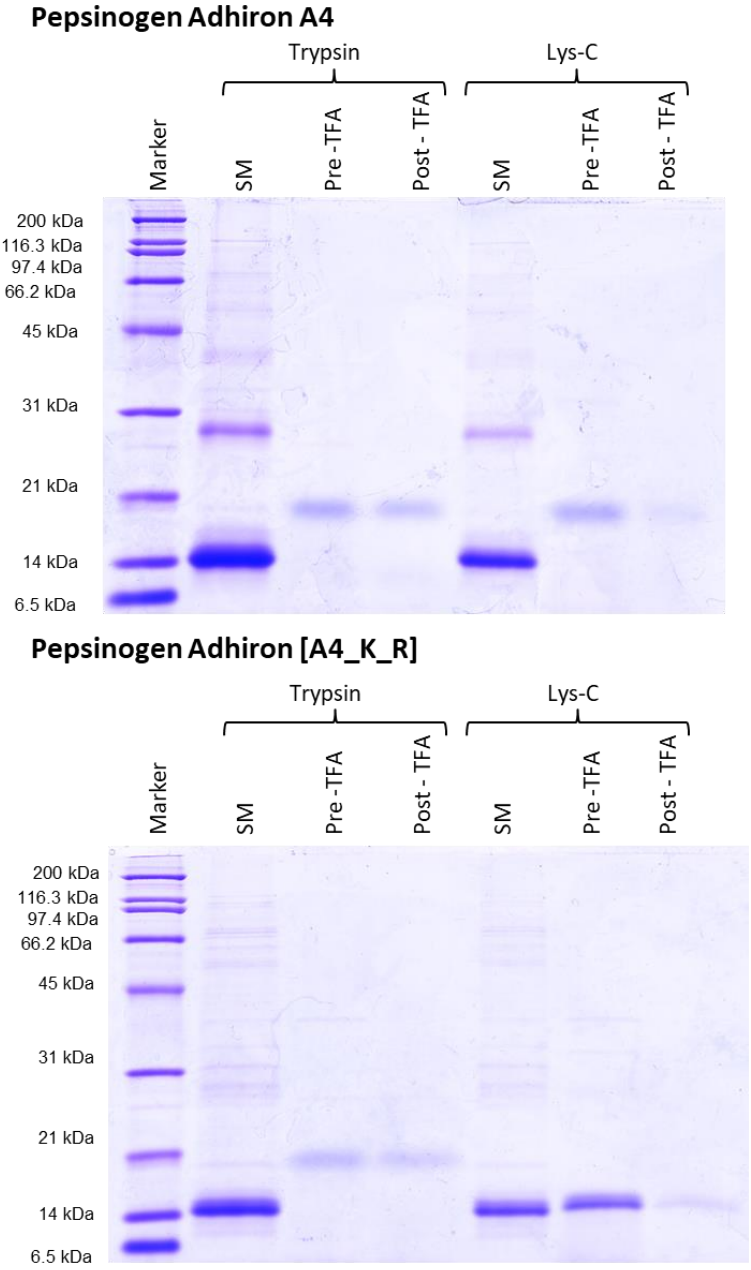


Figure 4.15 | SDS-PAGE analysis of pepsinogen Adhiron A4 and pepsinogen Adhiron [A4_K_R] trypsin and Lys-C proteolysis.

Top: Pepsinogen Adhiron A4 undergoes complete proteolysis when digested with trypsin and Lys-C. Bottom: Pepsinogen Adhiron [A4_K_R] undergoes complete proteolysis when digested with trypsin however remains intact when digested with Lys-C. SM – starting material. Pre-TFA – Adhiron after overnight digest before TFA addition. Post-TFA – Adhiron after overnight digest after TFA addition. The samples were ran on a 15 % gel and visualised with Coomassie plus stain.

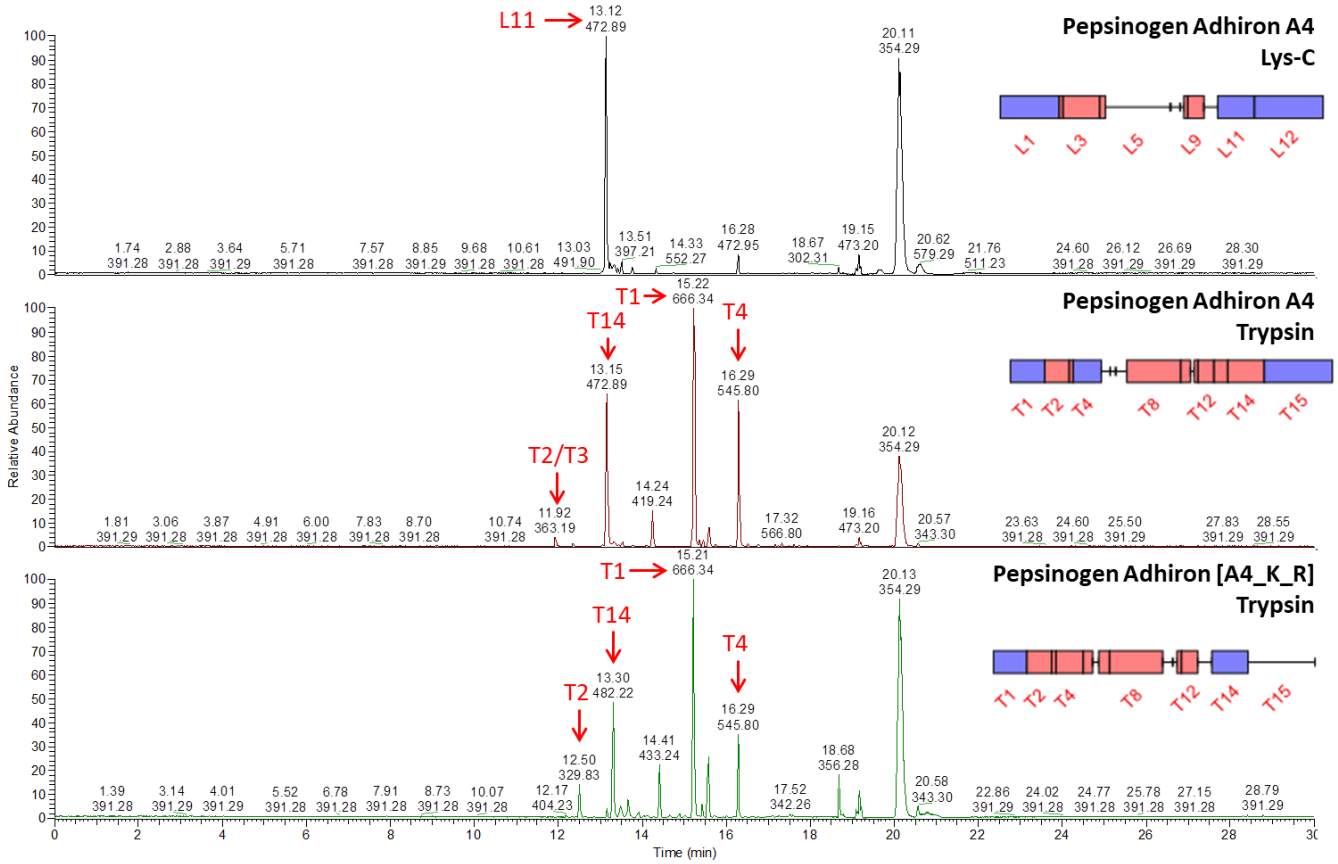


Figure 4.16 | BPI chromatograms and peptide map for pepsinogen Adhiron A4 and pepsinogen Adhiron [A4_K_R].

BPI chromatogram with main peaks labelled with corresponding peptide number. Blue peptides identified in mass spectrum with no missed cleavages sites, whereas red peptides were identified with one missed cleavage site. Black regions represent unidentified peptides. The peptide maps were generated using Peptide Mapper tool[234].

To investiagte whether the lysine to arginine substitions introduced into pepsinogen Adhiron A4 to generated pepsinogen Adhiron [A4_K_R] affects the stability of the protein, a proteolysis study was undertaken to examine the rate of proteolysis of the proteins. The protease used was chymotrypsin which cleaves at the aromatic amino residues of phenylalanine, tyrosine and tryptophan[44]. A total of 75 µg of both proteins were seperately incuabted with 1.5 µg chymotrypsin at 25 ^{o}C for 3 hours. Aliquots of sample were taken periodically to assess the degree of proteolysis and analysed by SDS-PAGE (Figure 4.17) and intact mass spectrometery (Figure 4.18). Proteolysis was stopped by the addition of either formic acid for MS analysis or by the addition of TCA for SDS-PAGE analysis followed by neutralisation with ammonium bicarbonate. A control sample of each protein was also incubated at 25 ^{o}C for 3 hours to confirm degredation from the protease and not from the temperature. SDS-PAGE analysis reveals that after a 3 hour incubation with chymotrypsin, neither pepsinogen Adhiron A4 or pepsinogen Adhiron [A4_K_R] underwent complete proteolysis. Lower molecular weight species were observed during the timecourse corresponding to an intermediate species forming. After 30 minutes, the protein band for the full length version of both proteins is not observed. Intact mass data supports the results with identification of fragments of 11421.8 Da and 11284.5 Da in pepsinogen Adhiron A4 and 11757.6 Da and 11620.6 Da for pepsinogen Adhiron [A4_K_R] corresponding to the loss of six and seven histidine residues (-HHHHHH and -HHHHHHH) in both proteins (Figure 4.18). An additional species was observed for pepsinogen Adhiron [A4_K_R] with the loss of 2 histidine residues at 2 and 10 minutes. The ratio of starting protein to intermediate species is slightly higher for pepsinogen Adhiron A4 compared to pepsinogen Adhiron [A4_K_R] at 30 minutes (Figure 4.18), however the rate of proteolysis is very comparable. The findings suggest that the mutations introduced into pepsinogen Adhiron A4 to form pepsinogen Adhiron [A4_K_R] do not alter the stability of the protein. Interestingly, histidine was not a preferential cleavage site of chymotrypsin, although histidine cleavage has been found to a less extent[235]. The free His-tag of the Adhirons probably faciliated histidine residue cleavage due to the ease of accessability. Lack of proteolysis of both Adhiron structures highlights the robust nature of the protein scaffold.

To further compare the stability of the two proteins, a collision induced unfolding (CIU) experiment was perfomed using ion mobility separation mass spectrometry (IMS- MS) as descibed in Chapter 2.23. CIU experiments were performed under native condtions and therefore the Adhirons were buffer exchanged into 500 mM ammonium acetate. The concentration of ammonium actetate used was typically less, around 20 mM, however it

was observed that when a lower concentration of ammmonium acetate was used, the Adhirons would precipitate from solution. The proteins were ionised by electrospray ionisation to produced multiple charged species. The number of charged states observed for both pepsinogen Adhiron A4 and pepsinogen Adhiron [A4_K_R] were fewer than when under denaturing conditions, due to the inaccessibility of the sites of protonation (data not shown). Furthermore, as the Adhirons did not undergo reduction prior to analysis, the dimer form of the Adhirons was also observed. For the CIU experiment, a single charge state was selected and subjected to collisional activation by CID. The activation collison energy (CE) was increased from 10 V to 26 V by 2 V increments. Comparison of the CCS distribution and CIU profiles of the two Adhirons revealed differences across the CE range (Figure 4.19). At low CE, both proteins adopted a similar cross sectional area of approximalty 12 nm^2. Both Adhirons required the same CE to induce unfolding however pepsinogen Adhiron [A4_K_R] required higher CE to initiate further unfolding compared to pepsinogen Adhiron A4 (22 V compared to 14 V). In addition, at higher CE (26 V) pesinogen Adhiron A4 adopted a single elongated conformation at 22 nm^2, whereas pepsinogen Adhiron [A4_K_R] adopted three distinct conformation at approxiamtely 16 nm^2, 19 nm^2, and 22 nm^2 (Figure 4.19). Although arginine and lysine are both positively charged basic residues, arginine has a higher pKa which means it generated more stable ionic interactions than lysine[236]. In addition, arginine residues contain a guanidinium group which increases the number of electrostatic interactions that can be made compared to lysine[237]. Therefore, the arginine residues may form an increased number of electrostatic interactions which may provide the protein with more stabilty. Stability studies on green fluorescent protein established that surface lysine to arginine mutations increased protein stabiltiy[238]. This could explain why a higher CE is needed to induce further unfolding of the pepsinogen Adhiron [A4_K_R] compared to pepsinogen Adhiron A4.

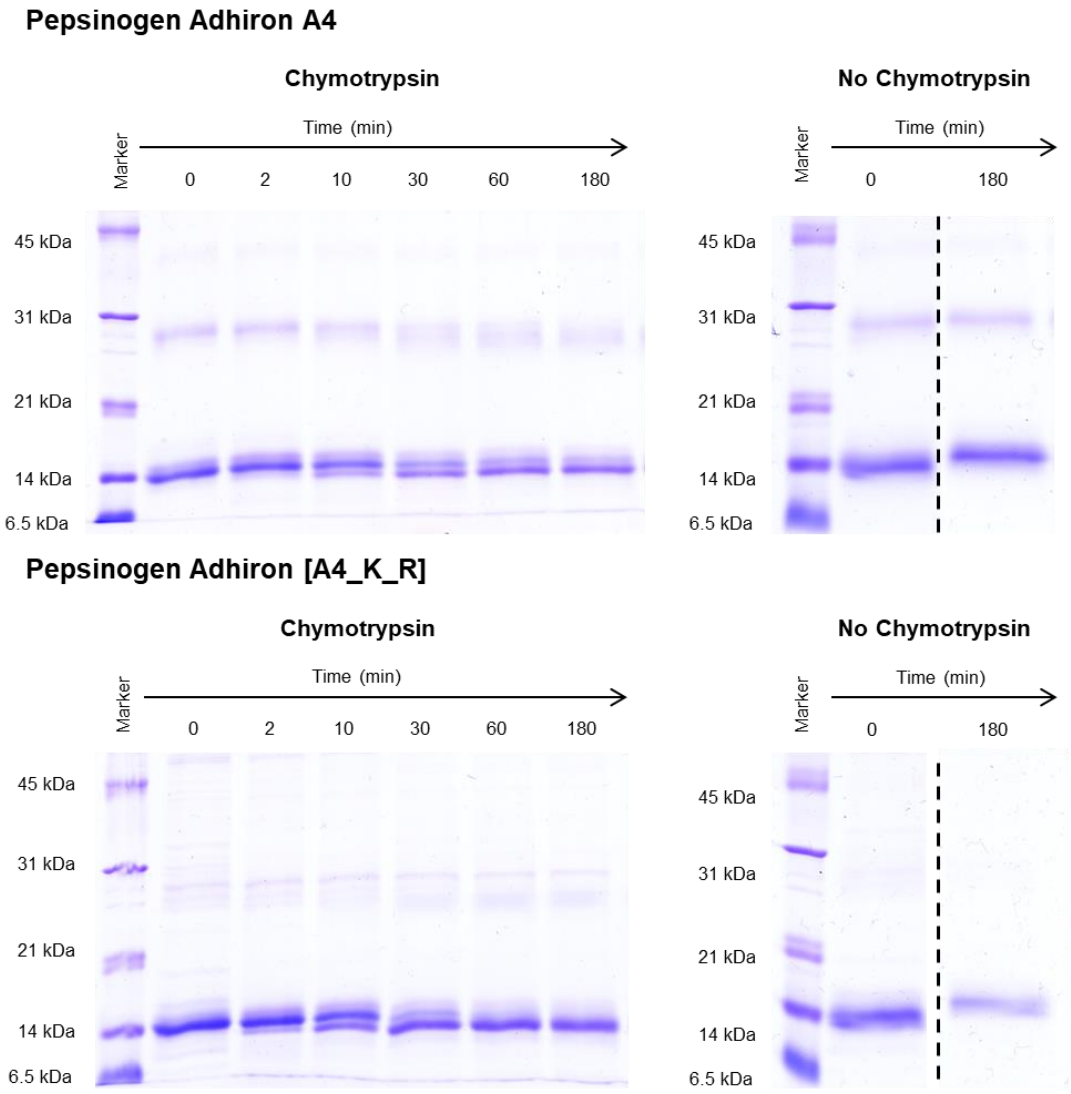


Figure 4.17 | SDS-PAGE analysis pepsinogen Adhiron [A4_K_R] and pepsinogen Adhiron A4 proteolysis time-course.

Top: Pepsinogen Adhiron A4. Bottom: Pepsinogen Adhiron [A4_K_R]. Proteins were incubated at 25 ^{o}C for 3 hours. Aliquots of samples were taken throughout the time-course for analysis. Both proteins incubated with chymotrypsin do not undergo complete proteolysis with a stable fragment forming after 3 hours. Control gels with no chymotrypsin demonstrate that both pepsinogen Adhiron A4 and pepsinogen Adhiron [A4_K_R] do not degrade after 180 minutes at 25 ^{o}C. Samples were ran on a 15 % gel and visualised with Coomassie plus stain.

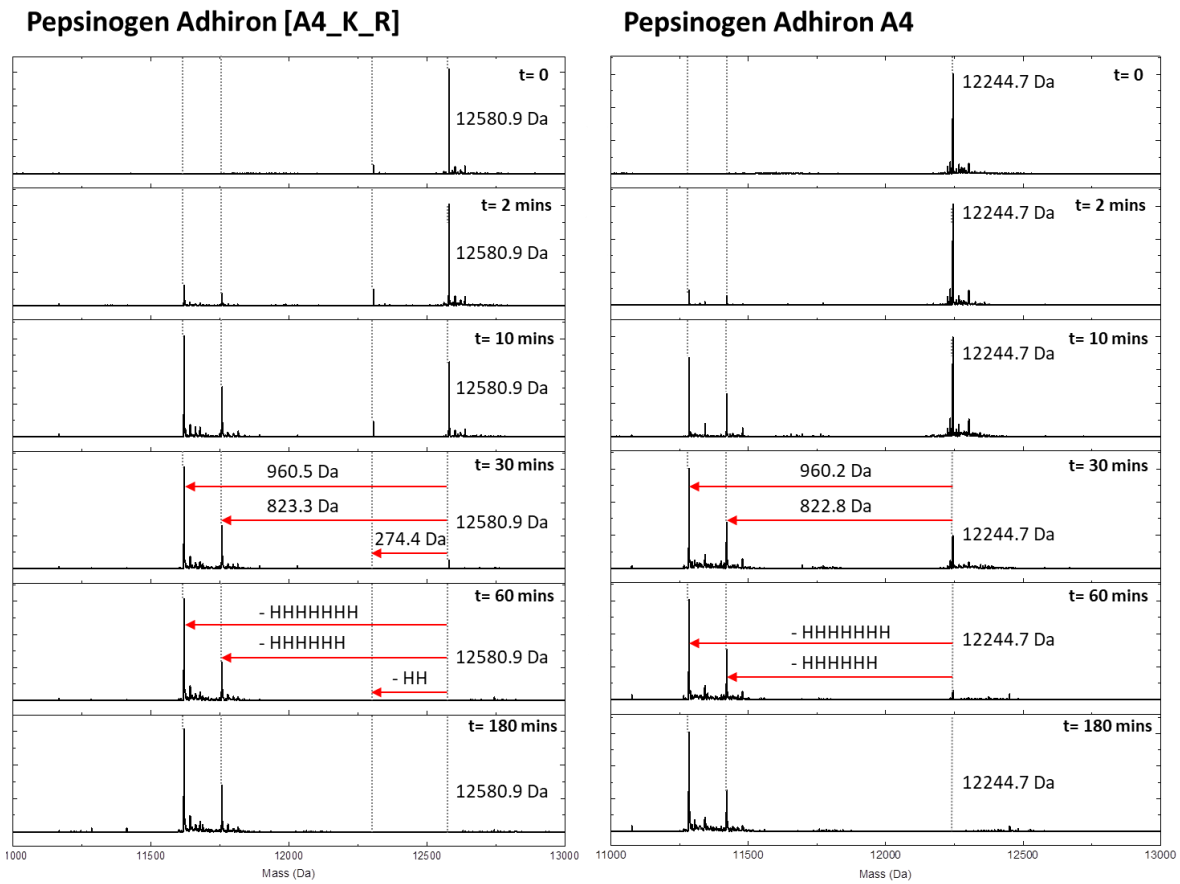


Figure 4.18 | ESI –MS analysis of pepsinogen Adhiron [A4_K_R] and pepsinogen Adhiron A4 proteolysis time-course.
Proteins were incubated at 25 ^{o}C for 3 hours. Aliquots of samples were taken throughout the time-course for analysis. Proteolysis was halted by the addition of TCA (5 % final concentration). Both proteins withstood complete proteolysis with intermediate species forming with the loss of six histidine residues and a further one histidine residues (seven total). Rate of proteolysis was comparable between proteins.

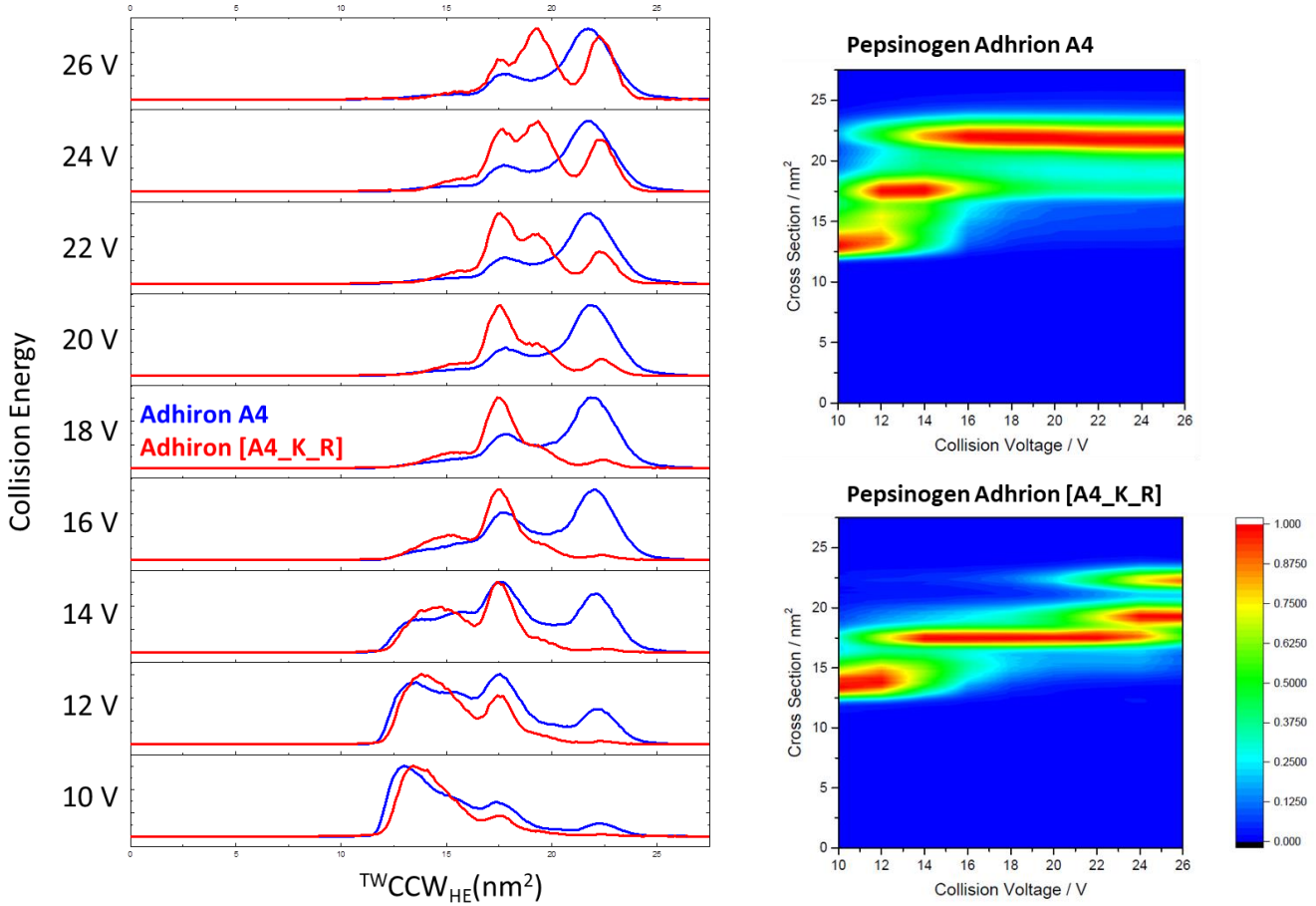


Figure 4.19 | CIU analysis of pepsinogen Adhiron A4 and pepsinogen Adhiron [A4_K_R] by ESI-IMS-MS.

Pepsinogen Adhiron A4 and pepsinogen Adhiron [A4_K_R] were analysed by ESI-IMS-MS on the Waters Synapt G2Si in ammonium acetate. CCS distribution and CIU profile of the two Adhirons is presented.

4.3.4 SulfoLink® Adhiron affinity purification of pepsinogen

Various immobilisation strategies were described and optimised in Chapter 3. As the Adhirons had a His-tag and a cysteine residue, numerous immobilisation techniques were available for the affinity purification. Of the immobilisation strategies described, the preferred method was the SulfoLink® resin which utilised the cysteine residue from covalent immobilisation. Firstly, Adhirons were immobilised to the SulfoLink® resin as described in Chapter 2.15. As a quick screen to assess pepsinogen capture by the Adhirons, all five Adhirons were analysed. To confirm the successful immobilisation of the Adhirons to the SulfoLink® resin, the remaining unbound Adhiron fraction was analysed by SDS-PAGE (Figure 4.20). Although the binding conditions and efficiency of Adhirons to the SulfoLink® had already been determined in Chapter 3, due to the position of the cysteine residue being before the His-tag, binding efficiency was assessed. Equal volumes of the Adhiron starting material and unbound fraction were loaded onto the gel and thus depletion of the Adhiron could be determined. The initial wash fraction was also analysed by SDS-PAGE to assess whether the Adhirons were washed from the resin. A very faint band at approximately 14 kDa was observed in the unbound fraction lane for all Adhirons indicating sufficient binding of the pepsinogen Adhirons to the SulfoLink® resin. No bands were observed in the wash fractions indicating that the Adhirons were not removed from the resin during the wash step and the lack of non-specific Adhiron binding.

After confirming immobilisation of the Adhirons, the resin-Adhiron complexes were blocked with free cysteine and then incubated with human pepsinogen. In total, 5 µg of human pepsinogen in PBS was incubated with the pepsinogen Adhirons resin complex. Recombinant pepsinogen was purchased from Cell Sciences (USA) and was the same recombinant protein that had been used in the phage display screens. As the pepsinogen Adhirons had not been validated for use in an affinity purification workflow, it was necessary to use recombinant protein from the same supplier as in the phage display screens. Pepsinogen Adhiron [A4_K_R] was also included in the analysis despite having the same variable regions as pepsinogen Adhiron A4. After incubation, the unbound pepsinogen fraction was removed and the resin washed with PBS. A stringent wash buffer was not initially used as the strength of the Adhiron-pepsinogen bond was unknown and to ensure the interaction was not diminished. After the final wash, PBS and sample buffer was added to the resin and then incubated for 5 minutes at 95 ^{o}C. The resin was re-suspended and analysed by SDS-PAGE (Figure 4.21). A faint band was visible at approximately 14 kDa corresponding to Adhiron for all six Adhiron samples. The Adhirons were irreversibly

covalently immobilised onto the SulfoLink® resin by a thioester bond and therefore even after incubation with reducing sample buffer, the Adhirons could not be eluted off the resin. However, it appears that a small amount of Adhiron has been removed from the resin. It was likely that the Adhiron removed from the resin was not covalently bound but instead non-specifically bound to either the resin or to other Adhirons that were covalently immobilised. At approximately 47 kDa, a faint band representing pepsinogen was observed for all pepsinogen Adhirons suggesting that the Adhirons successfully captured pepsinogen. However, despite a pepsinogen band present for all Adhirons samples, there is also a pepsinogen band present in the resin only lane. This indicates that pepsinogen was binding to the resin non-specifically and therefore capture of pepsinogen by the Adhirons couldnot be confirmed.

The wash step in the affinity purification was a PBS buffer which was used so that the pepsinogen interaction was not disrupted. As this wash was not sufficient in reducing non-specific interactions, the affinity purification was repeated with a more stringent wash buffer. As before, the resin was analysed by SDS-PAGE (Figure 4.21). Similar results were observed with a band at approximately 47 kDa representing pepsinogen observed in the Adhiron samples but also in the resin only lane. This again indicated that enrichment of pepsinogen could not be confirmed as in the Adhiron containing samples, pepsinogen could be binding to the resin instead of the Adhirons. Further analysis was not performed on protein bands visible at approximately 25 kDa, 70 kDa and 80 kDa due to time constraints. An in-gel digest of the bands to confirm the protein identity of the species would be the next step.

Due to time restrictions, further work developing the Adhiron SulfoLink® affinity purification of pepsinogen was not performed. However, various steps, outlined below, could be carried out with the aim to achieve a successful affinity purification using the SulfoLink® resin. Firstly, wash buffer optimisation is necessary. As it is unlikely that pepsinogen was binding covalently to the SulfoLink® resin as it contains no free cysteine residues, then increasing the stringency of the wash buffer should reduce the amount of non-specific binding of pepsinogen to the resin. However, there is a compromise between maintaining the Adhiron-pepsinogen interaction and reducing non-specific interactions. The five pepsinogen Adhirons selected as pepsinogen binders, were chosen based on ELISA data obtained at the University of Leeds. The Adhiron proteins used within this chapter were generated at the University of Liverpool. As the Adhirons were from a different preparation, confirming that they produce similar results obtained at the University of

Leeds by ELISA would further validate the pepsinogen Adhirons for binding. In addition, although all five pepsinogen Adhirons have the correct theoretical molecular weight, the proteins may not have adopted the same conformational shape as previously observed. Determination of correct protein folding should be carried out to ensure the correct confirmation of loop regions of the Adhirons.

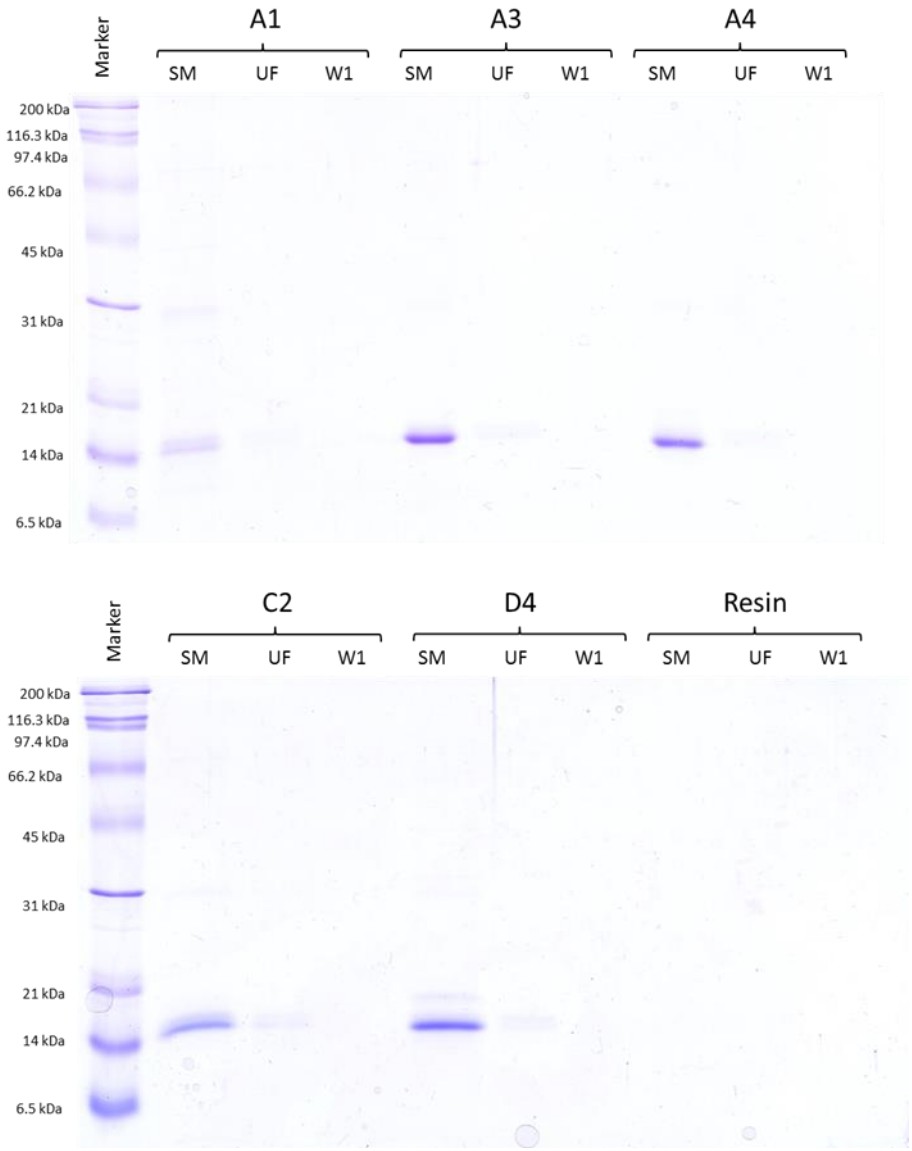
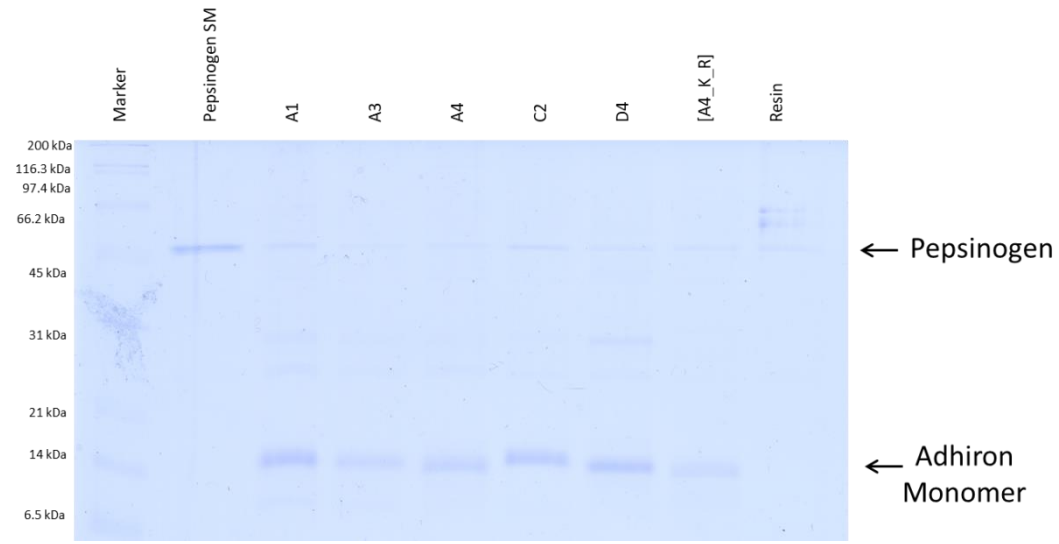


Figure 4.20 | SDS-PAGE analysis of unbound Adhiron fractions from affinity purification with SulfoLink® resin.

Equivalent volumes of starting material, unbound fraction and wash 1 loaded analysed by SDS-PAGE so depletion of the Adhiron due to immobilisation onto the SulfoLink resin can be observed. Faint band observed in the unbound fraction lanes for all Adhirons indicating that the majority of the Adhiron is immobilised onto the resin. SM – starting material, UF – unbound fraction, W1- wash 1. Samples were ran on a 15 % gel and visualised with Coomassie plus stain.

PBS Wash



pH 8 Wash

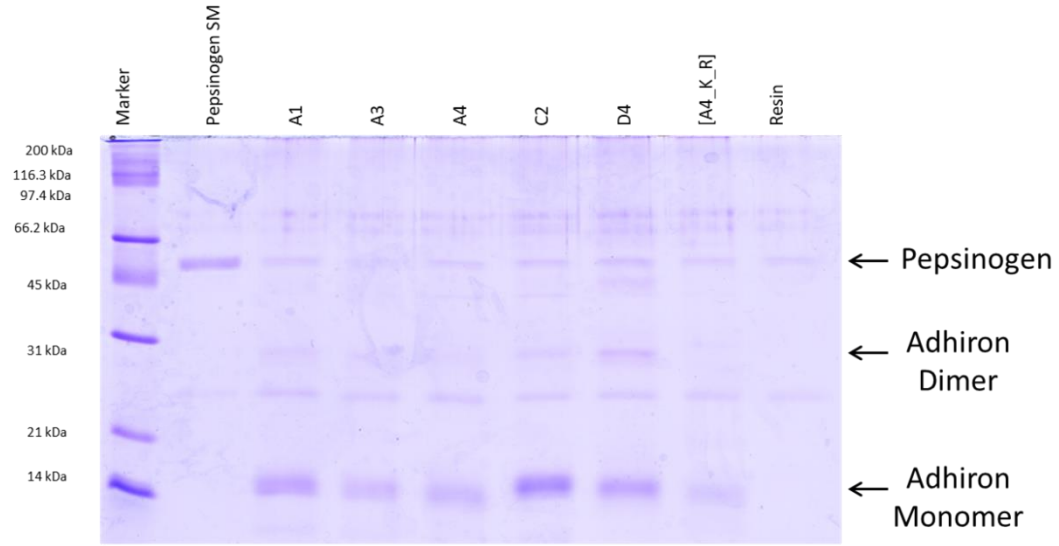


Figure 4.21 | SDS-PAGE analysis of pepsinogen affinity purification with SulfoLink resin.

Top: Resin washed with PBS only. Bottom: Resin washed with wash buffer pH 8. The resin was re-suspended with sample buffer and heated at 95 ^{o}C for 5 minutes. The resin-sample buffer mix was loaded directly onto the gel for analysis. Small amounts of non-specifically bound Adhiron eluting from the resin. Enrichment of pepsinogen observed by all Adhirons however, pepsinogen also identified in the negative control lane. Samples were ran on a 15 % gel and visualised with Coomassie plus stain.

4.4 Conclusions

A key aim of this chapter was to express and purify Adhiron binders that target human pepsinogen. The work undertaken has confirmed the successful expression and purification of five pepsinogen Adhirons. In addition, a key aim was develop a mutant Adhiron that was resistant to proteolysis. The substitution of all lysine residues to arginine residues within the pepsinogen Adhiron A4 sequence generated an Adhiron resistant to Lys-C proteolysis. Structural analysis by IMS-MS and a proteolysis time-course established that the mutations did not have a detrimental effect on protein stability. An affinity reagent resistant to Lys-C proteolysis will be beneficial in affinity purification mass spectrometry assays. The preliminary work demonstrates that as a proof of concept, Adhirons are amenable to further modification without affecting the structural stability of the protein.

The final aim of this chapter was to use pepsinogen Adhirons in a SulfoLink® affinity purification assay to capture human pepsinogen. Although the method was successfully developed in Chapter 3 using an Affimer designed to capture human IgG, enrichment of pepsinogen could not be confirmed due to pepsinogen binding non-specifically to the resin. This work demonstrates that a single affinity purification protocol cannot be applied to the analysis of all Adhirons. The workflow needs individualisation depending on the capture protein, the Adhiron and the biological background.

Chapter 5: Comparative Proteomic Analysis of Human Plasma from Patients with Sepsis

5.1 Introduction

Unbiased biomarker discovery studies are extremely important in clinical research to identify novel protein markers to improve disease diagnosis, monitor disease progression and to determine the effectiveness of therapeutics. As discussed in Chapter 3, the use of naive Affimer arrays in a large scale, multiplexed format provides great promise in biomarker discovery. The approach allows for the enrichment of target proteins using Affimers that contain randomised variable regions for an unbiased analysis of a proteome. Coupled with mass spectrometry for naive Affimer target protein identification, the potential of the technique to bring new insights into biomarker discovery is huge. However, despite the promise of naive Affimer technology, due to the current limitations outlined in Chapter 3, the methodology needs further development and optimisation to truly gain definitive identifications of naive Affimer protein targets. Therefore, developing an alternative method to identify putative protein markers is vital.

Figure 5.1 summarises the experimental workflow for both an Affimer- and a proteomics-based approach to reach the common goal of the identification of a panel of potential protein biomarkers. The outcomes of both approaches will guide the production of Type II Affimers, based on the cystatin A consensus sequence, that are the preferred scaffold choice due to the higher binding affinities observed to their intended protein target. The work described in this chapter will describe the proteomics approach taken to identify a panel of potential biomarkers of sepsis.

Sepsis is a systemic immune response caused by infection and is a major clinical problem across the world due to the high mortality rates. The symptoms of sepsis are extremely nonspecific as patients present with breathing difficulties, confusion, nausea and vomiting and fever. Without rapid therapeutic intervention, severe complications of sepsis can arise such as arterial hypotension or organ dysfunction significantly reducing patient survival rates[239]. There is an unprecedented need for an accurate and rapid diagnosis to improve patient outcomes. A proposed method for fast disease diagnosis is point-of-care testing (POCT). Typically, patient samples are sent to hospital laboratories, imposing a time delay on results and requiring highly skilled technicians. In contrast, POCT allows samples to be analysed immediately at the patient's location significantly increasing the rate of diagnosis.

Patients from low income and rural communities, where resources are limited would hugely benefit from a POCT kit for sepsis.

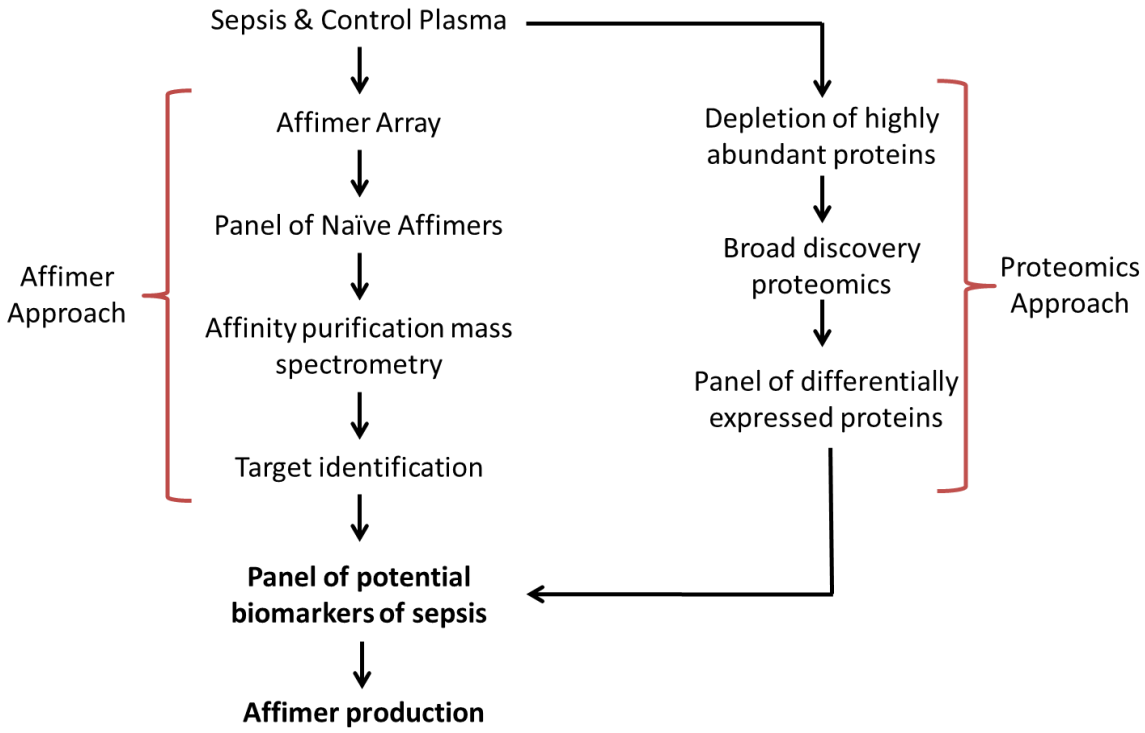


Figure 5.1 | Proposed experimental workflows to identify protein markers of sepsis.

Due to the lack of recognised biomarkers to diagnose sepsis, POCT kits are limited. Currently, POCT technology is available to measure lactate in the blood of patients with suspected sepsis[240] and to measure key acute phase proteins such as C-reactive protein[241] and procalcitonin[242]. Miniaturised multiplexed POCT devices that detect a panel of markers of sepsis have been described in the literature[243,244], however they are yet to be implemented in the clinic. The ability to generate highly stable Affimers to desired protein targets make them the ideal reagent for POCT devices. A preliminary study using Affimer reagents in a sandwich assay format has been reported[195,245] and with further technological advances, development and miniaturisation of a POCT device using Affimers is feasible. The work in this chapter will contribute to the ultimate goal of developing an Affimer-based POCT device for sepsis.

The first stage in the development of an Affimer-based POCT device is to identify sepsis plasma biomarkers. Plasma is the preferred biological matrix over serum[246] when studying sepsis due to the differences in sample processing; serum is obtained from blood that has

been allowed to clot whereas plasma is obtained from blood that has added anticoagulants. During whole blood coagulation, various proteins are secreted by cellular components of the process[247]. In addition, numerous proteins are removed during clot formation due to direct involvement in the coagulation or through non-specific interactions. As activation of the coagulation cascade, induced by toxins and the proinflammatory cascade occurs in patients with sepsis, to gain a true understanding of the proteomic expression in sepsis[248], it is vital that proteins involved in coagulation are not removed. Furthermore, plasma is considered a more reproducible matrix compared to serum[246]. The proteomic analysis of sepsis plasma by mass spectrometry is not a novel approach, with the methodology reported in numerous studies identifying many potential markers of sepsis including traditional markers such as C-reactive protein (CRP) [213,249]. Although procalcitonin has potential as a biomarker of sepsis, its value has been questioned in numerous studies as it is unsuccessful at distinguishing sepsis from other inflammatory diseases[239,250]. To date, an individual sepsis protein biomarker had not been discovered that can suitably diagnose and manage sepsis due to the lack of specificity of protein markers distinguishing sepsis from other inflammatory diseases. A new focus of identifying a panel of sepsis biomarkers is considered the best approach to accurately identify patients with sepsis.

As reviewed in Chapter 1, the complexity and large dynamic range of the human plasma proteome is a major challenge for LC-MS/MS discovery proteomics and has hindered biomarker discovery studies. The abundance of plasma proteins spans over 10 orders of magnitude[19] which is much greater than the dynamic range capabilities of current LC-MS/MS approaches. The peptides from highly abundant proteins can hinder the identification of peptides from lower abundance proteins by masking their signal from detection. Despite both high and low abundance proteins being clinically informative, it has been frequently reported that to gain novel insights into disease diagnosis and to identify potential disease biomarkers, the detection of low abundance proteins provides greater value. Low abundance plasma proteins are typically associated with disease specific protein secretions, leakage from tissues and cytokines making them ideal protein biomarkers[19].

The most abundant protein in plasma is albumin, with a normal clinical reference range of 35 – 50 mg/mL[251], which corresponds to approximately 50 % of the total plasma protein content[252]. Along with albumin, an additional 21 plasma proteins make up approximately 99 % of the total protein content of plasma[2]. Therefore, in order to detect and identify the remaining 1 % of plasma proteins, depletion strategies for the abundant proteins need to be applied prior to MS analysis.

In 2003, immunoaffinity depletion columns, containing antibodies to multiple protein targets were first described as a method to remove abundant proteins from plasma[66]. The work was a development from antibody-based protein purification methods where a single antibody was used per column [253]. The application of immunoaffinity depletion columns has become common practice in the analysis of plasma and serum, typically removing the top 12 or top 14 most abundant plasma proteins. However, a limitation of the columns is the high cost associated with the depletion columns and the risk of non-target protein loss.

An alternative approach to expensive immunoaffinity depletion methods is the targeted removal of albumin. As previously stated, albumin contributes to approximately 50 % of total plasma protein and therefore its removal would significantly reduce the dynamic range of the plasma proteome. The proposed method is based on a combined approach of TCA and organic solvent to precipitate protein[200], originally based on the Cohn's process[254] and later developed with the additional of TCA[255]. The two depletion strategies are yet to be compared.

5.2 Aims and Objectives

The work in this chapter focuses on the identification of a panel of candidate proteins implicated in sepsis for phage display production of Affimers. The key objectives of this chapter were to:

- Select a suitable method for plasma depletion and sample concentration.
- Extensively characterise the sepsis plasma proteome to determine changes in protein expression induced by sepsis.
- Select a panel of proteins implicated in sepsis for phage display Affimer production.

5.3 Results and Discussion

5.3.1. Selection of depletion method

As depletion of highly abundant plasma proteins is essential in order to increase proteome depth and coverage, two plasma depletion methods were compared to establish the ideal method for protein depletion. Human plasma was depleted using either the Thermo Scientific Pierce top 12 abundant protein depletion spin columns or the partial plasma depletion method as described in the methods section. The proteins depleted by the Pierce top 12 abundant protein depletion spin columns are listed in Table 5.1.

Table 5.1 | Top 12 plasma proteins depleted by Thermo Scientific Pierce Top 12 Abundant Depletion Spin Columns and plasma reference ranges.

α1- Acid Glycoprotein	60 – 120 mg/dL	**Fibrinogen**	200 – 400 mg/dL
α1-Antitrypsin	78 – 200 mg/dL	**Haptoglobin**	30 – 200 mg/dL
α2-Macroglobulin	106 – 279 mg/dL	**IgA**	70 – 400 mg/dL
Albumin	6.4 – 8.3 g/dL	**IgG**	700 – 1600 mg/dL
Apolipoprotein A-I	101 – 199 mg/dL	**IgM**	40 – 230 mg/dL
Apolipoprotein A-II	30 – 50 mg/dL	**Transferrin**	212 – 360 mg/dL

The depleted plasma samples, along with non-depleted plasma used to compare standard protein abundances, were digested in triplicate following the protocol listed in methods. Prior to LC-MS/MS analysis, plasma samples depleted using the depletion spin columns were concentrated using Vivaspin® centrifugal concentrators with a 3000 molecular weight cut off to reduce the total sample volume from 500 µL to 100 µL. A filter based method was selected for concentration so that the salt concentration within the sample did not increase. Preliminary LC-MS/MS analysis of the spin column depleted plasma samples revealed contamination by PEG. Insufficient washing of the filter concentrators prior to sample loading was responsible for the contaminant. To remove the PEG, an additional strong cation exchange step was performed, as described in the Chapter 2. Digested samples were analysed by LC-MS/MS over a two hour gradient and the raw data processed

using Proteome Discoverer and the resultant mgf file searched in MASCOT against a human database. The plasma digests were loaded onto the instrument to achieve base peak intensity (BPI) chromatograms of similar intensities as opposed to loading a set amount of protein. The total ion count (TIC) chromatogram was also assessed as they were more representative of the complete sample peptide content and revealed similar findings to the BPI chromatograms. The rationale behind this was that typically peptides from albumin limit the loading of the digested plasma. In the case of the depleted plasma digests, albumin and other highly abundant proteins had been removed allowing for more of the plasma digests to be loaded onto the instrument.

To evaluate the depletion workflows, the total number of proteins, peptides and PSMs identified from the digests of the different plasma preparation methods were compared (Figure 5.2.). The spin column depleted plasma achieved the best proteome coverage with the highest numbers of protein, peptide and PSM identifications. This result was expected as the highly abundant proteins have been depleted, more of the sample can be loaded onto the instrument resulting in the peptides from less abundant proteins brought into the detectable range of the instrument. The performance of TCA precipitation depleted plasma and non-depleted plasma have fairly comparable results with an average of 77 and 85 proteins identified respectively. Comparison of the BPI chromatograms reveal differences in the plasma sample complexity depending on depletion method; plasma samples depleted using the Pierce spin columns look far more complex than the non-depleted plasma and the partial TCA precipitation depleted plasma (Figure 5.3). BPI chromatograms of partial depleted and non-depleted plasma appear very similar suggesting inadequate removal of the highly abundant proteins from the partial depletion plasma samples. The results of total protein identifications further support this. The number of identifications achieved for each replicate of the same depletion method was similar as evident from the small standard deviations indicating low variability in depletion methods. For all three plasma preparation methods, peptides were distributed evenly across the gradient.

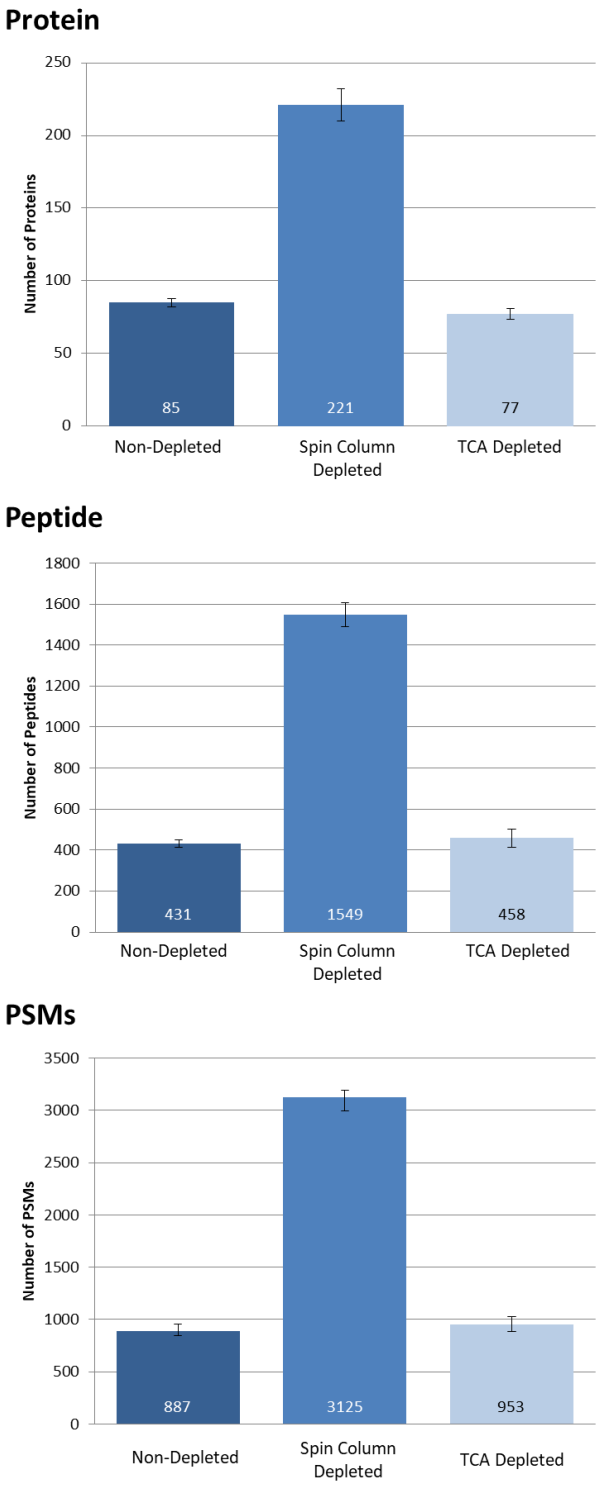


Figure 5.2 | Protein, peptide and PSMs identifications of different depletion methods.

Bars represent mean number of identifications of the 3 replicates with error bars representing standard deviation. Standard deviation is small for all samples indicating high reproducibility of the depletion methods. Spin column depleted plasma yielded the highest number of protein, peptide and PSM identifications of the methods. The numbers obtained for non-depleted and TCA plasma is fairly comparable.

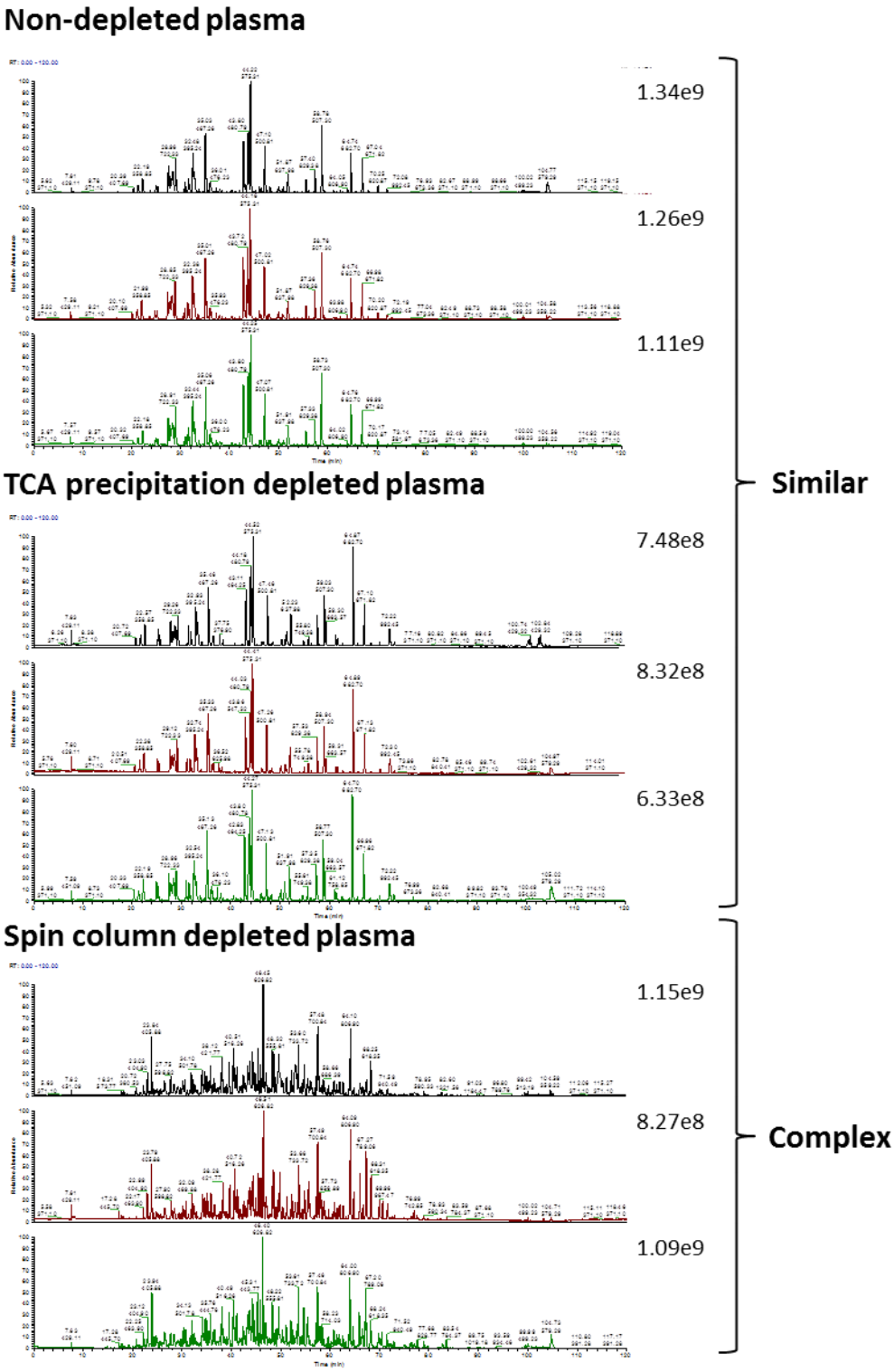


Figure 5.3 | Comparison of BPI chromatograms of depletion methods.

Digests analysed on Thermo Scientific QExactive mass spectrometer on a 2 hour gradient. Each sample was analysed in triplicate. Top panel: Non-depleted plasma. Middle panel: TCA precipitation partial plasma depletion. Bottom panel: Thermo Scientific Pierce depletion spin column plasma. BPI values are similar for each replicate.

To investigate the efficacy of the protein depletion methods at removing the highly abundant proteins, abundance values for 4 of the 12 most abundant plasma proteins were compared (Figure 5.4.). Due to the differences in sample loading, it was more informative to compare label-free protein abundances, calculated from the peak area of the top three most abundant peptides for a given protein, expressed as a percentage of total protein abundance instead of comparing raw protein abundance values. Albumin represented 2 %, 51 % and 53.8 % of the total plasma protein for spin column depleted, non-depleted and TCA precipitation partial depletion plasma, respectively. Albumin had thus been removed successfully from the spin column depleted plasma samples, whereas there has been no depletion of albumin from the partial depletion TCA precipitation plasma samples. The protein abundance values of haptoglobin, alpha-1 anti-trypsin and apoliporotein A-I for TCA precipitation partial depletion plasma and non-depleted plasma are also very similar demonstrating the lack of protein depletion.

As albumin contributed 50 % of the total protein content of plasma, its effective removal has a significant impact on the characterisation of plasma as demonstrated in the increase in the total proteins identified. The effect is also noted when comparing the protein abundance values of abundant plasma proteins that should be depleted (Figure 5.4). Albumin accounts for approximately 50 % of the protein abundance[19], for the TCA partially depleted plasma and the non-depleted plasma; this leaves the other proteins identified (76 proteins and 84 proteins, respectively) to share the remaining 50 % of the protein abundance signal. As a result, the protein abundance (%) of the other highly abundant proteins is relatively low compared to the spin column depleted plasma that has had the albumin effectively depleted. This is particularly evident with apolipoprotein A-I and alpha-1-antitrypsin where it appears that the proteins have not been depleted by the depletion spin columns when comparing the percentage protein abundance to non-depleted and TCA precipitation partial depleted plasma (Figure 5.4). Therefore, it may be more informative to compare the rank order of the most abundant proteins quantified in each method.

To further assess the effectiveness of the depletion methods, protein abundance plots were generated to evaluate the dynamic range of the plasma and the rank order of proteins quantified using the different depletion methods (Figure 5.5). The dynamic range of the TCA partial depleted plasma is very similar to non-depleted plasma with most proteins displaying a similar pattern in abundance. Albumin is the most abundant protein in the TCA partial depleted and non-depleted plasma, whereas in the spin column depleted plasma it is the 7th most abundant protein. The most abundant protein in the spin column depleted

plasma is hemopexin whereas in non-depleted and TCA partial depleted plasma it is the 18th and 13th most abundant protein respectively (Figure 5.5). In addition, the gradient of the dynamic range plot for the spin column depleted plasma is less steep at the highly abundant protein end of the plot compared to non-depleted plasma indicating effective depletion of abundant proteins.

The spin column depletion method allowed for approximately 120 times more of the plasma sample to be loaded onto the instrument than non-depleted plasma compared to the TCA partial depleted plasma which allowed for only 1.25 times more samples to be loaded. As fewer proteins were identified with this method, it is likely that the increase in loading was necessary to account for loss of protein during the depletion protocol. This significant increase in sample loading, as well as increase in the number of proteins identified can be attributed to protein depletion. Based on the findings from the analysis of the different plasma depletion methods, the spin column depletion method was selected as the depletion strategy throughout the rest of this study.

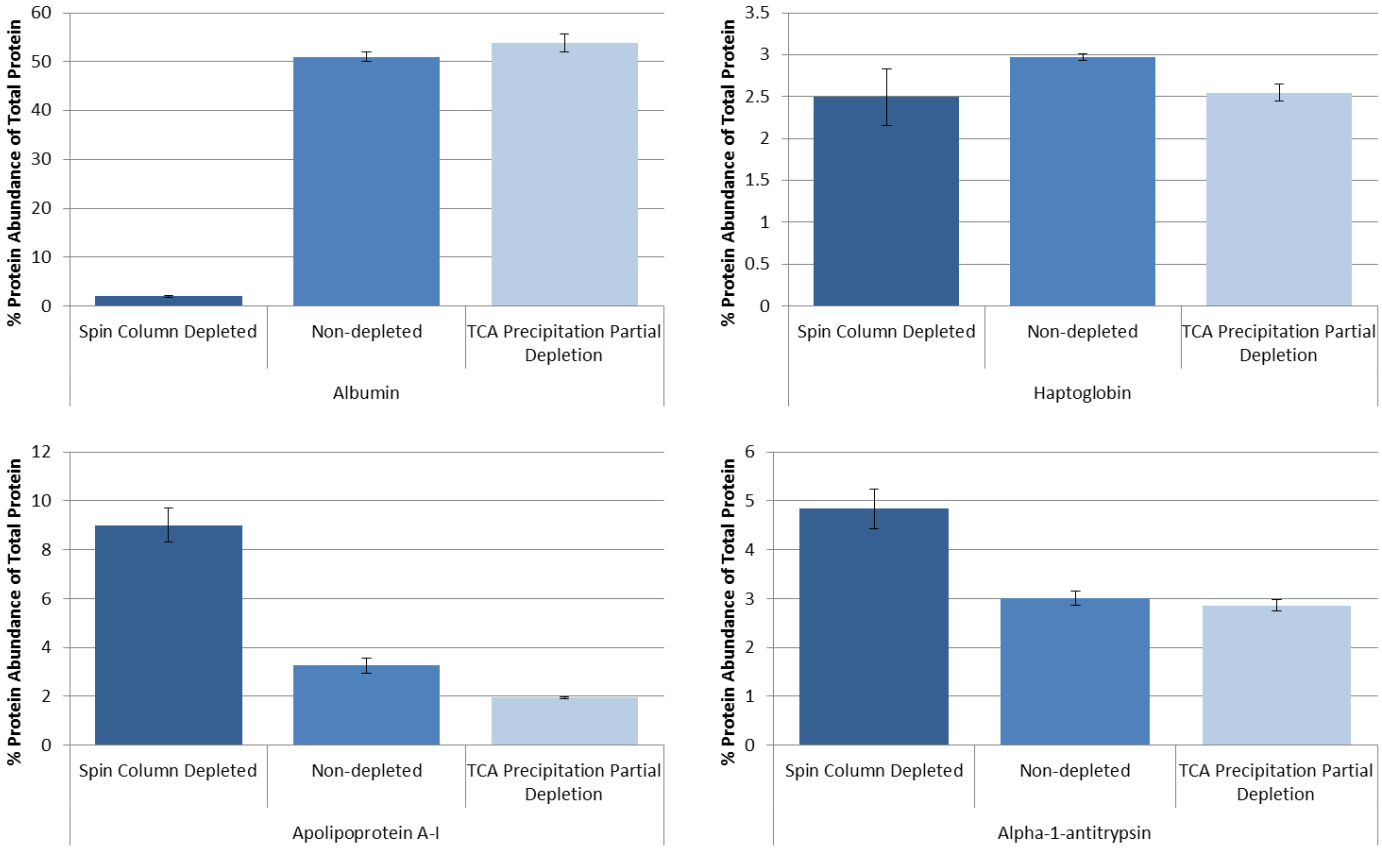


Figure 5.4 | Comparison of protein abundance values of four abundant plasma proteins.

Abundance values of albumin, haptoglobin, alpha-1-antitrypsin and apolipoprotein A-1 expressed as a percentage of total protein abundance. Comparison between spin column depleted, non-depleted and TCA precipitation partial plasma depletion methods. Error bars represent standard deviation.

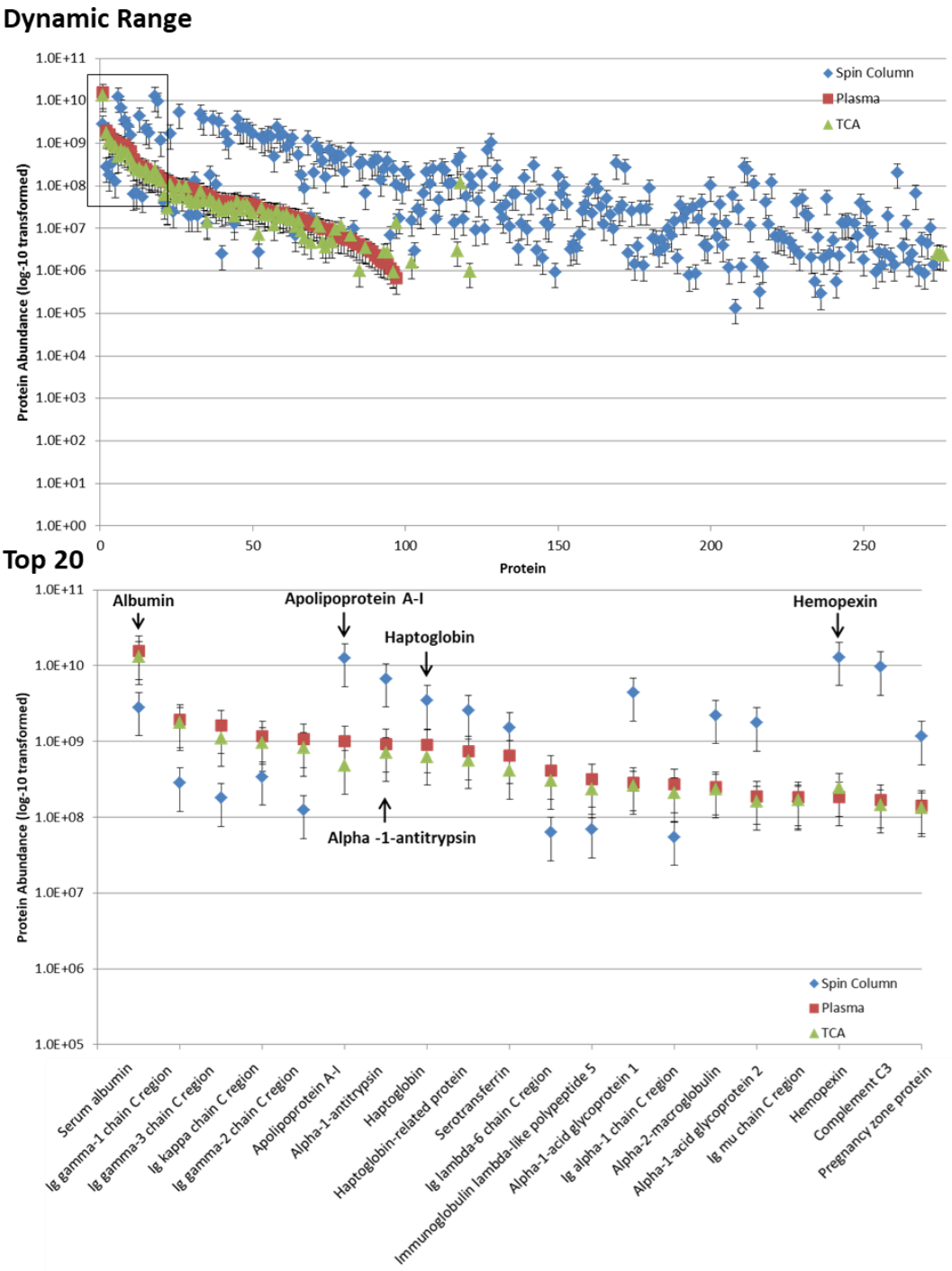


Figure 5.5 | Dynamic range comparison of depleted plasma proteome.

The mean normalised abundance value of each protein identified in spin column depleted, non-depleted and TCA precipitation partial depleted plasma was calculated in Proteome Discoverer using the peak area of the top 3 most abundant peptides for a given protein. Ordered by control abundance values. Error bars represent standard deviation. Top: Dynamic range plasma proteome. Lower: Top 20 most abundant proteins.

5.3.2. Selection of concentration method

5.3.2.1. Assessment of StataClean™ resin for protein concentration

In the preliminary experiment assessing plasma depletion protocols, a filter based method was used to concentrate the plasma sample following spin column depletion. However, sample recovery is a major concern when using protein concentrators with semi-permeable membranes such as the Vivaspin® filters. Proteins can adhere to the membrane resulting in low sample recovery rates. In addition, the membranes within the filtration devices are often coated with glycerol and humectants that can introduce contaminates into the sample if not washed properly.

To overcome this, a method to concentrate protein that is frequently used in our laboratory is a form of solid phase extraction (SPE) using StataClean™ resin[256]. StataClean™ resin was initially developed as an alternative to phenol based methods for DNA purification to remove proteins and enzymes[257]. The resin is coated with hydroxyl groups to which proteins bind, separating the DNA from proteins. The standard method used within our laboratory is to bind the proteins within a dilute sample and carry out an in-solution digest directly on the StrataClean™ resin. Direct digestion of the resin reduces sample handling steps and eliminates the need to elute the proteins from the resin.

To investigate the suitability of StrataClean™ resin to concentrate protein samples, the method was assessed to determine whether the resin binds all protein within a sample. Four separate human plasma samples, differing in protein concentration and volume (100 µg in 1 mL, 100 µg in 250 µL, 50 µg in 1 mL and 50 µg in 250 mL), were processed following the StrataClean™ concentration protocol as detailed in the methods. After the first round of sample incubation with Strataclean™ resin (labelled Strataclean 1), the unbound fraction was removed and incubated again with Stratclean™ resin (labelled Strataclean 2). TCA precipitation was performed on the unbound fraction from Strataclean™ resin 2. The Strataclean™ resins and protein pellet from the TCA precipitation were analysed by SDS-PAGE (Figure 5.6.). Multiple concentrations and volumes were investigated to determine if they affected performance of the Strataclean™ resin.

After the first round of incubating the sample with Strataclean™ resin, a large portion of the plasma proteins bound to the resin (Figure 5.6.). However, following the subsequent rounds of binding, protein bands were visible in Strataclean 2 and TCA precipitation lanes, indicating that even after two incubations with the StrataClean™ resin, all the protein

within the samples did not bind to the resin. Irrespective of the sample concentration and volume, similar results were observed with protein bands visible in the Strataclean 1 and 2 lanes of comparable intensity. In the 250 µL samples, more protein is present in the TCA precipitation lanes compared to the 1 mL samples. This may suggest that a larger sample volume allows for better mixing of the resin and protein sample, increasing the binding efficiency.

It was noted that the StrataClean™ resin displays similar behaviour to equalising beads, reducing difference in protein abundances. This effect is most notable for albumin as the resin appears to reach a saturation point for the protein. The largest band for albumin is in the TCA precipitation lane for all sample volumes and concentrations indicating that it remains mostly unbound. Despite this being an advantageous feature of potentially depleting albumin even further, the results demonstrates the inconsistencies in protein binding to StratClean™ resin.

Although the results from this preliminary investigation into Strataclean™ resin suggest the method is unsuitable for protein concentration, the results had not been directly compared to other concentration methods. Additionally, the assessment of StrataClean™ resin was initially performed on non-depleted plasma. To reproduce conditions more similar to the final application of the resin, it was hypothesised that when used to concentrate spin column depleted plasma, the StrataClean™ resin would bind protein in a more consistent and complete manner. Therefore, the method was investigated further to assess suitability for protein concentration.

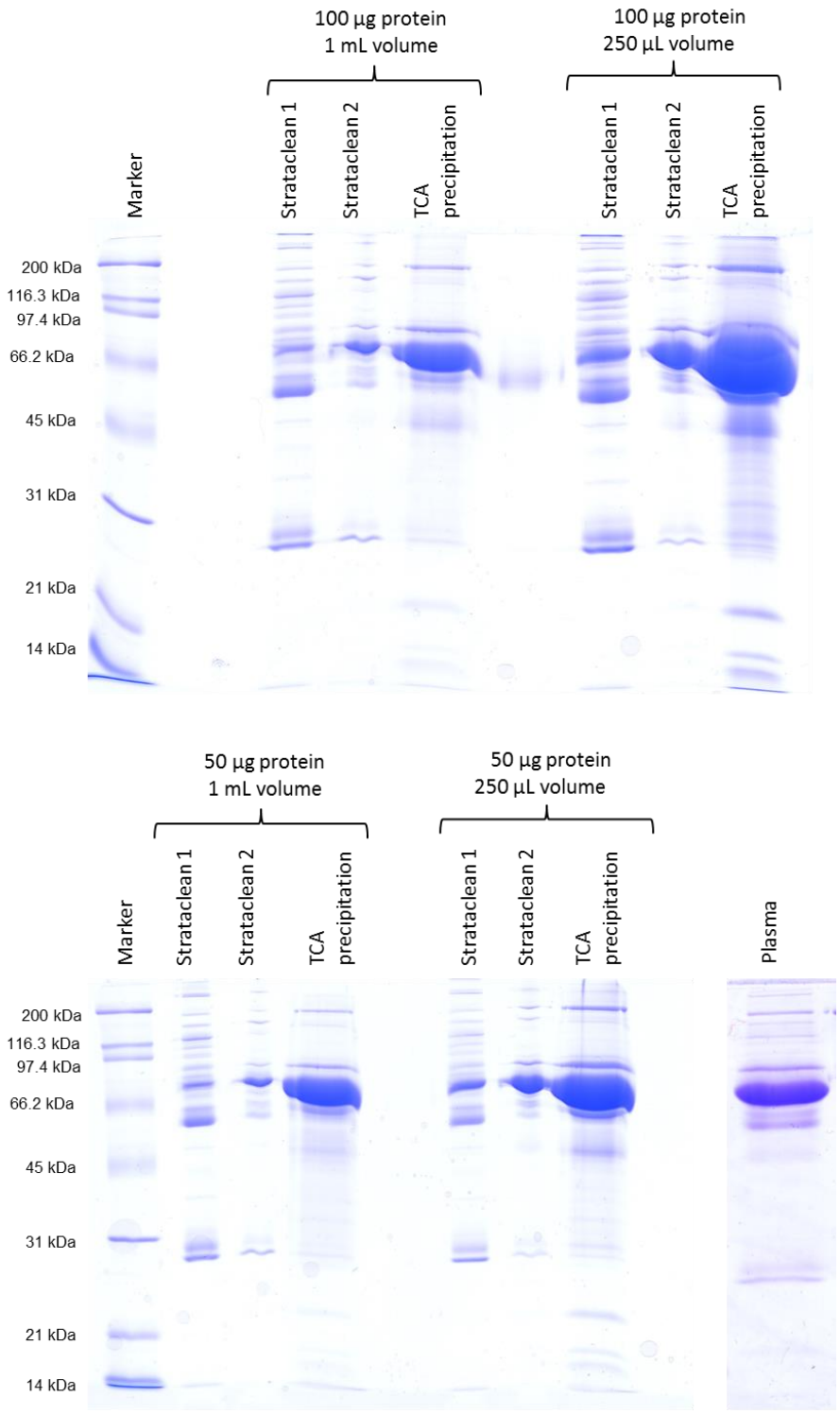


Figure 5.6 | SDS-PAGE analysis of human plasma samples concentrated using StrataClean™ resin.

Sample loading: 25 % of Strataclean 1 and 100 % of Strataclean 2 and TCA precipitation pellet. Four different starting concentrations and volumes of plasma were analysed: 100 µg n 1 mL, 100 µg in 250 µL, 50 µg in 1 mL and 50 µg in 250 mL. Comparable results were observed for all samples. StrataClean™ resin does not bind all protein within a sample. Samples ran on a 15 % polyacrylamide gel and stained with Coomassie blue.

5.3.2.2. Comparison of concentration method

To determine the optimal strategy for sample concentration, three concentration methods were compared by assessing the total number of proteins identified by LC-MS/MS. Human plasma (10 µL) was depleted using the Thermo Scientific Pierce depletion spin columns as described in the methods. The depleted human plasma was separated into three equal aliquots to be taken for protein concentration and digestion using StrataClean™ resin, FASP or TCA precipitation as detailed in methods section. The digests were analysed by LC-MS/MS on the Q Exactive Orbitrap mass spectrometer over a 2 hour gradient. To make a direct comparison between the methods, the final volume of each digest was made up to the same volume and equivalent amounts of each sample loaded onto the instrument. The analysis was performed on a single replicate to ensure that variability in the depletion method was not taken into account. The data was processed using Proteome Discoverer and searched in MASCOT against a human database with a 1 % FDR filter applied.

The number of protein, peptide, PSMs identified is shown in Table 5.2, where BPI chromatogram values were also noted. The FASP concentration method yielded the greatest number of proteins, peptide and PSM identifications (263, 3073, 6210 respectively) compared to the other techniques. In addition, the FASP method also achieved the highest BPI chromatogram value (Table 5.2). The StrataClean™ resin method produced the least number of protein identifications supporting the findings from the initial SDS-PAGE StrataClean™ analysis. Although it has been reported that FASP digestion can result in the loss of certain proteins/peptides due to them sticking to the filters, the results in this analysis indicate that the method is preferred for total protein identifications. The reduced number of sample handling steps in this method compared to StrataClean™ resin and TCA precipitation concentration methods may explain the improved number of protein, peptide and PSM identifications.

Table 5.2 | Protein, peptide and PSMs identifications of different concentration methods.

Samples	Number of Proteins Identified	Number of Peptides Identified	Number of PSMs	BPI
FASP	263	3073	6210	8.44 E8
StrataClean™	173	930	1466	2.78 E8
TCA Precipitation	200	1590	2562	3.67 E8

The chromatograms were visually compared to assess for similarities and differences in sample complexity (Figure 5.7). Overall, the FASP digestion and TCA precipitation chromatograms look fairly similar with peaks distributed evenly across the gradient. However, the FASP digestion method reached more than double the BPI than the TCA precipitation method suggesting better sample recovery for the FASP method. In the latter portion of the gradient for the StrataClean™ resin concentrated plasma, the intensity of the peptides eluting from the column decreases compared to the initial part of the gradient and when compared to the chromatograms from samples concentrated using the other methods. In reversed-phased chromatography, hydrophobic peptides are retained on the solid phase and elute later in the gradient. As there is a noted reduction in peptide intensity in this region of the gradient it may suggest that StrataClean™ resin disfavours binding of hydrophobic proteins and displays a preference to hydrophilic proteins.

Other filtered based concentration methods were not investigated. The FASP based method combines both sample concentration and digestion reducing the number of sample handling steps. The FASP concentration and digestion method was selected for subsequent sample preparation in this chapter.

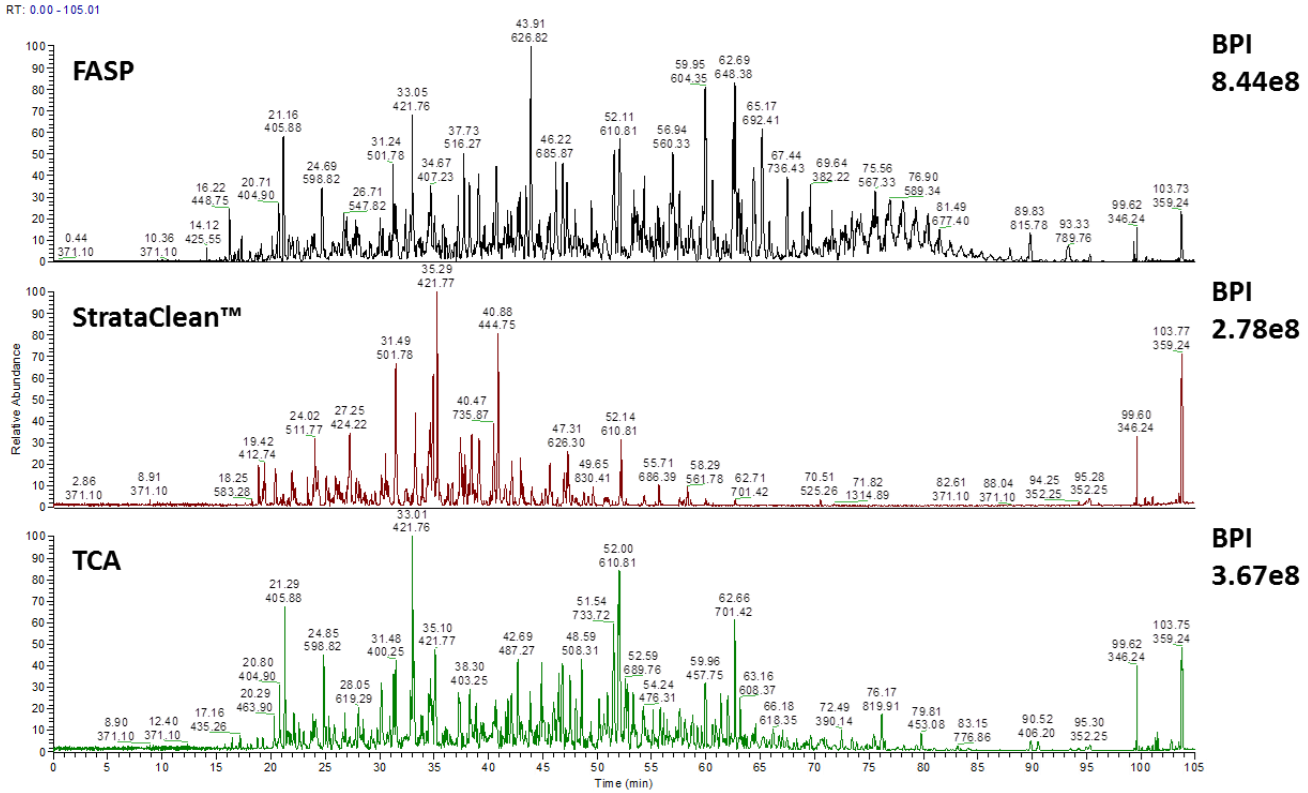


Figure 5.7 | Comparison of BPI chromatograms of concentration methods.

FASP, StrataClean™ resin and TCA precipitation method compared. Human plasma digests analysed on Thermo Scientific QExactive mass spectrometer on a 2 hour gradient. Chromatogram for FASP digestion is more complex and higher abundance compared to the other two methods.

5.3.3. Comparative proteomic analysis of sepsis plasma – Cohort 1

To identify a panel of candidate protein markers of sepsis, a comparative proteomics analysis of plasma was undertaken. The plasma from six sepsis and six hospital control patients were analysed in this study, named Cohort 1. The plasma from sepsis patients had been previously analysed for the bacterial diagnosis of sepsis. The control plasma was obtained from children admitted to the intensive care unit for elective cardiac surgery with no infection and who did not develop infection during their stay on the unit. Cohort 1 patient information is displayed in Table 3. All patient samples analysed in this chapter were obtained by Professor Enitan Carrol from Alder Hey Children's Hospital. Ethical approval was granted by the National Research Ethics Service and patient consent was received before sample collection (REC reference: 10/H1014/52).

Plasma samples were depleted, concentrated and digested using Thermo Scientific Pierce depletion spin columns and FASP digestion as described in the methods. Prior to the 2 hour LC-MS/MS sample analysis, each sample was analysed on a 30 minute gradient to ensure BPI chromatograms were of similar intensity. As each sample achieved similar BPI chromatogram intensities and peptides were subsequently analysed by LC-MS/MS on the Thermo Scientific QExactive HF on a 2 hour gradient with the same volume of sample loaded onto the instrument. The data was processed using Proteome Discoverer and the generated mgf file searched using MASCOT (Matrix Science) against a human database for peptide and protein identification. A 1% FDR filter was applied.

The number of protein, peptide and PSMs identifications for each sample was analysed (Figure 5.8.). Box plot analysis comparing the distribution of data between sepsis and control plasma revealed the number of protein, peptide and PSMs identified were greater in the sepsis plasma group compared to the control plasma group (Figure 5.9). To assess whether there was a significant statistical difference between the control and sepsis plasma groups in the number of protein, peptide and PSM identifications, a t-test was performed (significance level, $P < 0.05$). No statistical difference was observed between control and sepsis plasma in the number of peptide and PSM identifications. A significant difference ($P = 0.02657$) in the number of proteins identified between the sepsis and control plasma samples. As sepsis results in the initiation of an immune response and activation and secretion of immune proteins, this may explain the greater number of identifications in the sepsis group. Additionally, variability in the number of protein, peptide and PSMs was observed between the samples which could be explained by variability in protein depletion

or due to inherent differences in patient samples. BPI chromatograms of sepsis and control plasma digests were visually compared (Figure 5.10) and reveal the sepsis plasma chromatograms were typically more complex with more peaks of greater intensity along the base line compared to control plasma. This would suggest a higher degree of sample complexity compared to the control plasma which supports the data from total protein identifications.

To investigate the overlap in protein identification in control and sepsis plasma, a venn diagram of proteins identified in at least 2 or more samples of the same condition was prepared (Figure 5.8). A total of 154 proteins were identified across all conditions with 105 proteins identified in both conditions indicating high similarities between the control and sepsis plasma on a qualitative level. Six proteins were identified in control plasma only and 43 proteins were identified in sepsis plasma only. Functional annotation using DAVID of the proteins identified in the sepsis plasma only reveals proteins associated with receptor mediated endocytosis, positive regulation of B cell activation, innate immune response and acute phase responses. This finding was expected due to the systemic immune response induced by sepsis.

Table 5.3 | Patient demographic and clinical data.

F = female, M = male, Age (Years).

Group	Sample IDs	Bacterial Diagnosis	Age	Gender
Sepsis	60	*Streptococcus empyema*	2.31	F
	72		0.65	F
	104	*Streptococcus pyogenes*	2.28	M
	122		NA	NA
	133	*Meningococcal meningitis*	2.37	M
	134		1.76	M
Control	17	None	1.46	M
	43		1.11	F
	22		0.16	M
	7		10.36	M
	24		0.24	F
	42		3.43	F

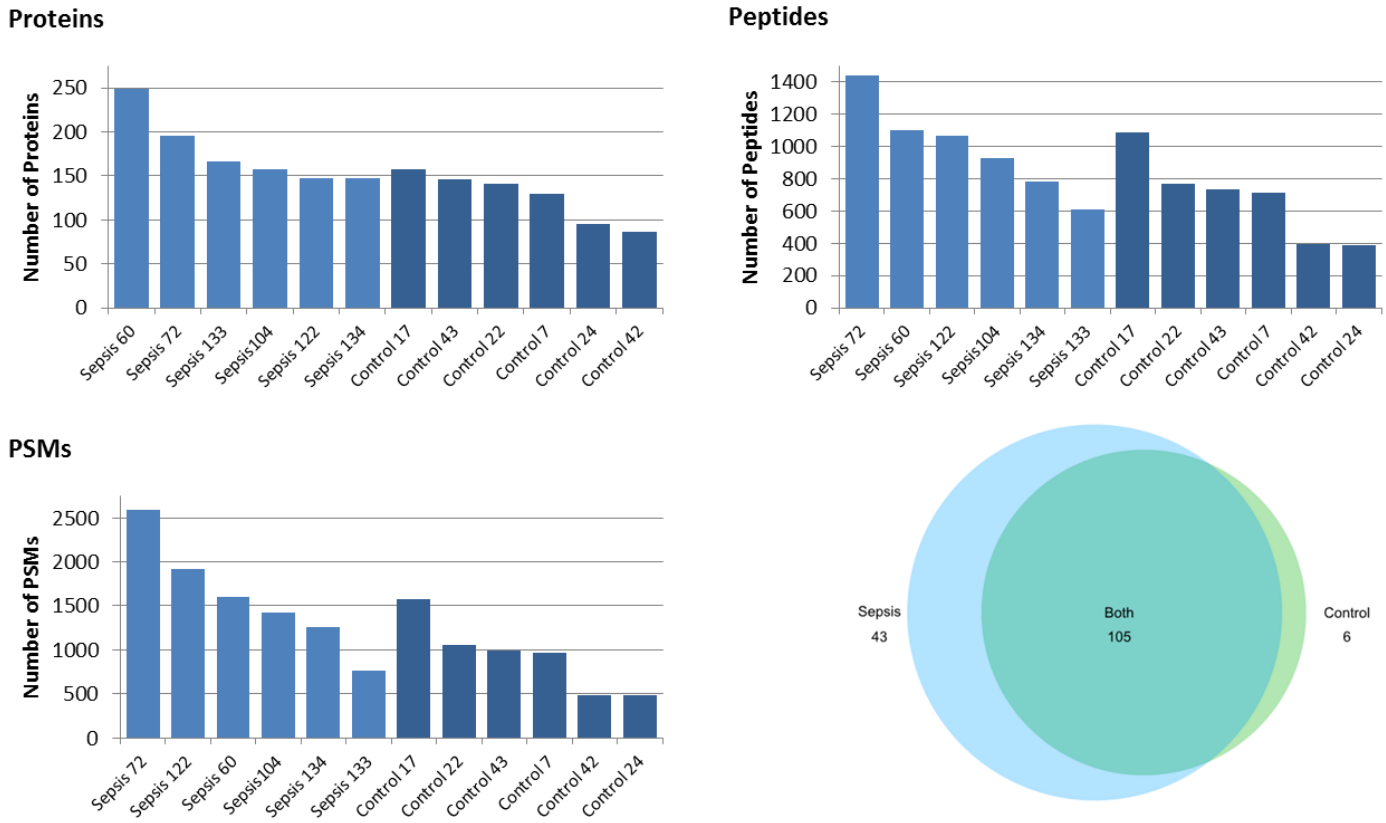


Figure 5.8 | Total number of protein, peptide and PSMs identified and overlap of protein identifications.

Data processed using Proteome Discoverer and searched in MASCOT against a human database with a 1 % FDR filter. Venn diagram: proteins identified in at least 2 samples of the same condition. 154 proteins identified in total. 105, 43 and 6 proteins identified in both conditions, sepsis only and control only respectively.

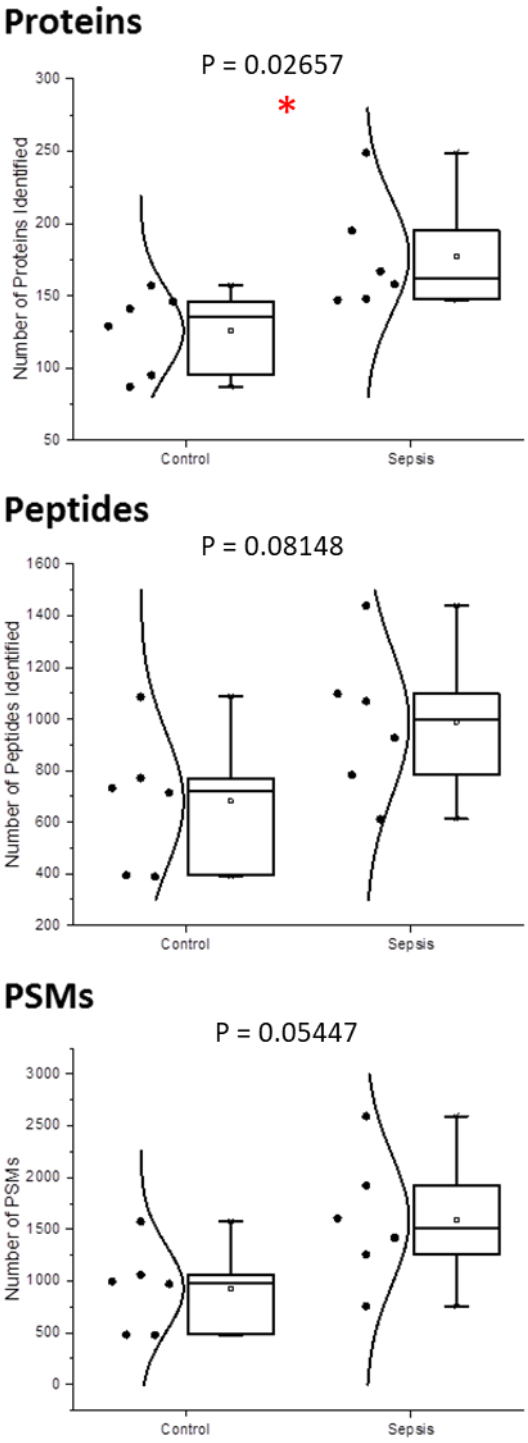


Figure 5.9 | Distribution of the number of protein, peptide and PSMs identified in control and sepsis plasma.

The overall number of protein, peptide and PSM identifications are highest in sepsis samples. Line represents median values. Whiskers represent minimum and maximum values. The median number of protein, peptide and PSMs identifications was highest in sepsis samples. Curve represents distribution. Statistical analysis (t-test) was performed to assess for differences between control and sepsis plasma ($P > 0.05$). Red star (*) denotes significance.

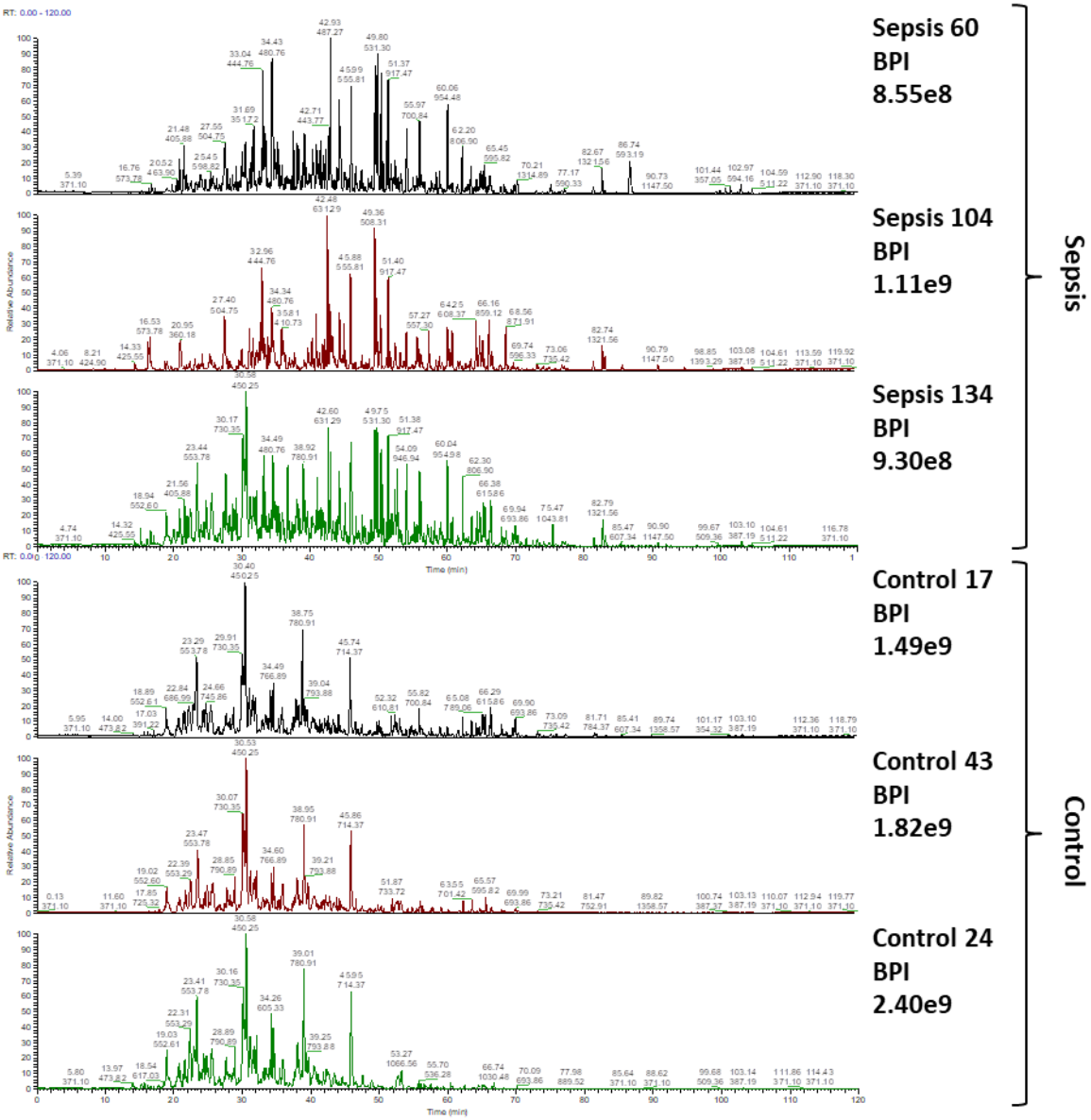


Figure 5.10 | Representative BPI chromatograms of sepsis and control plasma digests analysed by LC-MS/MS.

Depleted patient plasma FASP digests analysed on Thermo QExactive HF on a 2 hour gradient. Three BPI chromatograms were selected from sepsis and control samples that achieved highest, middle and lowest number of protein identifications.

To investigate quantitative difference in protein expression, label-free protein quantification was performed using Progenesis QI (Waters Corporation). Retention time alignment was carried out, with sepsis sample 134 selected as the alignment reference file. Alignment scores achieved 75% or greater signifying high similarities between retention time profiles. The normalisation scores for all samples were within the accepted range (Figure 5.11). The dynamic ranges of the sepsis and control plasma proteomes span seven orders of magnitude (Figure 5.12.). Typically, the dynamic range with non-depleted plasma is ten orders of magnitude[19] and the protein index curve has a steeper gradient towards the highly abundant proteins compared to the abundance data in this study. This suggests that the depletion strategy employed in this analysis has been successful in reducing the dynamic range of the plasma proteome. The dynamic range plot was ordered by control sample protein abundance and therefore proteins identified in the sepsis group that did not follow this trend were highlighted (Figure 5.12). Proteins that were more abundant in the sepsis group include haptoglobin, C-reactive protein, lipopolysaccharide binding protein, alpha-1-acid glycoprotein 1 and alpha-1-acid glycoprotein 2. As these proteins are involved in the acute-phase response, the higher abundance in the sepsis group is expected. The greatest variability in protein abundance between the control and sepsis groups was observed at the upper and lower ends of the dynamic range (Figure 5.13). Variability of protein abundances within the sepsis groups compared to the control group was also highest in these portions of the plot as evident by the larger error bars (Figure 5.13). As various factors affect protein expression such as stage of infection, age of patients and severity of the disease, this observation is expected.

In total, 297 proteins were quantified by relative quantification using non-conflicting peptides with a minimum of two peptides per protein and a protein score ≥ 20. The increase in the number of protein identifications with Progenesis compared to MASCOT searching only can be attributed to the way Progenesis handles the data. Progenesis works by analysing the data at the MS1 level to identify and then quantity peptide ion peaks that show differences between experimental conditions. The data is sent for database searching for protein identification. During data-dependent acquisition mode, not all peptide ions are triggered for fragmentation, especially low abundance peptides ions, which results in the loss of a significant amount of data. Additionally, further data is discarded during filtering steps due to peptide ions with low quality MS/MS spectra. This approach of merging the data is advantageous as peptide ions with no associated MS/MS spectra are not discarded from the analysis resulting in more data included in the expression change analysis

between the experimental groups. The result observed in this analysis would suggest that a large portion of the data acquired has low quality MS/MS spectra or that a large number of peptide ions were not triggered for fragmentation due to the large increase in protein identifications between the data searched in MASCOT only and the data searched in Progenesis.

To determine if samples could be separated based on biological differences, the dataset was subjected to principal component analysis of protein abundances. Clear separation between the two groups was observed with the greatest amount of variation explained by principal component 1 (at 35% of total variance compared to 21% for principal component 2) (Figure 5.14). The control plasma samples (blue dots) cluster closer to each other compared to the sepsis plasma samples (purple dots). This may be explained by the higher level of variability in the sepsis samples due to the varied bacterial diagnosis. A single sepsis plasma sample (sepsis sample 133) is an outlier separating more closely with the control samples. Again, this may be because of the high degree of biological variability of the sepsis plasma and low sample numbers. To further investigate separation of samples based on biological differences, protein abundance values were log10-transformed and analysed by hierarchical clustering (Figure 5.15). Clear discrimination between control and sepsis plasma was observed. Again, a single sepsis plasma sample outlier (sepsis sample 133) was identified. As this sample shared the same clinical diagnosis of meningococcal sepsis with sepsis sample 134, it is expected that these samples would cluster together; however, this was not the case. Multiple factors could explain this including patient demographic, stage of sepsis and if the patient had any other underlying clinical problems.

To investigate differentially expressed proteins between control and sepsis plasma, the dataset was filtered further; a P value of <0.05 and a q value of <0.05 cut-off was applied. In total, 40 proteins were differentially expressed between the two groups with a fold change greater than 1.5. Four proteins were quantified as having the highest mean abundances in the control group and 37 proteins quantified as having the highest mean abundance in the sepsis group (Table 5.4). As with the qualitative analysis of the data searched in MASCOT only, quantitative analysis reveals similar biological findings; differentially expressed proteins with the highest mean in sepsis plasma were strongly associated with immune response.

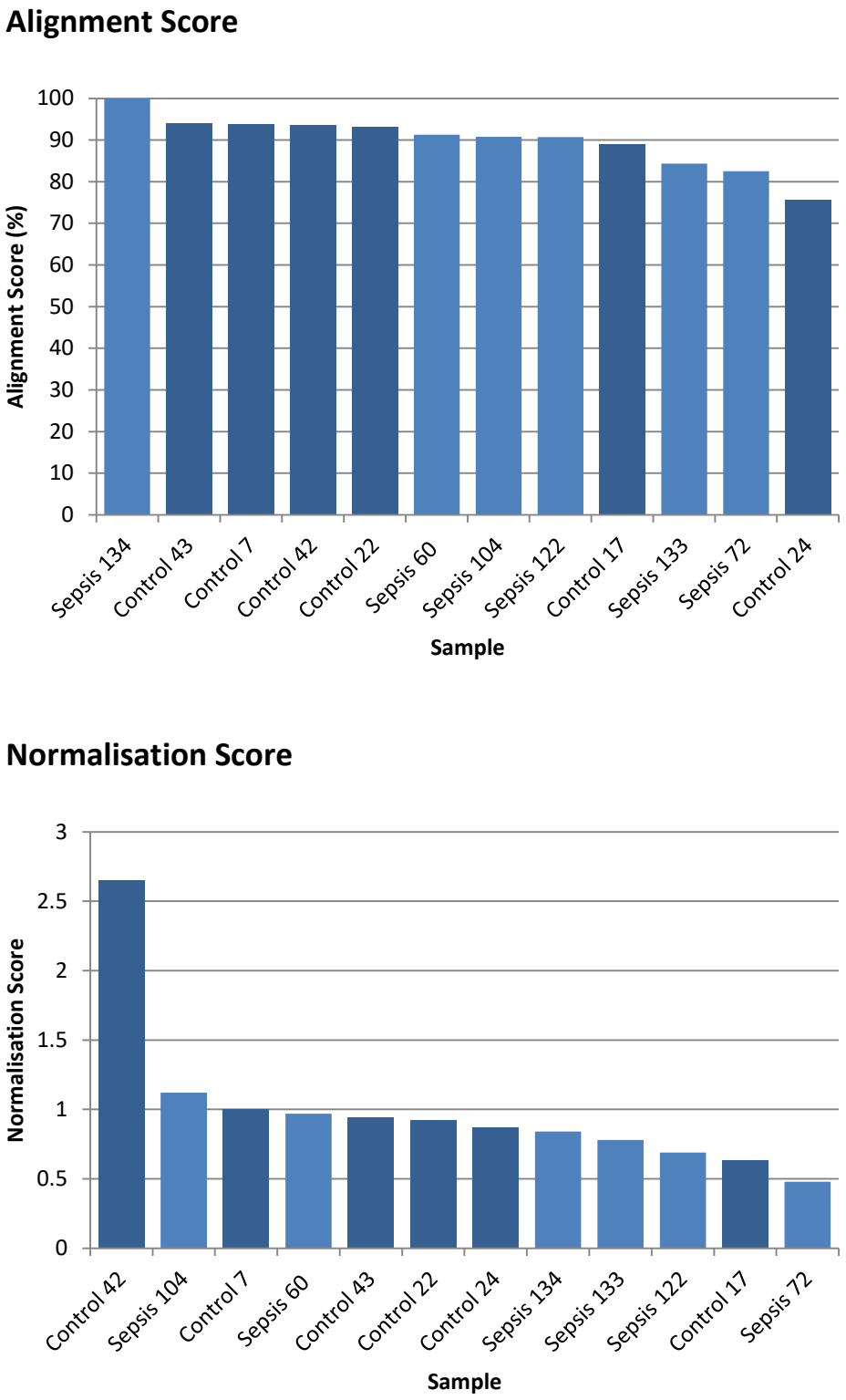


Figure 5.11 | Summary of alignment and normalisation scores of patient samples.

Top: Progenesis alignment scores. Sepsis sample 134 selected as the alignment reference file. Alignment scores achieved 75% or greater signifying high similarities between retention time profiles. Bottom: All progenesis normalisation scores were in the accepted range.

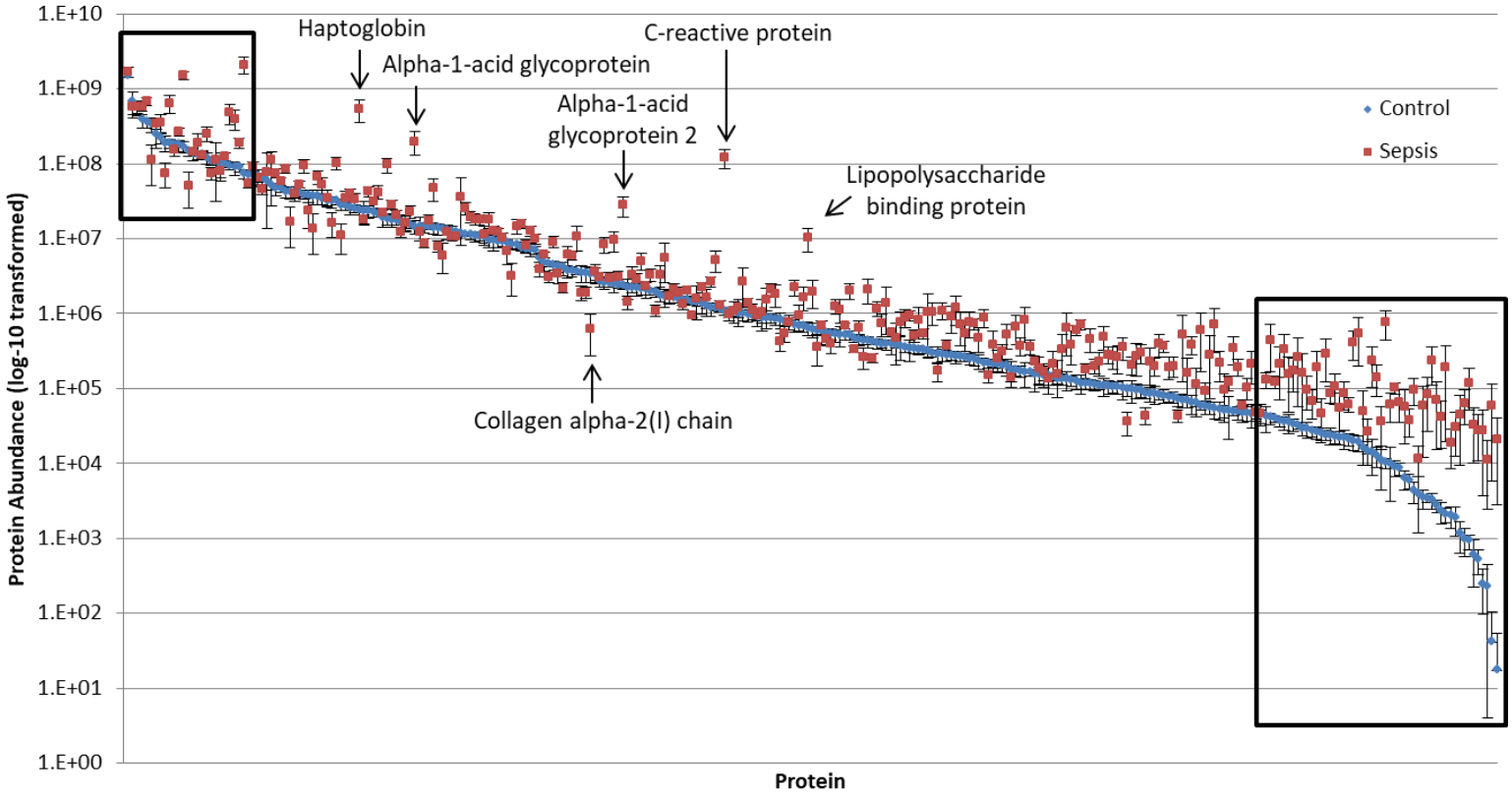


Figure 5.12 | Dynamic range of control and sepsis plasma proteome.

The mean normalised abundance value of each protein expressed in the control and sepsis plasma was calculated using Progenesis QI relative quantification using non-conflicting peptides. Ordered by control abundance values. Error bars represent standard error. Blue points correspond to control plasma and red points correspond to sepsis plasma. Most abundant and least abundant proteins highlighted in black boxes and examine further in Figure 5.13. Sepsis outlier proteins labelled.

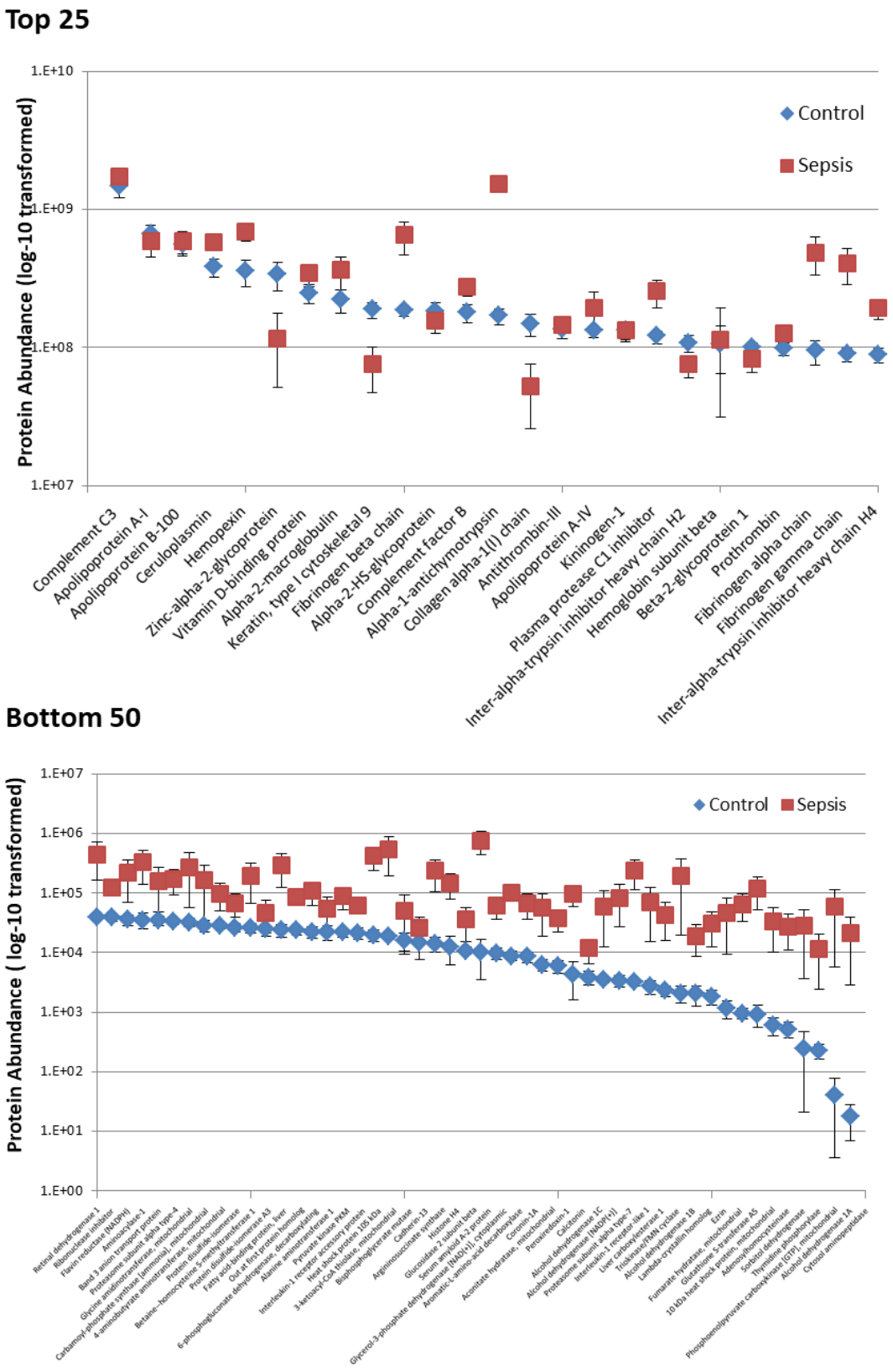


Figure 5.13 | Upper and Lower Dynamic range of control and sepsis plasma proteome.

Ordered by control abundance values. Top: Top 25 most abundant proteins. Bottom: 50 least abundant proteins. Error bars represent standard error.

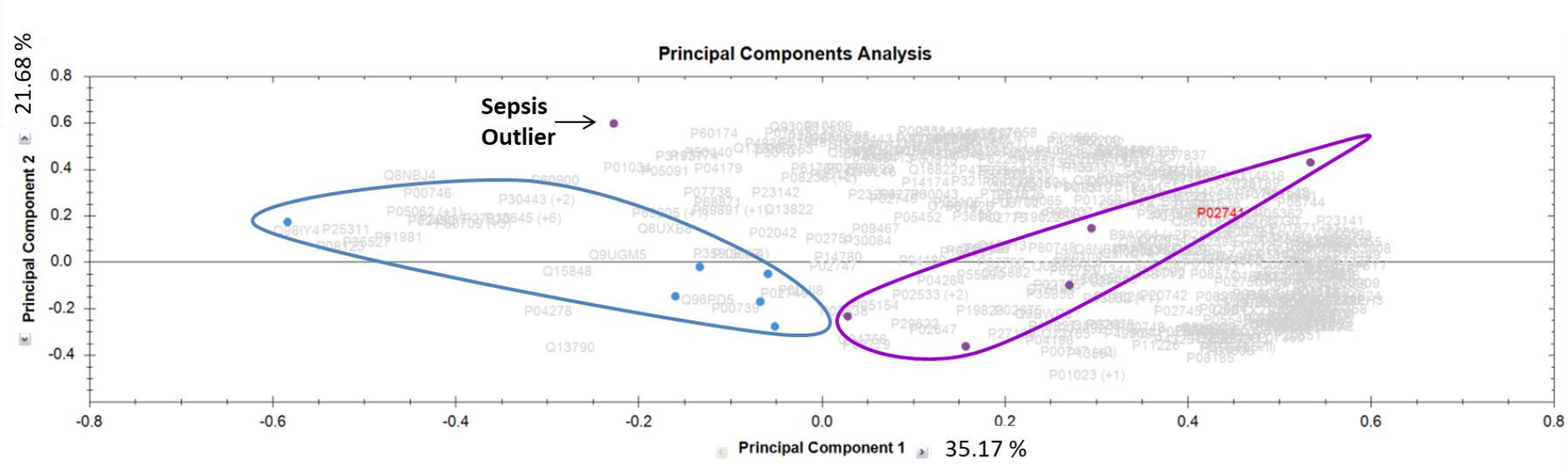


Figure 5.14 | PCA clustering of protein abundance values calculated in Progenesis.

Blue points represent control plasma samples and purple points represent sepsis plasma samples. Clear separation between control and sepsis samples observed. Sepsis sample 133 was identified as an outlier as it clustered with the control samples.

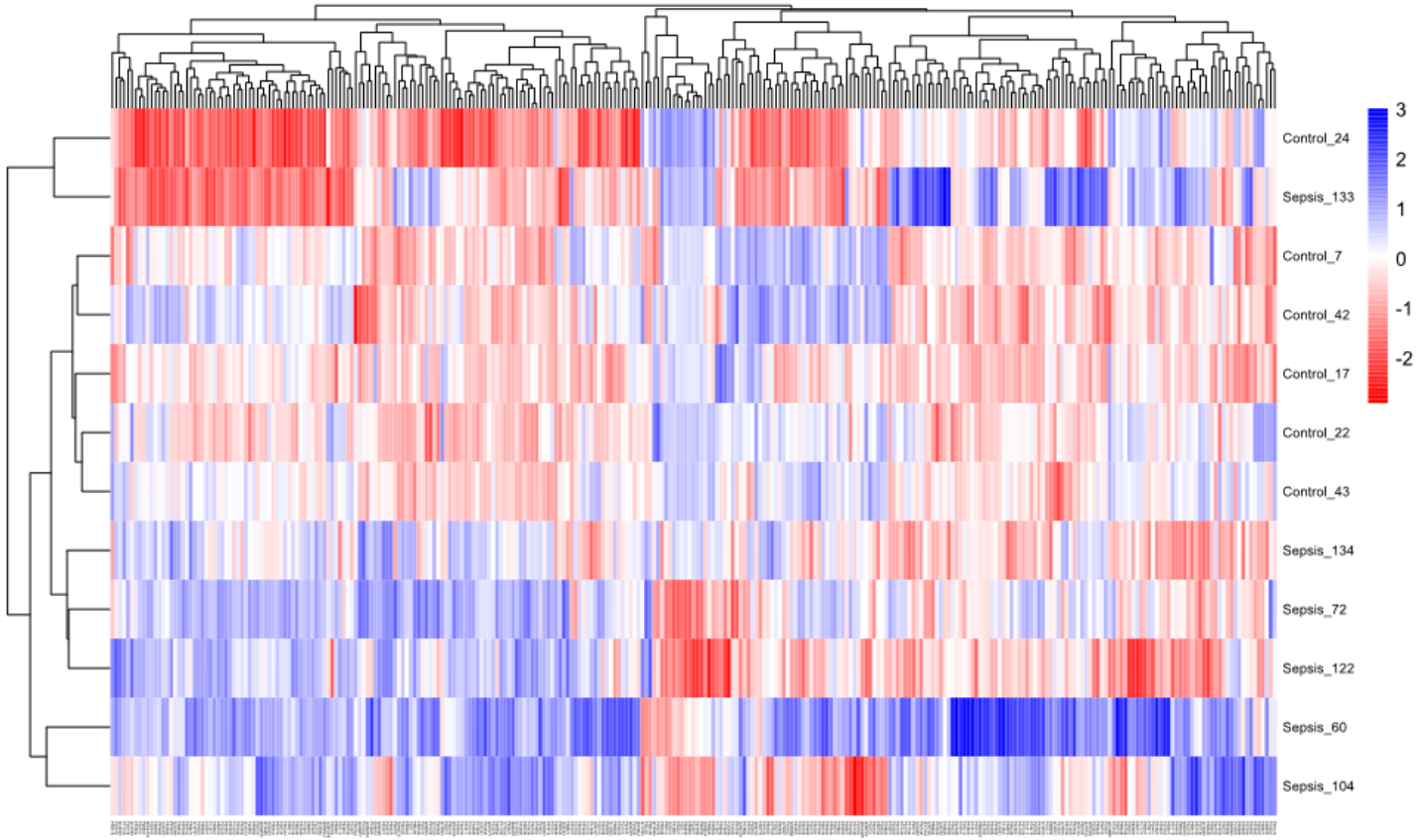


Figure 5.15 | Heatmap and hierarchical clustering of log10-transformed protein abundance data for comparative proteome analysis of patient cohort 1.

Normalised protein abundance values taken from Progenesis QI and log10- transformed. Samples clustered on protein abundance with high abundant proteins represented in blue and low abundant proteins represented in red.

Table 5.4 | Summary of differentially expressed proteins in cohort 1.

Proteins identified with a minimum of 2 peptides per protein, a protein score ≥ 20, a P value of <0.05, a q value of <0.05 cut-off and a fold change >1.5. Proteins ordered by maximum fold change.

Accession	Description	Unique peptides	Confidence score	Anova (p)	q Value	Max fold change	Highest mean condition
P02741	C-reactive protein	7	377.9	7.42E-08	1.24E-05	112.60	Sepsis
P0DJI9	Serum amyloid A-2 protein	2	287.46	0.00018	0.004808	75.40	Sepsis
Q01638	Interleukin-1 receptor-like 1	3	128.89	0.000659	0.011051	72.64	Sepsis
Q7RTV2	Glutathione S-transferase A5	2	66.49	0.005098	0.029342	67.20	Sepsis
P01009	Alpha-1-antitrypsin	36	2654.85	0.000551	0.010272	29.09	Sepsis
P01258	Calcitonin	2	158.85	0.004085	0.026033	22.49	Sepsis
P00738	Haptoglobin	16	2189.48	0.010482	0.044308	22.16	Sepsis
Q92598	Heat shock protein 105 kDa	3	134.82	0.002337	0.020414	21.31	Sepsis
P18428	Lipopolysaccharide-binding protein	11	716.36	0.000201	0.004808	16.20	Sepsis
P02763	Alpha-1-acid glycoprotein 1	7	773.92	0.003092	0.02358	13.82	Sepsis
P19652	Alpha-1-acid glycoprotein 2	8	712.86	0.003686	0.026033	12.05	Sepsis
P20711	Aromatic-L-amino-acid decarboxylase	2	84.06	3.11E-06	0.000174	11.69	Sepsis
Q03154	Aminoacylase-1	5	226.94	0.012584	0.046079	9.34	Sepsis
P01011	Alpha-1-antichymotrypsin	27	2214.44	7.17E-06	0.000301	8.98	Sepsis
P15144	Aminopeptidase N	5	180.83	0.000878	0.011337	6.14	Sepsis
P05362	Intercellular adhesion molecule 1	5	187.41	0.00035	0.007334	6.06	Sepsis
P02750	Leucine-rich alpha-2-glycoprotein	15	1062.94	0.002195	0.020414	5.42	Sepsis
P08123	Collagen alpha-2(I) chain	5	229.59	0.004593	0.027523	5.34	Control
Q8N6C8	Leukocyte immunoglobulin-like receptor subfamily A member 3	4	224.8	0.000824	0.011337	4.80	Sepsis

P33908	Mannosyl-oligosaccharide 1,2-alpha-mannosidase IA	6	280	0.000956	0.011452	4.66	Sepsis
P13796	Plastin-2	18	945.86	0.009294	0.043321	4.48	Sepsis
P06702	Protein S100-A9	4	189.73	0.008537	0.04213	4.22	Sepsis
P15151	Poliovirus receptor	3	102.17	0.004189	0.026033	4.06	Sepsis
P80188	Neutrophil gelatinase-associated lipocalin	4	167.98	0.012632	0.046079	4.05	Sepsis
Q15582	Transforming growth factor-beta-induced protein ig-h3	13	476.81	0.004058	0.026033	4.03	Sepsis
P01871	Ig mu chain C region	9	557.79	0.010826	0.044308	3.99	Sepsis
P06733	Alpha-enolase	8	609.01	0.001991	0.01965	3.88	Sepsis
Q86UD1	Out at first protein homolog	3	117.6	0.011355	0.045363	3.63	Sepsis
P01019	Angiotensinogen	16	1198.38	0.002433	0.020414	3.29	Sepsis
P55058	Phospholipid transfer protein	5	341.01	0.002775	0.02217	3.27	Sepsis
P24821	Tenascin	8	319.88	0.007839	0.039858	3.21	Sepsis
P13489	Ribonuclease inhibitor	2	78.36	0.00418	0.026033	3.17	Sepsis
Q8NBP7	Proprotein convertase subtilisin/kexin type 9	4	192.06	0.006086	0.032943	3.16	Sepsis
Q9UK55	Protein Z-dependent protease inhibitor	4	198.5	0.007034	0.036883	2.95	Sepsis
O00391	Sulfhydryl oxidase 1	7	332.16	0.001819	0.019075	2.79	Sepsis
Q96IY4	Carboxypeptidase B2	8	365.52	0.010787	0.044308	2.66	Control
P35527	Keratin, type I cytoskeletal 9	19	1143	0.014256	0.049889	2.52	Control
P12259	Coagulation factor V	20	724.91	0.009814	0.043336	2.29	Sepsis
P22792	Carboxypeptidase N subunit 2	12	755.14	0.014271	0.049889	2.17	Sepsis
Q96PD5	N-acetylmuramoyl-L-alanine amidase	9	577.76	0.011708	0.045399	1.61	Control

Summary

Of the 40 proteins identified as differentially expressed (fold change of 1.5 or greater), six were selected as possible candidate proteins based on previous work implicating their involvement in sepsis and immune response[239,258]. The six proteins were C-reactive protein, interleukin-1 receptor-like 1, glutathione S-transferase A5, calcitonin, protein S100A9 and neutrophil gelatinase-associated lipocalin. The role these proteins play in sepsis will be discussed in section 5.3.4. The normalised protein abundance plots for the candidate proteins are displayed Figure 5.16. To assess whether there was a significant statistical difference between the control and sepsis plasma groups, a t-test was performed ($P < 0.05$). A significant difference between the control and sepsis plasma was observed for all proteins (Figure 5.16). While the selected proteins show overall differences in expression between control and sepsis plasma, variation of protein abundance within the sepsis groups is relatively high. The sepsis samples in this study did not share a common bacterial diagnosis as there were three different bacterial causes; S*treptococcus empyema, S. pyogenes* and *Meningococcal meningitis*. Although the different bacteria all induced sepsis in the patients, the immune response may differ due to the different pathogens. Consequently, this may be reflected in the plasma proteome of the sepsis patients. Although biological variance is expected between patient samples, such noted difference in samples may have hindered candidate protein identification as protein expression may vary within the sepsis samples, reducing the statistical significance of the identified differentially expressed proteins.

To improve on the study design, selecting patients with a common bacterial diagnosis may result in a more robust dataset, potentially reducing the level of variance on protein expression and providing improved separation of samples in cluster analysis. In addition, increasing the number of samples in the study would provide greater statistical power to the analysis.

As the aims of this analysis were to identify a panel of proteins implicated in sepsis and not to identify a single novel sepsis biomarker, the initial findings of this analysis are promising. The results from cohort 1 can act as a validation dataset to provide further confidence to the results of the second comparative proteomics analysis.

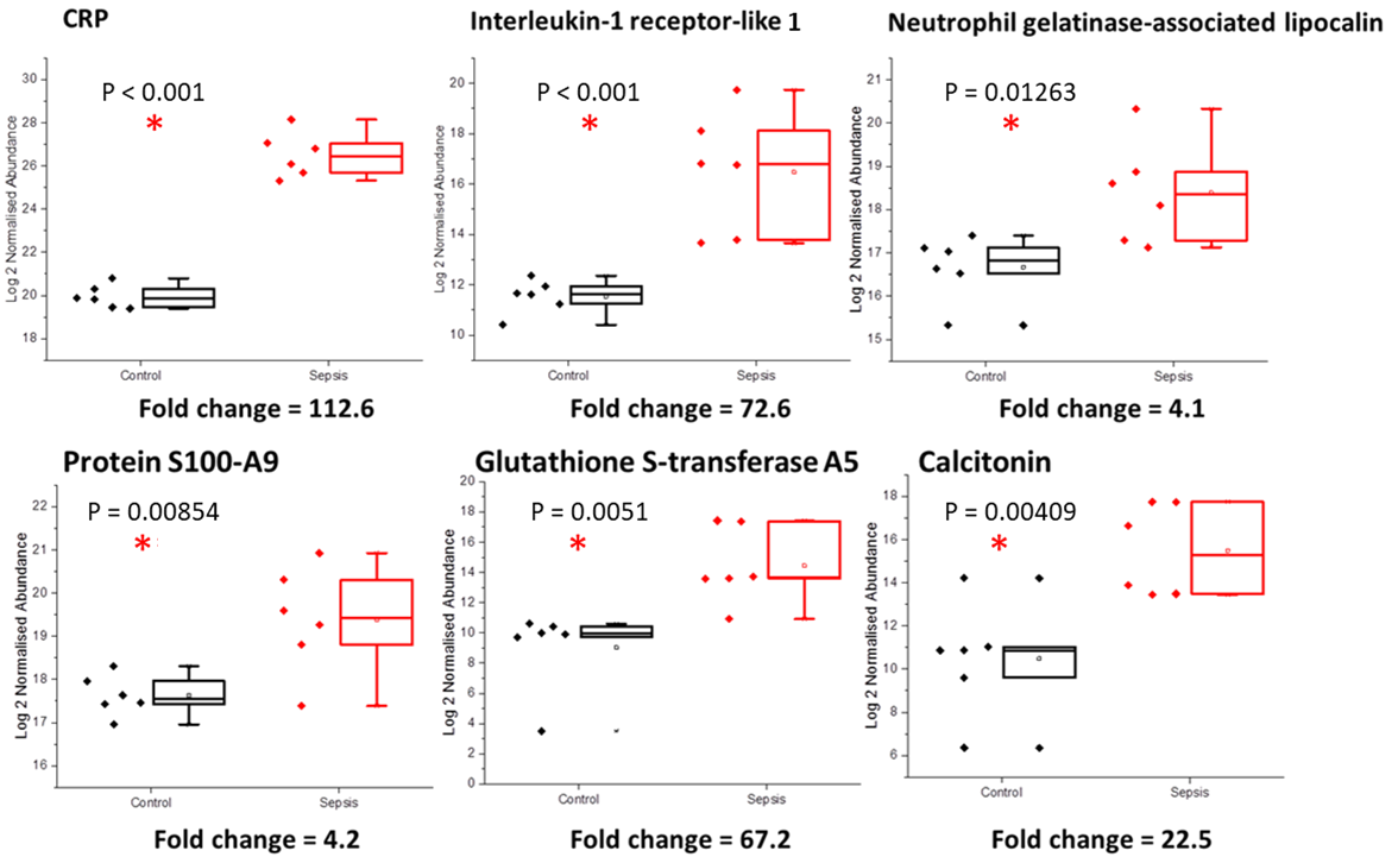


Figure 5.16 | Log 2 Progenesis QI normalised protein abundance plots of candidate proteins.

CRP, glutathione S-transferase A5, calcitonin, protein S100-A9, interleukin-1 receptor-like 1 and neutrophil gelatinase-associated lipocalin abundance plots. Plots display the log 2 normalised protein abundance values. Black plots represent control plasma samples and red plots represent sepsis plasma samples. Statistical analysis (t-test) was performed to assess for differences between control and sepsis plasma ($P > 0.05$). Red star (*) denotes significance.

5.3.4. Comparative proteomic analysis of sepsis plasma – Cohort 2

As discussed previously, to identify a panel of candidate markers of sepsis with greater statistical significance and confidence, a second cohort of 20 patient plasma samples was selected for comparative proteomics analysis. The plasma from 10 sepsis and 10 hospital control patients were analysed in this study, termed cohort 2. As before, the plasma from sepsis patients had been previously analysed for bacterial diagnosis and for this study were selected on the basis of meningococcal sepsis diagnosis. The control plasma was obtained from children admitted to the intensive care unit for elective cardiac surgery with no infection and who did not develop infection during their stay on the unit. Patient sample information is displayed in Table 5.5.

Plasma samples were prepared as discussed in section 5.3.3 and prior to the 2 hour LC-MS/MS sample analysis, each sample was analysed on a 30 minute gradient to check if the BPI chromatograms were of similar intensity. As the BPI chromatograms varied in intensity between samples, loading was adjusted to achieve similar intensities and peptides were subsequently analysed by LC-MS/MS on the Thermo Scientific QExactive HF on a 2 hour gradient with the same volume of sample loaded onto the instrument. The data was processed using Proteome Discoverer and the mgf file generated searched using MASCOT (Matrix Science) against a human database for peptide and protein identification with a 1% FDR filter. The number of proteins, peptides and PSMs identified in each sample were examined (Figure 5.17). When comparing the distribution of the data for the number of protein, peptide and PSMs identified between sepsis and control plasma, it was noted that sepsis samples generated the highest number of identifications (Figure 5.18) as observed in cohort 1. Similarities to cohort 1 were also observed in the spread of data, with high variability in the number of protein, peptide and PSMs identifications between samples. Variability in protein depletion of highly abundant proteins or inherent differences in patient samples could explain this. Due to the low number of protein identifications, sepsis sample 41 was identified as an outlier. To assess whether there was a significant statistical difference between the control and sepsis plasma groups in the number of protein, peptide and PSM identifications, a t-test was performed (significance level, $P < 0.05$). No statistical difference was observed between control and sepsis plasma in the number of protein, peptide and PSM identifications. BPI chromatograms of sepsis and control plasma digests were visually compared (Figure 5.19.); the sepsis plasma chromatograms were typically more complex, comparable with the samples in cohort 1.

To investigate the overlap in protein identification in control and sepsis plasma, a Venn diagram of proteins identified in at least 2 or more samples of the same condition were examined (Figure 5.17). A total of 209 proteins were identified with 178 proteins identified in both conditions demonstrating qualitative similarities between the control and sepsis plasma. Twenty-one proteins were identified in sepsis plasma only and 10 proteins were identified in control plasma only. Due to the low number of proteins identified in the sepsis plasma only, functional annotation was not carried out on this dataset.

Table 5.5 | Patient demographic and clinical data.

M = male, F = female, Age (Years).

Group	Sample IDs	Diagnosis	Age	Gender
Sepsis	330	*Meningococcal sepsis*	4.44	M
	370		1.78	M
	552		4.18	M
	204		0.67	M
	545		1.16	F
	549		1.2	M
	41		3.92	M
	429		2.63	M
	217		2.44	F
	487		6.99	F
Control	2	None	7.73	M
	32		0.24	F
	20		7.43	F
	6		0.87	F
	21		2.7	M
	51		4.37	M
	8		8.51	F
	48		6.05	M
	56		7.42	M
	33		7.75	M

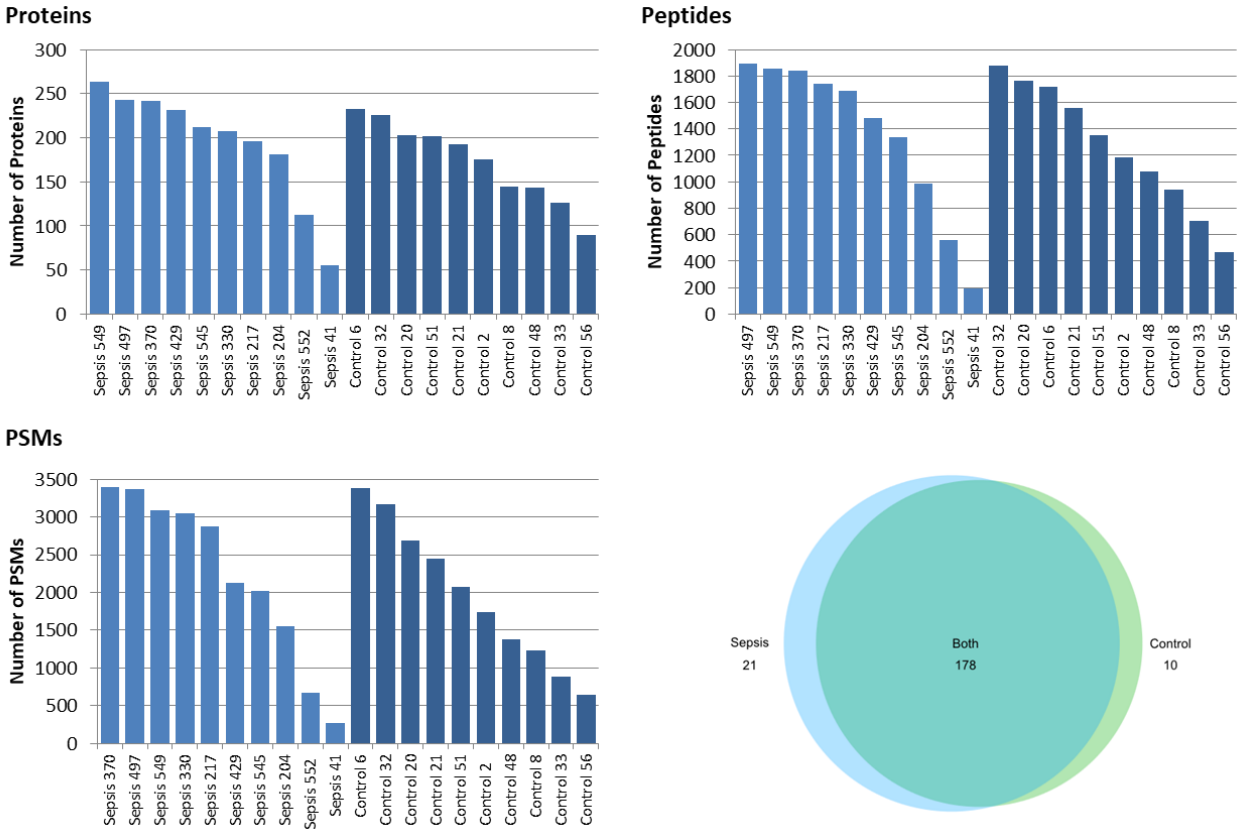


Figure 5.17 | Total number of protein, peptide and PSMs identified and overlap of protein identifications.

Data processed using Proteome Discoverer and searched in MASCOT against a human database with a 1 % FDR filter. Venn diagram: proteins identified in at least 2 samples of the same condition. 209 proteins identified in total. 178, 21 and 10 proteins identified in both conditions, sepsis only and control only respectively.

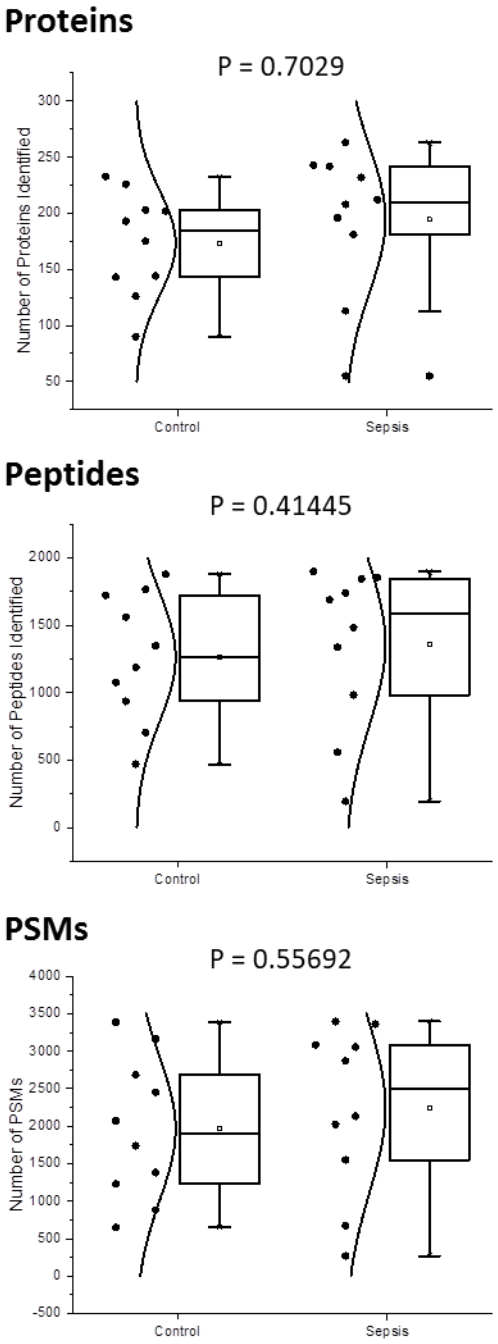


Figure 5.18 | Distribution of the number of protein, peptide and PSMs identified in control and sepsis plasma.

The overall number of protein, peptide and PSM identifications are highest in sepsis samples. Line represents median values. Whiskers represent minimum and maximum values. The median number of protein, peptide and PSMs identifications was highest in sepsis samples. Curve represents normal distribution. Statistical analysis (t-test) was performed to assess for differences between control and sepsis plasma ($P > 0.05$).

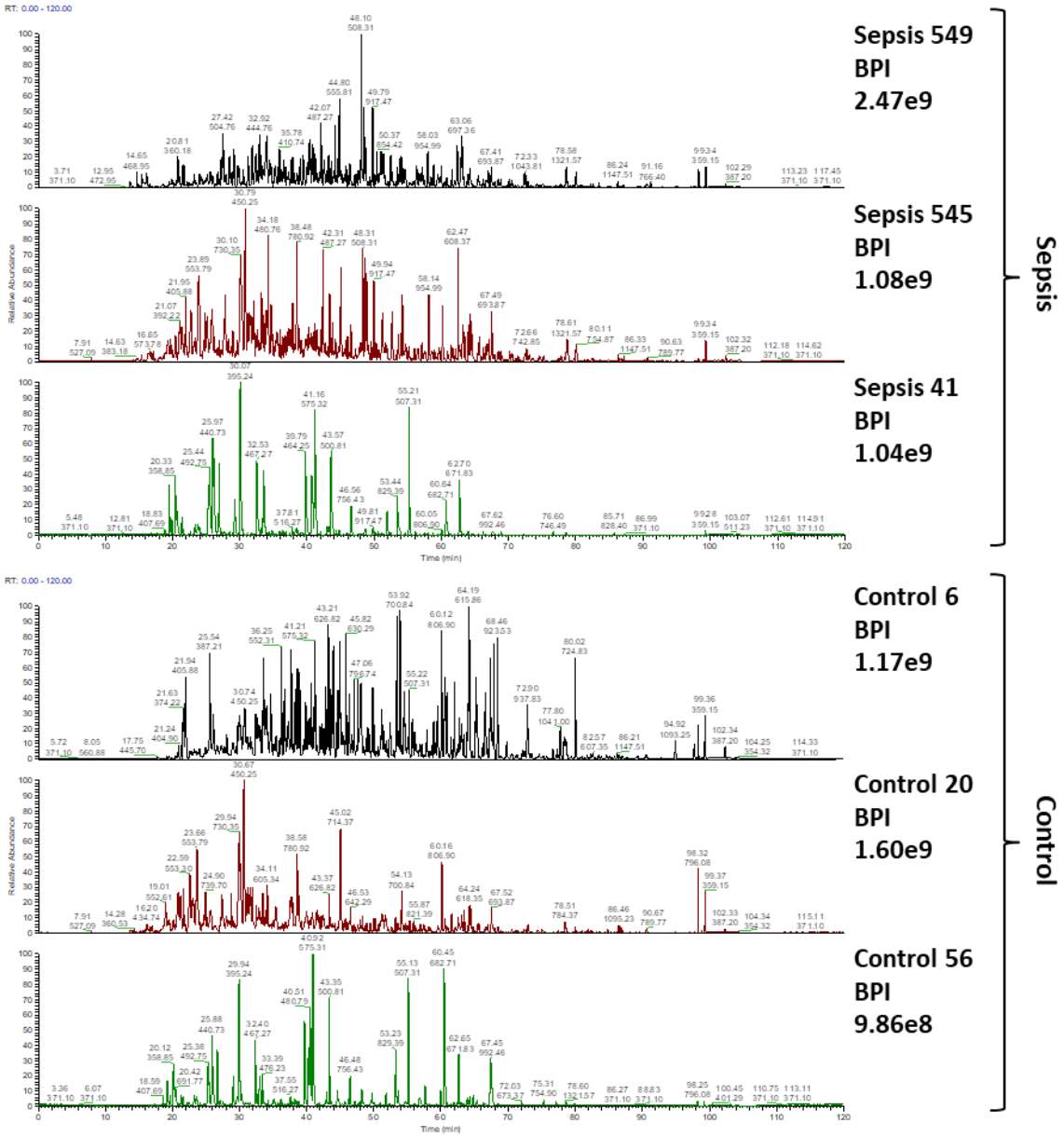


Figure 5.19 | Representative BPI chromatograms of sepsis and control plasma digests analysed by LC-MS/MS.

Depleted patient plasma FASP digested analysed on Thermo QExactive HF on a 2 hour gradient. Three BPI chromatograms were selected from sepsis and control samples that achieved highest, middle and lowest number of protein identifications.

To investigate quantitative difference in protein expression, label-free protein quantification was performed using Progenesis QI (Waters Corporation). Sepsis sample 545 was selected as the alignment reference file. The alignment scores achieved 69.6% or greater however control sample 56 failed to align (Figure 5.20). Alignment is an important feature of the Progenesis workflow to ensure accurate peak picking. Therefore, manual alignment of this sample was attempted to align the run correctly, however this also failed. When visually comparing the BPI chromatogram of this sample to the other samples in the cohort, large differences in peak intensities and distribution were observed (Figure 19). The BPI chromatogram for this sample looked more similar to non-depleted plasma suggesting a problem with the protein depletion spin column step of the protocol. The low number of proteins, peptides and PSMs identified would also support this explanation. The sample was therefore excluded from the Progenesis analysis. Apart from this sample, the alignment scores of the other samples indicate similar retention time profiles. Normalisation scores were also assessed (Figure 5.20), identifying sepsis sample 41 as an outlier with a normalisation score of 6.8. A total of 55 proteins were identified in this sample suggesting a technical problem with this sample. This sample was also excluded from the Progenesis analysis. A total of 307 proteins were quantified using relative quantification with a minimum of 2 unique peptides per protein and a protein score ≥ 20. As with cohort 1, an increase in the number of protein identification with Progenesis compared to MASCOT was observed.

The dynamic range of the sepsis and control plasma proteomes spans seven orders of magnitude (Figure 5.21). As observed in Cohort 1, the protein index curve has a less steep gradient towards the highly abundant proteins compared to non-depleted plasma again confirming depletion of highly abundant proteins. The dynamic range plot was ordered by control sample protein abundance and therefore proteins identified in the sepsis group that did not follow this trend were highlighted (Figure 5.22). Proteins that were more abundant in the sepsis group include haptoglobin, C-reactive protein, myoglobin, serum amyloid A-1 protein and leucine rich alpha-2 glycoprotein. An increase in the abundance of acute-phase proteins (haptoglobin, C-reactive protein and serum amyloid A-1 protein) in the sepsis group is expected. Leucine rich alpha-2 glycoprotein is expressed by granulocytes undergoing differentiation[259] and various studies have shown that expression levels in serum increase in bacterial infection and in inflammatory diseases[260-262]. Raised levels of myoglobin have been detected in patients with rhadbomyolysis, the breakdown in muscle tissue. While there are many causes of rhadbomyolysis, various infections can lead to

muscle breakdown and a recent study identified raised serum myoglobin as a predictive marker for increased sepsis severity[263]. As in Cohort 1, the most variability in protein abundance between the control and sepsis groups was observed at the upper and lower ends of the dynamic range plot (Figure 5.22). Variability of protein abundances within the sepsis groups compared to the control group was also highest in these portions of the plot as evident by the larger error bars (Figure 5.22).

The dataset was subjected to principal component analysis to assess the level of separation between control and sepsis plasma (Figure 5.23.). Clear discrimination was observed between control (blue dots) and sepsis (purple dots) plasma with the samples clustering into distinct groups. The control samples formed two separate clusters within the group. To further assess the separation of samples based on differences in protein expression, protein abundance values were log10-transformed and analysed by hierarchical clustering (Figure 5.24). Two distinct clusters of sepsis plasma and three distinct clusters of control plasma samples were observed.

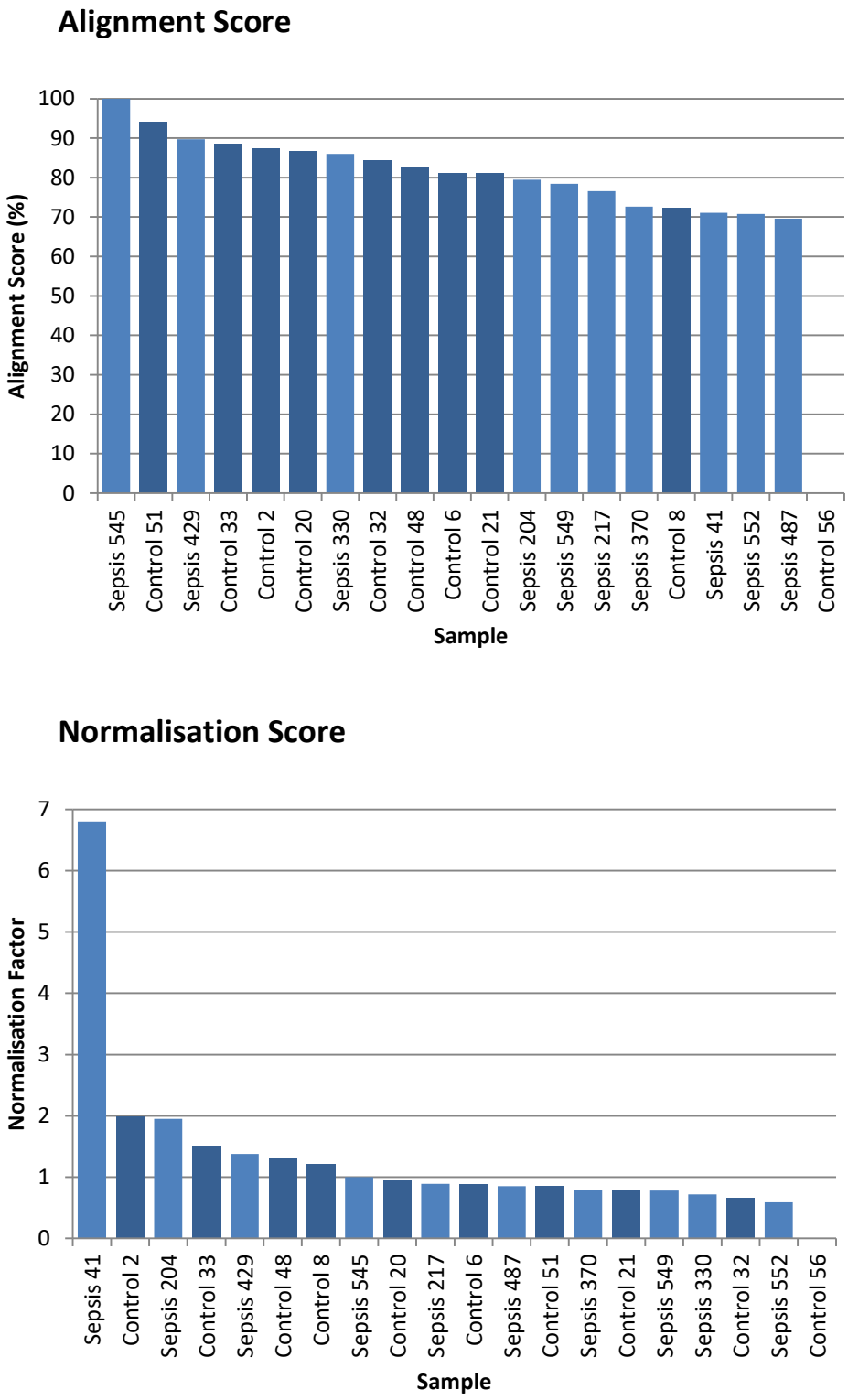


Figure 5.20 | Summary of alignment and normalisation scores of patient samples.

Top: Progenesis alignment scores. Sepsis sample 545 was selected as the alignment reference file. Control sample 56 failed to align. The other samples achieved an alignment scores of 69.6% or greater. Bottom: Progenesis normalisation scores were in the accepted range excluding sepsis sample 41 that had a normalisation score of over 6.5. Control sample 56 and sepsis sample 41 were excluded from further analysis.

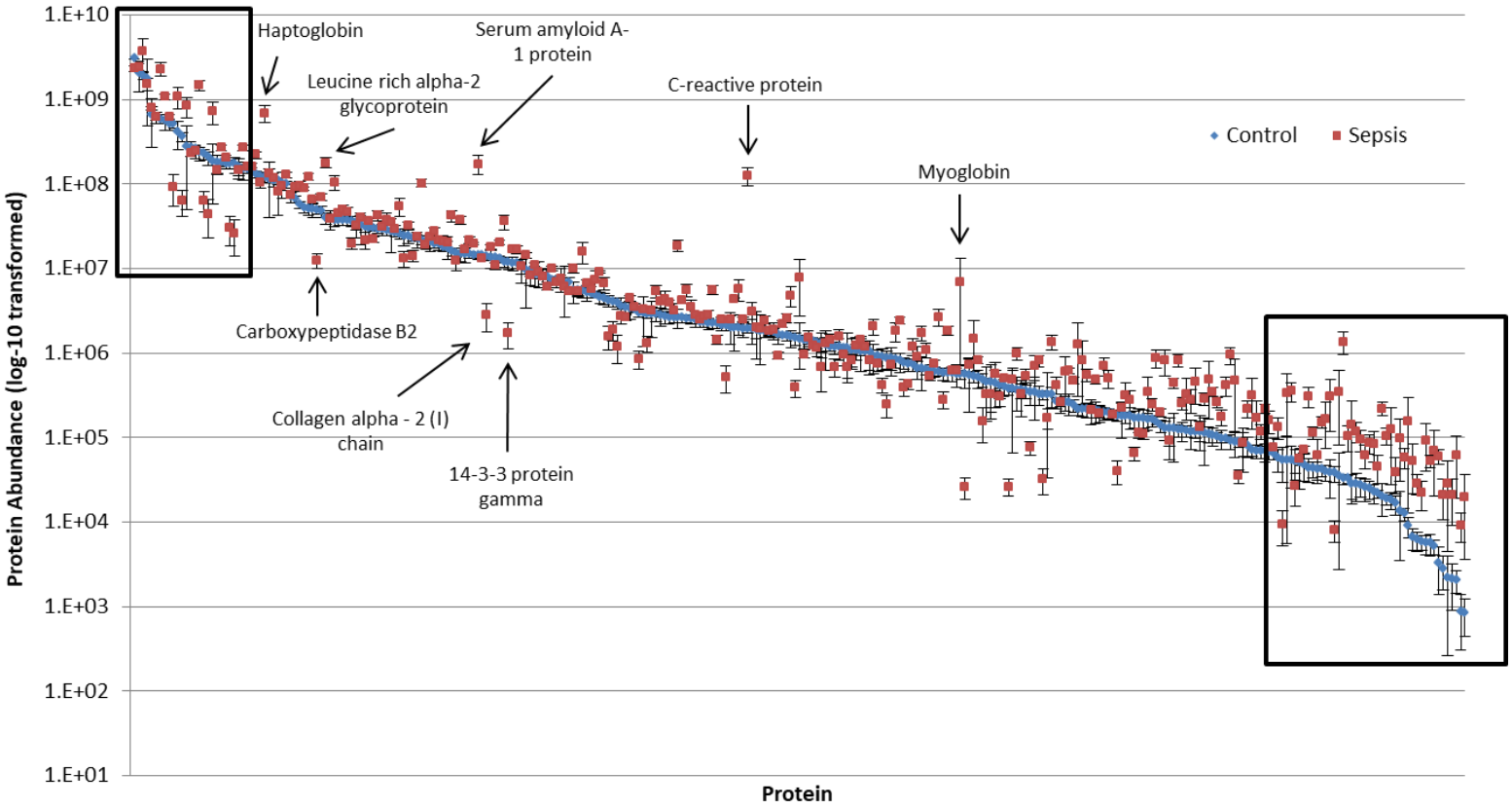


Figure 5.21 | Dynamic range of control and sepsis plasma proteome.

The mean normalised abundance value of each protein expressed in the control and sepsis plasma was calculated using Progenesis QI relative quantification using non-conflicting peptides. Ordered by control abundance values. Error bars represent standard error. Blue points correspond to control plasma and red points correspond to sepsis plasma. Most abundant and least abundant proteins highlighted in black boxes and examine further in Figure 5.22. Sepsis outlier proteins labelled.

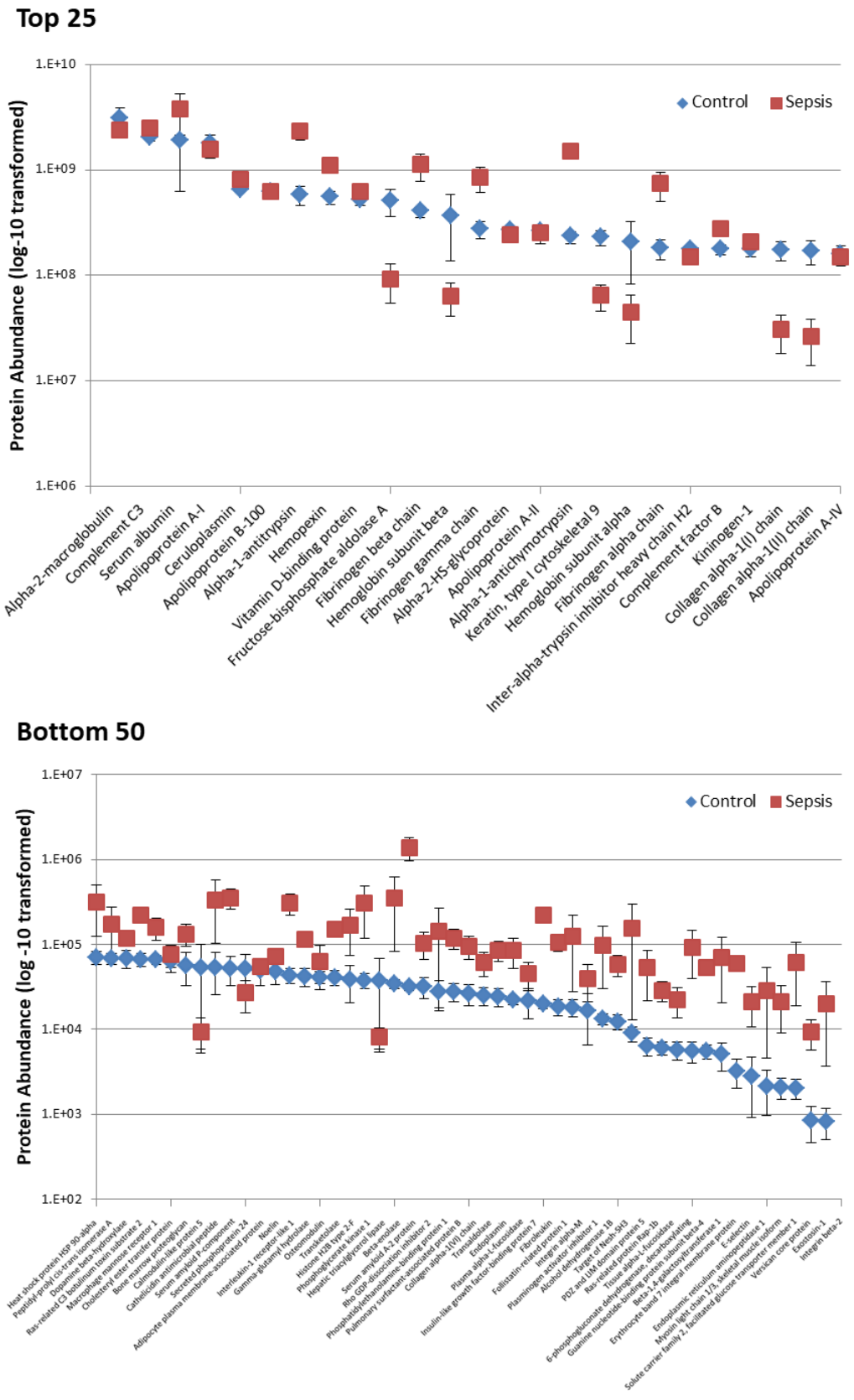


Figure 5.22 | Upper and lower dynamic range of control and sepsis plasma proteome.

Ordered by control abundance values. Top: Top 25 most abundant proteins. Bottom: 50 least abundant proteins. Error bars represent standard error.

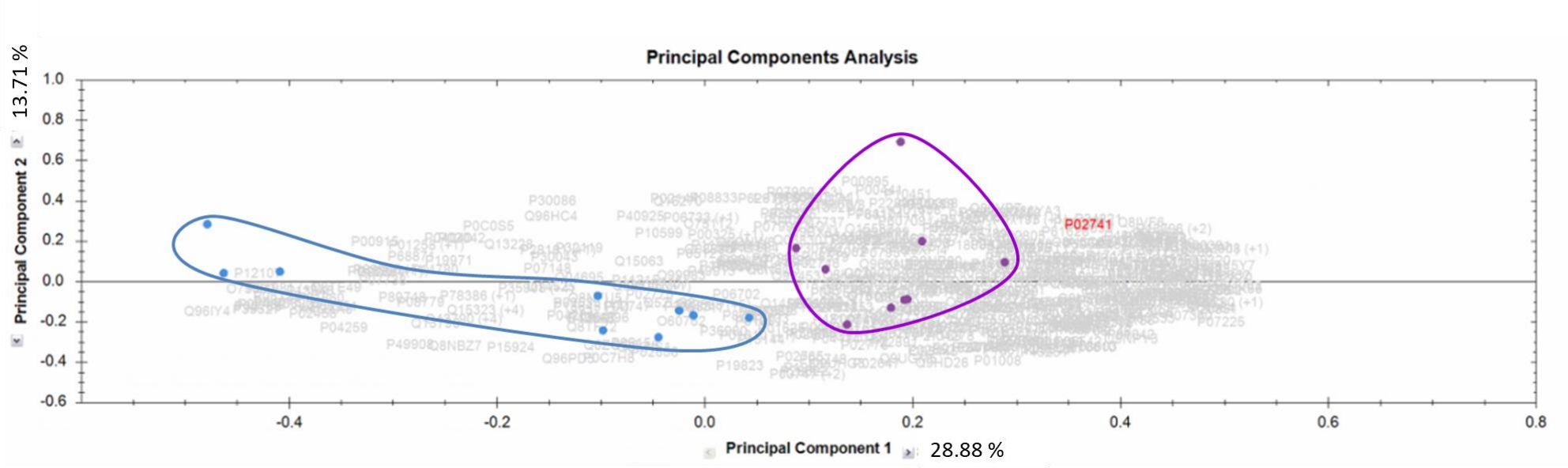


Figure 5.23 | PCA clustering of protein abundance values calculated in Progenesis QI.

Blue points represent control plasma samples and purple points represent sepsis plasma samples. Clear separation between control and sepsis samples observed.

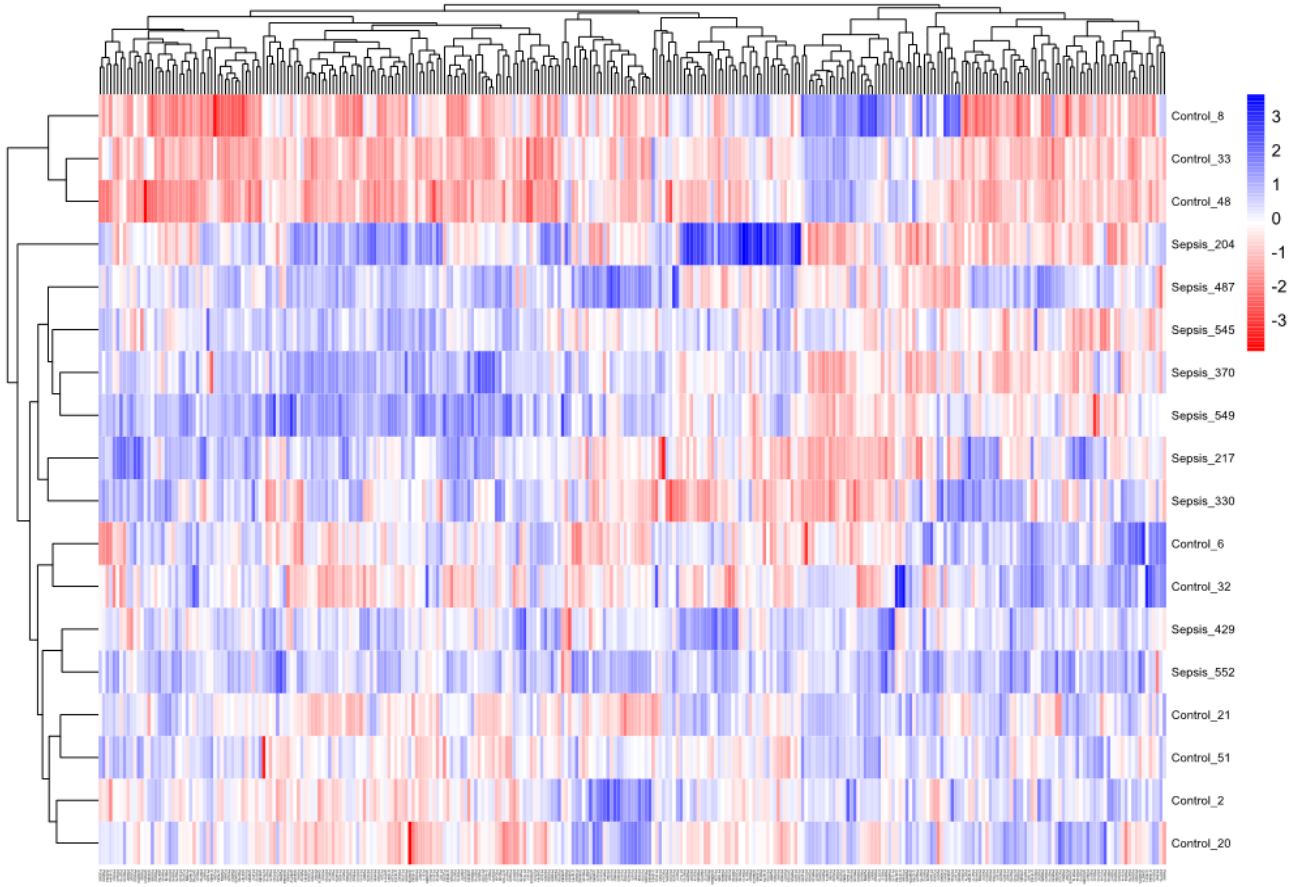


Figure 5.24 | Heatmap and hierarchical clustering of log10-transformed protein abundance data for comparative proteome analysis of patient cohort 2.

Normalised protein abundance values taken from Progenesis QI and log10- transformed. Samples clustered on protein abundance with high abundant proteins represented in blue and low abundant proteins represented in red.

To investigate differentially expressed proteins between control and sepsis plasma, the dataset was filtered further; a P value of <0.05 and a q value of <0.05 cut-off was applied. A total of 107 proteins were differentially expressed between the two groups with a fold change greater than 1.5. Twenty proteins were quantified as having the highest mean abundances in the control group and 87 proteins quantified as having the highest mean abundance in the sepsis group (Table 5.6). Functional annotation of the sepsis protein dataset was performed using the PANTHER classifications system[264] and DAVID[265,266] (Figure 5.25). Analysis of the biological processes associated with the identified proteins using PANTHER (Figure 5.25 A) highlighted 11 biological process categories with 144 process hits in total. Cellular process was the largest category with 43 proteins associated with this process. Further analysis of this category identified proteins strongly associated with cell communication, in particular signal transduction. As expected, cell signalling and communication is an important feature of the immune system in order to recruit and activate the appropriate effectors cells[267]. Interestingly, proteins associated with developmental processes were enriched in the sepsis plasma dataset. As the sepsis patients are neonates and children, identification of developmental proteins would be expected and as patient demographic is unknown for control patients it would suggest that sepsis patients are overall younger. Response to stimulus and immune system processes were also enriched categories as expected in response to sepsis.

KEGG pathway analysis using DAVID identified 5 KEGG terms with statistical significance (Bonferroni <0.05). The most enriched pathway was the complement and coagulation cascade (Bonferroni 1.37E-13). It is well known that sepsis activates the complement and coagulation cascade[268] and is thought to be responsible for the severe complications of disseminated intravascular coagulation that results in blood clot formation[269]. Gene ontology analysis of enriched biological processes using DAVID identified 16 GO terms with statistical significance. Enriched terms include acute-phase response, innate immune response and regulation of complement activation (Bonferroni of 3.27E-14, 0.001261 and 2.53E-04 respectively) signifying and confirming the systemic inflammatory and immune response associated with sepsis.

Table 5.6 | Summary of differentially expressed proteins in Cohort 2.

Total of 107 proteins identified with a minimum of 2 peptides per protein, a protein score ≥ 20, a P value of <0.05, a q value of <0.05 cut-off and a fold change >1.5. Proteins ordered by the maximum fold change.

Accession	Description	Unique peptides	Confidence score	Anova (p)	q Value	Max fold change	Highest mean condition
P02741	C-reactive protein	7	386.48	6.95E-09	1.39E-06	66.04	Sepsis
P0DJI9	Serum amyloid A-2 protein	2	374.35	2.62E-05	0.000464	42.82	Sepsis
P13611	Versican core protein	3	113.36	0.01016299	0.018992	30.56	Sepsis
P05107	Integrin beta-2	2	92.04	0.00410268	0.01038	23.89	Sepsis
P13646	Keratin, type I cytoskeletal 13	3	521.56	0.00039827	0.002156	21.28	Control
P16581	E-selectin	2	51.4	0.00096943	0.004212	18.42	Sepsis
Q9HAV0	Guanine nucleotide-binding protein subunit beta-4	2	64.52	0.01913902	0.028979	16.72	Sepsis
P0DJI8	Serum amyloid A-1 protein	3	727.95	1.94E-06	7.76E-05	12.37	Sepsis
Q14314	Fibroleukin	2	98.37	5.99E-06	0.000171	11.10	Sepsis
Q16394	Exostosin-1	2	41.72	0.00285504	0.00808	10.90	Sepsis
P26022	Pentraxin-related protein PTX3	7	440.16	0.00100148	0.004259	10.71	Sepsis
P11166	Solute carrier family 2, facilitated glucose transporter member 1	3	91.79	0.02455532	0.03432	10.05	Sepsis
P15291	Beta-1,4-galactosyltransferase 1	2	87.25	1.55E-05	0.000386	9.66	Sepsis
Q9NZ08	Endoplasmic reticulum aminopeptidase 1	2	58.14	0.00503163	0.012253	7.64	Sepsis
Q01638	Interleukin-1 receptor-like 1	2	82.95	0.00771611	0.015422	7.17	Sepsis
P18428	Lipopolysaccharide-binding protein	11	725.12	7.63E-08	6.45E-06	7.13	Sepsis
P61981	14-3-3 protein gamma	3	171.18	0.00020034	0.001437	6.81	Control
P02743	Serum amyloid P-component	2	125.83	0.00020715	0.001437	6.76	Sepsis

Q8N6C8	Leukocyte immunoglobulin-like receptor subfamily A member 3	5	226.5	4.71E-06	0.000157	6.72	Sepsis
P80188	Neutrophil gelatinase-associated lipocalin	5	197.94	0.00218279	0.007123	6.54	Sepsis
P02458	Collagen alpha-1(II) chain	2	75.35	0.00607307	0.013338	6.48	Control
P01011	Alpha-1-antichymotrypsin	26	2509.16	9.68E-08	6.45E-06	6.41	Sepsis
P00738	Haptoglobin	15	2281.35	8.77E-05	0.000877	5.81	Sepsis
Q12841	Follistatin-related protein 1	2	61.79	7.48E-05	0.000831	5.73	Sepsis
P02452	Collagen alpha-1(I) chain	10	594.98	0.00024787	0.001501	5.72	Control
P05451	Lithostathine-1-alpha	2	176.74	0.00382926	0.009939	5.71	Sepsis
P04075	Fructose-bisphosphate aldolase A	8	471.39	0.00127576	0.005	5.53	Control
P08123	Collagen alpha-2(I) chain	6	354.01	0.00734987	0.015144	4.92	Control
P02750	Leucine-rich alpha-2-glycoprotein	15	1214.56	3.42E-07	1.71E-05	4.87	Sepsis
Q7Z7G0	Target of Nesh-SH3	2	65.6	0.0024793	0.007508	4.74	Sepsis
P04066	Tissue alpha-L-fucosidase	2	94.52	0.00990651	0.018925	4.72	Sepsis
P13796	Plastin-2	17	703.55	1.87E-05	0.000414	4.62	Sepsis
P02763	Alpha-1-acid glycoprotein 1	7	849.4	8.11E-05	0.000853	4.58	Sepsis
Q15323	Keratin, type I cuticular Ha1	3	701.15	0.04106734	0.048282	4.51	Control
P07988	Pulmonary surfactant-associated protein B	4	157.44	0.00359662	0.009458	4.33	Sepsis
P07858	Cathepsin B	4	149.21	0.00896988	0.017576	4.33	Sepsis
P24821	Tenascin	16	724.62	0.00010548	0.000917	4.32	Sepsis
P02671	Fibrinogen alpha chain	43	3484.99	0.00440118	0.010996	4.03	Sepsis
P01009	Alpha-1-antitrypsin	35	2700.63	0.00038585	0.002156	4.00	Sepsis
P04259	Keratin, type II cytoskeletal 6B	2	866.8	0.00697573	0.014676	3.96	Control
Q96IY4	Carboxypeptidase B2	12	528.1	7.13E-05	0.000831	3.89	Control
Q9BTY2	Plasma alpha-L-fucosidase	4	104.7	0.01491552	0.023903	3.85	Sepsis
P29401	Transketolase	4	118.58	0.00215131	0.007123	3.71	Sepsis

P33908	Mannosyl-oligosaccharide 1,2-alpha-mannosidase IA	7	285.97	3.84E-05	0.000511	3.60	Sepsis
Q86VB7	Scavenger receptor cysteine-rich type 1 protein M130	8	248.03	0.00296502	0.008231	3.60	Sepsis
P35527	Keratin, type I cytoskeletal 9	28	1919.65	0.00042287	0.002179	3.58	Control
P12109	Collagen alpha-1(VI) chain	3	109.5	0.01397177	0.022889	3.58	Sepsis
P14625	Endoplasmin	3	173.23	0.002281	0.007123	3.55	Sepsis
P05362	Intercellular adhesion molecule 1	3	146.4	3.12E-05	0.00048	3.37	Sepsis
Q9UK55	Protein Z-dependent protease inhibitor	9	438.55	0.00018637	0.001433	3.29	Sepsis
P15153	Ras-related C3 botulinum toxin substrate 2	2	52.04	0.00021973	0.001464	3.29	Sepsis
P01860	Ig gamma-3 chain C region	4	610.9	0.0187447	0.028599	3.25	Control
Q13822	Ectonucleotide pyrophosphatase/phosphodiesterase family member 2	6	273	0.00257191	0.007559	3.18	Sepsis
P07451	Carbonic anhydrase 3	9	335.97	0.00039915	0.002156	3.10	Sepsis
P02679	Fibrinogen gamma chain	30	2632.55	0.00569326	0.013079	3.08	Sepsis
O00391	Sulfhydryl oxidase 1	12	578.25	2.20E-05	0.000439	3.03	Sepsis
P35443	Thrombospondin-4	10	727.86	0.01140923	0.020252	3.00	Sepsis
P12259	Coagulation factor V	27	1149.84	0.02400255	0.033784	2.96	Sepsis
P01019	Angiotensinogen	16	1573.09	0.00086327	0.003834	2.83	Sepsis
P26038	Moesin	4	138.73	0.00239502	0.007364	2.80	Sepsis
P07339	Cathepsin D	5	278.47	0.00170099	0.006071	2.73	Sepsis
Q92820	Gamma-glutamyl hydrolase	3	91.34	0.00638033	0.013861	2.73	Sepsis
P02675	Fibrinogen beta chain	32	3635.65	0.02290774	0.033177	2.73	Sepsis
P10451	Osteopontin	4	203.67	0.00264403	0.007659	2.72	Sepsis
P04275	von Willebrand factor	77	4355.68	0.00207616	0.007033	2.67	Sepsis
P20742	Pregnancy zone protein	6	1024.14	0.02887027	0.037227	2.64	Sepsis
P14543	Nidogen-1	8	322.88	0.00254078	0.007559	2.63	Sepsis
Q92496	Complement factor H-related protein 4	2	112.25	0.0076738	0.015422	2.61	Sepsis

P36222	Chitinase-3-like protein 1	3	92.74	0.00168644	0.006071	2.60	Control
P12111	Collagen alpha-3(VI) chain	8	346.22	0.01165302	0.020252	2.60	Sepsis
P19652	Alpha-1-acid glycoprotein 2	6	713.1	0.00117004	0.004772	2.57	Sepsis
P54108	Cysteine-rich secretory protein 3	3	103.74	0.00064637	0.003151	2.54	Sepsis
P08571	Monocyte differentiation antigen CD14	8	623.24	0.00011237	0.000936	2.53	Sepsis
P04004	Vitronectin	17	1141.87	0.00083412	0.003789	2.48	Sepsis
Q9Y5Y7	Lymphatic vessel endothelial hyaluronic acid receptor 1	5	222.91	0.00020847	0.001437	2.45	Sepsis
P13727	Bone marrow proteoglycan	2	54.95	0.01026255	0.018992	2.38	Sepsis
P19320	Vascular cell adhesion protein 1	8	302.94	0.000709	0.003374	2.36	Sepsis
P01033	Metalloproteinase inhibitor 1	5	219.2	0.01206942	0.020622	2.25	Sepsis
Q92954	Proteoglycan 4	11	523.16	0.00023382	0.001501	2.22	Sepsis
P55058	Phospholipid transfer protein	8	494.81	0.03585503	0.044236	2.21	Sepsis
P49747	Cartilage oligomeric matrix protein	15	872.88	0.00224811	0.007123	2.19	Sepsis
P02788	Lactotransferrin	9	325.9	0.0065201	0.014012	2.14	Control
P00746	Complement factor D	3	118.32	0.00766272	0.015422	2.11	Sepsis
P80748	Ig lambda chain V-III region LOI	2	154.81	0.01945139	0.02923	2.04	Control
O75636	Ficolin-3	7	492.95	0.00046716	0.002334	2.01	Sepsis
P07359	Platelet glycoprotein Ib alpha chain	3	118.01	0.00224516	0.007123	1.99	Sepsis
P02790	Hemopexin	28	2299.97	0.00143236	0.005505	1.96	Sepsis
P00739	Haptoglobin-related protein	10	1285.99	0.01387822	0.022889	1.92	Sepsis
P05154	Plasma serine protease inhibitor	7	317.29	0.02626571	0.035444	1.91	Control
P15169	Carboxypeptidase N catalytic chain	13	566.17	0.0068095	0.014479	1.85	Sepsis
Q14624	Inter-alpha-trypsin inhibitor heavy chain H4	44	3769.86	0.00039678	0.002156	1.84	Sepsis
P02656	Apolipoprotein C-III	4	312.89	0.03991473	0.047485	1.84	Control
O75882	Attractin	24	1358.53	0.02480577	0.034429	1.76	Control
P07195	L-lactate dehydrogenase B chain	7	375.61	0.03974569	0.047485	1.76	Sepsis

P04003	C4b-binding protein alpha chain	24	1459.54	0.0014738	0.005558	1.75	Sepsis
P49908	Selenoprotein P	6	299.94	6.57E-05	0.000821	1.72	Control
P05155	Plasma protease C1 inhibitor	21	1594.66	0.00287036	0.00808	1.71	Sepsis
P00742	Coagulation factor X	10	542.87	0.00176452	0.006187	1.67	Sepsis
P05109	Protein S100-A8	4	192.89	0.02824177	0.036892	1.67	Sepsis
P22105	Tenascin-X	10	512.23	0.01056853	0.019148	1.65	Sepsis
P02748	Complement component C9	24	1728.08	0.01153078	0.020252	1.65	Sepsis
P07358	Complement component C8 beta chain	21	1373.63	0.00390855	0.010015	1.58	Sepsis
P00751	Complement factor B	31	2526.62	0.00462938	0.011423	1.55	Sepsis
P22792	Carboxypeptidase N subunit 2	9	656.69	0.0088291	0.017472	1.54	Sepsis
P20851	C4b-binding protein beta chain	6	333.84	0.01271469	0.021536	1.52	Sepsis
Q06033	Inter-alpha-trypsin inhibitor heavy chain H3	20	1181.93	0.00561146	0.013041	1.51	Sepsis
Q13790	Apolipoprotein F	2	88.09	0.03830306	0.046373	1.50	Control

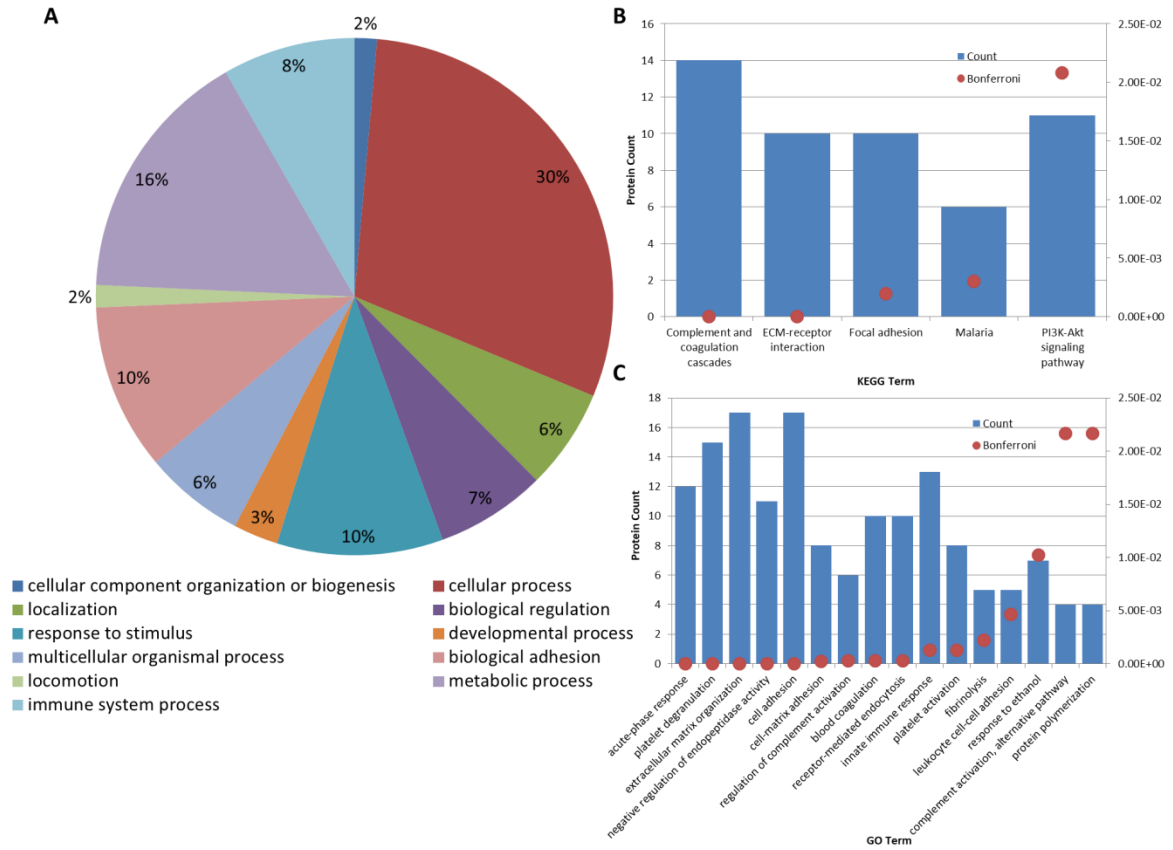


Figure 5.25| Functional annotation of up-regulated proteins in sepsis.

(A) Panther Classification – Biological Processes. 87 of the proteins up-regulated in sepsis associated with 144 biological processes (B) KEGG Terms 53 proteins covering 60.2% of dataset. (C) GoTerm Enrichment Analysis of Biological Processes. Bonferroni correction < 0.05, represented by red X KEGG and Go Enrichment analysis carried out using DAVID. Bonferroni correction adjusts the P-value to reduce the likelihood of false-positive results.

To determine the relationship between proteins quantified in both Cohort 1 and Cohort 2, the log 10 transformed mean protein abundance values for control and sepsis groups were compared (Figure 5.26). In total, 227 proteins were quantified in both studies and used to assess the relationship between cohorts. A strong positive linear correlation was observed between cohorts 1 and 2 for both sepsis and control proteins. A correlation R^2 value of 0.81 and 0.83 for the control and sepsis groups was achieved. This result suggests that the mean relative protein abundance values calculated for the two cohorts are in strong agreement demonstrating that the despite the independent studies analysing different samples, relative protein abundance values are similar.

The proteins identified and quantified as differentially expressed in both cohorts using Progenesis were also compared to determine characteristic proteome changes in sepsis (Table 5.7). A total of 25 proteins that met the q value filter of <0.05 in both studies were identified. Three proteins were identified as down-regulated in control plasma and the remaining 22 identified as up-regulated in sepsis.

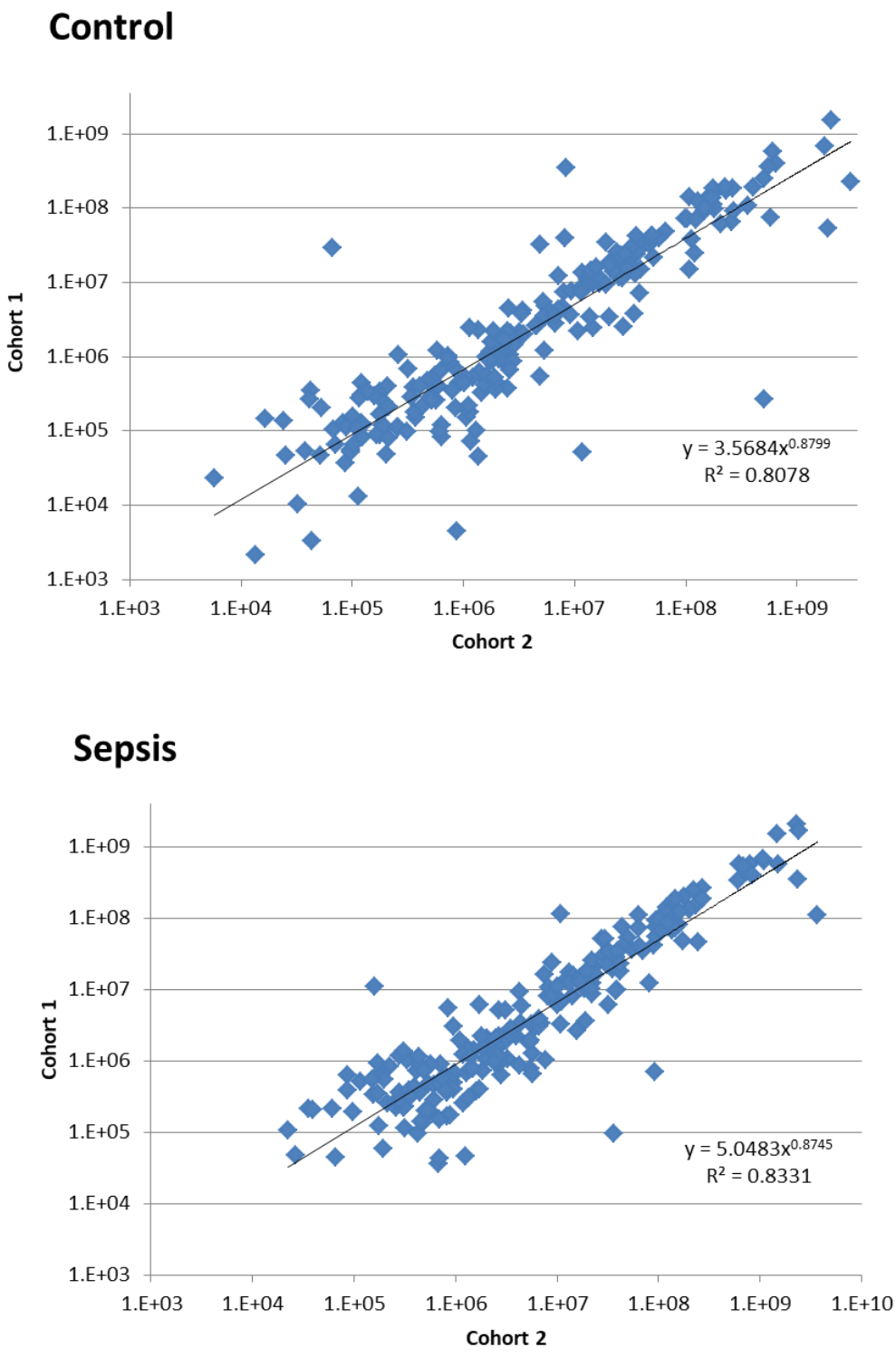


Figure 5.26 | Relationship of the average protein abundances between cohort 1 and cohort 2 measured using Progensis QI.

Top: Average protein abundance values for control sample. Bottom: Average protein abundance values for sepsis sample. The data was log10 transformed and linear relationship for both the control and sepsis data observed with a R^2 value of 0.81 and 0.83 respectively. Line shows linear regression.

Table 5.7 | Summary of proteins identified in both patient cohorts.

Proteins identified with a minimum of 2 peptides per protein, a protein score ≥ 20, a P value of <0.05, a q value of <0.05 cut-off and a fold change >1.5. Proteins ordered alphabetically.

			Cohort 1				Cohort 2			
Accession	**Description**	**Highest mean condition**	**Unique peptides**	**Anova (p)**	**q Value**	**Max fold change**	**Unique peptides**	**Anova (p)**	**q Value**	**Max fold change**
P02763	Alpha-1-acid glycoprotein 1	Sepsis	7	0.003092	0.02358	13.82	7	8.11E-05	0.00085283	4.58
P19652	Alpha-1-acid glycoprotein 2	Sepsis	8	0.003686	0.026033	12.05	6	0.00117	0.00477243	2.57
P01011	Alpha-1-antichymotrypsin	Sepsis	27	7.17E-06	0.000301	8.98	26	9.68E-08	6.45E-06	6.41
P01009	Alpha-1-antitrypsin	Sepsis	36	0.000551	0.010272	29.09	35	0.000386	0.00215613	4.00
P01019	Angiotensinogen	Sepsis	16	0.002433	0.020414	3.29	16	0.000863	0.00383417	2.83
Q96IY4	Carboxypeptidase B2	Control	8	0.010787	0.044308	2.66	12	7.13E-05	0.00083057	3.89
P22792	Carboxypeptidase N subunit 2	Sepsis	12	0.014271	0.049889	2.17	9	0.008829	0.01747154	1.54
P12259	Coagulation factor V	Sepsis	20	0.009814	0.043336	2.29	27	0.024003	0.03378355	2.96
P08123	Collagen alpha-2(I) chain	Control	5	0.004593	0.027523	5.34	6	0.00735	0.01514413	4.92
P02741	C-reactive protein	Sepsis	7	7.42E-08	1.24E-05	112.60	7	6.95E-09	1.39E-06	66.04
P00738	Haptoglobin	Sepsis	16	0.010482	0.044308	22.16	15	8.77E-05	0.0008769	5.81
P05362	Intercellular adhesion molecule 1	Sepsis	5	0.00035	0.007334	6.06	3	3.12E-05	0.00048021	3.37
Q01638	Interleukin-1 receptor-like 1	Sepsis	3	0.000659	0.011051	72.64	2	0.007716	0.01542178	7.17
P35527	Keratin, type I cytoskeletal 9	Control	19	0.014256	0.049889	2.52	28	0.000423	0.00217935	3.58
P02750	Leucine-rich alpha-2-glycoprotein	Sepsis	15	0.002195	0.020414	5.42	15	3.42E-07	1.71E-05	4.87
Q8N6C8	Leukocyte immunoglobulin-like receptor subfamily A member 3	Sepsis	4	0.000824	0.011337	4.80	5	4.71E-06	0.00015679	6.72
P18428	Lipopolysaccharide-binding protein	Sepsis	11	0.000201	0.004808	16.20	11	7.63E-08	6.45E-06	7.13

P33908	Mannosyl-oligosaccharide 1,2-alpha-mannosidase IA	Sepsis	6	0.000956	0.011452	4.66	7	3.84E-05	0.00051135	3.60
P80188	Neutrophil gelatinase-associated lipocalin	Sepsis	4	0.012632	0.046079	4.05	5	0.002183	0.0071233	6.54
P55058	Phospholipid transfer protein	Sepsis	5	0.002775	0.02217	3.27	8	0.035855	0.04423555	2.21
P13796	Plastin-2	Sepsis	18	0.009294	0.043321	4.48	17	1.87E-05	0.00041418	4.62
Q9UK55	Protein Z-dependent protease inhibitor	Sepsis	4	0.007034	0.036883	2.95	9	0.000186	0.00143264	3.29
P0DJI9	Serum amyloid A-2 protein	Sepsis	2	0.00018	0.004808	75.40	2	2.62E-05	0.00046393	42.82
O00391	Sulfhydryl oxidase 1	Sepsis	7	0.001819	0.019075	2.79	12	2.20E-05	0.00043905	3.03
P24821	Tenascin	Sepsis	8	0.007839	0.039858	3.21	16	0.000105	0.00091658	4.32

Summary

Five proteins were selected as candidates for phage display Affimer production. The selection criteria was that:

- The proteins were identified with a minimum of 2 peptides per protein and a protein score of ≥ 20.
- The proteins were identified as differentially expressed based on a P value of <0.05, a q value of <0.05 cut-off and a fold change >1.5.
- The proteins had the highest median abundance in the sepsis plasma.
- A recombinant protein was commercially available for the phage display screens.

The five proteins were C-reactive protein (CRP), Protein S100 A8/A9, cathepsin B, neutrophil gelatinase associated protein (NGAL) and interleukin-1 receptor-like 1 (Table 5.8). All five proteins selected met this criterion. Two proteins selected in cohort 1 were not included in the new panel; glutathione S-transferase A5 and calcitonin. Glutathione S-transferase A5 was not identified in cohort 2 and although calcitonin was identified in cohort 2, it had the highest mean abundance in control plasma. In addition to protein S100-A9, protein S100-A8 was also selected as a protein target. The proteins associates as a heterodimer and therefore phage display production can be targeted to the heterodimer. The normalised protein abundance plots for the candidate proteins are displayed Figure 5.27. To assess whether there was a significant statistical difference between the control and sepsis plasma groups, a t-test was performed ($P < 0.05$). A significant difference between the control and sepsis plasma was observed for all proteins (Figure 5.27).As the loading of each sample was adjusted so that similar BPI values were obtained, the normalised abundance values of the five selected proteins were adjusted to account for this (Figure 5.28). Although the difference between sepsis and control groups is less, the proteins selected still have a higher abundance in the sepsis group. To assess whether there was still a significant statistical difference between the control and sepsis plasma groups, a t-test was performed ($P < 0.05$). A significant difference between the control and sepsis plasma was observed for all proteins expect protein S100-A8 (Figure 5.28).

To determine the suitability of the five candidate proteins as biomarkers for the diagnosis of sepsis, receiver operating characteristic (ROC) curves were evaluated and the area under the curve (AUC) values assessed (Figure 5.29). For CRP, an AUC value of 1.0 was achieved

indicating that the protein has high sensitivity and specificity for discriminating the two groups of patients (control vs. sepsis). AUC values for intertelukin-1 receptor-like 1, cathepsin B, protein S100 A8 and neutrophil gelatinase-associated lipocalin were 0.84, 0.85, 0.63 and 0.85, respectively. The analysis would suggest that protein S100 A8 is a poor marker to discriminate between the two groups of patients, whereas intertelukin-1 receptor-like 1, cathepsin B and neutrophil gelatinase-associated lipocalins are relative sensitive and specific to the sepsis patients. Although the ROC curve analysis would suggest that the five proteins are relatively suitable as diagnostic biomarkers for sepsis, the proteins may merely be indicative of infection and less specific to sepsis. To assess this further, it would be beneficial to increase the patient cohort size and to include additional subclasses of patients with localised infection that has not yet spread to the blood stream.

The protein abundances in this study were calculated using label-free relative quantification. This approach is common practice for the analysis of complex samples for biomarker discovery and has resulted in the identification of several biomarkers[270,271]. The method is advantageous as there is no need for sample labelling and numerous clinical samples can be included in the study design providing increased statistical power. Label-free relative protein quantification requires the use of mass spectrometers with high resolving power and high mass accuracy meaning hybrid mass spectrometers are the instruments of choice. In this study, the Thermo Scientific QExactive HF mass spectrometer was used[272]. In addition, it is important that the stability of the instrument is monitored over the course of the study as during data processing, the raw LC-MS/MS data is compared. In this study, instrument performance was periodically assessed by the analysis of a standard *E.coli* digest.

To investigate the candidate proteins involvement in sepsis, the current literature on each protein was reviewed.

CRP: CRP is a well-established non-specific marker of an acute-phase response. Inflammatory cytokines induce CRP production in the liver and its key function is thought to induce phagocytosis through interacting with Fc receptors on phagocytic cells to remove apoptotic and necrotic cells[273,274]. The levels of CRP rise significantly higher than other acute phase proteins in response to inflammation which has meant the protein has been routinely used to confirm the presence of infection. In both cohorts, this was observed as CRP expression levels had the highest fold change between sepsis and control plasma. Despite CRP lacking specificity for sepsis, it does have clinical relevance; differentiating

infectious and non-infectious causes of localised pain and the screening of neonates for bacterial infection[275,276].

Cathepsin B: Cathepsin B is a cysteine protease involved in intracellular proteolysis[277]. Multiple studies report that cathepsin B over-expression is strongly associated with numerous cancer types including oral squamous cell carcinoma and hepatocellular carcinoma. However its association with sepsis is limited. In a study on patients with severe septic shock, cathepsin B activity measured in plasma, was elevated[278]. In another study analysing muscle protein degradation, cathepsin B activity was measured in muscle extracts from patients with sepsis and was 200 % higher than in controls[279]. Studies involving rats have also showed this association of increased cathepsin B activity and sepsis[280,281]. Inhibition of cathepsin B activity has also shown to reduce CSF leukocyte count and severity of pneumococci meningitis[282] . Therefore, the potential involvement of cathepsin B in immune response may indicate its suitability as a marker of sepsis.

Protein S100 A8/A9: Protein S100 A8 and protein S100 A9 belong to the S100 protein family. Together, they form a non-covalent heterodimer. The role of protein S100 A8 and A9 in infection and immune response is well known and they are considered important to maintain homeostasis of the immune system; the proteins are expressed by neutrophils, and macrophages and monocytes in damaged tissues. Conflicting studies report S100 A8 and S100 A9 as having roles in both pro-inflammatory[283] and anti-inflammatory role in sepsis[284]. An investigation into mice models with endotoxin-induced lethal shock and sepsis demonstrated that the S100A8/A9 complex is involved in increasing the activation of phagocytes in sepsis[285]. In addition, a patent was filed in 2009 describing a method to measure the S100A8/A9 heterodimer in plasma as a prognostic marker for patients with severe sepsis [286].

Neutrophil gelatinase-associated lipocalin (NGAL): NGAL belongs to the lipocalin protein family and plays a role are preventing bacterial growth. Multiple studies report the involvement of NGAL as a marker of sepsis in plasma and urine[249,271,287,288]. In a proteomics analysis on patients with early septic shock elevated levels of NGAL were observed compared with control patients[249]. However, a reduction in NGAL in sepsis patients was linked with patient non-survival suggesting that elevated NGAL could be used as a prognostic marker. Another study on 120 paediatric patients with suspected sepsis analysed serum NGAL levels using a commercial ELISA kit[289]. They also identified NGAL to be

a useful diagnostic marker of sepsis. However in contrast to the previously discussed study, a significantly elevated level of NGAL was considered to correlate with patient death.

Interleukin-1 receptor-like 1: Interleukin-1 receptor-like 1, also known as suppression of tumourgenecity 2 (ST2) is a member of the IL-1 receptor family and is expressed on many cells of the immune system including lymphocytes, natural killer T cells and monocytes. In tissue damage, the ligand for Il1RL1, IL-33, is released from endothelial and epithelial cells and binds with IL1RL1 inducing numerous immune cell types to release pro-inflammatory and anti-inflammatory mediators. Numerous studies have identified IL1RL1 as a potential biomarker of sepsis with elevated levels measured in patient's serum[290,291]. One study reported that elevated IL1RL1 levels, measured using a commercial ELISA kit, in sepsis patients was shown to correlate with patient mortality[292].

As previously discussed, the application of a single sepsis biomarker is not suitable for disease diagnosis due to the lack of specificity of individual protein markers[258]. The five proteins identified in this study have been previously implicated in sepsis or play a role in immune response. Therefore, the combined analysis of these candidate proteins may provide an initial foundation the characterisation and detection of sepsis.

Table 5.8 | Summary of proteins selected for phage display Affimer production.

In cohort 1, protein S100-A8 and cathepsin B and in cohort 2, protein S100-A9 were identified with 2 or more peptides per protein and a protein score of >20 however they did not pass the q value cut–off score. Note: The phage display will target Protein S100 A8/A9 as a heterodimer.

			Cohort 1				Cohort 2			
Accession	**Description**	**Highest mean condition**	**Unique peptides**	**Anova (p)**	**q Value**	**Max fold change**	**Unique peptides**	**Anova (p)**	**q Value**	**Max fold change**
P02741	C-reactive protein	Sepsis	7	7.42E-08	1.24E-05	112.60	7	6.95E-09	1.39E-06	66.04
P05109	Protein S100-A8	Sepsis	4	0.016102	0.050978	2.91	4	0.02824177	0.036892	1.67
P06702	Protein S100-A9	Sepsis	4	0.008537	0.04213	4.22	2	0.822054	0.0553	1.31
P07858	Cathepsin B	Sepsis	5	0.046641	0.083577	13.829	4	0.00896988	0.017576	4.33
P80188	Neutrophil gelatinase-associated lipocalin	Sepsis	4	0.012632	0.046079	4.05	5	0.002183	0.0071233	6.54
Q01638	Interleukin-1 receptor-like 1	Sepsis	3	0.000659	0.011051	72.64	2	0.007716	0.0154217	7.17

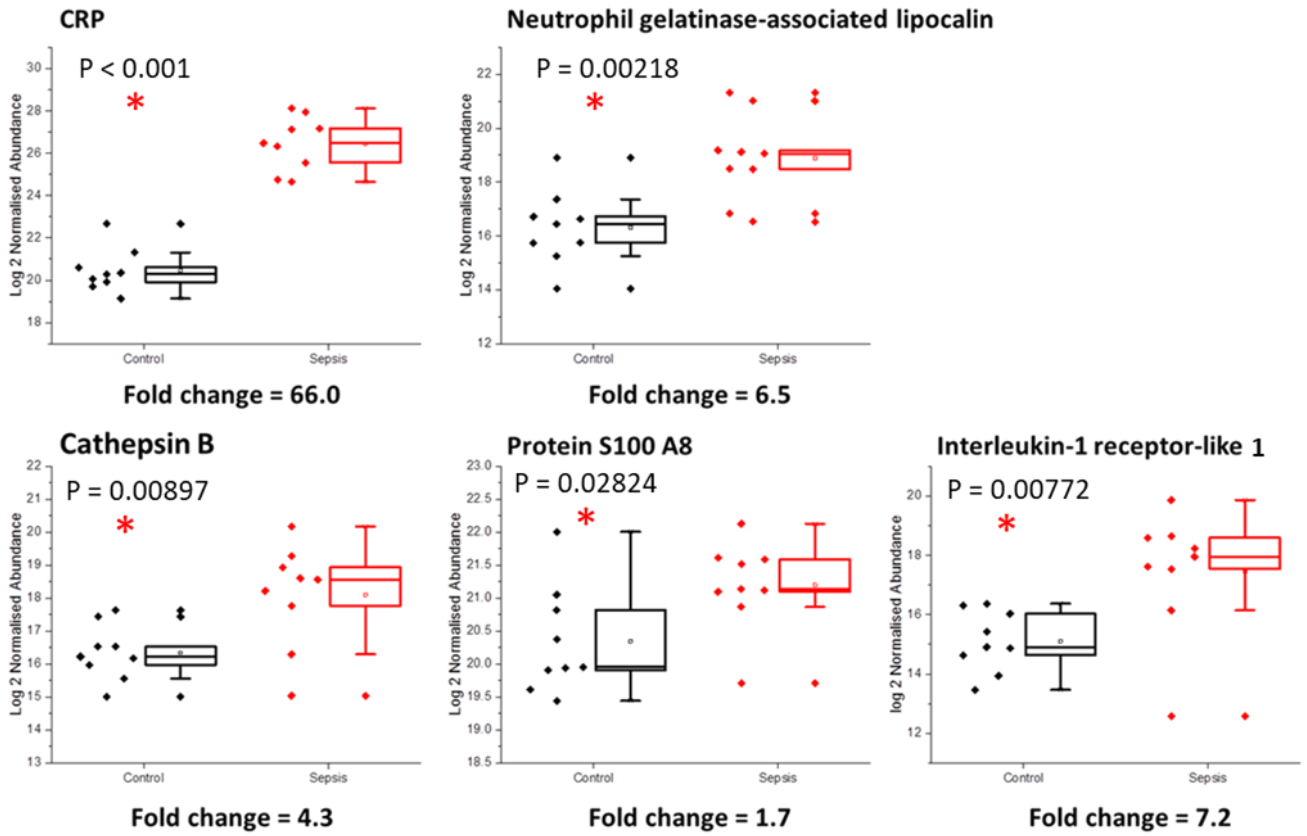


Figure 5.27 | Log 2 normalised abundance plots for the five candidate proteins for Cohort 2

CRP, protein s100-A8, cathepsin B interleukin-1 receptor-like 1 and neutrophil gelatinase-associated lipocalin abundance plots. Plots display the log 2 normalised abundances of the normalised protein abundance values. Black plots are control samples and red plots are sepsis samples. Statistical analysis (t-test) was performed to assess for differences between control and sepsis plasma ($P > 0.05$). Red star (*) denotes significance.

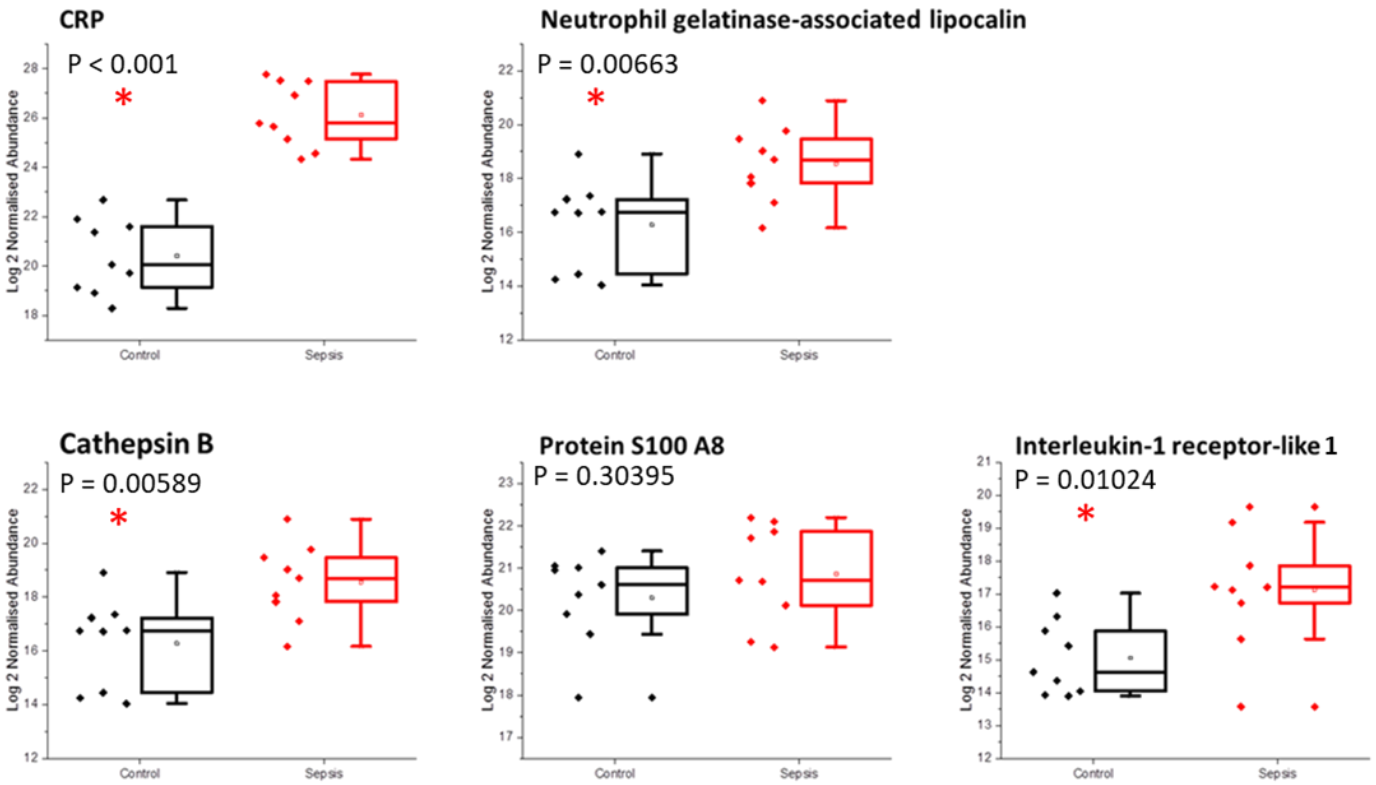


Figure 5.28 | Log 2 normalised abundance plots for the five candidate proteins for Cohort 2 adjusted for sample loading.

CRP, protein S100-A8, cathepsin B, interleukin-1 receptor-like 1 and neutrophil gelatinase-associated lipocalin abundance plots. Plots display the log 2 normalised abundances of the normalised protein abundance values adjusted for sample loading. Black plots are control samples and red plots are sepsis samples. Statistical analysis (t-test) was performed to assess for differences between control and sepsis plasma ($P > 0.05$). Red star (*) denotes significance.

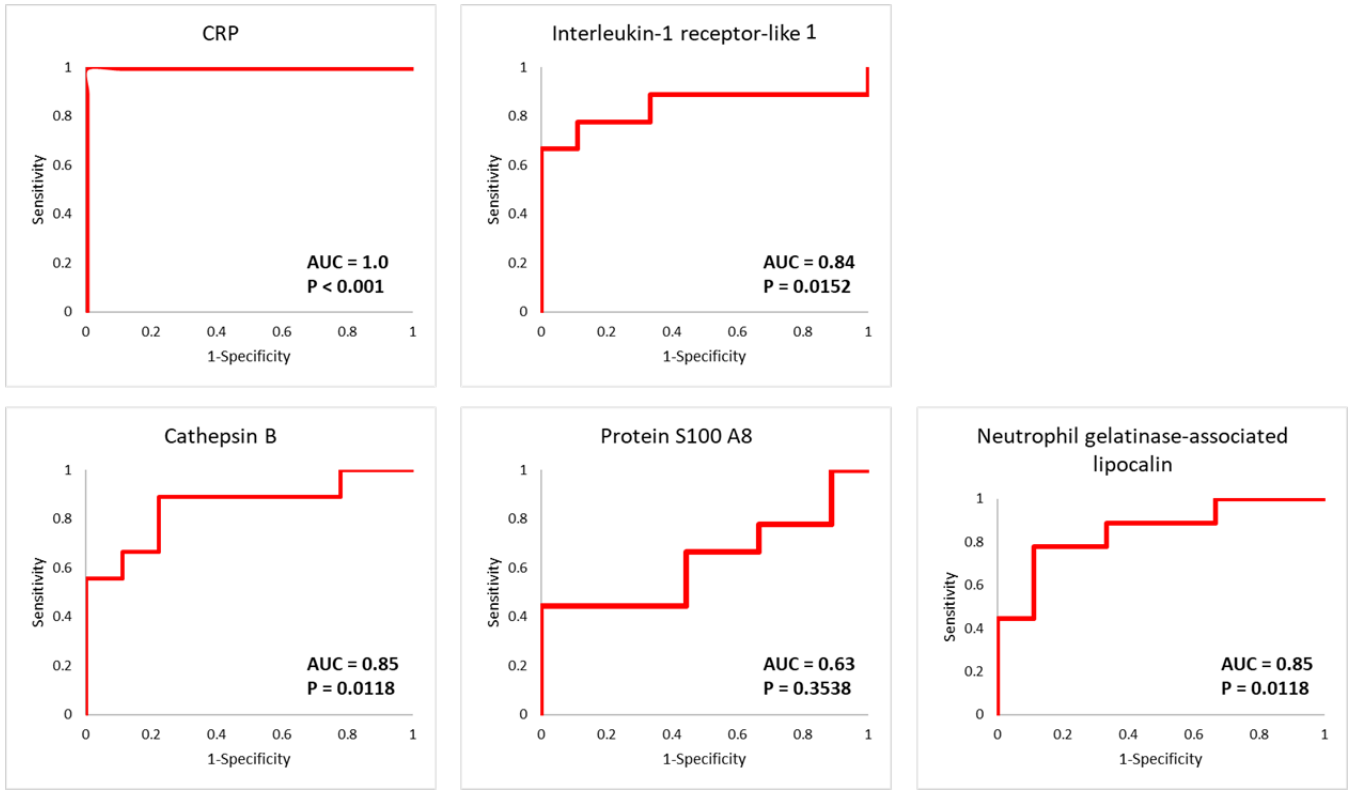


Figure 5.29 | ROC curves for five candidate proteins for Cohort 2.

ROC curve analysis for the normalised protein abundance values adjusted for sample loading of CRP, S100-A8, cathepsin B, interleukin-1 receptor-like 1 and neutrophil gelatinase-associated lipocalins measured in Cohort 2 to assess suitability of the proteins as biomarkers for sepsis diagnosis. The AUC and significance p values for each protein are displayed.

5.3.5 Characterisation of Phage Display Adhirons

After selection of the panel of target proteins implicated in sepsis, Adhiron reagents were generated. The Adhirons were produced by Thomas Taylor and Christian Tiede in the BioScreening Technology Group at the University of Leeds following the standard protocol[111]. Recombinant protein for the selected proteins was purchased for the phage display screenings. Initially, five Adhiron binders for each protein were selected at random based on validation ELISA data to confirm target binding. The selected Adhiron binders were sub-cloned into plasmids for the addition of a cysteine residue required for Adhiron immobilisation. However, a portion of these Adhirons did not express and therefore a total of 21 Adhirons were received (Table 5.9).

The Adhiron sequences were aligned using CLC Sequence Viewer and a phylogenetic tree created to compare divergence in sequences (Figure 5.30). The variable loop regions of the Adhiron protein scaffold are evident, located at residue positions 39 to 48 and 72 to 82 for loop 1 and loop 2, respectively. A total of 20 unique Adhiron sequences were generated; NGAL 10 and NGAL 78 had identical sequences. In total, 11 Adhirons did not contain peptide insertions in the loop 2 region of the scaffold. As with the pepsinogen Adhirons, there is a greater distribution of the hydrophobic amino acid residues in the loop regions. Phylogenetic analysis interestingly revealed that the Affimers were sub-divided based on the target protein, suggesting the Affimers produced for a single target share similar properties.

To confirm the protein product of Adhiron expression and purification, ESI-MS analysis was performed of the intact Adhiron proteins as detailed in Chapter 2.21. As the Adhirons were purified into a buffer that did not contain reducing agent, the Adhrions were first incubated with DTT to obtain the monomer form of the protein. Prior to ESI-MS analysis, an additional clean-up step was carried out using a C4 trap to remove the glycerol and salts from the Adhirons samples. For each Adhiron, the predominant species within each mass spectrum corresponded to the theoretical mass of the monomer form of the protein minus the initiating methionine (Figure 5.31 – 5.35, Table 5.10). Due to the N-terminal sequence, cleavage of the initiating methionine is expected [293].

The results confirm the correct protein product had expressed for all Adhirons and had been purified without contaminants. To confirm that the Adhirons bind their intended target protein, future work should ensure that the Adhrions are validated in an affinity purification format. The SulfoLink® affinity purification protocol described in Chapter 2.15

would be suitable validate target binding using recombinant proteins. Following this, the next step would be to confirm that the Adhirons bind endogenous target protein in plasma.

Table 5.9 | Phage Display Adhirons.

Concentration of the 21 Adhirons received from the BioScreening Technology Group at the University of Leeds.

Target	Binder	Concentration (mg/mL)
Cathepsin B	C20	1.3
	C2	1.6
	C3	0.9
	C45	1.0
Neutrophil gelatinase-associated lipocalin	NGAL 10	1.6
	NGAL 39	1.1
	NGAL 42	0.8
	NGAL 78	1.6
Protein S100 A8/A9	S7	0.9
	S20	1.1
	S28	1.2
Interleukin-1 receptor-like 1	ST 15	0.5
	ST 3	0.7
	ST 45	0.3
	ST 55	0.5
	ST 59	1.2
CRP	CRP 16	1.3
	CRP 21	1.3
	CRP 22	1.4
	CRP 30	1.3
	CRP 3	1.2

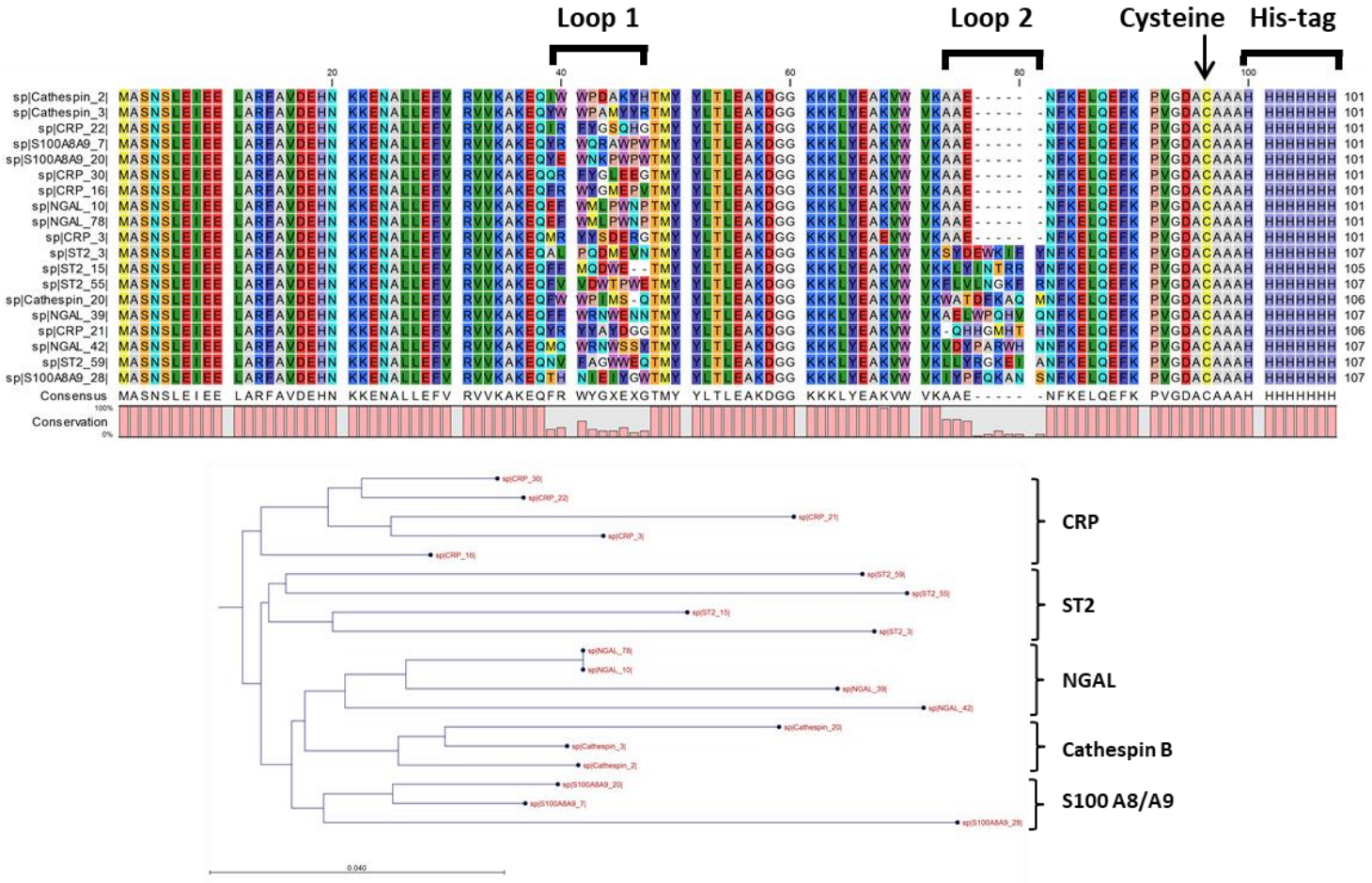


Figure 5.30 | Amino acid multiple sequence alignment and phylogenetic tree of Phage Display Adhirons.

Protein sequences were aligned and visualised using CLC Sequence Viewer. Residues were coloured using default Rasmol colours. Eleven Adhirons do not contain peptide insertions in loop 2 of the scaffold. Phylogenetic tree indicates Adhirons that target the same protein are similar in amino acid sequences.

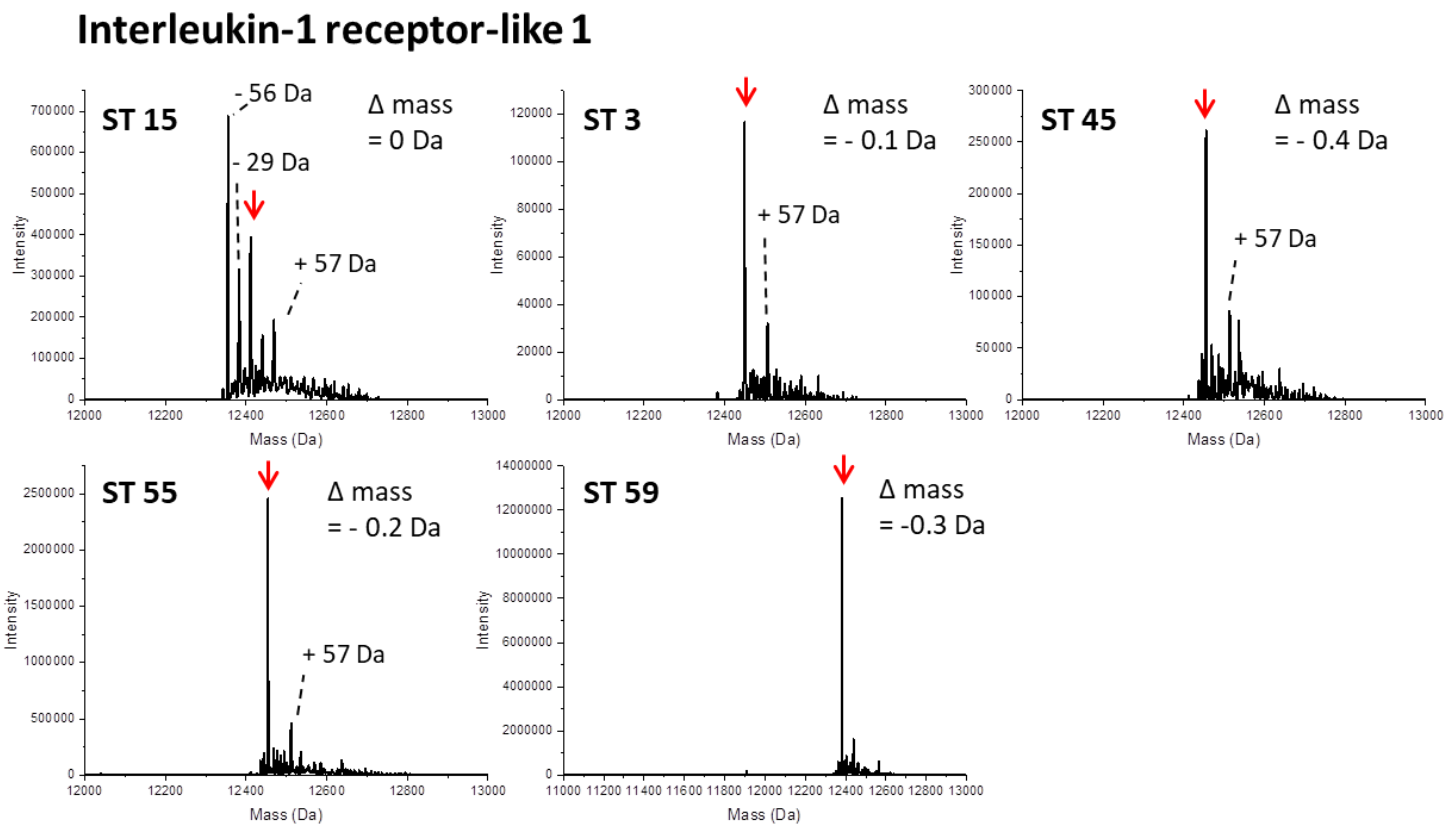


Figure 5.31 | ESI–MS analysis of Phage Display Adhirons targeting Interleukin-1 receptor-like 1.

Adhirons were prepared using a C4 trap to remove glycerol and salts. Samples were incubated with DTT and diluted to 1 pmol/µL in 3% acetonitrile, 0.1% formic acid. 2 µL of each sample was loaded onto a C4 desalting trapping column. Samples were analysed by ESI-MS on the Waters G2 mass spectrometer. Multiply charged protein envelope was deconvoluted using Waters MAXENT 1 algorithm to determine average mass of intact proteins. Observed masses of Adhirons correspond to their theoretical masses minus the loss of the initiating methionine (red arrow).

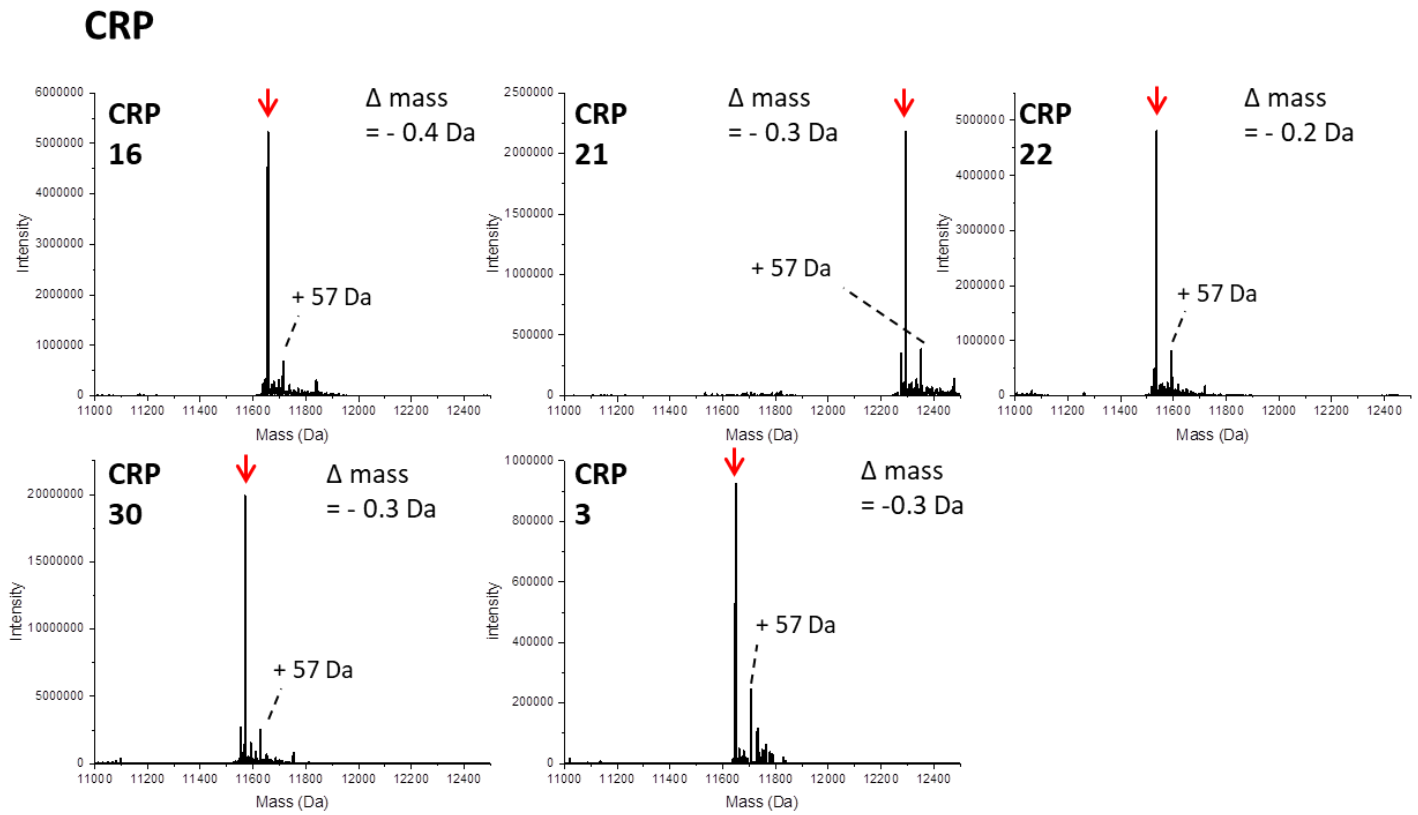


Figure 5.32 | ESI–MS analysis of Phage Display Adhirons targeting CRP.

Adhirons were prepared using a C4 trap to remove glycerol and salts. Samples were incubated with DTT and diluted to 1 pmol/µL in 3% acetonitrile, 0.1% formic acid. 2 µL of each sample was loaded onto a C4 desalting trapping column. Samples were analysed by ESI-MS on the Waters G2 mass spectrometer. Multiply charged protein envelope was deconvoluted using Waters MAXENT 1 algorithm to determine average mass of intact proteins. Observed masses of Adhirons correspond to their theoretical masses minus the loss of the initiating methionine (red arrow).

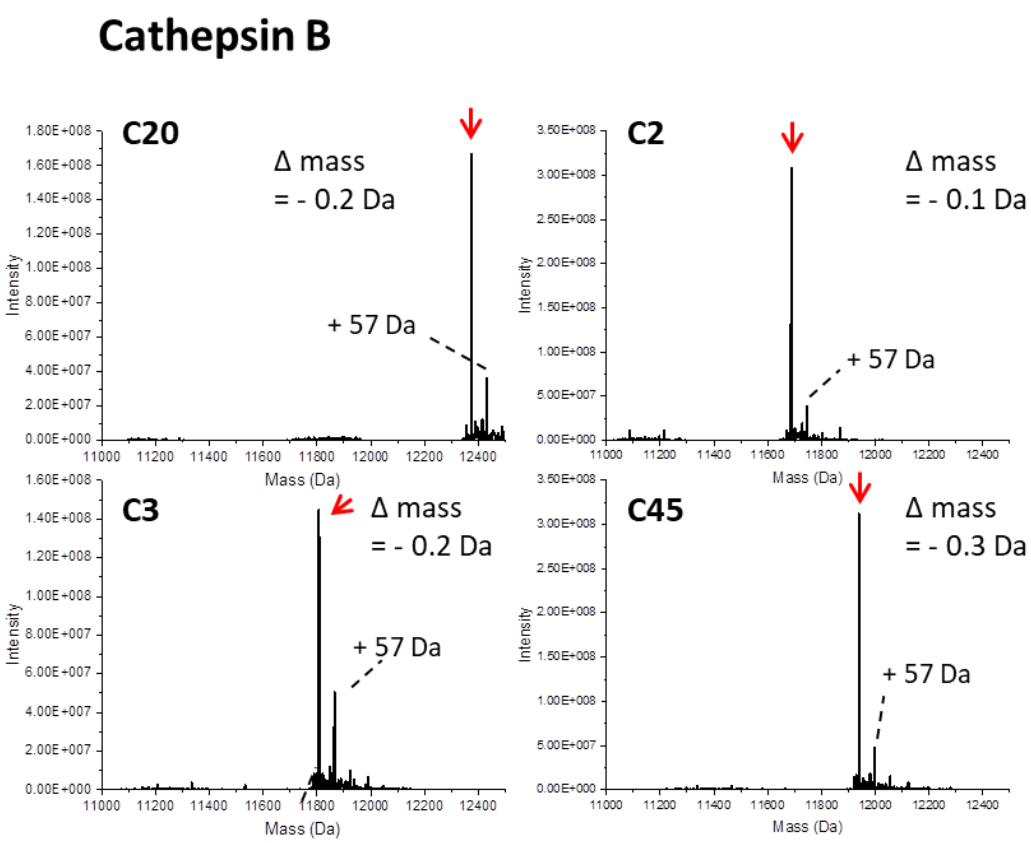


Figure 5.33 | ESI–MS analysis of Phage Display Adhirons targeting Cathepsin B.

Adhirons were prepared using a C4 trap to remove glycerol and salts. Samples were incubated with DTT and diluted to 1 pmol/µL in 3% acetonitrile, 0.1% formic acid. 2 µL of each sample was loaded onto a C4 desalting trapping column. Samples were analysed by ESI-MS on the Waters G2 mass spectrometer. Multiply charged protein envelope was deconvoluted using Waters MAXENT 1 algorithm to determine average mass of intact proteins. Observed masses of Adhirons correspond to their theoretical masses minus the loss of the initiating methionine (red arrow).

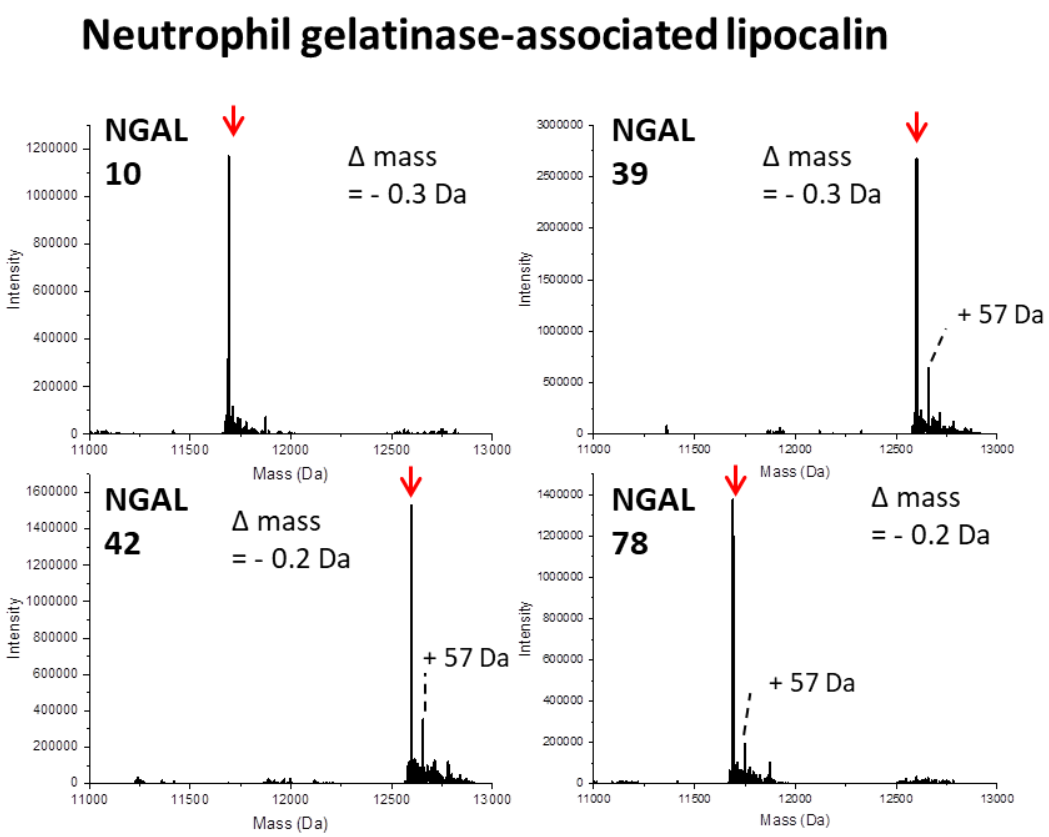


Figure 5.34 | ESI–MS analysis of Phage Display Adhirons targeting neutrophil gelatinase-associated lipocalin.

Adhirons were prepared using a C4 trap to remove glycerol and salts. Samples were incubated with DTT and diluted to 1 pmol/µL in 3% acetonitrile, 0.1% formic acid. 2 µL of each sample was loaded onto a C4 desalting trapping column. Samples were analysed by ESI-MS on the Waters G2 mass spectrometer. Multiply charged protein envelope was deconvoluted using Waters MAXENT 1 algorithm to determine average mass of intact proteins. Observed masses of Adhirons correspond to their theoretical masses minus the loss of the initiating methionine (red arrow).

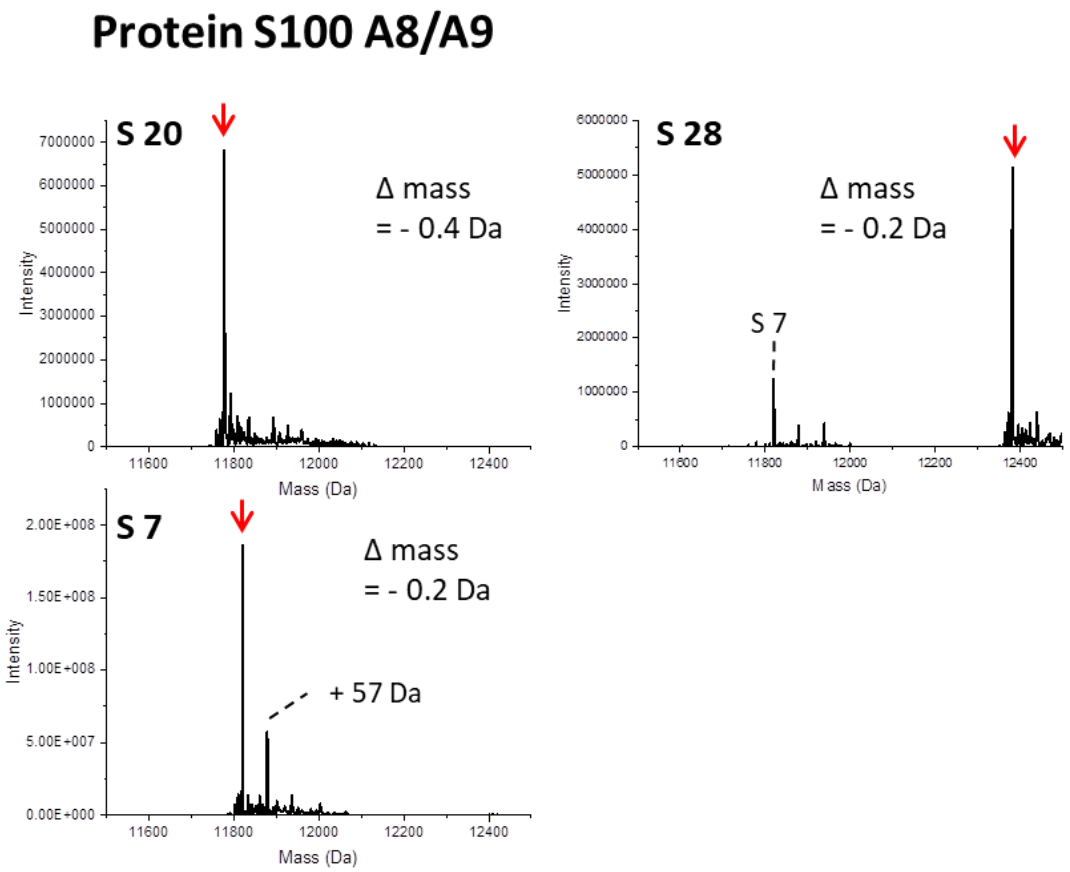


Figure 5.35 | ESI–MS analysis of Phage Display Adhirons targeting protein S100 A8/A9.

Adhirons were prepared using a C4 trap to remove glycerol and salts. Samples were incubated with DTT and diluted to 1 pmol/µL in 3% acetonitrile, 0.1% formic acid. 2 µL of each sample was loaded onto a C4 desalting trapping column. Samples were analysed by ESI-MS on the Waters G2 mass spectrometer. Multiply charged protein envelope was deconvoluted using Waters MAXENT 1 algorithm to determine average mass of intact proteins. Observed masses of Adhirons correspond to their theoretical masses minus the loss of the initiating methionine (red arrow).

Table 5.10 | Comparison between theoretical and observed intact mass analysis of Phage Display Adhrions.

Theoretical mass (Da) calculated minus the initiating methionine.

Adhiron		Observed Mass (Da)	Theoretical Mass (Da)	Difference (Da)
Cathespin B	**C20**	12373.9	12374.1	-0.2
	C2	11687.1	11687.2	-0.1
	C3	11807.2	11807.4	-0.2
	C45	11939.4	11939.7	-0.3
Neutrophil gelatinase-associated lipocalin	**NGAL 10**	11691.0	11691.3	-0.3
	NGAL 39	12601.9	12602.2	-0.3
	NGAL 42	12597.0	12597.2	-0.2
	NGAL 78	11691.1	11691.3	-.02
Protein S100 A8/A9	**S7**	11820.2	11820.4	-0.2
	S20	11776.9	11777.3	-0.4
	S28	12381.8	12382.0	-0.2
Interleukin-1 receptor-like 1	**ST 15**	12411.1	12411.1	0
	ST 3	12449.0	12449.1	-0.1
	ST 45	12454.9	12455.3	-0.4
	ST 55	12454.0	12454.2	-0.2
	ST 59	12380.8	12381.1	-0.3
CRP	**CRP 16**	11655.9	11656.3	-0.4
	CRP 21	12293.5	12293.8	-0.3
	CRP 22	11535.8	11536.0	-0.2
	CRP 30	11569.8	11570.1	-0.3
	CRP 3	11648.8	11649.1	-0.3

5.4. Conclusions

The primary aim of this chapter was to identify a panel of proteins implicated in sepsis in order to guide phage display production of Affimers. Comparative proteomic analysis of two patient cohorts using LC-MS/MS identified differentially expressed proteins and a panel of five candidate protein markers of sepsis were selected. The proteins selected met strict data filters to obtain a high level of confidence in candidate selection. In addition, candidate proteins displayed the same direction of change in expression in both cohorts and recombinant forms of the proteins were available. This was an essential criteria of protein selection due to the phage display process requiring recombinant protein. Overall, proteins identified as up-regulated in sepsis plasma were implicated in inflammation and characteristic of an acute immune response. The proteins identified in this analysis were in agreement with previously published data.

As discussed, a major challenge of plasma proteomics is the large dynamic range and in particular, the highly abundant protein, albumin. Two key methods to reduce sample complexity included sample fractionation and depletion strategies. Routinely, complex protein samples are digested into peptides and subjected to pre-fractionation prior to LC-MS/MS analysis (2D LC-MS/MS) to add an additional dimension to peptide separation[78]. However, this technique was not employed in this study due to the increase in the number of samples requiring analysis and additional data analysis steps needed. In this analysis, an antibody-based depletion strategy was employed successfully to remove abundant plasma proteins. Proteins that would typically go un-detected were identified in this study demonstrating the necessity for depletion strategies prior to MS analysis to improve proteome coverage.

The mass spectrometer used in this study for the discovery proteomics was a hybrid instrument with a quadrupole-Orbitrap mass analyser and is the ideal instrument for the analysis of complex proteomes due to the high scanning speed, mass accuracy and resolution[272]. The fragmentation spectra produced is of high quality and enables the correct assignment of product ions for accurate peptide sequencing. However, despite the application of protein depletion methodologies and high quality instrumentation, mass spectrometers still display a bias towards the most abundant peptide ions within a sample due to the selection of the top *n* ions selected for fragmentation. Therefore, low abundance proteins can still remain undetected. The use of strategies such as Progenesis QI that

applies protein quantification prior to identification increases low abundance protein identification across all samples.

Although the overall aim of the chapter was met, there were limitations in the study. Two samples from cohort 2 were excluded from protein quantification as they did not meet the alignment and normalisation cut-off in Progenesis QI. Low numbers of protein identifications due to insufficient depletion of highly abundant proteins was suggested as a reason for this. Variability of the depletion spin columns to remove highly abundant proteins was not assessed in this analysis. Ideally, variability in the depletion method should be analysed. If variability in protein depletion is high, a large negative effect on protein quantification and identification of differentially expressed proteins could be observed. In addition, increasing the number of samples analysed in this study would generate a more robust dataset. Furthermore, patient samples were not analysed in triplicate. To gain greater confidence in protein identifications and abundance values it would desirable to analyse each patient sample in triplicate, although it should be noted that the data has undergone very stringent filtering to generate reliable protein abundance values. Obviously, factors such as cost, through-put and availability of patient samples limit repeated sample analysis.

The key outcome of this proteomic study of sepsis plasma is the production of Affimers to the five proteins identified as having increased expressed in sepsis patients. Intact mass analysis of the Affimers suggested expression and purification of the correct protein product with the addition of a cysteine residue necessary for immobilisation. Future work should include validation of the Adhirons to confirm they bind both the recombinant and endogenous form of the protein targets they were generated against with high affinity and specificity.

The work in this chapter forms part of a collaborative project to use Affimer reagents in a multiplexed ELISA format. Following extensive optimisation and development of the multiplexed ELISA, the ultimate goal is to develop a miniaturised POCT device for sepsis diagnosis and detection. The POCT device will allow rapid detection of sepsis which will lead to earlier therapeutic intervention improving patient prognosis. While extensive validation is necessary to confirm the candidate proteins are suitable for use as a diagnostic panel for sepsis, their differential expression in this pilot study demonstrate they are of interest in sepsis.

Chapter 6: Darcin as an Antibody Alternative Protein Scaffold

6.1 Introduction

Lipocalins are small, secreted proteins that encompass a large family of proteins across many different species[173]. Sequence homology between the different lipocalins is extremely low however certain members of the protein family, known as prototypic lipocalins, have high homology in their secondary structure[172]. A primary property of proteins in this family is molecular recognition including ligand binding, complex formation and receptor binding[173]. Furthermore, the proteins have a compact, rigid structure attributed to the β-barrel core of the proteins[173]. These features make lipocalins ideal proteins for engineering as protein scaffolds. Their development as such has been described in numerous studies and has resulted in the commercialisation of lipocalin based alternative affinity reagents[107,171,172]. Despite many lipocalin proteins undergoing development as protein scaffolds, the protein group known as major urinary proteins (MUPs) are yet to be studied for development as potential protein scaffolds. MUPs are produced in the urine of rodents typically mice and rats and play a role in chemical communication[294]. As the MUPs are deposited in the urine and left for communication with other rodents, it is vital that the MUPs are highly stable. A particular MUP, darcin, has demonstrated enhanced stability compared to the other MUPs (Figure 6.1). A study examining the effects of high urea concentration (8M) on darcin has shown that the protein undergoes little denaturation, demonstrating the robust nature of darcin[197]. Additionally, darcin displays abnormally high mobility on SDS-PAGE [295], a characteristic that can be explained by its compact structure preserved by disulfide bonds enabling the protein to pass through the gel more freely. This feature of high stability and its robust nature suggests that darcin is suitable for development as a protein scaffold.

The work in Chapter 4 demonstrated that the pepsinogen Adhiron was able to tolerate lysine to arginine mutations to produce a non-digestible scaffold, when Lys-C was the protease. Therefore, it was proposed whether darcin could tolerate similar mutations without compromising the stability of the protein. A non-digestible scaffold would be advantageous for affinity purification MS applications. Furthermore, it was also investigated whether the disulphide bond was necessary to maintain the rigid structure of darcin.

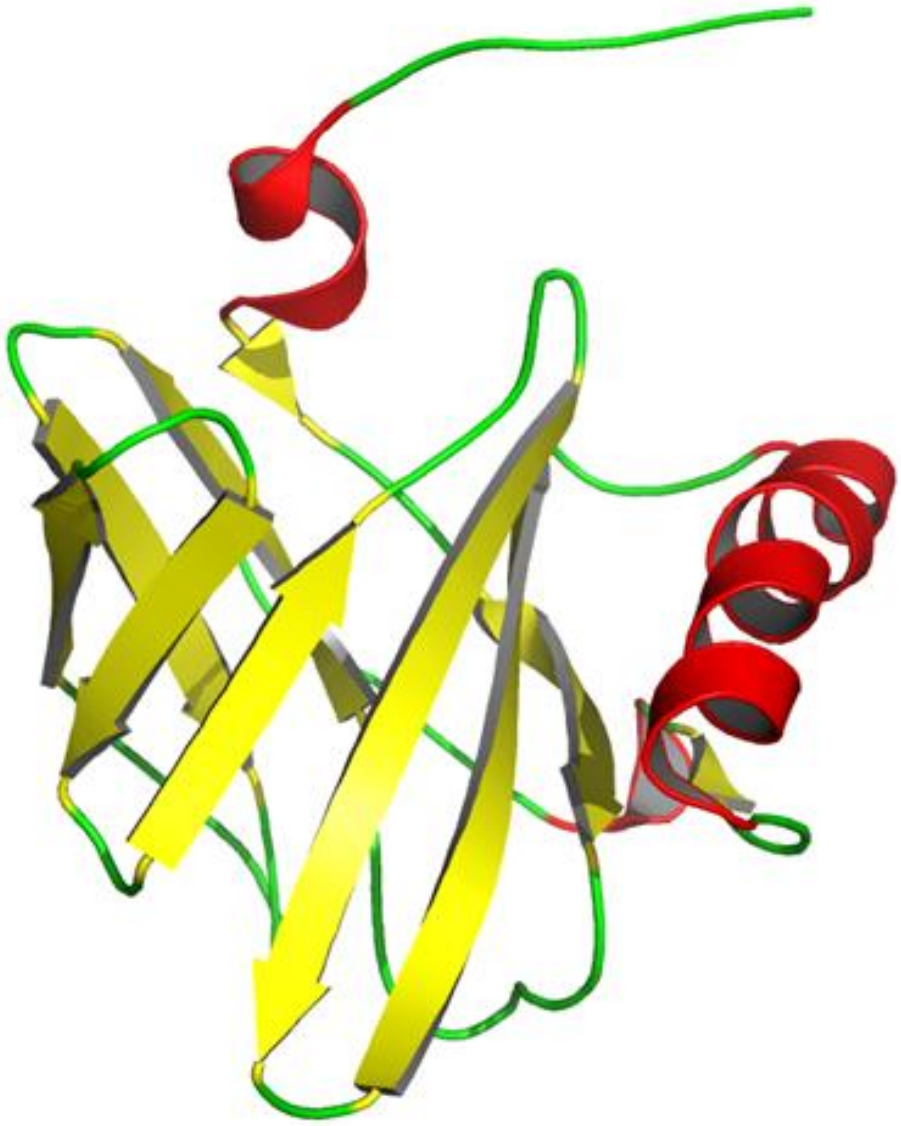


Figure 6.1 | 3D tertiary structure of darcin.

Ribbon diagram of the 3D tertiary structure of darcin, a mouse MUP. β-sheets highlighted in yellow form the central β-barrel and highlighted in red are the α-helices. Structure visualised using PyMol[191].

6.2 Aims and Objectives

The work in this chapter will focus on the preliminary development of darcin, as an alternative affinity reagent protein scaffold. The key objectives were to:

- Express and purify recombinant darcin.
- Express and purify a mutant darcin lacking a disulfide bond.
- Analyse the role of the disulfide bond in protein stability.
- Compare non-specific binding of proteins to the mutant darcin lacking a disulfide bond and darcin.
- Design darcin variants with other features required for protein scaffolds.

6.3 Results and Discussion

6.3.1 Design of darcin lacking a disulfide bond

While all MUPs have a single disulfide bond which may explain their high stability[296], it may be a detrimental feature when designing a protein scaffold. Previous work has shown that a recombinant form of darcin expresses at high levels and forms with a disulfide bond producing a fully folded protein[197,297]. However, modifying the protein to make it more suitable to function as a protein scaffold, for example with the addition of variable loop regions, may affect protein expression and folding. Therefore to ensure production of a reliable, fully folded, consistent protein scaffold it may be advantageous that the scaffold lacks the disulfide bond. In addition, engineered removal of the disulfide bond may increase protein flexibility, a potentially useful feature when developing a protein scaffold as accessibility to potential variable interaction sites may increase. However, it is necessary to evaluate whether darcin lacking a disulfide bond results in the loss of high stability.

To assess the role of the disulfide bond in protein stability a darcin mutant, termed darcin [$C_{78}S$, $C_{171}S$] v 1.0, was engineered so that the two cysteine residues at positions 64 and 171 were replaced for serine residues, preventing disulfide bond formation (Figure 6.2). The amino acid serine was chosen as the preferred residue substitution for cysteine as it has the most similar structure and properties.

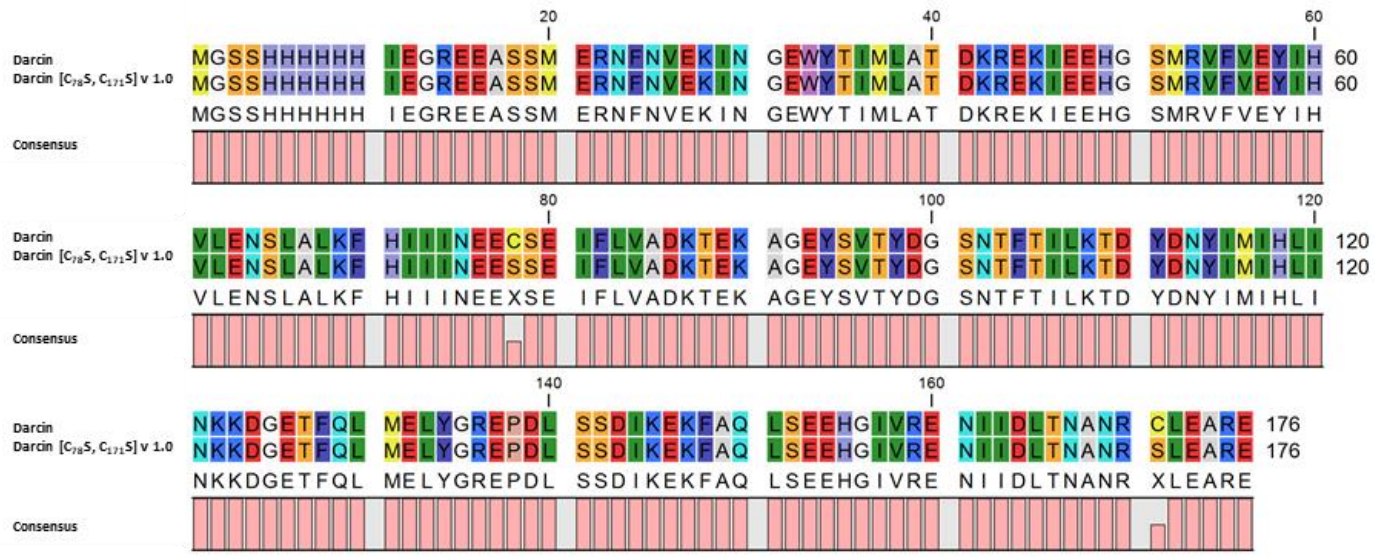


Figure 6.2 | Amino acid sequence alignment of darcin and darcin [$C_{78}S$, $C_{171}S$] v 1.0.

Protein sequences were aligned and visualised using CLC Viewer. Amino acid residues were coloured using the default Rasmol colours. Difference in two amino acid residues; cysteine residues at positions 78 and 171 have substitutions for serine residues. Theoretical molecular weight of darcin and darcin [$C_{78}S$, $C_{171}S$] v 1.0 20403 Da (accounting for -2 Da due to disulfide bond formation) and20373 Da respectively.

6.3.2. Expression of darcin and darcin [$C_{78}S$, $C_{171}S$] v 1.0

Darcin and darcin [$C_{78}S$, $C_{171}S$] v 1.0 were expressed following standard protocol in Chapter 2.1. Both genes were codon optimised for enhanced expression in *E.coli* and cloned into a pET28b + plasmid. A glycerol stock of BL21 (λDE3) *E.coli* cells containing a pET28b plasmid for darcin had been previously generated and was used to streak a kanamycin containing agar plate and grown overnight. The darcin [$C_{78}S$, $C_{171}S$] v 1.0 plasmid was transformed into BL21 *E.coli* cells following the transformation protocol. Due to the antibiotic resistance gene to kanamycin within the plasmid, only transformed cells containing the plasmid were selectively grown. A single colony for both darcin and darcin [$C_{78}S$, $C_{171}S$] v 1.0 were selected from the overnight agar plate cultures and used to inoculate a small overnight LB culture. The overnight broth was used to make glycerol stocks of darcin [$C_{78}S$, $C_{171}S$] v 1.0. For larger scale protein production, the overnight cultures were used to inoculate 200 mL of LB broth containing kanamycin following the protein expression protocol. The culture growth rate and protein expression post IPTG induction was monitored by removal of culture solution at time points during the trial induction and analysed by SDS-PAGE. Pre- and post – induction protein expression gels for darcin and darcin [$C_{78}S$, $C_{171}S$] v 1.0 and growth curves were analysed to assess for protein expression (Figure 6.3).

For the darcin construct, after IPTG induction, a protein band was visible on SDS-PAGE at approximately 20 kDa, which was not present before induction (Figure 6.3). This is the approximate expected mobility for darcin. For darcin [$C_{78}S$, $C_{171}S$] v 1.0, after IPTG induction, doublet protein bands were visible on SDS-PAGE at approximately 20 kDa and 16 kDa, both of which were not present prior to induction (Figure 6.3). Again, 20 kDa is the approximate molecular weight for darcin [$C_{78}S$, $C_{171}S$] v 1.0, however the lower molecular weight band suggests that either a truncated form of the protein was expressed or that the band is a degradation product of the full length protein. To determine the products of protein expression, an in-gel digestion was carried out on the three protein bands discussed above as described in Chapter 2.16 (Figure 6.4) and the peptides analysed by MALDI MS. Peptides corresponding to darcin were identified for all three protein bands suggesting protein expression. However, due to low sequence coverage it was not possible to confirm expression of the full length proteins. Although the results suggest that darcin and darcin [$C_{78}S$, $C_{171}S$] v 1.0 have expressed, it was still unclear why darcin [$C_{78}S$, $C_{171}S$] v 1.0 expressed as a doublet band. Therefore, to further evaluate the products of protein expression, the proteins were purified and analysed by intact mass spectrometry.

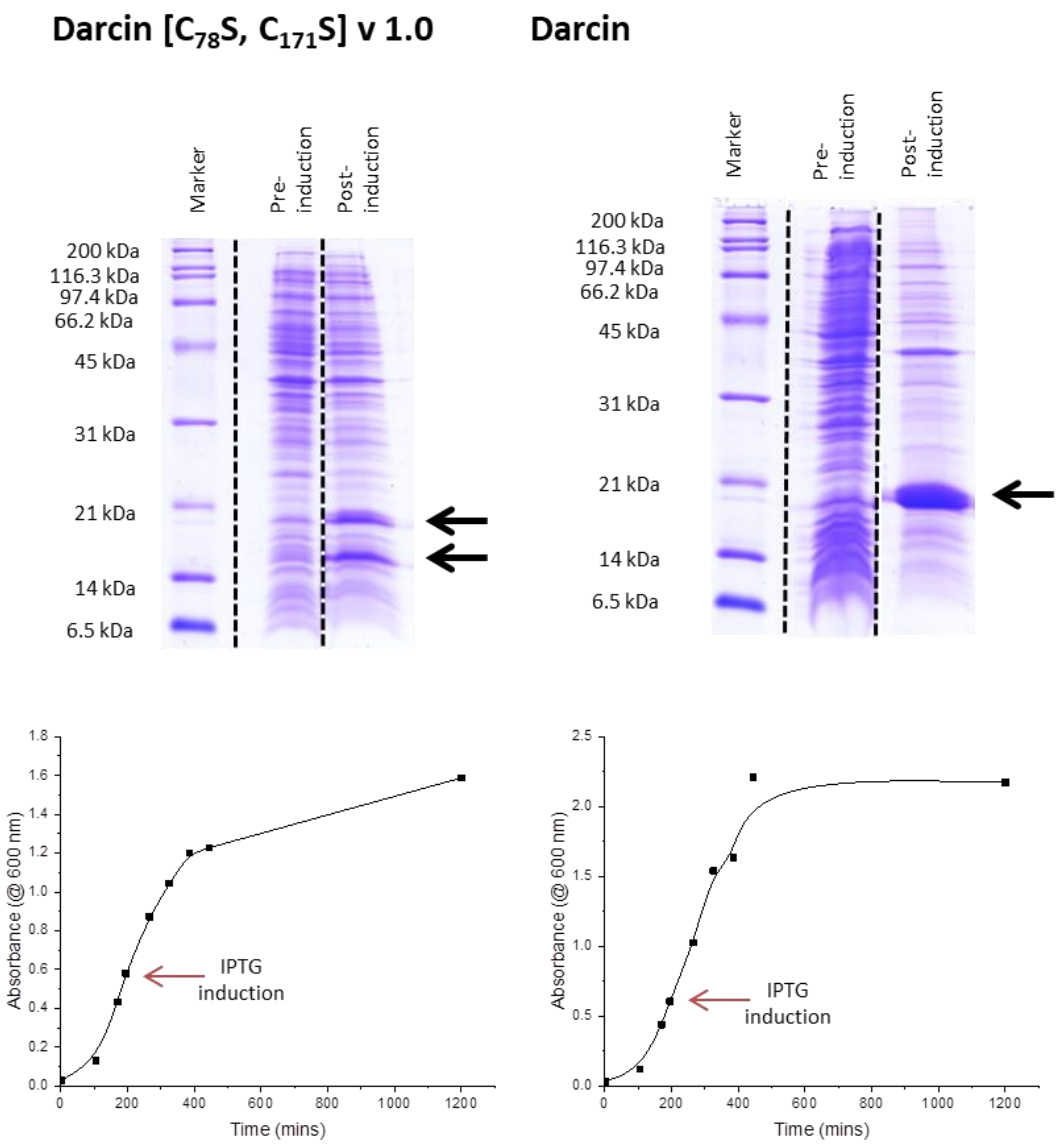


Figure 6.3 I *E.coli* culture and expression of darcin [$C_{78}S$, $C_{171}S$] v 1.0 and darcin.

SDS-PAGE analysis of pre-induction and post-induction timepoints and growth curves for darcin and Darcin [$C_{78}S$, $C_{171}S$] v 1.0 expression. Protein expression was induced with IPTG at an OD = 600 nm and protein left to accumulate for 16 hours. Arrows represent protein bands appearing after IPTG induction. A doublet band is visible at approximately 20 kDa and 16 kDa for darcin [$C_{78}S$, $C_{171}S$] v 1.0 and a single band at approximately 20 kDa for darcin. Samples were ran on a 15 % gel and visualised with Coomassie plus stain.

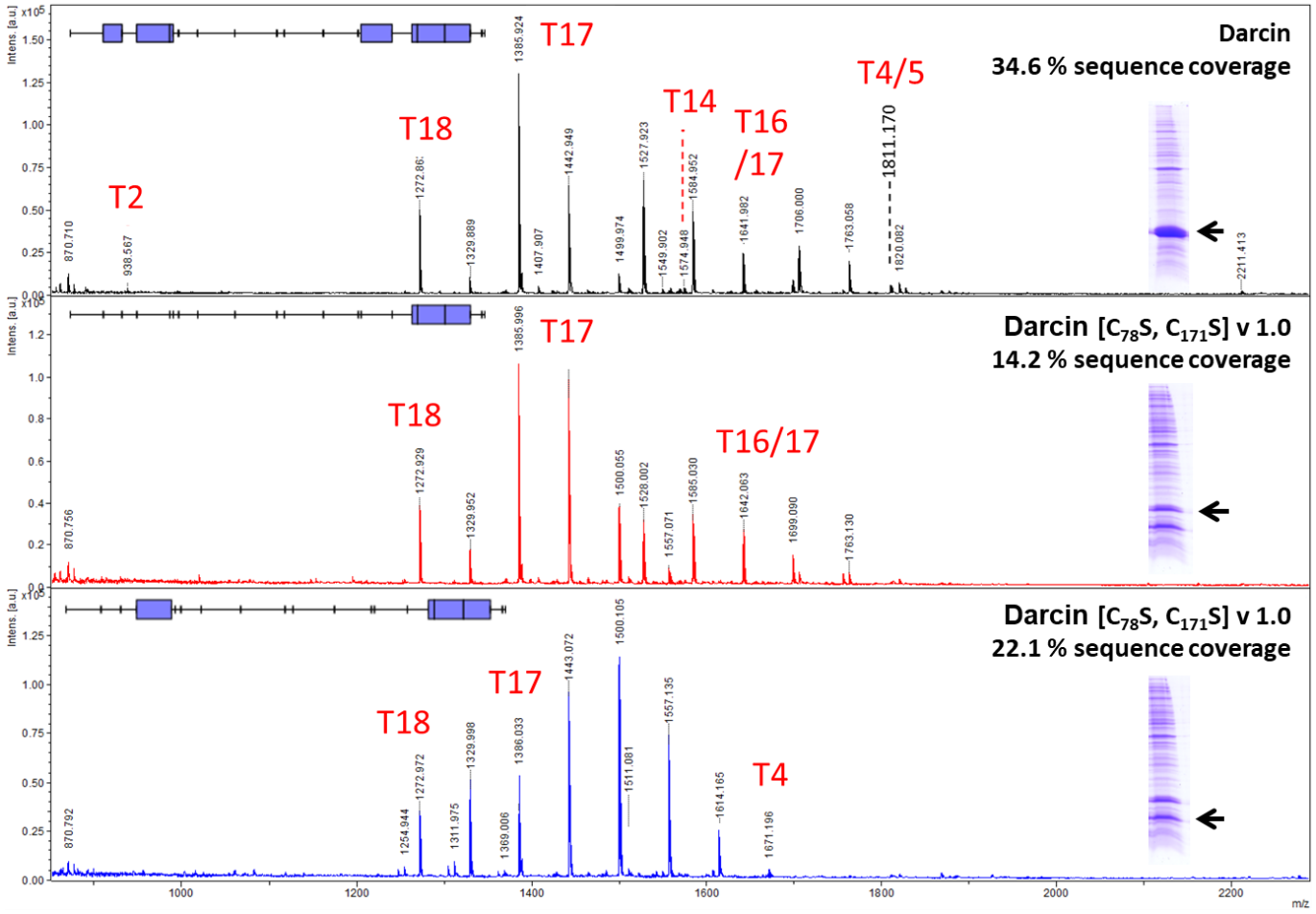


Figure 6.4 | MALDI spectra of in-gel digest of the products of darcin and darcin [$C_{78}S$, $C_{171}S$] v 1.0 protein expression.

Significant peaks labelled with peptide number. Peptide maps demonstrate low sequence coverage. Peptide maps generated using peptide mapper[234].

Prior to purification, the cells were lysed by sonication. Following centrifugation, the soluble fraction and insoluble inclusion body were analysed by SDS-PAGE to determine the location of the protein; both darcin and darcin [$C_{78}S$, $C_{171}S$] were expressed in the soluble fraction of the lysate (Figure 6.5). The proteins were purified using the immobilised metal affinity chromatography protocol using the AKTA start and fractions analysed by SDS-PAGE (Figure 6.5). Darcin eluted in fractions 4 to 7 and was present at high levels, whereas darcin [$C_{78}S$, $C_{171}S$] v 1.0 eluted predominantly in fractions 5 to 7. Although both doublet bands of darcin [$C_{78}S$, $C_{171}S$] v 1.0 contained peptides for this protein by in-gel digestion and MALDI analysis, the lower molecular weight band did not bind to the purification column and was present in the flow through fraction. This further supports the idea that degradation of darcin [$C_{78}S$, $C_{171}S$] v 1.0 has occurred, with this portion of the protein lacking the His-tag. Elution fraction 6 for darcin [$C_{78}S$, $C_{171}S$] v 1.0 and elution fraction 5 and 6 for darcin were dialysed into 25 mM ammonium bicarbonate and the protein concentration determined by Coomassie plus protein assay.

To confirm full length protein expression and the absence of contaminant proteins following purification, darcin and darcin [$C_{78}S$, $C_{171}S$] v 1.0 were analysed by intact mass spectrometry following the method in Chapter 2.21 (Figure 6. 6). The predominant species within the mass spectrum for darcin has a mass of 20403.29 Da which corresponded to the theoretical mass of darcin with the loss of the N-terminal initiating methionine and the presence of a disulfide bond (- 2 Da). An additional species, with a mass of 20581 Da was also observed which corresponds to a modification of + 178 Da. Work by Geoghegan *et al* first reported the observation of a + 178 Da and a + 258 Da mass adduct to recombinant proteins expressed in *E.coli* containing N-terminal His-tags[298]. Thus, the modification of + 178 Da was suspected as gluconoylation. Although the modification results in a heterogeneous population of the protein, it is thought that it does not affect protein folding due to its location on the His-tag portion of the protein. The predominant species within the mass spectrum for darcin [$C_{78}S$, $C_{171}S$] v 1.0 has a mass of 20373.30 Da which corresponds to the theoretical mass of the protein with the loss of the initiating N-terminal methionine.

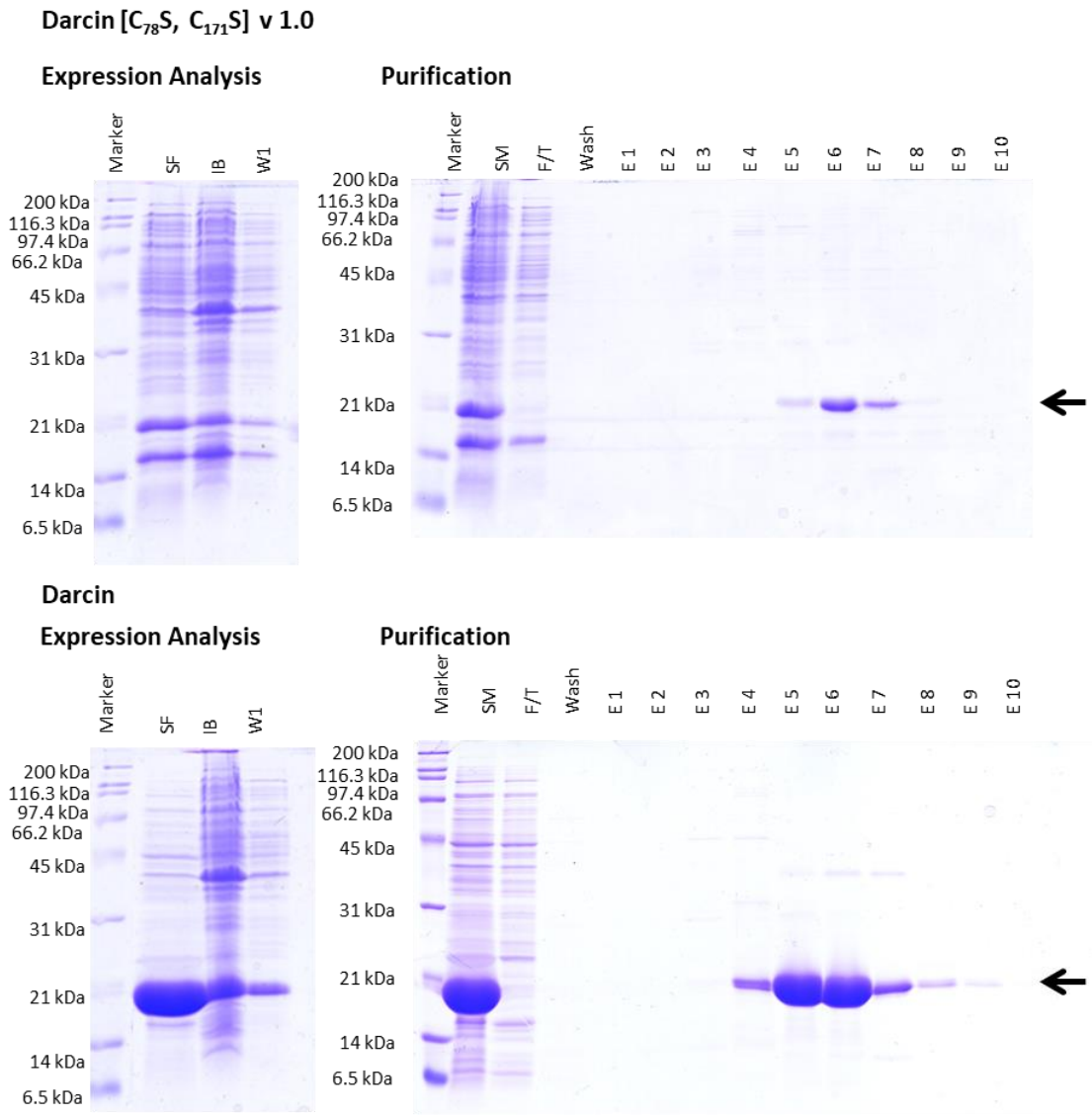


Figure 6.5 | Expression analysis and purification of darcin and darcin [$C_{78}S$, $C_{171}S$] v 1.0.

Top and bottom left: Cell pellet generated from cultures were lysed by sonication and centrifuged. The supernatant (SF) was collected and the remaining cell pellet washed by re-suspension. Sample was centrifuged at supernatant collected (W1). The cell pellet (inclusion body (IB)) was re-suspended for analysis. Darcin and darcin [$C_{78}S$, $C_{171}S$] v 1.0. predominantly expresses as soluble protein. Top and bottom right: His-tag proteins were purified using GE Healthcare NiNTA affinity columns on the AKTA start using a gradient elution of 0 % buffer B (elution buffer) to 100 % buffer B over 20 minutes. Arrows at approx. 20 kDa represent protein eluting from column. 10 µl of each fraction was ran on 15 % gels and visualised with Coomassie plus stain.

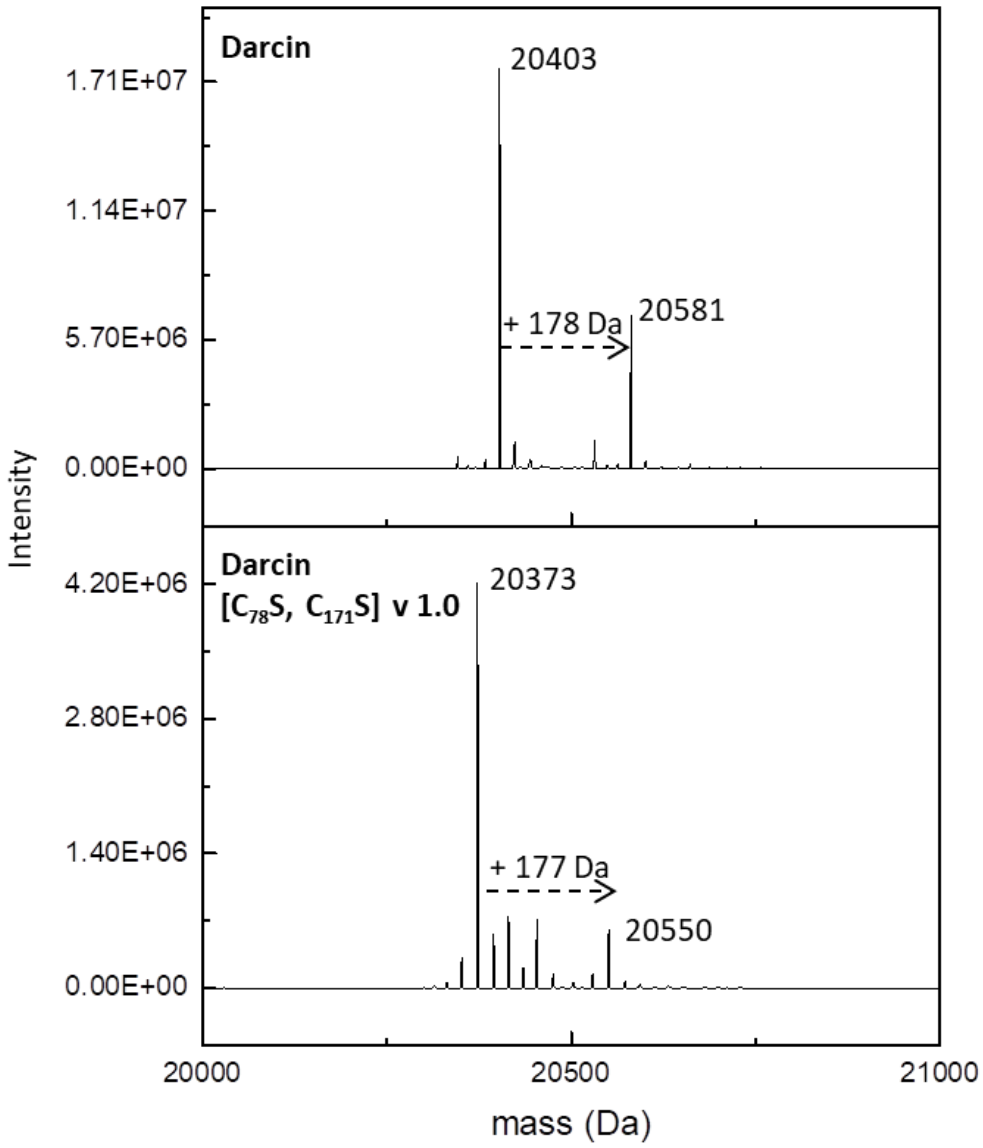


Figure. 6.6 | Deconvoluted ESI-MS analysis of darcin and darcin [C_{78}S, C_{171}S] v 1.0.

Protein samples were diluted to 1 pmol/µL in 3% acetonitrile, 0.1% formic acid. 2 µL of each sample was loaded onto a C4 desalting trapping column. Samples were analysed by ESI-MS on the Waters G2 mass spectrometer. Multiply charged protein envelope was deconvoluted using Waters MAXENT 1 algorithm to determine average mass of intact proteins. Observed masses of 20403 Da and 20581 Da for darcin corresponds to the loss of the initiating methionine and modification of gluconoylation respectively. Masses take into account disulfide bond (-2 Da). Observed mass for darcin [C_{78}S, C_{171}S] v 1.0 of 20373 Da and 20550 Da corresponds with the loss of the initiating methionine and modification of gluconoylation respectively.

6.3.3 Expression of darcin [$C_{78}S$, $C_{171}S$] v 2.0

The expression level of darcin [$C_{78}S$, $C_{171}S$] v 1.0 was considerably lower than that of darcin protein expression, as well as the obvious issues with truncation of some of the protein product of darcin [$C_{78}S$, $C_{171}S$] v 1.0. This could be explained due to two amino acid substitutions introduced into the darcin sequence resulting in two cysteine residues substituted for serine residues. As it was hoped that a cysteine to serine substitution would have minimal effect on protein expression, this result was disappointing.

Therefore, to deduce the cause of the altered expression level of the proteins, the DNA and protein sequences of darcin and darcin [$C_{78}S$, $C_{171}S$] v 1.0 were analysed further (Figure 6.7). Despite the protein sequences of darcin and darcin [$C_{78}S$, $C_{171}S$] v 1.0 differing in only two amino acid residues, at a DNA level, the sequences differed by 67 base residues. This could explain the low expression levels and truncation of darcin [$C_{78}S$, $C_{171}S$] v 1.0 compared to darcin. Codon frequency analysis was performed and revealed the gene was suitability optimised for expression in *E.coli,* so it was unexpected that darcin [$C_{78}S$, $C_{171}S$] v 1.0 did not express in a similar way to darcin. To assess this hypothesis, a second darcin mutant construct was designed, termed darcin [$C_{78}S$, $C_{171}S$] v 2.0. In this version the DNA sequence was identical to that of the original darcin construct, apart from the six DNA bases coding for the cysteine residues mutated to DNA bases that code for serine residues.

Figure 6.7 | DNA sequence alignment of darcin and darcin [$C_{78}S$, $C_{171}S$] v 1.0.

Protein sequences were aligned and visualised using CLC Viewer. Amino acid residues were coloured using the default Rasmol colours. Despite the amino acid sequences differing in two residues, there were 67 DNA bases differences.

The plasmid for darcin [$C_{78}S$, $C_{171}S$] v 2.0 was transformed into BL21 (λDE3) *E.coli* cells and expressed as described previously. Time points of the culture solution were taken during expression and analysed by SDS-PAGE. After IPTG induction a single protein band at the approximate molecular weight for darcin [$C_{78}S$, $C_{171}S$] v 2.0 (20 kDa) was visible, suggesting expression of the full length protein (Figure 6.8). When comparing protein expression levels between darcin [$C_{78}S$, $C_{171}S$] v 1.0 and darcin [$C_{78}S$, $C_{171}S$] v 2.0, increased protein expression levels were observed for darcin [$C_{78}S$, $C_{171}S$] v 2.0. This result demonstrates that, for this particular protein, keeping the original gene optimisation is the ideal strategy for yielding high protein expression when creating a variant with minor modifications to a pre-existing gene.

The darcin [$C_{78}S$, $C_{171}S$] v 2.0 cell pellet was sonicated and centrifuged. The protein was expressed in the soluble fraction, determined by SDS-PAGE analysis of the soluble fraction and insoluble inclusion body (Figure 6.9). Darcin [$C_{78}S$, $C_{171}S$] v 2.0 was purified as described in Chapter 2.3 and eluted in fractions 5 to 8. Elution fractions 6 and 7 were dialysed against 25 mM ammonium bicarbonate and the protein concentration determined using the Coomassie plus protein assay. Expression levels of darcin [$C_{78}S$, $C_{171}S$] v 2.0 were comparable to darcin. To confirm full length protein expression, darcin [$C_{78}S$, $C_{171}S$] v 2.0 was analysed by intact mass spectrometry as described in Chapter 2.21 (Figure 6.9). The main species within the mass spectrum had a mass of 20373 Da corresponding to the theoretical mass of darcin [$C_{78}S$, $C_{171}S$] v 2.0 minus the initiating N-terminal methionine. An additional species, with a mass of 20551 Da was also observed which corresponds to a modification of + 178 Da. This is the same modification observed with darcin corresponding to gluconoylation.

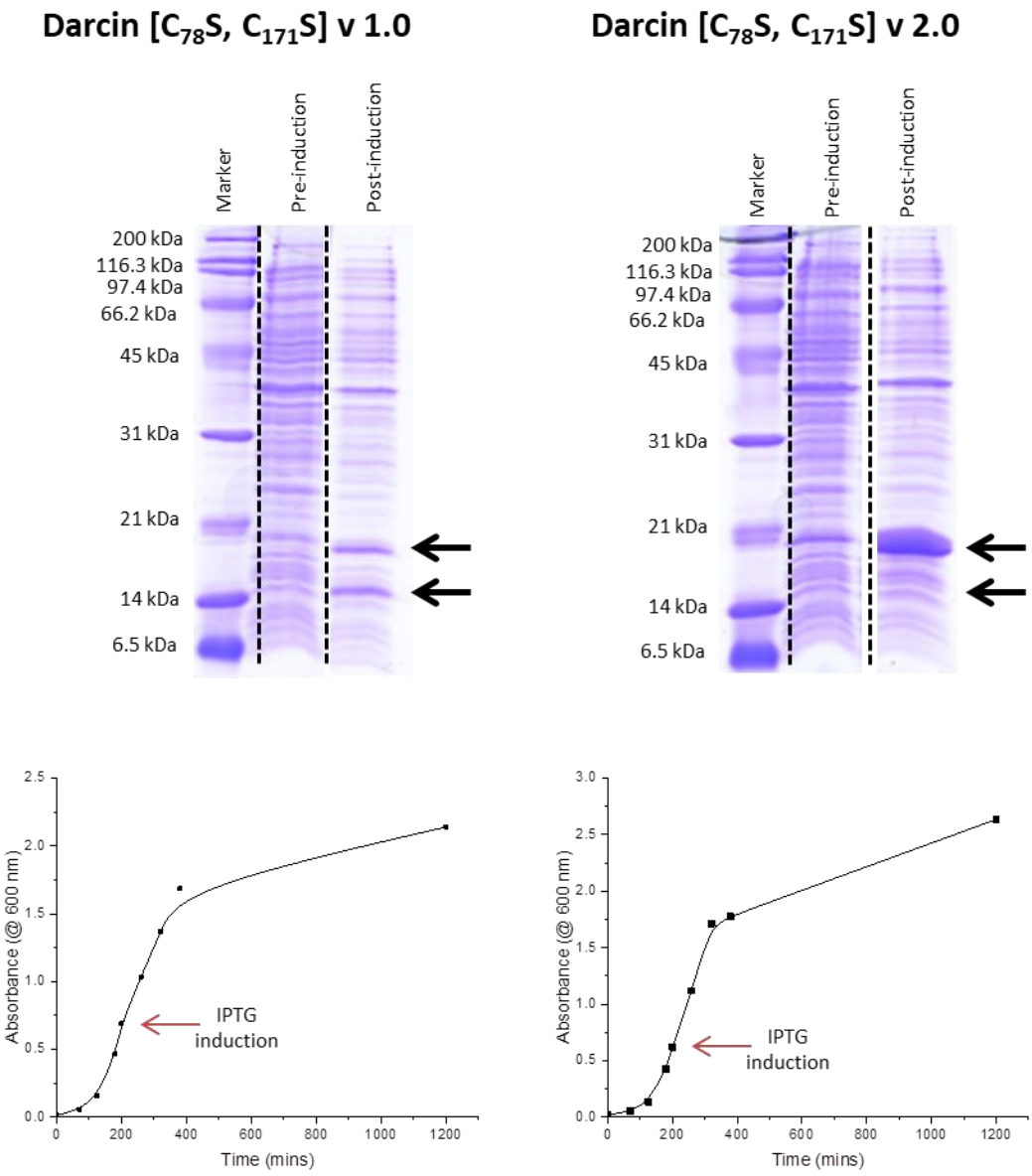


Figure 6.8 | *E.coli* culture and expression of darcin [$C_{78}S$, $C_{171}S$] v 1.0 and darcin [$C_{78}S$, $C_{171}S$] v 2.0.

SDS-PAGE analysis of pre-induction and post-induction timepoints and growth curves for darcin [$C_{78}S$, $C_{171}S$] v 1.0 and darcin [$C_{78}S$, $C_{171}S$] v 2.0 expression. Protein expression was induced with IPTG at an OD = 600 nm and protein left to accumulate for 16 hours. Arrows represent protein bands appearing after IPTG induction. A doublet band is visible at approximately 20 kDa and 16 kDa for darcin [$C_{78}S$, $C_{171}S$] v 1.0 and a single band at approximately 20 kDa for darcin [$C_{78}S$, $C_{171}S$] v 2.0. Samples ran on a 15 % gel and visualised with Coomassie plus stain.

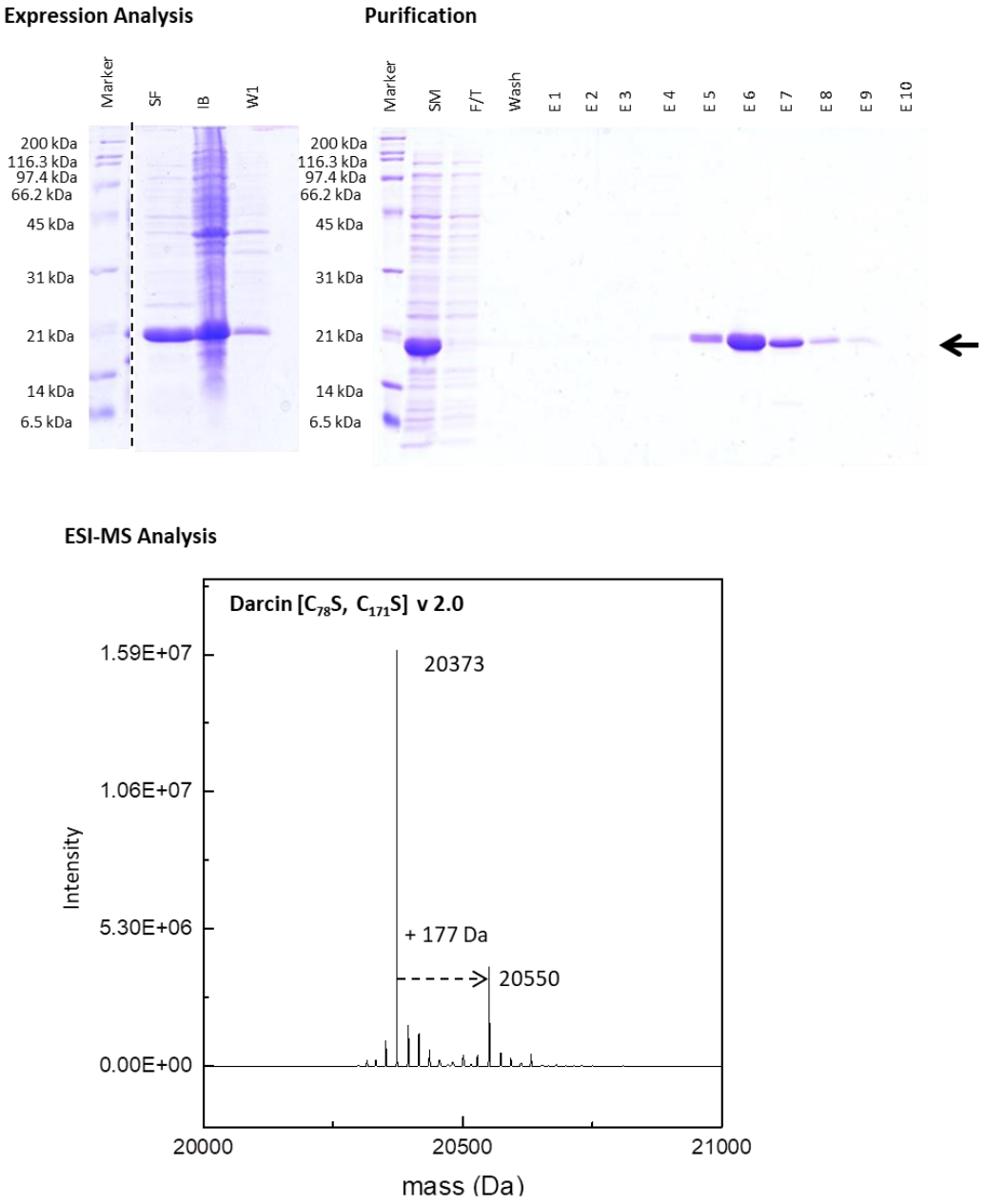


Figure 6.9 | Expression, purification and ESI-MS analysis of darcin [C_{78}S, C_{171}S] v 2.0.

Top: Cell pellet generated from culture was lysed by sonication and centrifuged. The supernatant (SF) was collected and the remaining cell pellet washed by re-suspension. Sample was centrifuged at supernatant collected (W1). The cell pellet (inclusion body (IB)) was re-suspended for analysis. Darcin [C_{78}S, C_{171}S] v 2.0. predominantly expresses as soluble protein. His-tag darcin [C_{78}S, C_{171}S] v 2.0. was purified using GE Healthcare Ni-NTA affinity columns on the AKTA start. Arrows at approx. 20 kDa represent protein eluting from column. Bottom: Protein samples were diluted to 1 pmol/µL in 3% acetonitrile, 0.1% formic acid. 2 µL of protein was loaded onto a C4 desalting trapping column and analysed by ESI-MS on the Waters G2 mass spectrometer. Multiply charged protein envelope was deconvoluted using Waters MAXENT 1 algorithm to determine average mass of intact proteins. Observed masses of 20373 Da and 20551 Da correspond to the loss of the initiating methionine and modification of gluconoylation respectively.

6.3.4 Role of the disulfide bond in stability of darcin

To explore whether the disulfide bond in darcin helped to provide the high stability, a proteolysis study was undertaken to examine the rate of degradation of darcin and darcin [$C_{78}S$, $C_{171}S$] v 2.0, the variant lacking the disulfide bond. It was hypothesised that darcin [$C_{78}S$, $C_{171}S$] v 2.0 would be more susceptible to degradation than darcin and therefore undergo proteolysis at a quicker rate. A total of 200 µg of both proteins were separately incubated with 1.25 µg trypsin at 37 ^{o}C for 3 hours. Samples were taken periodically to assess the rate of proteolysis by SDS-PAGE and MS. Proteolysis was stopped by the addition of TCA and neutralised with ammonium bicarbonate for SDS-PAGE analysis (Figure 6.10) or by formic acid for analysis by mass spectrometry (Figure 6.11). A control sample of each protein was also incubated at 37 ^{o}C for 3 hours to confirm that the trypsin was responsible for degradation and not the temperature. SDS-PAGE analysis revealed that after a 3 hour incubation with trypsin, darcin does not undergo complete proteolysis. Lower molecular weight species become visible during the time-course with a consistent intermediate species forming after 10 minutes at approximately 14 kDa. The protein band for full length darcin disappeared after 60 minutes. Intact mass spectrometry analysis of proteolysis time-point supports this result as a stable fragment species at 17973 Da becomes the main species from 15 minutes onwards with the full length darcin signal disappearing after 45 minutes. Conversely, darcin [$C_{78}S$, $C_{171}S$] v 2.0 undergoes complete and rapid proteolysis after 3 hour incubation with trypsin indicated by smearing on SDS-PAGE, with partial proteolysis occurring within the first 2 minutes of the time-course. Intact mass analysis of the time-course fractions supports this result with no species detectable after 5 minutes due to extensive proteolysis of the starting material (Figure 6.11). Analysis of control gels reveal that a temperature of 37^{o}C does not cause either protein to degrade after 3 hour incubation (Figure 6.10).

These findings suggest that removal of the disulfide bond in darcin to form darcin [$C_{78}S$, $C_{171}S$] v 2.0 reduces the protein stability and resistance to proteolysis. The disulfide bond allows the three dimensional structure of the protein to remain stable and compact. Removal of the bond causes the protein structure to unfold more readily, increasing the number of potential proteolytic sites available to interact with the protease enzyme, thus increasing the rate of degradation. As stability is a highly desired feature of a protein scaffold, the results suggest that darcin [$C_{78}S$, $C_{171}S$] v 2.0 is unsuitable for development whereas the usual disulfide bond containing darcin may be a more suitable candidate for development as a protein scaffold.

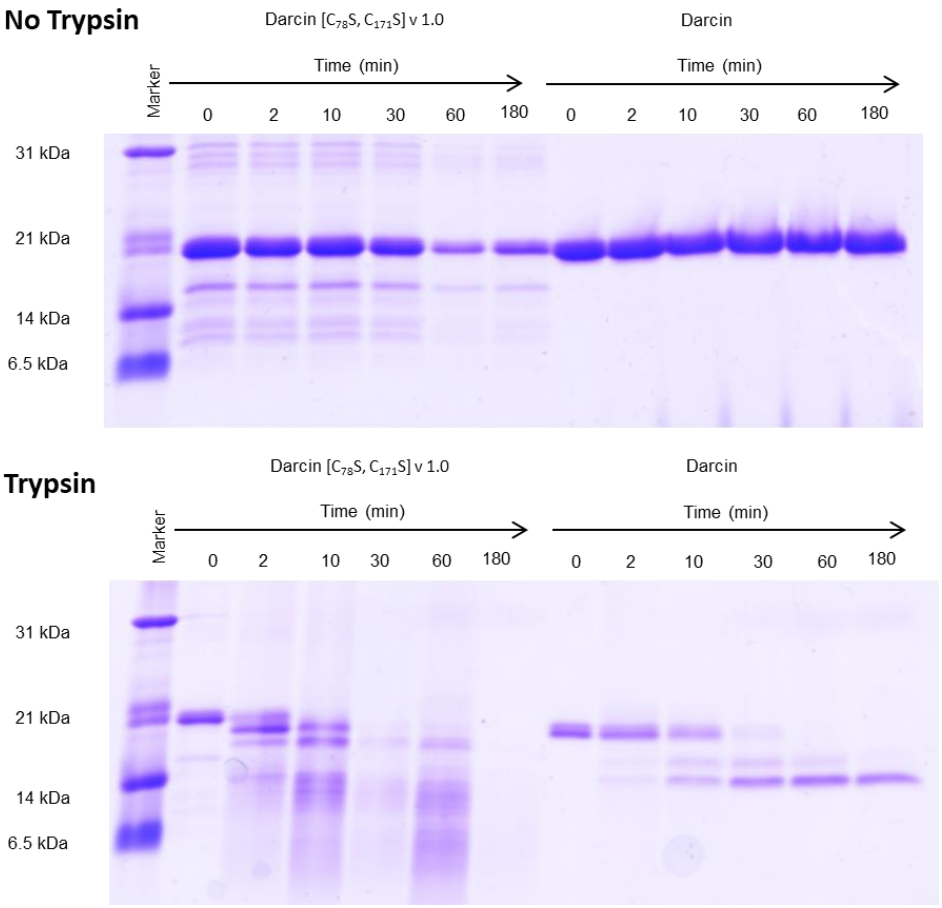


Figure 6.10 | SDS-PAGE analysis of darcin and darcin [$C_{78}S$, $C_{171}S$] v 2.0 proteolysis time-course.

Recombinant darcin and darcin [$C_{78}S$, $C_{171}S$] v 2.0 were incubated at 37 ^{o}C for 3 hours. Aliquots of sample were taken throughout the time-course for analysis. Top gel: Control gel of protein incubated without trypsin indicates darcin and darcin [$C_{78}S$, $C_{171}S$] v 2.0 remain intact at 37 ^{o}C. Bottom gel: Protein incubated with trypsin. Darcin [$C_{78}S$, $C_{171}S$] v 2.0 undergoes rapid and complete proteolysis represented by smearing on gel. Darcin forms a stable proteolytic fragment at 3 hours and is resistant to complete proteolysis. Samples ran on a 15 % gel and visualised with Coomassie plus stain.

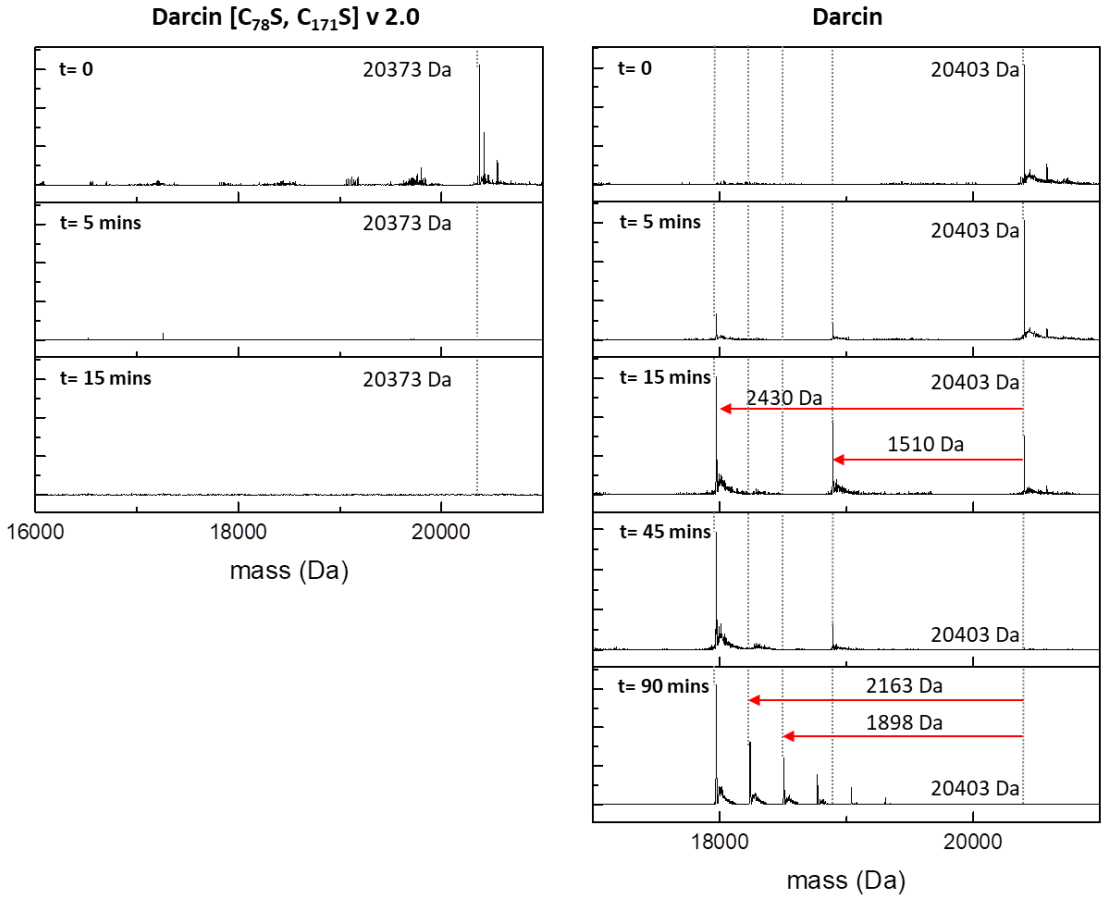


Figure 6.11 | ESI–MS analysis of darcin and darcin [$C_{78}S$, $C_{171}S$] v 2.0 proteolysis time-course.

Recombinant darcin and darcin [$C_{78}S$, $C_{171}S$] v 2.0 were incubated at 37 ^{o}C for 3 hours. Aliquots of sample were taken throughout the time-course for analysis. Proteolysis was halted by the addition of TCA (5 % final concentration). No detectable species were measurable for darcin [$C_{78}S$, $C_{171}S$] v 2.0 at 5 minutes onwards indicating rapid proteolysis. Darcin withstood complete proteolysis with an intermediate stable species forming. 1510.5 Da initially removed corresponding to the His-tag (GSSHHHHHHIEGR) followed by a further 920 Da loss corresponding to more of the N-terminal (EEASSMER).

6.3.5 Assessment of non-specific background in darcin affinity purification assay

As previously discussed, a protein scaffold should be biologically neutral, meaning that it has low or no binding affinity for proteins when in a complex biological matrix and that it exhibits no biological activity. Therefore, to explore protein interactions with darcin, an affinity purification assay using darcin in place on an affinity reagent was performed as described in Chapter 2.12. To assess whether removal of the disulfide bond in darcin alters protein interactions and affinity of matrix proteins to darcin, darcin [$C_{78}S$, $C_{171}S$] v 2.0 was also analysed in the assay.

Darcin and darcin [$C_{78}S$, $C_{171}S$] v 2.0 were immobilised onto Ni-NTA magnetic agarose beads via the His-tag. The bead-darcin complexes were incubated with either yeast lysate or human plasma, separated from the biological background using a magnetic rack and then washed to remove non-specific background. To identify proteins that bind to the Ni-NTA beads, a bead control with no darcin or darcin [$C_{78}S$, $C_{171}S$] v 2.0 was exposed to either yeast lysate or human plasma and analysed as described above. The samples were divided into two for SDS-PAGE or LC-MS/MS analysis.

SDS-PAGE analysis revealed that with either darcin or darcin [$C_{78}S$, $C_{171}S$] v 2.0 bound to the beads, the amount of non-specific proteins from both human plasma (Figure 6.12) and yeast lysate (Figure 6.13) observed is less than beads only exposed to biological matrix. This was more evident with the human plasma than the yeast. It is likely that both darcin and darcin [$C_{78}S$, $C_{171}S$] v 2.0 block the available interaction sites on the bead surface preventing additional proteins binding. Additionally, the beads had selective binding properties, displaying preferences to certain proteins. This is evident by the notable differences observed between the plasma only lane and lanes where the beads have been incubated with plasma; these lanes do not look like the plasma lane. Beads incubated with yeast lysate also appear to share this selective binding property. Non-specific binding of both human plasma and yeast lysate was similar in darcin and darcin [$C_{78}S$, $C_{171}S$] v 2.0.

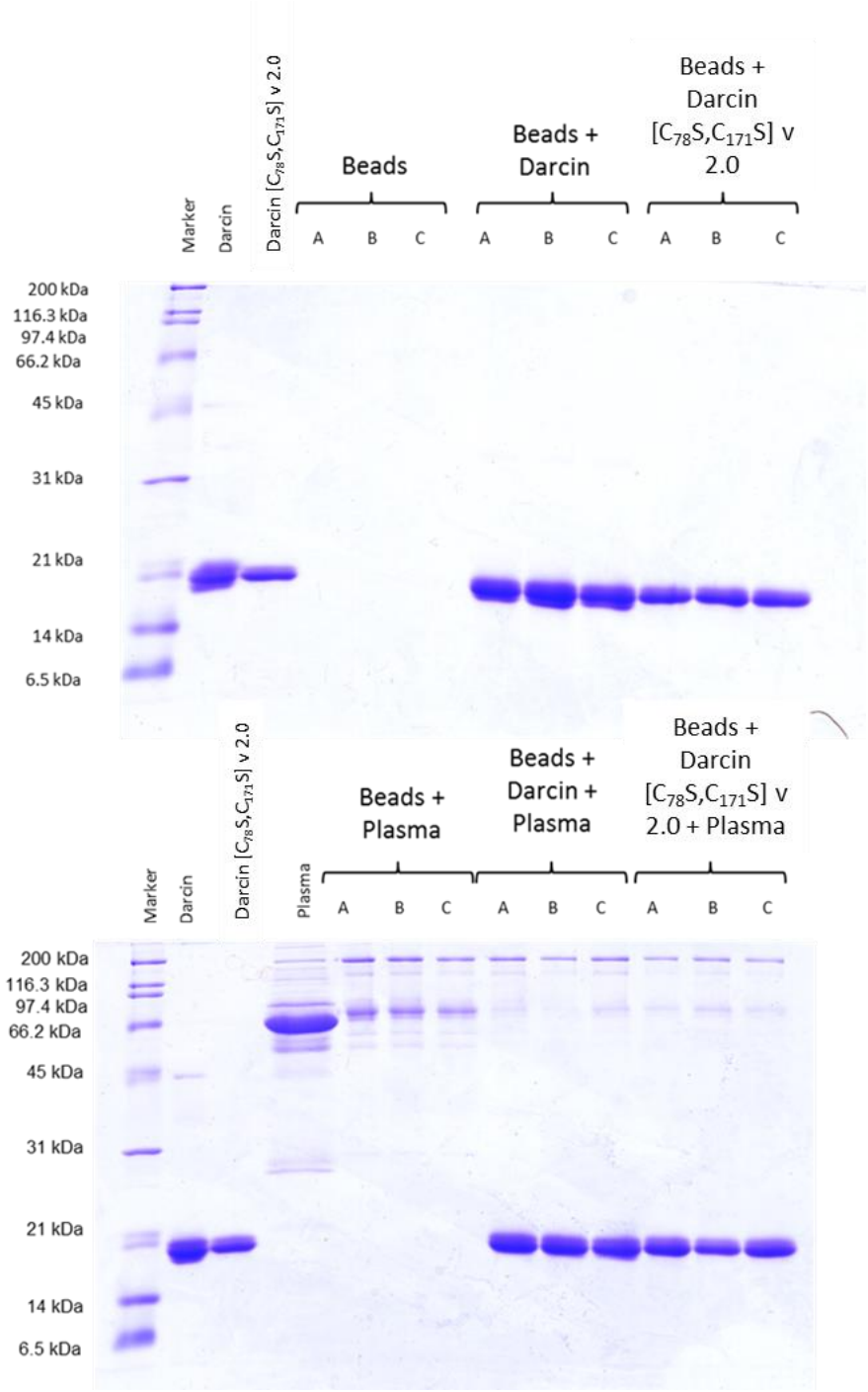


Figure 6.12 | SDS-PAGE analysis of darcin and darcin [C_{78}S, C_{171}S] v 2.0 background analysis of plasma.

Beads incubated with sample buffer and loaded directly onto the gel. Top: Control gel with beads only, beads & darcin and beads and darcin [C_{78}S, C_{171}S] v 2.0. Bottom: Beads & plasma, beads, darcin & plasma and beads darcin [C_{78}S, C_{171}S] v 2.0 and plasma. All samples analysed in triplicate as indicated by A, B and C. Samples were ran on 15 % gels and visualised with Coomassie plus stain.

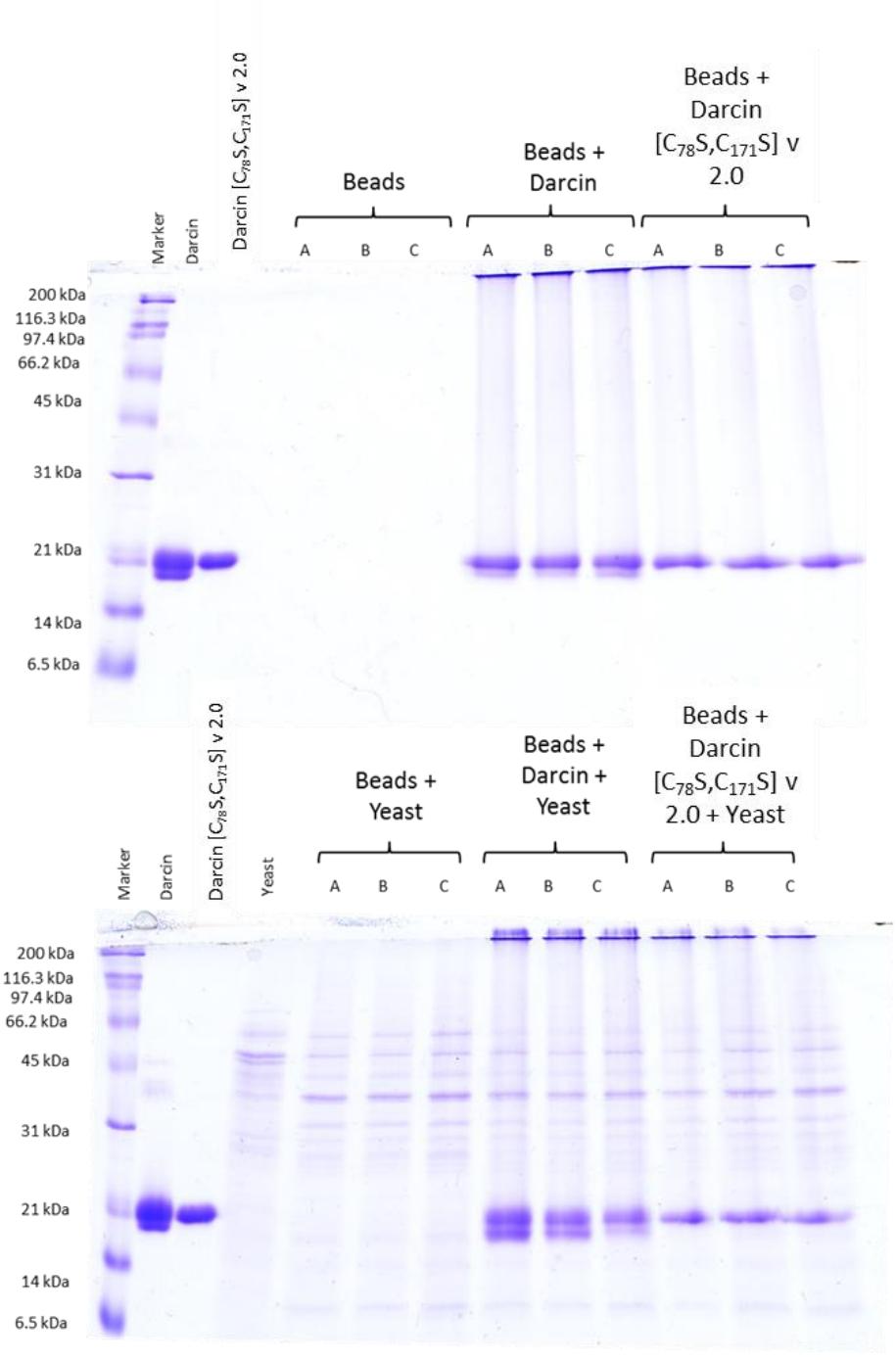


Figure 6.13 | SDS-PAGE analysis of darcin and darcin [$C_{78}S$, $C_{171}S$] v 2.0 background analysis of yeast lysate.

Beads incubated with sample buffer and loaded directly onto the gel. Top: Control gel with beads only, beads & darcin and beads and darcin [$C_{78}S$, $C_{171}S$] v 2.0. Bottom: Beads & yeast, beads, darcin & yeast and beads darcin [$C_{78}S$, $C_{171}S$] v 2.0 and yeast. All samples analysed in triplicate as indicated by A, B and C. Samples were ran on 15 % gels and visualised with Coomassie plus stain.

To prepare samples for LC-MS/MS analysis, an on-bead digest was carried out as described in Chapter 2.17. Peptides were analysed by the Thermo Scientific QExactive HF mass spectrometer on a 30 minute gradient. The raw data was processed using Proteome Discoverer and the generated mgf file searched in MASCOT against a database contain the two darcin sequences (sequences in Figure 6.2) and either a human or yeast database for the plasma or yeast lysate samples respectively. A 1% FDR filter was applied and proteins had to be identified by two or more peptides. As samples were analysed in triplicate, the average number of protein identified was reported for each condition (Figure 6.14). Label-free quantification data was not obtained for this experiment as it was a preliminary investigation into the non-specific binding. As observed with SDS-PAGE results, the plasma only and yeast only beads achieved the highest number of protein identifications. Due to the inherent 'stickiness' of the agarose beads, this is expected[216]. Furthermore, as the beads had no darcin or darcin [$C_{78}S$, $C_{171}S$] v 2.0 coated on them prior to exposure to biological sample, binding sites on the bead surface and the nickel sites would be available to interact with plasma and yeast proteins. More proteins were identified in the beads exposed to both human plasma and yeast lysate that had darcin [$C_{78}S$, $C_{171}S$] v 2.0 compared to darcin suggesting that the increased flexibility in the structure of darcin [$C_{78}S$, $C_{171}S$] v 2.0 increases the amount of non-specific binding to the protein. In the beads only, darcin only and darcin [$C_{78}S$, $C_{171}S$] v 2.0 only samples, an average of 4, 3 and 3 proteins were identified, respectively. Other than darcin and darcin [$C_{78}S$, $C_{171}S$] v 2.0, the proteins identified were keratins. Despite precautions and care taken during the affinity purification and sample digest steps, keratin contamination is extremely common and nearly unavoidable during proteomic analysis. A separate search of the yeast data against a human database did reveal keratin identification. Darcin and darcin [$C_{78}S$, $C_{171}S$] v 2.0 was identified in all digests where it had been pre-bound to the beads however, due to the high sequence homology between darcin and darcin [$C_{78}S$, $C_{171}S$] v 2.0, the two proteins could not be discriminated.

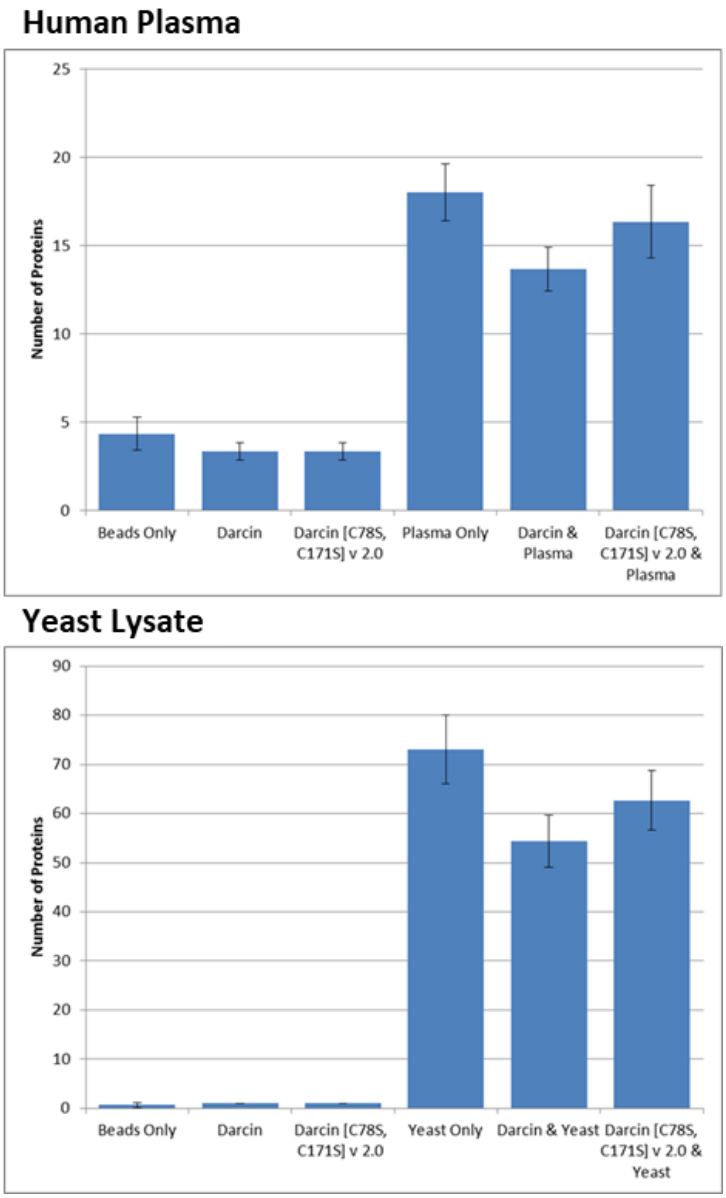


Figure 6.14 | Average number of protein identifications in background binding analysis comparison.

Top: Human plasma samples. Bottom: Yeast lysate samples. Error bars represent standard deviation. Proteins identified with 2 or more peptides.

To further investigate the non-specific proteins binding to darcin and darcin [$C_{78}S$, $C_{171}S$] v 2.0, the proteins identified in the different conditions were compared. Proteins identified in at least two of the replicates were included in the analysis. The MASCOT score and number of peptides identified per protein were reported for the plasma background binding (Table 6.1) and yeast background binding (Table 6.2) analysis. In total, 22 and 83 unique proteins were identified in the plasma and yeast lysate background experiment respectively.

Histidine-rich glycoprotein was identified in all samples where plasma was added. As indicated by the name, histidine-rich glycoprotein contains a large number of histidine residues. As the magnetic nickel beads using in the affinity purification bind proteins through the his-tag, it is likely that histidine-rich glycoprotein binds to the beads. Apolipoprotein C-III and apolipoprotein E are the only two proteins identified in the darcin & plasma and darcin [$C_{78}S$, $C_{171}S$] v 2.0 & plasma that were not identified in the plasma only sample. This may suggest that these two proteins bind non-specifically to both darcin and darcin [$C_{78}S$, $C_{171}S$] v 2.0 instead of to the magnetic beads. Apolipoprotein D and Ig kappa chain C region are the only two proteins identified in solely darcin [$C_{78}S$, $C_{171}S$] v 2.0 & plasma samples. This again suggests that the more flexible structure of darcin [$C_{78}S$, $C_{171}S$] v 2.0 results in more non-specific binding of plasma proteins. Most proteins identified in the plasma only control are also identified in both the darcin & plasma and darcin [$C_{78}S$, $C_{171}S$] v 2.0 & plasma samples. This suggests there is a common subset of proteins that bind non-specifically to the Ni-NTA magnetic agarose beads. Of these proteins identified, most have been previously identified as contaminates of affinity purifications[299]. When analysing the yeast data, the same pattern was observed with most proteins identified in the yeast only samples also identified in the darcin & yeast and darcin [$C_{78}S$, $C_{171}S$] v 2.0 & yeast samples. Significantly more proteins were identified in the yeast background binding analysis compared to the human plasma experiments. Although the same amount of protein was used in both studies, the dynamic range of a yeast lysate is lower than human plasma[19] which could explain the lower number of protein identifications in the human plasma background binding analysis. A large number of ribosomal proteins (46) were identified in the yeast samples but as they form complexes, this was expected[300]. Darcin was identified in both the bead only samples for the yeast and plasma experiments, despite not being incubated with beads. Therefore, it was probably identified because of carryover from the previous sample analysed in the mass spectrometer, despite blank samples being ran in-between each sample run to try and reduce carryover.

Table 6.1 | Summary of the MASCOT score and number of peptides observed for proteins identified in plasma background binding analysis.

Protein identified in at least two of the three replicates. The highest MASCOT score reported. Proteins ordered alphabetically.

	Mascot Score						Number of Peptides					
Protein	Beads Only	Darcin	Darcin [$C_{78}S$, $C_{171}S$] v 2.0	Plasma Only	Darcin & Plasma	Darcin [$C_{78}S$, $C_{171}S$] v 2.0 & Plasma	Beads Only	Darcin	Darcin [$C_{78}S$, $C_{171}S$] v 2.0	Plasma Only	Darcin & Plasma	Darcin [$C_{78}S$, $C_{171}S$] v 2.0 & Plasma
Actin, cytoplasmic 1	63			73			2			2		
Alpha-2-macroglobulin				1067	859	836				34	29	32
Apolipoprotein A-I				301	361	339				7	9	8
Apolipoprotein C-III					198	155					2	2
Apolipoprotein D						50						2
Apolipoprotein E					127	114					2	3
Clusterin				200	218	228				5	6	5
Darcin	158	788	889	88	940	1399	3	13	13	3	13	13
Fibrinogen alpha chain				239		147				7		3
Fibrinogen beta chain				170		97				5		3
Fibrinogen gamma chain				85						2		
Hemopexin				205	114	130				8	3	6
Histidine-rich glycoprotein				271	272	218				9	8	5
Ig gamma-1 chain C region				104	88	101				3	2	2
Ig kappa chain C region						107						2
Keratin, type I cytoskeletal 10		335	320	514	440	387		12	11	14	15	11
Keratin, type I cytoskeletal 9	217						6					
Keratin, type II cytoskeletal 1	228	396	350	444	439	446	8	11	12	14	13	12
Kininogen-1				153	157	147				4	4	4
Selenoprotein P				71	79					2	3	
Serum albumin				370	482	579				16	21	21
Serum amyloid P-component				164						6		

Table 6.2 | Summary of the MASCOT score and number of peptides observed for proteins identified in yeast lysate background binding analysis.

Protein identified in at least two of the three replicates. The highest MASCOT score reported. Proteins ordered alphabetically.

	Mascot Score						Number of Peptides					
Protein	Beads Only	Darcin	Darcin [$C_{78}S$, $C_{171}S$] v 2.0	Yeast Only	Darcin & Yeast	Darcin [$C_{78}S$, $C_{171}S$] v 2.0 &Yeast	Beads Only	Darcin	Darcin [$C_{78}S$, $C_{171}S$] v 2.0	Yeast Only	Darcin & Yeast	Darcin [$C_{78}S$, $C_{171}S$] v 2.0 &Yeast
40S ribosomal protein S0-A					157						4	
40S ribosomal protein S0-B				59		168				3		5
40S ribosomal protein S11-A				141	104	128				4	3	5
40S ribosomal protein S13				122	162	143				4	5	4
40S ribosomal protein S14-A				137	190	186				3	4	4
40S ribosomal protein S15					70	82					2	2
40S ribosomal protein S16-A					102	134					3	4
40S ribosomal protein S17-A				127	138	201				4	2	4
40S ribosomal protein S18-A				206	218	233				6	7	8
40S ribosomal protein S1-A				159	147	193				5	5	8
40S ribosomal protein S2				92	169	139				4	6	4
40S ribosomal protein S20				213	115	128				6	4	6
40S ribosomal protein S23-A						75						4
40S ribosomal protein S25-B				64						2		
40S ribosomal protein S27-A				60	79	80				2	3	3
40S ribosomal protein S3				373	381	374				12	13	13
40S ribosomal protein S4-A				176	174	188				9	11	8
40S ribosomal protein S5				120	235	259				4	7	7
40S ribosomal protein S6-A				170	192	179				4	3	4
40S ribosomal protein S7-A				45		81				2		4
40S ribosomal protein S8-A				190	212	261				3	3	5
40S ribosomal protein S9-B				100	169	161				3	7	6
6,7-dimethyl-8-ribityllumazine synthase				74						2		
60S acidic ribosomal protein P0				79	168	152				3	5	5
60S ribosomal protein L10					77	109					3	4

60S ribosomal protein L11-A				53	56					2	3		
60S ribosomal protein L12-A				114						2			
60S ribosomal protein L15-A				149	70	96				5	3	5	
60S ribosomal protein L16-B				192	151	201				6	4	7	
60S ribosomal protein L18-A				232	202	213				8	6	7	
60S ribosomal protein L19-A					87	103					3	3	
60S ribosomal protein L20-A						108						3	
60S ribosomal protein L21-A				88	86	76				4	4	3	
60S ribosomal protein L23-A				53						2			
60S ribosomal protein L26-A				37						2			
60S ribosomal protein L27-A				122	173	156				3	5	4	
60S ribosomal protein L28				101						2			
60S ribosomal protein L2-A				59	66	100				2	3	4	
60S ribosomal protein L3				287	360	346				9	10	9	
60S ribosomal protein L32				68	59					3	4		
60S ribosomal protein L33-A						101						2	
60S ribosomal protein L37-A				90	122	87				3	4	3	
60S ribosomal protein L4-A					122	187					4	5	
60S ribosomal protein L6-A				69						3			
60S ribosomal protein L7-A				114	139	123				5	3	5	
60S ribosomal protein L8-A					107						3		
60S ribosomal protein L9-A					100	88					2	3	
Actin				136	204	193				4	6	6	
Alcohol dehydrogenase 1				509	367	399				12	11	12	
Asparagine synthetase [glutamine-hydrolyzing] 2				122						4			
Aspartate--tRNA ligase, cytoplasmic				94						3			
ATP synthase subunit alpha, mitochondrial				109						3			
ATP-dependent 6-phosphofructokinase subunit alpha				162						4			
ATP-dependent 6-phosphofructokinase subunit beta				182	90	114				5	2	3	
ATP-dependent molecular chaperone HSP82				159	162	254				9	4	7	
ATP-dependent RNA helicase eIF4A						88						3	
Carboxypeptidase S				49						2			
Darcin	302	5209	3761	421	4474	5262	6	16	13	4	15	17	
Elongation factor 1-alpha				192	154	137				7	6	6	
Elongation factor 2				124						2			
Elongation factor 3A				103		141				2		8	
Enolase 2				423	403	492				10	12	14	

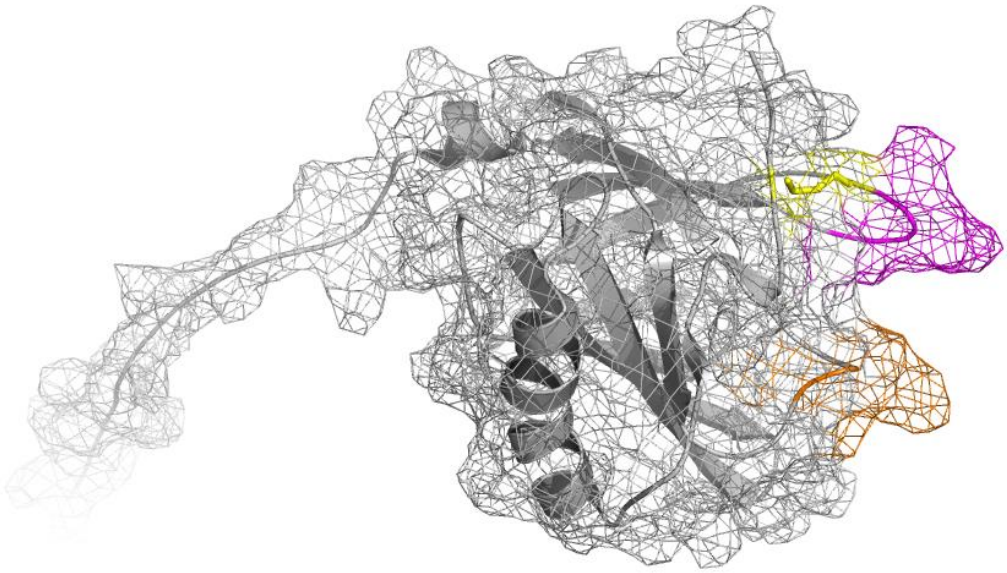
Enolase 2				423	403	492				10	12	14
Fructose-bisphosphate aldolase				89	166	176				2	2	3
Glucose-6-phosphate isomerase				111						3		
Glyceraldehyde-3-phosphate dehydrogenase 3				669	665	696				12	14	15
Guanine nucleotide-binding protein subunit beta-like protein				212	221	205				6	7	7
Heat shock protein 26					61	78					2	2
Heat shock protein SSA1				139	149	166				5	5	5
Heat shock protein SSC1, mitochondrial				345	232	360				8	7	8
Mannose-1-phosphate guanyltransferase				42						2		
Nicotinamidase				91	116	154				3	2	4
Phosphoglycerate kinase					179	165					4	3
Phosphoglycerate mutase 1				262	313	315				9	12	13
Plasma membrane ATPase 1				53						2		
Protein BMH1				128	134	112				5	4	3
Pyruvate decarboxylase isozyme 1				234	361	322				7	7	7
Pyruvate kinase 1				470	345	430				11	8	10
Serine hydroxymethyltransferase, cytosolic				230						5		
Thioredoxin-2				76						2		
Transposon Ty1-DR5 Gag-Pol polyprotein				78	67	122				2	2	3
Ubiquitin-40S ribosomal protein S31						94						2
UTP--glucose-1-phosphate uridylyltransferase				395	193	317				14	6	11
Vacuolar aminopeptidase 1				87						2		

6.3.6 Engineered darcin

Although it has been determined that darcin is extremely stable and that the disulfide bond is vital in preserving this stability, it is necessary that a protein scaffold possesses additional features to make them suitable for use[147], as discussed in Chapter 1. Therefore the design of various engineered darcin proteins will be discussed.

Variable loop regions

The ability to engineer variable loop regions into the protein to facilitate binding is a key feature of a protein scaffold. Therefore, to identify regions of the protein suitable for variable peptide loop insertions, the 3D structure of darcin was visualised and examined using PyMol. The protein structure was loaded into PyMol from the Protein Data Bank archive and represented as both a ribbon and mesh structure. The mesh view was more informative as it provides a view of the reach of the proteins interactions. Areas of the protein that had 'natural' loop regions and clefts were first considered for selection. Figure 6.15 presents the 3D structure of darcin and highlights the two regions of the protein hypothesised for loop selection. The two proposed loop regions naturally form loops in darcin and therefore could be suitable for further engineering. Loop 1 consists of residues 46 - 49 and loop 2 consists of residues 73-77, separated by 24 residues. To evaluate the suitability of the selected loops regions further, a multiple sequence alignment analysis of the MUP proteins was performed (Figure 6.16). The MUP proteins have high sequence homology and therefore by analysing whether the proposed loop regions are conserved across the MUPs may provide insight into the role of the selected regions. The protein sequence of loop 1 is fairly consistent across all MUP whereas the sequence of loop 2 is more variable. The high variability of loop 2 may suggest that this portion of the protein does not play a role in the proteins stability or function, indicating its suitability as a variable loop.



MGSSHHHHHHIEGREEASSMERNFNVEKINGEWYTIMLATDK
REKIEEHGSMRVFVEYIHVLENSLALKFHIIINEECSEIFLV
ADKTEKAGEYSVTYDGSNTFTILKTDYDNYIMIHLINKKDGE
TFQLMELYGREPDLSSDIKEKFAQLSEEHGIVRENIIDLTNA
NRCLEARE

Figure 6.15 | 3D structure of darcin with proposed loops.

Orange region represents loop 1. Purple region represents loop 2. Yellow residues represent cysteine residues involved in disulfide bond formation. Visualised in PyMol [191].

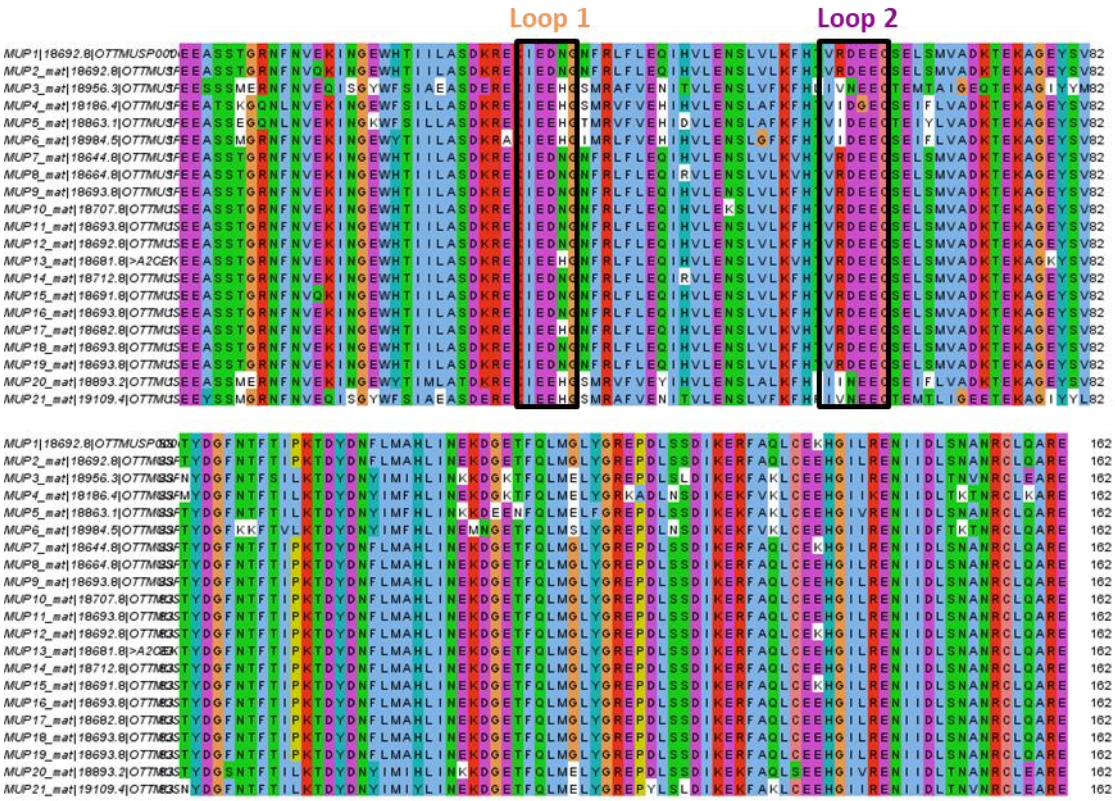


Figure 6.16 | Multiple amino acid sequence alignment of MUPs.

Sequences were aligned using Clustal X[301], a tool for multiple sequence alignment and visualised in Jalview[302]. Amino acid residues were coloured using the default Clustal X colours and highlights differences in residues. High sequence homology observed between MUPs. Black boxes represent part of the protein suggested for loop insertion.

'Non-digestible' darcin

As with the 'non-digestible' pepsinogen Adhiron, a darcin mutant that can withstand proteolysis would be beneficial for use in affinity purifications. As discussed in Chapter 3, a major problem with affinity purifications and co-immunoprecipitations is the large signal in mass spectrometry assays from the tryptic peptides of the reagent involved in the purification and enrichment. The signals of peptides of interest from enriched or purified proteins are usually too low to detect. Generating a 'non-digestible' affinity reagent would therefore be beneficial, allowing the reagent to be separated from peptides prior to MS analysis. A mutant darcin was designed that had all lysine residues substituted to arginine residues. Therefore, when the mutant darcin and its payload are digested using the proteases Lys-C, that cleaves at lysine residues, darcin will remain intact and its protein payload will undergo proteolysis, assuming the proteins contain lysine residues. Although trypsin is the standard protease used in proteomics, numerous other enzymes are used, for example in *de novo* sequencing protocols. A lysine to arginine substitution is considered the most favourable, as discussed previously in Chapter 4.

Cleavable His-tag darcin

The addition of a His-tag to recombinant darcin is necessary for simple protein purification from other proteins and cellular components. In addition, for use as a protein scaffold, the His-tag is beneficial for affinity purifications for protein immobilisation. However, the His-tag may affect the stability, conformation and activity of darcin. The His-tag is also fairly long and flexible which could interfere with the interaction sites. His-tag cleavage is easy to achieve using proteases. Therefore a mutant darcin was designed containing a protease cleavage site. The tobacco etch virus (TEV) protease recognises the amino acid sequence ENLYFQ↓G (cleavage site annotated with ↓) and was chosen due to its high specificity. To save time, both mutations of the lysine to arginine substitutions and TEV cleavage sites were combined into one modified protein, termed darcin [KtoR], as shown below. The red portion represents the His-tag, blue portion represents the TEV cleavage site, green residues represent arginine substitutions and the orange and purple portions represent loop 1 and loop 2, respectively.

MGSSHHHHHHIEGREENLYFQGEASSMERNFNVERINGEWYTIMLATDRRERIEEH
GSMRVFVEYIHVLENSLALRFHIIINEECSEIFLVADRTERAGEYSVTYDGSNTFT
ILRTDYDNYIMIHLINRRDGETFQLMELYGREPDLSSDIRERFAQLSEEHGIVREN
IIDLTNANRCLEARE

6.3.7 Confirmation of darcin [KtoR] expression

It was decided that a darcin variant with a cleavable His-tag and with a resistance to Lys-C proteolysis would be the first variant to express. To generate a plasmid for protein expression, the desired sequence was sent to Eurofins Scientific, a biotechnology company that provides genomic services. The typical workflow for plasmid generation first requires the optimised DNA sequence for the desired protein to be inserted into a vector. This vector is usually a cloning vector as they allow for the easy insertion of DNA fragments. To generate a protein product the vector requires promoter regions which requires an expression vector. However, a full length clone in an *E.coli* expression vector could not be generated and instead a cloning vector was supplied. To overcome this and as part of a collaboration, the wheat germ cell-free protein expression system was proposed for darcin [KtoR] protein expression. Victoria Harman from the Centre for Proteome Research, alongside collaborators at Ehime University in Japan carried out the work to express and purify the protein[303]. Due to requirements of the other proteins expressed in the workflow, darcin [KtoR] was expressed as heavy labelled.

To identify the product of protein expression and purification, an in-solution digest was carried out following the standard protocol as described in Chapter 2.17. However, it was suspected that the protein concentration was fairly low so the amount of trypsin added was adjusted. Peptides were analysed on the Thermo Scientific QExactive mass spectrometer on a 30 minute gradient and the raw data processed using Proteome Discoverer. The resulting mgf file was searched in mascot against a darcin database with the additional constant modification of [$^{13}C_6$][$^{15}N_4$] arginine due to the heavy expression. Based on MS/MS identifications, 50 % protein coverage was achieved (Figure 6.17). Extracted ion chromatograms of precursor ions were able to increase protein coverage to 89.6% (Figure 6.17). A large number of peptides were identified with missed cleavage sites. Previous work has demonstrated that darcin is difficult to digest fully due to the protein structure being fairly resistant to reduction and thus remains tightly folded. The high number of missed cleavage sites could also suggest that darcin [KtoR] is also fairly robust. The region of the protein with the TEV cleavage site was identified with a peptide with high quality MS/MS spectra and there was sufficient coverage of the other mutated residues to confirm the expression and purification of darcin [KtoR] [$^{13}C_6$][$^{15}N_4$].

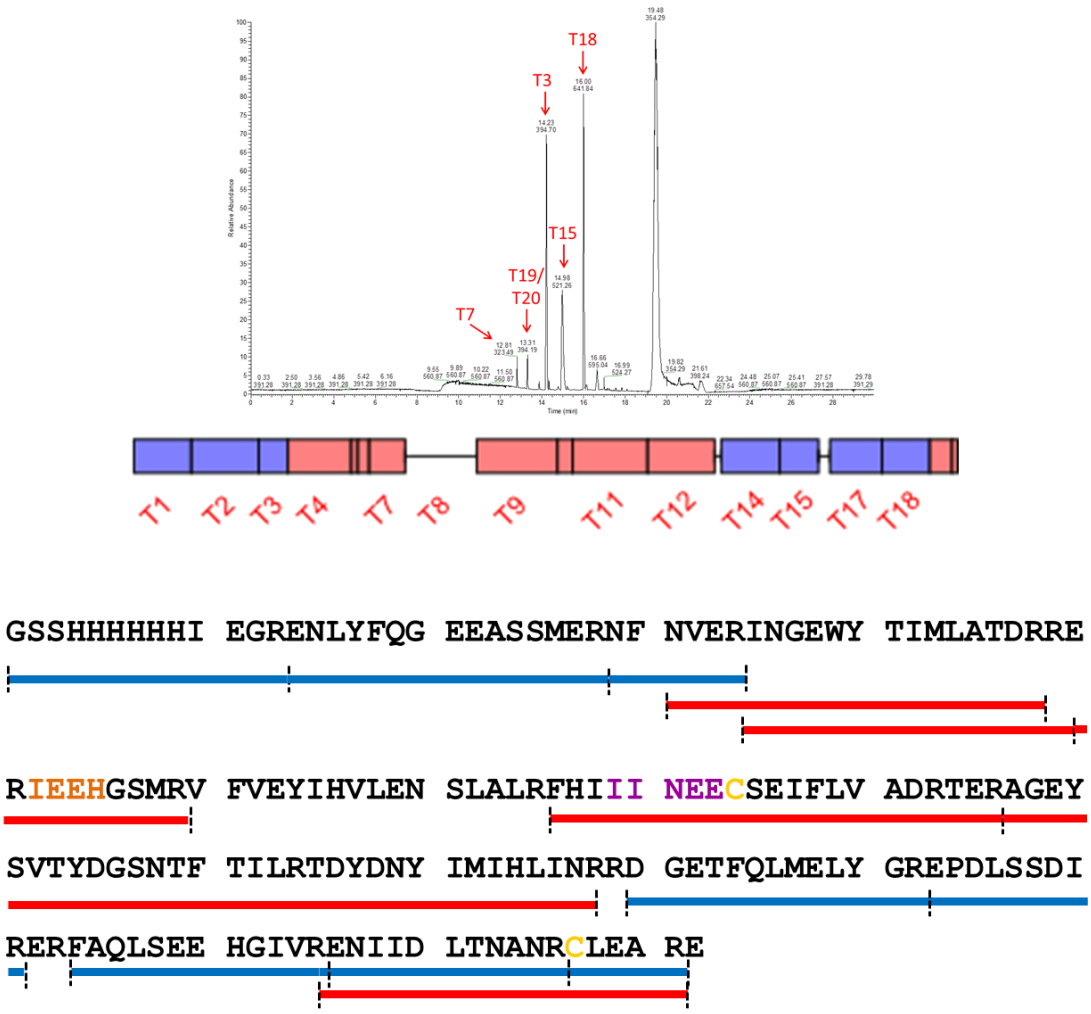


Figure 6.17 | Tryptic peptide map of darcin [KtoR].

89.6% sequence coverage achieved. Top: BPI chromatogram with the main peaks labelled with the corresponding peptide number. Middle: Peptide map of darcin [KtoR] generated using Peptide Mapper[234]. Blue peptides identified in mass spectrum with no missed cleaves, whereas red peptides were identified with one missed cleavage site. Bottom: Darcin [KtoR] protein sequence. Black residues are unidentified regions of the protein.

Summary

In total, three darcin constructs have been designed to further develop darcin as putative protein scaffold. A mutant darcin, termed darcin [KtoR] combining two of the three designs has been expressed successfully using cell-free methodology. Due to time constraints, it was not possible to investigate the effects on the stability and structure of darcin [KtoR] compared to darcin. Future work should confirm that the protein cannot be digested using Lys-C and that the his-tag can be cleaved from the protein using the TEV protease. If these mutations introduced into the protein do not have a detrimental effect on the stability of the protein then future work should also establish whether the introduction of loop regions into the protein sequence alter the stability of the protein.

6.4 Conclusion

The main aim of this work was to establish whether darcin was suitable for development as a protein scaffold. The robust nature and highly stable structure of darcin is an ideal feature of a protein scaffold. The preliminary developmental work has demonstrated that darcin is suitable and amenable for modifications. This work has confirmed that the disulfide bond is vital in providing the protein with stability. Although a darcin protein scaffold lacking a disulfide bond may express with greater ease and may provide more flexibility for the variable loop regions, all future designs should include the cysteine residues for disulfide bond formation. In an affinity purification application, darcin exhibits low levels of non-specific binding of plasma and yeast proteins. Despite this, the challenge with all affinity purification mass spectrometry workflows, of non-specific binding of proteins to the solid support and the overriding signal of the affinity reagent still exist. The successful expression of a darcin variant without lysine residues should address the issue of the signal of the protein scaffold by the application of the protease, Lys-C. Due to time restrictions, further work characterising darcin [KtoR] was not carried out however assessing the stability of the new variant is vital in determining its suitability as a scaffold.

Future work should investigate whether darcin can tolerate the insertion of peptide fragments into the proposed loop regions of the protein. Furthermore, the addition of a single cysteine at either the N- or C-terminus of the protein would be useful for covalent immobilisation and also allow for labelling, increasing the functionality of the scaffold. With further development and implementation of the discussed improvements, the work demonstrates that darcin has potential to serve as a novel protein scaffold.

Chapter 7: General Discussion and Conclusions

7.1 Summary

The aim of this thesis was to characterise next generation affinity reagents, with a focus on Affimer technology. The primary aims, described in each chapter were:

- Develop an affinity purification workflow for naïve Affimer target identification by MS.
- Characterise Affimers that target human pepsinogen.
- Identify differentially expressed proteins in sepsis plasma using LC-MS/MS as candidates for Affimer generation.
- Investigate whether darcin was suitable for development as a novel protein scaffold.

In each chapter, the underlying rationale was to develop approaches that address and overcome proteome complexity and the associated analytical challenges. Whilst this goal was partly achieved, because of the challenging nature of affinity purifications, alternative strategies to Affimer-based proteomics were employed for complex proteome analysis.

This thesis presents novel approaches to improve on current affinity purification methods and reagents. A novel proof-of-concept strategy to remove the affinity reagent from the affinity purification following enzymatic digestion was developed by generating both an Affimer and Adhiron that was resistant to Lys-C proteolysis. An affinity purification mass spectrometry method to identify Affimer array protein binders was developed as a novel strategy to identify differentially expressed proteins. Finally, a proteomic analysis of plasma was undertaken that identified an original panel of five proteins implicated in sepsis for Affimer development.

The work presented in the first chapter highlights the many approaches available for the immobilisation of Affimers for affinity purifications. Furthermore, the challenges of affinity purification method development, particularly when using low affinity binding reagents are also presented. A method for covalent immobilisation of Affimers through a free cysteine was developed and used successfully for the enrichment of both recombinant IgG and endogenous IgG from human plasma using a high affinity Affimer for IgG, produced using

phage display. However, the affinity purifications using naïve Affimers selected from the Affimer arrays were unsuccessful in enriching for interacting proteins.

Despite failing to identify binding partners of naïve Affimers, knowledge of the common non-specific background was obtained. A frequently reported problem of affinity purification assays coupled with mass spectrometry is determining a 'true' interaction from non-specific background[215]. Although numerous immobilisation chemistries and solid supports were investigated, background contamination of non-specific proteins continued to be problematic. Higher stringency washes could be applied to reduce background contamination however this could compromise low affinity interactions. These findings could be useful for future reference by adding to the previously identified non-specific contaminant protein repertoire, helping in the identification of 'true' affinity purification interactions.

Although for differential expression analysis, the naïve Affimer array overcomes the limitations of typical bottom-up LC-MS/MS as the method displays no bias towards the most abundant species, the methodology presents additional challenges for subsequent target identification. When the Affimers are immobilised on the array, changes in the protein conformation may occur that may differ when bound to the His-tag or SulfoLink® resin. This may explain the lack of Affimer target identification due to the inability to bind protein targets, although it has previously been established that Affimers are able to retain their recognition function when immobilised onto a solid surface. It is more likely that target identification was not possible due to the weak binding affinities of the naïve Affimers, an inherent feature of combinatorial peptides that do not undergo multiple rounds of panning to enrich for high affinity binders.

An approach to overcome the challenges of affinity purifications was considered and involves direct interrogation of the array using MALDI-MS, however due to the small amount of Affimer immobilised onto the array, the limit of detection of the instrument would not have been low enough for detection of the Affimer let alone the protein targets. It would be interesting to investigate whether, with increased amounts of Affimer immobilised on the array, direct interrogation of the array using MALDI-MS would allow protein detection. MALDI does not produce highly charged species during ionisation, and therefore limitations in the mass range of the mass analyser mean that high molecular proteins would not be detected. Therefore, proteolysis of the Affimer-target complex immobilised on the array would be required.

Affimer production can be untargeted with unknown binding partners such as with the naïve Affimers or targeted to interact with specific proteins using phage display technology[111,192]. In Chapter 4, Affimers were produced to target the human protein pepsinogen to address the limitations and availability of existing affinity reagents. The challenges of affinity purifications were again highlighted with pepsinogen interacting with the SulfoLink® resin preventing the confirmation of enrichment with the pepsinogen Adhirons.

To overcome this issue, other methods to assess binding could be employed such as isothermal titration calorimetry (ITC), surface plasmon resonance (SPR) and size-exclusion multi-angled light scattering (SEC-MALS). This would provide information regarding on the strength of binding, something that is not feasible with an affinity purification assay, allowing for the selection of the highest affinity binder. Gaining an insight into the specific sites of binding of an Affimer with its protein targets would be extremely beneficial and increase the applications of Affimers. For example, if the Affimers bound to two different epitopes on the target proteins surface, the Affimers could be used for simultaneous immobilisation and detection completely removing the need for antibodies in a sandwich ELISA formats. A limitation to these methods are that extremely pure proteins are needed for analysis and therefore recombinant or purified endogenous protein is require to explore protein interactions. Whilst this is possible for proteins with known binding targets such as those produced by phage display, analysis of the naïve Affimers could not be performed using this technique without earlier target identification. It would be interesting to investigate native MS of the Adhiron-pepsinogen complex to determine if binding is observed in the gas phase.

The preliminary work in Chapters 4 and 6 developing non-digestible protein scaffolds through specific engineering has great potential for future applications in affinity purification MS. It is common that the signal from the affinity reagents themselves suppress the detection and identification of any protein interactors. A non-digestible protein scaffold would eliminate this issue by allowing a pre-digestion step using Lys-C to remove captured proteins followed by the selective removal of the undigested protein scaffold using either precipitation techniques or a molecular weight cut-off. Unfortunately, this method would not eliminate non-specifically bound proteins from analysis. Alternatively, if the variable loops in the binding region were unknown, then the protein could be retained for MS/MS sequencing.

Although the challenges of Affimer affinity purification failed to allow the identification of naive Affimer targets, an alternative approach was proposed. LC-MS/MS is a complementary approach to Affimer arrays for the differential proteomic analysis of complex proteomes. Therefore, a comparative proteomics analysis using LC-MS/MS was implemented for the identification of proteins differentially expressed in sepsis for directed Affimer generation. Following depletion of highly abundant proteins using antibody spin columns, samples were digested and analysed by LC-MS/MS. Label-free relative quantification revealed a consistent set of proteins differentially expressed in two separate cohorts. Five proteins (CRP, protein S100 A8/A9, cathepsin B, interleukin-1 receptor-like 1 and neutrophil gelatinase-associated lipocalin) were nominated as targets for phage display Affimer production. The work in thesis characterised the phage display Affimers and on-going work is aiming to identify the Affimers that have the highest affinity for their protein targets.

To further enhance this study, several analyses could be carried out. Firstly, validation of the Affimers generated to the five proteins selected from the comparative study could be undertaken to confirm binding of the recombinant target proteins in a complex mixture. This could be carried out using the SulfoLink® resin affinity purification method optimised in this study. As the Affimers were produced using phage display, it is expected that the Affimer have high binding affinities and therefore will enrich for target. Furthermore, it would be interesting to assess whether enrichment of the five proteins from the plasma samples analysed in the comparative study could be achieved using the Adhirons generated.

Although Affimers were originally intended to be used from start to finish in the biomarker pipeline from discovery, validation and clinical detection, they still have value despite proving to be inadequate for biomarker discovery. The five proteins identified from the comparative proteomics analysis of sepsis plasma will be used for an on-going study to develop and identify protein biomarkers of sepsis. The Affimers will be validated using multiplexed ELISA technology to determine if they have improved sensitivity and specificity to the currently available antibodies.

The final Chapter in this thesis investigated the suitability of darcin as a proteins scaffold. To gain a greater insight into the potential of darcin as a protein scaffold, studies examining the effect of loop insertions of protein stability would be necessary. As the loop regions had

been selected, if time had allowed for the expression and stability studies of a darcin scaffold with variable loops, this would have been valuable addition to the chapter.

7.2 Key conclusions

- Affinity purification method development using Affimers is challenging, particularly when using naïve Affimers and when coupled with MS for target identification.
- The selective mutation of Affimers to prevent Lys-C proteolysis may be an appropriate method to overcome challenges of affinity purifications.
- Differential proteomic analysis of sepsis plasma using LC-MS/MS identified a panel of five protein markers of sepsis; CRP, neutrophil-gelatinase associated protein, Protein S100 A8/A9, interleukin-1 receptor-like 1, cathepsin B.
- Preliminary investigations of darcin as a novel protein scaffold are promising.

7.3 Future perspectives

Whilst Affimer technology is promising, the findings in this thesis suggest significant further development of Affimer reagents is needed for them to compete with antibody reagents, particularly for capture and enrichment applications. The high levels of non-specific background, low binding affinities and low specificity would need addressing before the reagents have any real future potential. In other areas of research, Affimers have shown promise[112] however it is unlikely that the scientific community will choose a well characterised antibody reagent over a novel Affimer. A shift in attitudes amongst the research community will be necessary for the wide-spread and favoured use of antibody alternative affinity reagents over antibodies. It should be noted however, that Affimer technology was not developed to replace antibodies, instead to complement the current affinity reagent toolbox available. The revolutionary and fundamental role of antibodies is invaluable in biological research and will undoubtedly remain the key affinity tool in the future. The development of Affimers to protein targets with no affinity reagents currently available may be the ideal application for Affimers. The beneficial features such as speed of production, high stability and extensive potential targets suggest that, with further development, Affimers do have a potential future as renewable affinity reagents in the biological research.

The study of complex proteomes requires the use of enrichment and fractionation strategies in order to overcome the associated dynamic range and complexity challenges. Affinity based methods provide a selective and targeted approach for protein enrichment

thereby reducing complexity issues. This therefore means that enrichment is highly reliant on high quality binding reagents that target the specific protein of interest.

The concept of developing an anti-proteome, in which affinity reagents are available that target a complete proteome, is in principle possible. However, this is only realised if an anti-proteome does not take into account the different proteoforms generated from a single gene. The development of an anti-proteome for the human proteome has made significant progress due to the huge collaborative effort of various consortia and projects including as The Human Protein Atlas Project[304], ProteomeBinders[305], AffinityProteome[306], and The Renewable Protein Binder Working Group[306]. The latest release of Antibodypeida, a database containing validated antibodies against human proteins, currently contains reviews for antibodies that target proteins for 19142 gene products corresponding to 94 % of all human genes[307,308]. However, whilst this is impressive coverage, the dynamic and variable expression of proteins means that full proteome coverage is still a long way from completion. With the estimated number of highly transient PTMs at 200[309], this may be very difficult to achieve. It is doubtful that affinity reagents that target every proteoform will ever exist.

A concern of achieving a complete anti-proteome relates to the availability of target proteins necessary for binder production and validation. Furthermore, to be confident in the quality and specificity of the binder, additional validation is needed for the intended application of the binder, increasing the amount of target protein required. Whilst sequencing of the human genome has theoretically made it possible to generate recombinant proteins to any protein target, certain proteins such as membrane proteins are notoriously difficult to express[310]. Therefore, if recombinant proteins are not available, binder generation is not possible. Fortunately, full length proteins are not always needed for affinity reagents generation as peptides, partial protein domains and protein epitope signature tag (PrEST)[311] can be used. In addition, cell-free methods can be employed for recombinant protein production that readily expresses typically temperamental proteins[312]. Furthermore, whilst an anti-proteome may be a possibility for the human proteome, it is unlikely that an anti-proteome will be developed for other species. The huge cost and time implications associated with binder production do not yet warrant the benefits or demand for anti-proteomes for other organisms or species. However, if improvements to the speed, cost and amount of target protein needed for affinity reagent production were made then, it may become feasible.

Whilst MS methods may be improving in terms of their sensitivity and achieving greater proteome coverage, it is important to note that accuracy and confidence in protein identifications and quantification may still not be satisfactory. Therefore, the availability of affinity reagents for the targeting of every protein within a proteome is still needed despite these advancements in MS technology. The scale and challenges of developing a 'binder' for an entire proteome, although daunting, will ultimately lead to advancements in both basic and translation research. Particularly, the development of affinity reagents that target the complete human proteome will enhance clinical research and investigations improving patient diagnosis, prognosis and treatment.

8. References

1. Wasinger VC, Cordwell SJ, Cerpa-Poljak A, et al. (1995) Progress with gene-product mapping of the mollicutes: Mycoplasma genitalium. *Electrophoresis*. ;16(7):1090-1094.

2. Wilkins MR, Sanchez JC, Gooley AA, et al. (1996) Progress with proteome projects: Why all proteins expressed by a genome should be identified and how to do it. *Biotechnol Genet Eng Rev*. ;13:19-50.

3. Ezkurdia I, Juan D, Rodriguez JM, et al. (2014) Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Hum Mol Genet*. ;23(22):5866-5878.

4. Wilkins MR, Sanchez JC, Williams KL, Hochstrasser DF. (1996) Current challenges and future applications for protein maps and post-translational vector maps in proteome projects. *Electrophoresis*. ;17(5):830-838.

5. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*. ;40(12):1413-1415.

6. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. (2015) A global reference for human genetic variation. *Nature*. ;526(7571):68-74.

7. Meacham CE, Morrison SJ. (2013) Tumor heterogeneity and cancer cell plasticity. *Nature*. ;501(7467):328-337.

8. Lee MV, Topper SE, Hubler SL, et al. (2011) A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. *Molecular Systems Biology*. ;7:514-514.

9. Alberts B, Johnson A, Lewis J, et al. An overview of gene control. In: *Molecular biology of the cell.* 4th edition ed. New York: Garland Science; 2002.

10. Levine M, Tjian R. (2003) Transcription regulation and animal diversity. *Nature*. ;424(6945):147-151.

11. Greenbaum D, Colangelo C, Williams K, Gerstein M. (2003) Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol*. ;4(9):117-117.

12. Pratt JM, Petty J, Riba-Garcia I, et al. (2002) Dynamics of protein turnover, a missing dimension in proteomics. *Mol Cell Proteomics*. ;1(8):579-591.

13. Greenbaum D, Colangelo C, Williams K, Gerstein M. (2003) Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol*. ;4(9):117-117.

14. Nagaraj N, Wisniewski JR, Geiger T, et al. (2011) Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol*. ;7:548.

15. Beck M, Schmidt A, Malmstroem J, et al. (2011) The quantitative proteome of a human cell line. *Mol Syst Biol*. ;7:549.

16. Gholami AM, Hahne H, Wu Z, et al. (2013) Global proteome analysis of the NCI-60 cell line panel. *Cell Rep*. ;4(3):609-620.

17. Zubarev RA. (2013) The challenge of the proteome dynamic range and its implications for in-depth proteomics. *Proteomics*. ;13(5):723-726.

18. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B. (2007) Quantitative mass spectrometry in proteomics: A critical review. *Analytical and Bioanalytical Chemistry*. ;389(4):1017-1031.

19. Anderson NL, Anderson NG. (2002) The human plasma proteome: History, character, and diagnostic prospects. *Mol Cell Proteomics*. ;1(11):845-867.

20. Geiger T, Wehner A, Schaab C, Cox J, Mann M. (2012) Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Molecular & Cellular Proteomics : MCP*. ;11(3):M111.014050.

21. Picotti P, Bodenmiller B, Mueller LN, Domon B, Aebersold R. (2009) Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell*. ;138(4):795-806.

22. Ghaemmaghami S, Huh WK, Bower K, et al. (2003) Global analysis of protein expression in yeast. *Nature*. ;425(6959):737-741.

23. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B. (2007) Quantitative mass spectrometry in proteomics: A critical review. *Analytical and Bioanalytical Chemistry*. ;389(4):1017-1031.

24. O'Farrell PH. (1975) High resolution two-dimensional electrophoresis of proteins. *J Biol Chem*. ;250(10):4007-4021.

25. Fey SJ, Larsen PM. 20012D or not 2D. *Current Opinion in Chemical Biology*. ;5(1):26-33. doi: https://doi-org.liverpool.idm.oclc.org/10.1016/S1367-5931(00)00167-8.

26. Hoving S, Voshol H, van Oostrum J. (2000) Towards high performance two-dimensional gel electrophoresis using ultrazoom gels. *Electrophoresis*. ;21(13):2617-2621.

27. Magdeldin S, Enany S, Yoshida Y, et al. (2014) Basics and recent advances of two dimensional-polyacrylamide gel electrophoresis. *Clinical proteomics*. ;11(1):16.

28. Unlu M, Morgan ME, Minden JS. (1997) Difference gel electrophoresis: A single gel method for detecting changes in protein extracts. *Electrophoresis*. ;18(11):2071-2077.

29. Ramachandran N, Srivastava S, LaBaer J. (2008) Applications of protein microarrays for biomarker discovery. *Proteomics.Clinical applications*. ;2(10-11):1444-1459.

30. Stoevesandt O, Taussig MJ, He M. (2009) Protein microarrays: High-throughput tools for proteomics. *Expert Rev Proteomics*. ;6(2):145-157.

31. Orchekowski R, Hamelinck D, Li L, et al. (2005) Antibody microarray profiling reveals individual and combined serum proteins associated with pancreatic cancer. *Cancer Res*. ;65(23):11193-11202.

32. Davies DH, Liang X, Hernandez JE, et al. (2005) Profiling the humoral immune response to infection by using proteome microarrays: High-throughput vaccine and diagnostic antigen discovery. *Proceedings of the National Academy of Sciences of the United States of America*. ;102(3):547-552.

33. Jeong JS, Jiang L, Albino E, et al. (2012) Rapid identification of monospecific monoclonal antibodies using a human proteome microarray. *Mol Cell Proteomics*. ;11(6):O111.016253.

34. Zhu H, Bilgin M, Bangham R, et al. (2001) Global analysis of protein activities using proteome chips. *Science*. ;293(5537):2101-2105.

35. Zhu H, Snyder M. (2001) Protein arrays and microarrays. *Curr Opin Chem Biol*. ;5(1):40-45.

36. Larman HB, Zhao Z, Laserson U, et al. (2011) Autoantigen discovery with a synthetic human peptidome. *Nat Biotechnol*. ;29(6):535-541.

37. Li QZ, Zhou J, Wandstrat AE, et al. (2007) Protein array autoantibody profiles for insights into systemic lupus erythematosus and incomplete lupus syndromes. *Clin Exp Immunol*. ;147(1):60-70.

38. Catherman AD, Skinner OS, Kelleher NL. (2014) Top down proteomics: Facts and perspectives. *Biochem Biophys Res Commun*. ;445(4):683-693.

39. Smith LM, Kelleher NL. (2013) Proteoform: A single term describing protein complexity. *Nat Meth*. ;10(3):186-187.

40. Tran JC, Zamdborg L, Ahlf DR, et al. (2011) Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature*. ;480(7376):254-258.

41. Hayter JR, Robertson DH, Gaskell SJ, Beynon RJ. (2003) Proteome analysis of intact proteins in complex mixtures. *Mol Cell Proteomics*. ;2(2):85-95.

42. Zhang Y, Fonslow BR, Shan B, Baek MC, Yates JR,3rd. (2013) Protein analysis by shotgun/bottom-up proteomics. *Chem Rev*. ;113(4):2343-2394.

43. Steen H, Mann M. (2004) The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol*. ;5(9):699-711.

44. Olsen JV, Ong S, Mann M. (2004) Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Molecular & Cellular Proteomics*. ;3(6):608-614.

45. Olsen JV, Ong S, Mann M. (2004) Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Molecular & Cellular Proteomics*. ;3(6):608-614.

46. Wenner BR, Lynn BC. 2004Factors that affect ion trap data-dependent MS/MS in proteomics. *Journal of the American Society for Mass Spectrometry*. ;15(2):150-157. doi: https://doi.org/10.1016/j.jasms.2003.10.006.

47. Michalski A, Cox J, Mann M. (2011) More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J Proteome Res*. ;10(4):1785-1793.

48. Hoopmann MR, Finney GL, MacCoss MJ. (2007) High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. *Anal Chem*. ;79(15):5620-5632.

49. Houel S, Abernathy R, Renganathan K, Meyer-Arendt K, Ahn NG, Old WM. (2010) Quantifying the impact of chimera MS/MS spectra on peptide identification in large scale proteomics studies. *Journal of proteome research*. ;9(8):4152-4160.

50. Venable JD, Dong MQ, Wohlschlegel J, Dillin A, Yates JR. (2004) Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat Methods*. ;1(1):39-45.

51. Martins-de-Souza D, Faca VM, Gozzo FC. (2017) DIA is not a new mass spectrometry acquisition method. *Proteomics*. ;17(7):10.1002/pmic.201700017. Epub 2017 Mar 3.

52. Gillet LC, Navarro P, Tate S, et al. (2012) Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: A new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics*. ;11(6):O111.016717.

53. Silva JC, Gorenstein MV, Li GZ, Vissers JP, Geromanos SJ. (2006) Absolute quantification of proteins by LCMSE: A virtue of parallel MS acquisition. *Mol Cell Proteomics*. ;5(1):144-156.

54. Hebert AS, Richards AL, Bailey DJ, et al. (2013) The one hour yeast proteome. *Molecular & Cellular Proteomics : MCP*. ;13(1):339-347.

55. Burniston JG, Connolly J, Kainulainen H, Britton SL, Koch LG. (2014) Label-free profiling of skeletal muscle using high-definition mass spectrometry. *Proteomics*. ;14(20):2339-2344.

56. Doerr A. (2015) DIA mass spectrometry. *Nat Meth*. ;12(1):35-35.

57. Ting YS, Egertson JD, Bollinger JG, et al. (2017) PECAN: Library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nat Methods*. ;14(9):903-908.

58. Holman SW, Sims PF, Eyers CE. (2012) The use of selected reaction monitoring in quantitative proteomics. *Bioanalysis*. ;4(14):1763-1786.

59. Shi T, Su D, Liu T, et al. (2012) Advancing the sensitivity of selected reaction monitoring-based targeted quantitative proteomics. *Proteomics*. ;12(8):1074-1092.

60. Rabilloud T. (2002) Two-dimensional gel electrophoresis in proteomics: Old, old fashioned, but it still climbs up the mountains. *Proteomics*. ;2(1):3-10.

61. Sampath G. (2017) A minimalist approach to protein identification. *bioRxiv*. . doi: 10.1101/187013.

62. Hong P, Koza S, Bouvier ESP. (2012) A review size-exclusion chromatography for the analysis of protein biotherapeutics and their aggregates. *J Liq Chromatogr Rel Technol*. ;35(20):2923-2950.

63. Chen X, Ge Y. (2013) Ultra-high pressure fast size exclusion chromatography for top-down proteomics. *Proteomics*. ;13(17):10.1002/pmic.201200594.

64. Millea KM, Krull IS, Cohen SA, Gebler JC, Berger SJ. (2006) Integration of multidimensional chromatographic protein separations with a combined "top-down" and "bottom-up" proteomic strategy. *J Proteome Res*. ;5(1):135-146.

65. Roth MJ, Parks BA, Ferguson JT, Boyne MT, Kelleher NL. (2008) "Proteotyping"□: Population proteomics of human leukocytes using top down mass spectrometry. *Anal Chem*. ;80(8):2857-2866.

66. Pieper R, Su Q, Gatlin CL, Huang ST, Anderson NL, Steiner S. (2003) Multi-component immunoaffinity subtraction chromatography: An innovative step towards a comprehensive survey of the human plasma proteome. *Proteomics*. ;3(4):422-432.

67. Granger J, Siddiqui J, Copeland S, Remick D. (2005) Albumin depletion of human plasma also removes low abundance proteins including the cytokines. *Proteomics*. ;5(18):4713-4718.

68. Kay R, Barton C, Ratcliffe L, et al. (2008) Enrichment of low molecular weight serum proteins using acetonitrile precipitation for mass spectrometry based proteomic analysis. *Rapid Commun Mass Spectrom*. ;22(20):3255-3260.

69. Mahn A, Ismail M. (2011) Depletion of highly abundant proteins in blood plasma by ammonium sulfate precipitation for 2D-PAGE analysis. *J Chromatogr B Analyt Technol Biomed Life Sci*. ;879(30):3645-3648.

70. Colantonio DA, Dunkinson C, Bovenkamp DE, Van Eyk JE. (2005) Effective removal of albumin from serum. *Proteomics*. ;5(15):3831-3835.

71. Righetti PG, Boschetti E, Lomas L, Citterio A. (2006) Protein equalizer technology : The quest for a democratic proteome. *Proteomics*. ;6(14):3980-3992.

72. Guerrier L, Thulasiraman V, Castagna A, et al. (2006) Reducing protein concentration range of biological samples using solid-phase ligand libraries. *Journal of chromatography.B, Analytical technologies in the biomedical and life sciences*. ;833(1):33-40.

73. Guerrier L, Righetti PG, Boschetti E. (2008) Reduction of dynamic protein concentration range of biological extracts for the discovery of low-abundance proteins by means of hexapeptide ligand library. *Nature protocols*. ;3(5):883-890.

74. Rivers J, Hughes C, McKenna T, et al. (2011) Asymmetric proteome equalization of the skeletal muscle proteome using a combinatorial hexapeptide library. *PLOS ONE*. ;6(12):e28902.

75. Fonslow BR, Carvalho PC, Academia K, et al. (2011) Improvements in proteomic metrics of low abundance proteins through proteome equalization using ProteoMiner prior to MudPIT. *Journal of proteome research*. ;10(8):3690-3700.

76. Thulasiraman V, Lin S, Gheorghiu L, et al. (2005) Reduction of the concentration difference of proteins in biological liquids using a library of combinatorial ligands. *Electrophoresis*. ;26(18):3561-3571.

77. Hsieh EJ, Bereman MS, Durand S, Valaskovic GA, MacCoss MJ. (2012) Effects of column and gradient lengths on peak capacity and peptide identification in nanoflow LC-MS/MS of complex proteomic samples. *J Am Soc Mass Spectrom*. ;24(1):148-153.

78. Stoll DR, Li X, Wang X, Carr PW, Porter SEG, Rutan SC. (2007) Fast, comprehensive two-dimensional liquid chromatography. *Journal of chromatography.A*. ;1168(1-2):3-2.

79. Wolters DA, Washburn MP, Yates JR,3rd. (2001) An automated multidimensional protein identification technology for shotgun proteomics. *Anal Chem*. ;73(23):5683-5690.

80. Washburn MP, Wolters D, Yates JR,3rd. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol*. ;19(3):242-247.

81. Nagaraj N, Alexander Kulak N, Cox J, et al. (2011) System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top orbitrap. *Molecular & Cellular Proteomics : MCP*. ;11(3):M111.013722.

82. Cohen P. (2000) The regulation of protein function by multisite phosphorylation--a 25 year update. *Trends Biochem Sci*. ;25(12):596-601.

83. Apweiler R, Hermjakob H, Sharon N. (1999) On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim Biophys Acta*. ;1473(1):4-8.

84. Villen J, Gygi SP. (2008) The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. *Nature protocols*. ;3(10):1630-1638.

85. Cantin GT, Shock TR, Park SK, Madhani HD, Yates JR. (2007) Optimizing TiO(2)-based phosphopeptide enrichment for automated multidimensional liquid chromatography coupled to tandem mass spectrometry. *Anal Chem*. ;79(12):4666-4673.

86. Urh M, Simpson D, Zhao K. 2009Chapter 26 affinity chromatography: General methods. *Methods in Enzymology*. ;463(Supplement C):417-438. doi: https://doi-org.liverpool.idm.oclc.org/10.1016/S0076-6879(09)63026-3.

87. Mertins P, Qiao JW, Patel J, et al. (2013) Integrated proteomic analysis of post-translational modifications by serial enrichment. *Nature methods*. ;10(7):634-637.

88. Sacco F, Silvestri A, Posca D, et al. (2016) Deep proteomics of breast cancer cells reveals that metformin rewires signaling networks away from a pro-growth state. *Cell Syst*. ;2(3):159-171.

89. Deshmukh AS, Murgia M, Nagaraj N, Treebak JT, Cox J, Mann M. (2015) Deep proteomics of mouse skeletal muscle enables quantitation of protein isoforms, metabolic pathways, and transcription factors. *Mol Cell Proteomics*. ;14(4):841-853.

90. Wang H, Yang Y, Li Y, et al. (2015) Systematic optimization of long gradient chromatography mass spectrometry for deep analysis of brain proteome. *J Proteome Res*. ;14(2):829-838.

91. Kim MS, Pinto SM, Getnet D, et al. (2014) A draft map of the human proteome. *Nature*. ;509(7502):575-581.

92. Wilhelm M, Schlegl J, Hahne H, et al. (2014) Mass-spectrometry-based draft of the human proteome. *Nature*. ;509(7502):582-587.

93. Angeloni S, Ridet JL, Kusy N, et al. (2005) Glycoprofiling with micro-arrays of glycoconjugates and lectins. *Glycobiology*. ;15(1):31-41.

94. Fry S, Afrough B, Leathem A, Dwek M. (2012) Lectin array-based strategies for identifying metastasis-associated changes in glycosylation. *Methods Mol Biol*. ;878:267-272.

95. Fry S,A., Afrough,Babak, Lomax-Browne H,J., Timms J,F., Velentzis L,S., Leathem A,J.C. (2011) Lectin microarray profiling of metastatic breast cancers. *Glycobiology*. ;21(8):1060-1070.

96. Nelson RW, Krone JR, Bieber AL, Williams P. (1995) Mass spectrometric immunoassay. *Anal Chem*. ;67(7):1153-1158.

97. Anderson NL, Anderson NG, Haines LR, Hardie DB, Olafson RW, Pearson TW. (2004) Mass spectrometric quantitation of peptides and proteins using stable isotope standards and capture by anti-peptide antibodies (SISCAPA). *J Proteome Res*. ;3(2):235-244.

98. Sherma ND, Borges CR, Trenchevska O, et al. (2014) Mass spectrometric immunoassay for the qualitative and quantitative analysis of the cytokine macrophage migration inhibitory factor (MIF). *Proteome Science*. ;12(1):52.

99. Gallagher RI, Espina V. (2014) Reverse phase protein arrays: Mapping the path towards personalized medicine. *Molecular Diagnosis & Therapy*. ;18(6):619-630.

100. Conti A, Espina V, Chiechi A, et al. (2014) Mapping protein signal pathway interaction in sarcoma bone metastasis: Linkage between rank, metalloproteinases turnover and growth factor signaling pathways. *Clin Exp Metastasis*. ;31(1):15-24.

101. Kohler G, Milstein C. (1975) Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature*. ;256(5517):495-497.

102. Smith GP. (1985) Filamentous fusion phage: Novel expression vectors that display cloned antigens on the virion surface. *Science*. ;228(4705):1315-1317.

103. Tuerk C, Gold L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*. ;249(4968):505-510.

104. Ellington AD, Szostak JW. (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature*. ;346(6287):818-822.

105. Colas P, Cohen B, Jessen T, Grishina I, McCoy J, Brent R. (1996) Genetic selection of peptide aptamers that recognize and inhibit cyclin-dependent kinase 2. *Nature*. ;380(6574):548-550.

106. Koide A, Bailey CW, Huang X, Koide S. (1998) The fibronectin type III domain as a scaffold for novel binding proteins. *J Mol Biol*. ;284(4):1141-1151.

107. Beste G, Schmidt FS, Stibora T, Skerra A. (1999) Small antibody-like proteins with prescribed ligand specificities derived from the lipocalin fold. *Proc Natl Acad Sci U S A*. ;96(5):1898-1903.

108. Binz HK, Stumpp MT, Forrer P, Amstutz P, Pluckthun A. (2003) Designing repeat proteins: Well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. *J Mol Biol*. ;332(2):489-503.

109. Woodman R, T-H Yeh J, Laurenson S, Ko Ferrigno P. *Design and validation of a neutral protein scaffold for the presentation of peptide aptamers.* Vol 352. ; 2005:1118-33. 10.1016/j.jmb.2005.08.001.

110. Silverman J, Liu Q, Bakker A, et al. (2005) Multivalent avimer proteins evolved by exon shuffling of a family of human receptor domains. *Nat Biotechnol*. ;23(12):1556-1561.

111. Tiede C, Tang AA, Deacon SE, et al. (2014) Adhiron: A stable and versatile peptide display scaffold for molecular recognition applications. *Protein Eng Des Sel*. ;27(5):145-155.

112. Tiede C, Bedford R, Heseltine SJ, et al. (2017) Affimer proteins are versatile and renewable affinity reagents. *Elife*. ;6:10.7554/eLife.24903.

113. Schroeder H,W., Cavacini,Lisa. (2010) Structure and function of immunogloulins. *The Journal of allergy and clinical immunology*. ;125(202):S41-S52.

114. Fleischman JB, Pain RH, Porter RR. (1962) Reduction of gamma-globulins. *Arch Biochem Biophys*. ;Suppl 1:174-180.

115. Kirkham PM, Schroeder HW,Jr. (1994) Antibody structure and the evolution of immunoglobulin V gene segments. *Semin Immunol*. ;6(6):347-360.

116. Janeway CJ, Travers P, Walport M, et al. The structure of a typical antibody molecule. In: *Immunobiology: The immune system in health and disease.* 5th edition ed. New York: Garland Science; 2001. http://www.ncbi.nlm.nih.gov/books/NBK27144/.

117. Leenaars M, Hendriksen CF. (2005) Critical steps in the production of polyclonal and monoclonal antibodies: Evaluation and recommendations. *ILAR J*. ;46(3):269-279.

118. Nelson PN, Reynolds GM, Waldron EE, Ward E, Giannopoulos K, Murray PG. (2000) Demystified â€¦: Monoclonal antibodies. *Molecular Pathology*. ;53(3):111-117.

119. Ouchterlony O. (1949) Antigen-antibody reactions in gels. *Acta Pathol Microbiol Scand*. ;26(4):507-515.

120. Yalow RS, Berson SA. (1959) Assay of plasma insulin in human subjects by immunological methods. *Nature*. ;184 (Suppl 21)(Suppl 21):1648-1649.

121. Nakane PK, Pierce GB,Jr. (1967) Enzyme-labeled antibodies for the light and electron microscopic localization of tissue antigens. *J Cell Biol*. ;33(2):307-318.

122. Wide L, Porath J. 1966Radioimmunoassay of proteins with the use of sephadex-coupled antibodies. *Biochimica et Biophysica Acta (BBA) - General Subjects*. ;130(1):257-260. doi: https://doi.org/10.1016/0304-4165(66)90032-8 ".

123. Shepard HM, Phillips GL, D Thanos C, Feldmann M. (2017) Developments in therapy with monoclonal antibodies and related proteins. *Clin Med (Lond)*. ;17(3):220-232.

124. Bradbury A, Pluckthun A. (2015) Reproducibility: Standardize antibodies used in research. *Nature*. ;518(7537):27-29.

125. Helsby MA, Fenn JR, Chalmers AD. (2013) Reporting research antibody use: How to increase experimental reproducibility. *F1000Res*. ;2:10.12688/f1000research.2-153.v2.

126. Helsby MA, Leader PM, Fenn JR, et al. (2014) CiteAb: A searchable antibody database that ranks antibodies by the number of times they have been cited. *BMC Cell Biol*. ;15:6-2121-15-6.

127. Roncador G, Engel P, Maestre L, et al. (2015) The european antibody network's practical guide to finding and validating suitable antibodies for research. *MAbs*. ;8(1):27-36.

128. McCafferty J, Griffiths AD, Winter G, Chiswell DJ. (1990) Phage antibodies: Filamentous phage displaying antibody variable domains. *Nature*. ;348(6301):552-554.

129. Frenzel A, Kugler J, Wilke S, Schirrmann T, Hust M. (2014) Construction of human antibody gene libraries and selection of antibodies by phage display. *Methods Mol Biol*. ;1060:215-243.

130. Lloyd C, Lowe D, Edwards B, et al. (2009) Modelling the human immune response: Performance of a 1011 human antibody repertoire against a broad panel of therapeutically relevant antigens. *Protein Eng Des Sel*. ;22(3):159-168.

131. Foote J, Eisen HN. (2000) Breaking the affinity ceiling for antibodies and T cell receptors. *Proc Natl Acad Sci U S A*. ;97(20):10679-10681.

132. Frenzel A, Hust M, Schirrmann T. (2013) Expression of recombinant antibodies. *Frontiers in Immunology*. ;4:217.

133. Hamers-Casterman C, Atarhouch T, Muyldermans S, et al. (1993) Naturally occurring antibodies devoid of light chains. *Nature*. ;363(6428):446-448.

134. Greenberg AS, Avila D, Hughes M, Hughes A, McKinney EC, Flajnik MF. (1995) A new antigen receptor gene family that undergoes rearrangement and extensive somatic diversification in sharks. *Nature*. ;374(6518):168-173.

135. Hassanzadeh-Ghassabeh G, Devoogdt N, De Pauw P, Vincke C, Muyldermans S. (2013) Nanobodies and their potential applications. *Nanomedicine (Lond)*. ;8(6):1013-1026.

136. De Genst E, Silence K, Decanniere K, et al. (2006) Molecular basis for the preferential cleft recognition by dromedary heavy-chain antibodies. *Proceedings of the National Academy of Sciences of the United States of America*. ;103(12):4586-4591.

137. Stijlemans B, Conrath K, Cortez-Retamozo V, et al. (2004) Efficient targeting of conserved cryptic epitopes of infectious agents by single domain antibodies: AFRICAN TRYPANOSOMES AS PARADIGM. *Journal of Biological Chemistry*. ;279(2):1256-1261.

138. Sundberg EJ, Mariuzza RA. (2002) Molecular recognition in antibody-antigen complexes. *Adv Protein Chem*. ;61:119-160.

139. van der Linden RH, Frenken LG, de Geus B, et al. (1999) Comparison of physical chemical properties of llama VHH antibody fragments and mouse monoclonal antibodies. *Biochim Biophys Acta*. ;1431(1):37-46.

140. Stoltenburg R, Reinemann C, Strehlitz B. (2007) SELEX--a (r)evolutionary method to generate high-affinity nucleic acid ligands. *Biomol Eng*. ;24(4):381-403.

141. Gold L, Polisky B, Uhlenbeck O, Yarus M. (1995) Diversity of oligonucleotide functions. *Annu Rev Biochem*. ;64:763-797.

142. Gragoudas ES, Adamis AP, Cunningham ET,Jr, Feinsod M, Guyer DR, VEGF Inhibition Study in Ocular Neovascularization Clinical Trial Group. (2004) Pegaptanib for neovascular age-related macular degeneration. *N Engl J Med*. ;351(27):2805-2816.

143. Ohuchi S. (2012) Cell-SELEX technology. *BioResearch Open Access*. ;1(6):265-272.

144. Carlson B. (2007) Aptamers : The new frontier in drug development? *Biotechnol Healthc*. ;4(2):31-36.

145. Colas P, Cohen B, Jessen T, Grishina I, McCoy J, Brent R. (1996) Genetic selection of peptide aptamers that recognize and inhibit cyclin-dependent kinase 2. *Nature*. ;380(6574):548-550.

146. LaVallie ER, DiBlasio EA, Kovacic S, Grant KL, Schendel PF, McCoy JM. (1993) A thioredoxin gene fusion expression system that circumvents inclusion body formation in the E. coli cytoplasm. *Biotechnology (N Y)*. ;11(2):187-193.

147. Binz HK, Amstutz P, Pluckthun A. (2005) Engineering novel binding proteins from nonimmunoglobulin domains. *Nat Biotechnol*. ;23(10):1257-1268.

148. Ladner RC. (1995) Constrained peptides as binding entities. *Trends Biotechnol*. ;13(10):426-430.

149. Smith GP, Petrenko VA. (1997) Phage display. *Chem Rev*. ;97(2):391-410.

150. Binz HK, Amstutz P, Pluckthun A. (2005) Engineering novel binding proteins from nonimmunoglobulin domains. *Nat Biotechnol*. ;23(10):1257-1268.

151. Nygren PA, Skerra A. (2004) Binding proteins from alternative scaffolds. *J Immunol Methods*. ;290(1-2):3-28.

152. Skerra A. (2003) Imitating the humoral immune response. *Curr Opin Chem Biol*. ;7(6):683-693.

153. Stadler LK, Hoffmann T, Tomlinson DC, et al. (2011) Structure-function studies of an engineered scaffold protein derived from stefin A. II: Development and applications of the SQT variant. *Protein Eng Des Sel*. ;24(9):751-763.

154. Schlehuber S, Skerra A. (2005) Lipocalins in drug discovery: From natural ligand-binding proteins to "anticalins". *Drug Discov Today*. ;10(1):23-33.

155. Koide A, Bailey CW, Huang X, Koide S. (1998) The fibronectin type III domain as a scaffold for novel binding proteins. *J Mol Biol*. ;284(4):1141-1151.

156. Nord K, Nilsson J, Nilsson B, Uhlen M, Nygren PA. (1995) A combinatorial library of an alpha-helical bacterial receptor domain. *Protein Eng*. ;8(6):601-608.

157. Nord K, Gunneriusson E, Ringdahl J, Stahl S, Uhlen M, Nygren PA. (1997) Binding proteins selected from combinatorial libraries of an alpha-helical bacterial receptor domain. *Nat Biotechnol*. ;15(8):772-777.

158. Dennis MS, Herzka A, Lazarus RA. (1995) Potent and selective kunitz domain inhibitors of plasma kallikrein designed by phage display. *J Biol Chem*. ;270(43):25411-25417.

159. Ebersbach H, Fiedler E, Scheuermann T, et al. 2007Affilin–Novel binding molecules based on human γ-B-crystallin, an all β-sheet protein. *Journal of Molecular Biology*. ;372(1):172-185. doi: https://doi.org/10.1016/j.jmb.2007.06.045.

160. Nilsson B, Moks T, Jansson B, et al. (1987) A synthetic IgG-binding domain based on staphylococcal protein A. *Protein Eng*. ;1(2):107-113.

161. Wikman M, Rowcliffe E, Friedman M, et al. (2006) Selection and characterization of an HIV-1 gp120-binding affibody ligand. *Biotechnol Appl Biochem*. ;45(Pt 2):93-105.

162. Nord K, Gunneriusson E, Uhlen M, Nygren PA. (2000) Ligands selected from combinatorial libraries of protein A for use in affinity capture of apolipoprotein A-1M and taq DNA polymerase. *J Biotechnol*. ;80(1):45-54.

163. Gronwall C, Sjoberg A, Ramstrom M, et al. (2007) Affibody-mediated transferrin depletion for proteomics applications. *Biotechnol J*. ;2(11):1389-1398.

164. Baum RP, Prasad V, Muller D, et al. (2010) Molecular imaging of HER2-expressing malignant tumors in breast cancer patients using synthetic 111In- or 68Ga-labeled affibody molecules. *J Nucl Med*. ;51(6):892-897.

165. Cheng Z, De Jesus OP, Kramer DJ, et al. (2010) 64Cu-labeled affibody molecules for imaging of HER2 expressing tumors. *Mol Imaging Biol*. ;12(3):316-324.

166. Li J, Mahajan A, Tsai MD. (2006) Ankyrin repeat: A unique motif mediating protein-protein interactions. *Biochemistry*. ;45(51):15168-15178.

167. Binz HK, Amstutz P, Kohl A, et al. (2004) High-affinity binders selected from designed ankyrin repeat protein libraries. *Nat Biotechnol*. ;22(5):575-582.

168. Zahnd C, Kawe M, Stumpp MT, et al. (2010) Efficient tumor targeting with high-affinity designed ankyrin repeat proteins: Effects of affinity and molecular size. *Cancer Res*. ;70(4):1595-1605.

169. Martin-Killias P, Stefan N, Rothschild S, Pluckthun A, Zangemeister-Wittke U. (2011) A novel fusion toxin derived from an EpCAM-specific designed ankyrin repeat protein has potent antitumor activity. *Clin Cancer Res*. ;17(1):100-110.

170. Schilling J, Schoppe J, Pluckthun A. (2014) From DARPins to LoopDARPins: Novel LoopDARPin design allows the selection of low picomolar binders in a single round of ribosome display. *J Mol Biol*. ;426(3):691-721.

171. Gebauer M, Skerra A. (2012) Anticalins small engineered binding proteins based on the lipocalin scaffold. *Methods Enzymol*. ;503:157-188.

172. Skerra A. (2000) Lipocalins as a scaffold. *Biochim Biophys Acta*. ;1482(1-2):337-350.

173. Flower DR. (1996) The lipocalin protein family: Structure and function. *Biochemical Journal*. ;Aug 15(318 (Pt 1)):1-14.

174. Schonfeld D, Matschiner G, Chatwell L, et al. (2009) An engineered lipocalin specific for CTLA-4 reveals a combining site with structural and conformational features similar to antibodies. *Proc Natl Acad Sci U S A*. ;106(20):8198-8203.

175. Richter A, Skerra A. (2017) Anticalins directed against vascular endothelial growth factor receptor 3 (VEGFR-3) with picomolar affinities show potential for medical therapy and in vivo imaging. *Biol Chem*. ;398(1):39-55.

176. Litvinovich SV, Ingham KC. (1995) Interactions between type III domains in the 110 kDa cell-binding fragment of fibronectin. *J Mol Biol*. ;248(3):611-626.

177. Lipovsek D. (2011) Adnectins: Engineered target-binding protein therapeutics. *Protein Eng Des Sel*. ;24(1-2):3-9.

178. Xu L, Aha P, Gu K, et al. (2002) Directed evolution of high-affinity antibody mimics using mRNA display. *Chem Biol*. ;9(8):933-942.

179. Ramamurthy V, Krystek S, Bush A, et al. (2012) Structures of adnectin/protein complexes reveal an expanded binding footprint. *Structure*. ;20(2):259-269.

180. Le Nguyen D, Heitz A, Chiche L, et al. (1990) Molecular recognition between serine proteases and new bioactive microproteins with a knotted structure. *Biochimie*. ;72(6-7):431-435.

181. Gracy J, Le-Nguyen D, Gelly J, Kaas Q, Heitz A, Chiche L. (2008) KNOTTIN: The knottin or inhibitor cystine knot scaffold in 2007. *Nucleic Acids Res*. ;36:D314-D319.

182. Gelly JC, Gracy J, Kaas Q, Le-Nguyen D, Heitz A, Chiche L. (2004) The KNOTTIN website and database: A new information system dedicated to the knottin scaffold. *Nucleic Acids Res*. ;32(Database issue):D156-9.

183. Kimura RH, Cheng Z, Gambhir SS, Cochran JR. (2009) Engineered knottin peptides: A new class of agents for imaging integrin expression in living subjects. *Cancer Res*. ;69(6):2435-2442.

184. Christmann A, Walter K, Wentzel A, Kratzner R, Kolmar H. (1999) The cystine knot of a squash-type protease inhibitor as a structural scaffold for escherichia coli cell surface display of conformationally constrained peptides. *Protein Eng*. ;12(9):797-806.

185. Wentzel A, Christmann A, Kratzner R, Kolmar H. (1999) Sequence requirements of the GPNG beta-turn of the ecballium elaterium trypsin inhibitor II explored by combinatorial library screening. *J Biol Chem*. ;274(30):21037-21043.

186. Baggio R, Burgstaller P, Hale SP, et al. (2002) Identification of epitope-like consensus motifs using mRNA display. *J Mol Recognit*. ;15(3):126-134.

187. Smith GP, Patel SU, Windass JD, Thornton JM, Winter G, Griffiths AD. (1998) Small binding proteins selected from a combinatorial repertoire of knottins displayed on phage. *J Mol Biol*. ;277(2):317-332.

188. Lehtio J, Teeri TT, Nygren PA. (2000) Alpha-amylase inhibitors selected from a combinatorial library of a cellulose binding domain scaffold. *Proteins*. ;41(3):316-322.

189. Silverman AP, Levin AM, Lahti JL, Cochran JR. (2008) Engineered cystine-knot peptides that bind $\alpha_v\beta_3$ integrin with antibody-like affinities. *J Mol Biol*. ;385(4):1064-1075.

190. Hoffmann T, Stadler LK, Busby M, et al. (2010) Structure-function studies of an engineered scaffold protein derived from stefin A. I: Development of the SQM variant. *Protein Eng Des Sel*. ;23(5):403-413.

191. Schrodinger L. 2015The PyMOL molecular graphics system, Version~1.8. .

192. Song Q, Stadler LK, Peng J, Ko Ferrigno P. (2011) Peptide aptamer microarrays: Bridging the bio-detector interface. *Faraday Discuss*. ;149:79-92; discussion 137-57.

193. Rifai N, Gillette MA, Carr SA. (2006) Protein biomarker discovery and validation: The long and uncertain path to clinical utility. *Nat Biotech*. ;24(8):971-983.

194. Mandel P, Metais P. (1948) Not available. *C R Seances Soc Biol Fil*. ;142(3-4):241-243.

195. Straw S, Ferrigno PK, Song Q, Tomlinson D, Galdo FD. (2013) Proof of concept study to identify candidate biomarkers of fibrosis using high throughput peptide aptamer microarray and validate by enzyme linked immunosorbant assay. *Journal of Biomedical Science and Engineering*. ;Vol.06No.08:11.

196. Trinick R, Johnston N, Dalzell AM, McNamara PS. (2012) Reflux aspiration in children with neurodisability--a significant problem, but can we measure it? *J Pediatr Surg*. ;47(2):291-298.

197. Phelan MM, McLean L, Armstrong SD, Hurst JL, Beynon RJ, Lian LY. (2014) The structure, stability and pheromone binding of the male mouse protein sex pheromone darcin. *PLoS One*. ;9(10):e108415.

198. Yamashita M, Fenn J. (1984) Electrospray ion source. another variation on the free-jet theme. *J Phys Chem*. ;88(20):4451-4459.

199. Laemmli UK. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*. ;227(5259):680-685.

200. Liu G, Zhao Y, Angeles A, Hamuro LL, Arnold ME, Shen JX. (2014) A novel and cost effective method of removing excess albumin from plasma/serum samples and its impacts on LC-MS/MS bioanalysis of therapeutic proteins. *Anal Chem*. ;86(16):8336-8343.

201. Link AJ, LaBaer J. (2011) Trichloroacetic acid (TCA) precipitation of proteins. *Cold Spring Harb Protoc*. ;2011(8):993-994.

202. Porath J, Carlsson J, Olsson I, Belfrage G. (1975) Metal chelate affinity chromatography, a new approach to protein fractionation. *Nature*. ;258(5536):598-599.

203. Hochuli E, Dobeli H, Schacher A. (1987) New metal chelate adsorbent selective for proteins and peptides containing neighbouring histidine residues. *J Chromatogr*. ;411:177-184.

204. Cuatrecasas P. (1970) Protein purification by affinity chromatography: DERIVATIZATIONS OF AGAROSE AND POLYACRYLAMIDE BEADS. *Journal of Biological Chemistry*. ;245(12):3059-3065.

205. Diamandis EP, Christopoulos TK. (1991) The biotin-(strept)avidin system: Principles and applications in biotechnology. *Clin Chem*. ;37(5):625-636.

206. Livnah O, Bayer EA, Wilchek M, Sussman JL. (1993) Three-dimensional structures of avidin and the avidin-biotin complex. *Proc Natl Acad Sci U S A*. ;90(11):5076-5080.

207. Groff K, Brown J, Clippinger AJ. (2015) Modern affinity reagents: Recombinant antibodies and aptamers. *Biotechnol Adv*. ;33(8):1787-1798.

208. Davis JJ, Tkac J, Humphreys R, Buxton AT, Lee TA, Ko Ferrigno P. (2009) Peptide aptamers in label-free protein detection: 2. chemical optimization and detection of distinct protein isoforms. *Anal Chem*. ;81(9):3314-3320.

209. Gauthier MS, Perusse JR, Awan Z, et al. (2015) A semi-automated mass spectrometric immunoassay coupled to selected reaction monitoring (MSIA-SRM) reveals novel relationships between circulating PCSK9 and metabolic phenotypes in patient cohorts. *Methods*. ;81:66-73.

210. Kiernan UA, Phillips DA, Trenchevska O, Nedelkov D. (2011) Quantitative mass spectrometry evaluation of human retinol binding protein 4 and related variants. *PLoS One*. ;6(3):e17282.

211. Liu Y, Eisenberg D. (2002) 3D domain swapping: As domains continue to swap. *Protein Sci*. ;11(6):1285-1299.

212. Kardar GA, Oraei M, Shahsavani M, et al. (2012) Reference intervals for serum immunoglobulins IgG, IgA, IgM and complements C3 and C4 in iranian healthy children. *Iranian Journal of Public Health*. ;41(7):59-63.

213. Shen Z, Want EJ, Chen W, et al. (2006) Sepsis plasma protein profiling with immunodepletion, three-dimensional liquid chromatography tandem mass spectrometry, and spectrum counting. *J Proteome Res*. ;5(11):3154-3160.

214. Zhang J, Xin L, Shan B, et al. (2012) PEAKS DB: De novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol Cell Proteomics*. ;11(4):M111.010587.

215. Mellacheruvu D, Wright Z, Couzens AL, et al. (2013) The CRAPome: A contaminant repository for affinity purification mass spectrometry data. *Nature methods*. ;10(8):730-736.

216. Trinkle-Mulcahy L, Boulon S, Lam YW, et al. (2008) Identifying specific protein interaction partners using quantitative mass spectrometry and bead proteomes. *J Cell Biol*. ;183(2):223-239.

217. Rees JS, Lilley KS, Jackson AP. (2014) The chicken B-cell line DT40 proteome, beadome and interactomes. *Data in Brief*. ;3:29-33.

218. Tu H, Sun L, Dong X, et al. (2015) Temporal changes in serum biomarkers and risk for progression of gastric precancerous lesions: A longitudinal study. *Int J Cancer*. ;136(2):425-434.

219. Huang YK, Yu JC, Kang WM, et al. (2015) Significance of serum pepsinogens as a biomarker for gastric cancer and atrophic gastritis screening: A systematic review and meta-analysis. *PLoS One*. ;10(11):e0142080.

220. Habibullah CM, Mujahid Ali M, Ishaq M, Prasad R, Pratap B, Saleem Y. (1984) Study of duodenal ulcer disease in 100 families using total serum pepsinogen as a genetic marker. *Gut*. ;25(12):1380-1383.

221. Maury CP,J., RÃ¤sÃ¤nen,V.E.S.A., Teppo,Anna-Maija, Helve,Tapani, Wegelius,Otto. Serum pepsinogen i in rheumatic diseases. reduced levels in sjÃ¶gren's syndrome. .

222. Samuels TL, Johnston N. (2010) Pepsin as a marker of extraesophageal reflux. *Ann Otol Rhinol Laryngol*. ;119(3):203-208.

223. Kageyama T. (2002) Pepsinogens, progastricsins, and prochymosins: Structure, function, evolution, and development. *Cell Mol Life Sci*. ;59(2):288-306.

224. Gremel G, Wanders A, Cedernaes J, et al. (2015) The human gastrointestinal tract-specific transcriptome and proteome as defined by RNA sequencing and antibody-based profiling. *J Gastroenterol*. ;50(1):46-57.

225. Ponten F, Jirstrom K, Uhlen M. (2008) The human protein atlas--a tool for pathology. *J Pathol*. ;216(4):387-393.

226. Richter C, Tanaka T, Yada RY. (1998) Mechanism of activation of the gastric aspartic proteinases: Pepsinogen, progastricsin and prochymosin. *Biochem J*. ;335 (Pt 3)(Pt 3):481-490.

227. James MN, Sielecki AR. (1986) Molecular structure of an aspartic proteinase zymogen, porcine pepsinogen, at 1.8 A resolution. *Nature*. ;319(6048):33-38.

228. Neubert H, Gale J, Muirhead D. (2010) Online high-flow peptide immunoaffinity enrichment and nanoflow LC-MS/MS: Assay development for total salivary pepsin/pepsinogen. *Clin Chem*. ;56(9):1413-1423.

229. Pelay-Gimeno M, Glas A, Koch O, Grossmann TN. (2015) Structure-based design of inhibitors of protein-protein interactions: Mimicking peptide binding epitopes. *Angew Chem Int Ed Engl*. ;54(31):8896-8927.

230. Hirel PH, Schmitter MJ, Dessen P, Fayat G, Blanquet S. (1989) Extent of N-terminal methionine excision from escherichia coli proteins is governed by the side-chain length of the penultimate amino acid. *Proceedings of the National Academy of Sciences*. ;86(21):8247-8251.

231. Vagenende V, Yap MG, Trout BL. (2009) Mechanisms of protein stabilization and prevention of protein aggregation by glycerol. *Biochemistry*. ;48(46):11084-11096.

232. Han JC, Han GY. (1994) A procedure for quantitative determination of tris(2-carboxyethyl)phosphine, an odorless reducing agent more stable and effective than dithiothreitol. *Anal Biochem*. ;220(1):5-10.

233. Jurczak P, Groves P, Szymanska A, Rodziewicz-Motowidlo S. (2016) Human cystatin C monomer, dimer, oligomer, and amyloid structures are related to health and disease. *FEBS Lett*. ;590(23):4192-4201.

234. Beynon RJ. (2005) A simple tool for drawing proteolytic peptide maps. *Bioinformatics*. ;21(5):674-675.

235. Lackie JM. C. In: Lackie JM, ed. *The dictionary of cell & molecular biology (fifth edition).* Boston: Academic Press; 2013:88-161. https://doi.org/10.1016/B978-0-12-384931-1.00003-9.

236. Turunen O, Vuorio M, Fenel F, Leisola M. (2002) Engineering of multiple arginines into the ser/thr surface of trichoderma reesei endo-1,4-beta-xylanase II increases the thermotolerance and shifts the pH optimum towards alkaline pH. *Protein Eng*. ;15(2):141-145.

237. Donald JE, Kulp DW, DeGrado WF. (2011) Salt bridges: Geometrically specific, designable interactions. *Proteins*. ;79(3):898-915.

238. Sokalingam S, Raghunathan G, Soundrarajan N, Lee S. (2012) A study on the effect of surface lysine to arginine mutagenesis on protein stability and structure using green fluorescent protein. *PLoS ONE*. ;7(7):e40410.

239. Kibe,Savitri, Adams,Kate, Barlow,Gavin. (2011) Diagnostic and prognostic biomarkers of sepsis in critical care. *Journal of Antimicrobial Chemotherapy*. ;66(suppl_2):ii33-ii40.

240. Ward MJ, Self WH, Singer A, Lazar D, Pines JM. (2016) Cost-effectiveness analysis of early point-of-care lactate testing in the emergency department. *J Crit Care*. ;36:69-75.

241. Joshi A, Perin DP, Gehle A, Nsiah-Kumi PA. (2013) Feasibility of using C-reactive protein for point-of-care testing. *Technol Health Care*. ;21(3):233-240.

242. Oshita H, Sakurai J, Kamitsuna M. (2010) Semi-quantitative procalcitonin test for the diagnosis of bacterial infection: Clinical use and experience in japan. *J Microbiol Immunol Infect*. ;43(3):222-227.

243. Buchegger P, Sauer U, Toth-Szekely H, Preininger C. (2012) Miniaturized protein microarray with internal calibration as point-of-care device for diagnosis of neonatal sepsis. *Sensors (Basel)*. ;12(2):1494-1508.

244. Kemmler M, Sauer U, Schleicher E, Preininger C, Brandenburg A. Biochip point-of-care device for sepsis diagnostics. *Sensors and Actuators B: Chemical*. ;192:205-215. doi: http://dx.doi.org.liverpool.idm.oclc.org/10.1016/j.snb.2013.10.003.

245. Johnson A, Song Q, Ko Ferrigno P, Bueno PR, Davis JJ. (2012) Sensitive affimer and antibody based impedimetric label-free assays for C-reactive protein. *Anal Chem*. ;84(15):6553-6560.

246. Omenn GS. (2007) THE HUPO human plasma proteome project. *Proteomics Clin Appl*. ;1(8):769-779.

247. Tuck MK, Chan DW, Chia D, et al. (2009) Standard operating procedures for serum and plasma collection: Early detection research network consensus statement standard operating procedure integration working group. *Journal of proteome research*. ;8(1):113-117.

248. Reinhart K, Bauer M, Riedemann NC, Hartog CS. (2012) New approaches to sepsis: Molecular diagnostics and biomarkers. *Clin Microbiol Rev*. ;25(4):609-634.

249. DeCoux A, Tian Y, DeLeon-Pennell K, et al. (2015) Plasma glycoproteomics reveals sepsis outcomes linked to distinct proteins in common pathways. *Crit Care Med*. ;43(10):2049-2058.

250. Liu D, Su L, Han G, Yan P, Xie L. (2015) Prognostic value of procalcitonin in adult patients with sepsis: A systematic review and meta-analysis. *PLoS One*. ;10(6):. doi:10.1371/journal.pone.0129450.

251. Ballinger A. *Essentials of kumar and clark's clinical medicine E-book.* Vol 5th ed. Edinburgh: Saunders; 2011. https://liverpool.idm.oclc.org/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edsebk&AN=973465&site=eds-live&scope=site.

252. Quinlan GJ, Martin GS, Evans TW. (2005) Albumin: Biochemical properties and therapeutic potential. *Hepatology*. ;41(6):1211-1219.

253. Wheatley JB. 1992Multiple ligand applications in high-performance immunoaffinity chromatography. *Journal of Chromatography A*. ;603(1):273-278. doi: http://dx.doi.org/10.1016/0021-9673(92)85371-Y.

254. COHN EJ, HUGHES WL,Jr, WEARE JH. (1947) Preparation and properties of serum and plasma proteins; crystallization of serum albumins from ethanol water mixtures. *J Am Chem Soc*. ;69(7):1753-1761.

255. Chen YY, Lin SY, Yeh YY, et al. (2005) A modified protein precipitation procedure for efficient removal of albumin from serum. *Electrophoresis*. ;26(11):2117-2127.

256. Ziegler J, Vogt T, Miersch O, Strack D. (1997) Concentration of dilute protein solutions prior to sodium dodecyl sulfate-polyacrylamide gel electrophoresis. *Anal Biochem*. ;250(2):257-260.

257. Sauve DM, Ho DT, Roberge M. (1995) Concentration of dilute protein for gel electrophoresis. *Anal Biochem*. ;226(2):382-383.

258. Ludwig KR, Hummon AB. (2017) Mass spectrometry for the discovery of biomarkers of sepsis. *Mol Biosyst*. ;13(4):648-664.

259. O'Donnell LC, Druhan LJ, Avalos BR. (2002) Molecular characterization and expression analysis of leucine-rich alpha2-glycoprotein, a novel marker of granulocytic differentiation. *J Leukoc Biol*. ;72(3):478-485.

260. Shirai R, Hirano F, Ohkura N, Ikeda K, Inoue S. (2009) Up-regulation of the expression of leucine-rich alpha(2)-glycoprotein in hepatocytes by the mediators of acute-phase response. *Biochem Biophys Res Commun*. ;382(4):776-779.

261. Bini L, Magi B, Marzocchi B, et al. (1996) Two-dimensional electrophoretic patterns of acute-phase human serum proteins in the course of bacterial and viral diseases. *Electrophoresis*. ;17(3):612-616.

262. Kharbanda AB, Rai AJ, Cosme Y, Liu K, Dayan PS. (2012) Novel serum and urine markers for pediatric appendicitis. *Acad Emerg Med*. ;19(1):56-62.

263. Yao L, Liu Z, Zhu J, Li B, Chai C, Tian Y. (2016) Higher serum level of myoglobin could predict more severity and poor outcome for patients with sepsis. *Am J Emerg Med*. ;34(6):948-952.

264. Mi,Huaiyu, Huang,Xiaosong, Muruganujan,Anushya, et al. (2016) PANTHER version 11: Expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research*. ;45(D1):D183-D189.

265. Huang da W, Sherman BT, Lempicki RA. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. ;4(1):44-57.

266. Huang da W, Sherman BT, Lempicki RA. (2009) Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. ;37(1):1-13.

267. Newton K, Dixit VM. (2012) Signaling in innate immunity and inflammation. *Cold Spring Harb Perspect Biol*. ;4(3):10.1101/cshperspect.a006049.

268. Laudes IJ, Chu JC, Sikranth S, et al. (2002) Anti-c5a ameliorates coagulation/fibrinolytic protein changes in a rat model of sepsis. *Am J Pathol*. ;160(5):1867-1875.

269. Guo RF, Ward PA. (2005) Role of C5a in inflammatory responses. *Annu Rev Immunol*. ;23:821-852.

270. Liu NQ, Stingl C, Look MP, et al. (2014) Comparative proteome analysis revealing an 11-protein signature for aggressive triple-negative breast cancer. *J Natl Cancer Inst*. ;106(2):djt376.

271. Welberry Smith MP, Zougman A, Cairns DA, et al. (2013) Serum aminoacylase-1 is a novel biomarker with potential prognostic utility for long-term outcome in patients with delayed graft function following renal transplantation. *Kidney Int*. ;84(6):1214-1225.

272. Scheltema RA, Hauschild JP, Lange O, et al. (2014) The Q exactive HF, a benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field orbitrap analyzer. *Mol Cell Proteomics*. ;13(12):3698-3708.

273. Mold C, Gresham HD, Du Clos TW. (2001) Serum amyloid P component and C-reactive protein mediate phagocytosis through murine fc gamma rs. *J Immunol*. ;166(2):1200-1205.

274. Mold C, Baca R, Du Clos TW. (2002) Serum amyloid P component and C-reactive protein opsonize apoptotic cells for phagocytosis through fcgamma receptors. *J Autoimmun*. ;19(3):147-154.

275. Flood RG, Badik J, Aronoff SC. (2008) The utility of serum C-reactive protein in differentiating bacterial from nonbacterial pneumonia in children: A meta-analysis of 1230 children. *Pediatr Infect Dis J*. ;27(2):95-99.

276. Hengst JM. (2003) The role of C-reactive protein in the evaluation and management of infants with suspected sepsis. *Adv Neonatal Care*. ;3(1):3-13.

277. Turk V, Stoka V, Vasiljeva O, et al. (2012) Cysteine cathepsins: From structure, function and regulation to new frontiers. *Biochim Biophys Acta*. ;1824(1):68-88.

278. Assfalg-Machleidt I, Jochum M, Klaubert W, Inthorn D, Machleidt W. (1988) Enzymatically active cathepsin B dissociating from its inhibitor complexes is elevated in blood plasma of patients with septic shock and some malignant tumors. *Biol Chem Hoppe Seyler*. ;369 Suppl:263-269.

279. Klaude M, Mori M, Tjader I, Gustafsson T, Wernerman J, Rooyackers O. (2012) Protein metabolism and gene expression in skeletal muscle of critically ill patients with sepsis. *Clin Sci (Lond)*. ;122(3):133-142.

280. Ruff RL, Secrist D. (1984) Inhibitors of prostaglandin synthesis or cathepsin B prevent muscle wasting due to sepsis in the rat. *J Clin Invest*. ;73(5):1483-1486.

281. Hummel RP,3rd, James JH, Warner BW, Hasselgren PO, Fischer JE. (1988) Evidence that cathepsin B contributes to skeletal muscle protein breakdown during sepsis. *Arch Surg*. ;123(2):221-224.

282. Hoegen T, Tremel N, Klein M, et al. (2011) The NLRP3 inflammasome contributes to brain injury in pneumococcal meningitis and is activated through ATP-dependent lysosomal cathepsin B release. *J Immunol*. ;187(10):5440-5451.

283. van Zoelen MA, Vogl T, Foell D, et al. (2009) Expression and role of myeloid-related protein-14 in clinical and experimental sepsis. *Am J Respir Crit Care Med*. ;180(11):1098-1106.

284. Achouiti A, Vogl T, Urban CF, et al. (2012) Myeloid-related protein-14 contributes to protective immunity in gram-negative pneumonia derived sepsis. *PLoS Pathog*. ;8(10):e1002987.

285. Vogl T, Tenbrock K, Ludwig S, et al. (2007) Mrp8 and Mrp14 are endogenous activators of toll-like receptor 4, promoting lethal, endotoxin-induced shock. *Nat Med*. ;13(9):1042-1049.

286. Payen DLGD, Lukaszewicz AC. 2010Methods and kits for the rapid determination of patients at high risk of death during severe sepsis and septic shock. .

287. Wang B, Chen G, Zhang J, Xue J, Cao Y, Wu Y. (2015) Increased neutrophil gelatinase-associated lipocalin is associated with mortality and multiple organ dysfunction syndrome in severe sepsis and septic shock. *Shock*. ;44(3):234-238.

288. Pynn JM, Parravicini E, Saiman L, Bateman DA, Barasch JM, Lorenz JM. (2015) Urinary neutrophil gelatinase-associated lipocalin: Potential biomarker for late-onset sepsis. *Pediatr Res*. ;78(1):76-81.

289. Saleh NY, Abo El Fotoh WMM, El-Hawy MA. (2017) Serum neutrophil gelatinase-associated lipocalin: A diagnostic marker in pediatric sepsis. *Pediatr Crit Care Med*. ;18(6):e245-e252.

290. Brunner M, Krenn C, Roth G, et al. (2004) Increased levels of soluble ST2 protein and IgG1 production in patients with sepsis and trauma. *Intensive Care Med*. ;30(7):1468-1473.

291. Hur M, Kim H, Kim HJ, et al. (2015) Soluble ST2 has a prognostic role in patients with suspected sepsis. *Ann Lab Med*. ;35(6):570-577.

292. Hoogerwerf JJ, Tanck MW, van Zoelen MA, Wittebole X, Laterre PF, van der Poll T. (2010) Soluble ST2 plasma concentrations predict mortality in severe sepsis. *Intensive Care Med*. ;36(4):630-637.

293. Frottin F, Martinez A, Peynot P, et al. (2006) The proteomics of N-terminal methionine cleavage. *Molecular & Cellular Proteomics*. ;5(12):2336-2349.

294. Hurst JL, Payne CE, Nevison CM, et al. (2001) Individual recognition in mice mediated by major urinary proteins. *Nature*. ;414(6864):631-634.

295. Armstrong SD, Robertson DH, Cheetham SA, Hurst JL, Beynon RJ. (2005) Structural and functional differences in isoforms of mouse major urinary proteins: A male-specific protein that preferentially binds a male pheromone. *Biochem J*. ;391(Pt 2):343-350.

296. Bocskei Z, Groom CR, Flower DR, et al. (1992) Pheromone binding to two rodent urinary proteins revealed by X-ray crystallography. *Nature*. ;360(6400):186-188.

297. Roberts SA, Simpson DM, Armstrong SD, et al. (2010) Darcin: A male pheromone that stimulates female memory and sexual attraction to an individual male's odour. *BMC Biol*. ;8:75-7007-8-75.

298. Geoghegan KF, Dixon HB, Rosner PJ, et al. (1999) Spontaneous alpha-N-6-phosphogluconoylation of a "his tag" in escherichia coli: The cause of extra mass of 258 or 178 da in fusion proteins. *Anal Biochem*. ;267(1):169-184.

299. Mellacheruvu D, Wright Z, Couzens AL, et al. (2013) The CRAPome: A contaminant repository for affinity purification-mass spectrometry data. *Nat Meth*. ;10(8):730-736.

300. Hanson CL, Videler H, Santos C, Ballesta JPG, Robinson CV. (2004) Mass spectrometry of ribosomes from saccharomyces cerevisiae: IMPLICATIONS FOR ASSEMBLY OF THE STALK COMPLEX. *Journal of Biological Chemistry*. ;279(41):42750-42757.

301. Larkin MA, Blackshields G, Brown NP, et al. (2007) Clustal W and clustal X version 2.0. *Bioinformatics*. ;23(21):2947-2948.

302. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. (2009) Jalview version 2â€"a multiple sequence alignment editor and analysis workbench. *Bioinformatics*. ;25(9):1189-1191.

303. Takemori N, Takemori A, Tanaka Y, et al. (2016) High-throughput production of a stable isotope-labeled peptide library for targeted proteomics using a wheat germ cell-free synthesis system. *Mol Biosyst*. ;12(8):2389-2393.

304. Berglund L, Bjorling E, Oksvold P, et al. (2008) A genecentric human protein atlas for expression profiles based on antibodies. *Mol Cell Proteomics*. ;7(10):2019-2027.

305. Taussig MJ, Stoevesandt O, Borrebaeck CA, et al. (2007) ProteomeBinders: Planning a european resource of affinity reagents for analysis of the human proteome. *Nat Methods*. ;4(1):13-17.

306. Stoevesandt O, Taussig MJ. (2012) European and international collaboration in affinity proteomics. *N Biotechnol*. ;29(5):511-514.

307. Alm T, von Feilitzen K, Lundberg E, Sivertsson A, Uhlen M. (2014) A chromosome-centric analysis of antibodies directed toward the human proteome using antibodypedia. *J Proteome Res*. ;13(3):1669-1676.

308. Bjorling E, Uhlen M. (2008) Antibodypedia, a portal for sharing antibody and antigen validation data. *Mol Cell Proteomics*. ;7(10):2028-2037.

309. Khoury GA, Baliban RC, Floudas CA. (2011) Proteome-wide post-translational modification statistics: Frequency analysis and curation of the swiss-prot database. . ;1:90.

310. Bernaudat F, Frelet-Barrand A, Pochon N, et al. (2011) Heterologous expression of membrane proteins: Choosing the appropriate host. *PLOS ONE*. ;6(12):e29191.

311. Zeiler M, Straube WL, Lundberg E, Uhlen M, Mann M. (2012) A protein epitope signature tag (PrEST) library allows SILAC-based absolute quantification and multiplexed determination of protein copy numbers in cell lines. *Mol Cell Proteomics*. ;11(3):O111.009613.

312. Carlson ED, Gan R, Hodgman CE, Jewett MC. (2011) Cell-free protein synthesis: Applications come of age. *Biotechnol Adv*. ;30(5):1185-1194.