Author/s:
Doughty, Larissa

Title:
Discovering inhibitors of cell surface receptor function as the basis for novel therapeutics to treat cancer

Date:
2020

Persistent Link:
https://hdl.handle.net/11343/265748

Discovering inhibitors of cell surface receptor function as the basis for novel therapeutics to treat cancer

I. **CD33 as a target for treating acute myeloid leukemia**

II. **CD151 as a target for inhibiting metastatic prostate cancer**

Larissa Doughty

ORCID: 0000-0002-4929-528X

Submitted in total fulfilment of the requirements of the degree of Doctor of Philosophy

October 2020

Department of Biochemistry and Molecular Biology

Faculty of Medicine, Dentistry and Health Sciences

The University of Melbourne

ABSTRACT

As our understanding of the molecular changes that trigger and potentiate cancer increases, cancer therapies are becoming increasingly refined and specifically targeted to increase efficacy and reduce side effects. Structure-based rational drug design has become a common method to identify lead compounds that can be further optimised as potential drug candidates. Cell-surface receptors that undergo changes in expression and activity in cancer and can be manipulated to elicit a therapeutic effect by a binding ligand that affects protein function make ideal targets.

CD33 is a transmembrane protein from the sialic acid-binding immunoglobulin like lectin (siglec) family. It is expressed on the leukemic blasts from the majority of patients with acute myeloid leukemia (AML) but is not expressed on normal stem cells. CD33 also undergoes endocytosis when bound by CD33 antibodies. This expression pattern and the ability to potentially internalise a cytotoxic compound into the cancer cell implicates CD33 as a target for therapeutic intervention.

Another cell surface receptor, CD151, from the tetraspanin family of proteins, has been shown to be over expressed in a variety of cancers. CD151 and its binding partner integrin, are key signalling proteins in cell motility and adhesion, which is a driver for cancer metastasis. Prostate cancer, which is only lethal once metastasised, could be effectively treated if metastasis could be controlled. Targeting the specific interaction between CD151 and integrin by inhibiting the unique CD151-integrin binding site has been shown to inhibit cell motility in cell based assays and is a promising target for small-molecule inhibitors.

The specific aims of this thesis are to explore structural and functional aspects of CD33 and CD151, that are validated targets for certain cancers. Insights into the structure of these proteins will be used to identify small molecules that modulate protein function to develop novel strategies to treat common cancers in humans.

DECLARATION

This is to certify that:

I. This thesis comprises only my original work towards the PhD except where indicated in the preface.
II. Due acknowledgement has been made in the text to all other material used.
III. The thesis is fewer than 100,000 words in length.

Larissa Doughty

PREFACE

This thesis is divided into separate chapters that outline a multidisciplinary approach to rationally, and selectively, targeting CD33 and CD151 for pharmacological intervention to develop novel strategies to treat common cancers.

Chapter 2 provides an overview of general methods that are widely utilised in this thesis.

Chapter 3 describes the development of an SPR-based small molecule screen, which was employed to successfully identify novel small molecule ligands for CD33. This chapter contains 90% my own work. Protein was recombinantly expressed and purified as indicated by Dr Stefan Hermans and Jasmina Markulić. Structural modelling and virtual small molecule screening was performed by Dr Tracy Nero.

Chapter 4 describes a bioinformatical study based on structural insights of CD151, and the tetraspanin superfamily more widely, and is 98% my own work. Homology modelling was performed by Dr Tracy Nero.

Chapter 5 outlines the many different strategies used to produce recombinant CD151 LEL protein, and discusses the successes and failures, advantages and disadvantages of each method and is 97% my own work. MS-TOF analysis of insect cell protein was performed by Sam Issah. Water's Vion IMS QToF Ion Mobility Quadrupole Mass Spectrometry experiments were performed by Dr Ching-Seng Ang and Dr Nick Williamson.

Chapter 6 details trials to crystallise recombinant CD151 LEL for X-ray crystallography studies and is 100% my own work.

Chapter 7 expands the SPR methodology described in Chapter 3 to develop a fragment based screen against the CD151 LEL and is 100% my own work.

Chapter 8 provides a general discussion of the results of the preceding experimental chapters, and their significance to the study of CD33 and CD151.

This research was undertaken with the support of a National Health and Medical Research Council Dora Lush Scholarship APP1075540

LIST OF ABBREVIATIONS

ADC	Antibody drug conjugate
Ab	Antibody
ADT	Androgen deprivation therapy
ALL	Acute lymphocytic leukaemia
AML	Acute myeloid leukaemia
BLAST	Basic local alignment search tool
BSA	Bovine serum albumin
CD	Circular dichroism
CHAPS	3-((3-cholamidopropyl) dimethylammonio)-1-propanesulfonate
CLL	Chronic lymphocytic leukaemia
CM	Carboxymethyldextran
CMC	Critical micellar concentration
CMD	Carboxymethyldextran hydrogel
CML	Chronic myeloid leukaemia
CMV	Cytomegalovirus
CryoEM	Cryo-electron microscopy
CV	Column volume
Da	Dalton
DIY	Do it yourself
DLS	Dynamic light scattering
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DSF	Differential scanning fluorimetry
DTT	Dithiothreitol
ECM	Extracellular matrix
EDC	1-ethyl-3-(3-dimethylaminopropyl)-carbodiimide
EDTA	Ethylene-diamine-tetraacetic acid
EGFR	Epidermal growth factor receptor
ER	Endoplasmic reticulum
ESI-MS	Electrospray ionisation mass spectrometry
ESI-TOF	Electrospray ionisation time-of-flight mass spectrometry
FBDD	Fragment based drug discovery
FDA	U. S. Food and drug administration
FF	Fast flow
GE	General Electric
GFP	Green fluorescent protein
Glc	Glucose
GO	Gemtuzumab ozogamicin
GRAVY	Grand average of hydropathicity
GST	Glutathione S-transferase
HBM	Honey bee mellitin
HBS-P	0.01 M HEPES pH 7.4, 0.15 M NaCl, 0.005% v/v TWEEN®20
HBS-P+	0.01 M HEPES pH 7.4, 0.15 M NaCl, 0.05% v/v TWEEN®20

HCl	Hydrochloric acid
HCV	Hepatitis C virus
HEK	Human embryonic kidney
HEPES	N-(2-Hydroxyethyl)piperazine-N'-(2-ethanesulfonic acid)
HF	High fidelity
His	Histidine
HP	High performance
HPLC	High performance liquid chromatography
HTS	High throughput screening
ID	Identification
IMAC	Immobilised metal ion affinity chromatography
IPTG	Isopropyl β-D-I-thiogalactopyranoside
ITC	Isothermal titration calorimetry
ITIM	Immunoreceptor tyrosine-based inhibitory motif
JBS	Jena bioscience
JCSG	Joint Centre for Structural Genomics
K_D	Equilibrium dissociation constant
LB	Luria-Bertani broth
LC	Liquid chromatography
LC-MS	Liquid chromatography-Mass Spectrometry
LEL	Large extracellular loop
LM	Laminin
mAb	Monoclonal antibody
MBP	Maltose-binding protein
MCSG	Midwest Centre for Structural Genomics
MES	2-(N-Morpholino)ethanesulfonic acid hemisodium salt
MFP	Monash fragment platform
mg	Milligram
ml	Millilitre
MMP	Matrix metalloproteinases
MOI	Multiplicity of infection
MST	Microscale thermophoresis
Mw	Molecular weight
NADPH	Nicotinamide adenine dinucleotide phosphate
NCBI	National Centre for Biotechnology Information
NCI	National Cancer Institute
NDSB	3-(1-Pyridinio)-1-propanesulfonate
NHS	N-hydroxysuccinimide
NID	NTA derivatised carboxymethyldextran hydrogel
NIHC	Poly - NTA derivatised linear polycarboxylate hydrogel with high affinity
NMR	Nuclear Magnetic Resonance Spectroscopy
NTA	Nitrilotriacetic acid
PAGE	Polyacrylamide gel electrophoresis
PAINS	Pan assay interference compounds
PBS	0.14 M NaCl, 10 mM NaH_2PO_4, 10 mM Na_2HPO_4 pH 7.4

PCa	Prostate cancer
PCR	Polymerase chain reaction
PCT	Pre-crystallisation test
PDB	Protein Data Bank
PDEA	2-(2-pyridinyldithio)ethanolamine
PKC	Protein kinase C
PMSF	Phenylmethanesulfonyl fluoride
PPI	Protein-protein interaction
PROSPER	Protease specificity prediction server
PROSS	Protein repair one-stop shop
PSA	Prostate-specific antigen
PSMA	Prostate membrane specific antigen
PVDF	Polyvinylidene fluoride
RCSB	Research Collaboratory for Structural Bioinformatics
RFU	Relative fluorescence units
RM	Rich media
RMSD	Root-mean-square deviation
RNA	Ribonucleic acid
RNR	Ribonucleotide reductase
RPM	Revolutions per minute
RT	Room temperature (23 +/- 1 C)
RU	Response unit
SAHC	Streptavidin, immobilised in a linear polycarboxylate hydrogel
SAR	Structure-activity relationship
SBDD	Structure-based drug design
SDS	Sodium dodecyl sulfate
SEC	Size-exclusion chromatography
SEL	Small extracellular loop
SER	Surface entropy reduced mutants
SFM	Serum free media
SOC	Outgrowth media
SPR	Surface plasmon resonance
SUMO	Small Ubiquitin-like Modifier
TBS	50 mM Tris-Cl, pH 7.5, 150 mM NaCl
TEV	Tobacco etch virus
TFA	Trifluoroacetic acid
TM	Transmembrane
TOF	Time-of-flight
TRAMP	Transgenic adenocarcinoma of the mouse prostate
TSP	Tetraspanin
TTD	Therapeutic target database
ULP	Ubiquitin-like-specific protease
USDA	United States Department of Agriculture
UV	Ultraviolet

LIST OF AMINO ACIDS

Name	3-letter symbol	1-letter symbol
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Hydroxyproline	Hyp	O
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

UNITS OF MEASUREMENT

Symbol	Measurement
k_a	association rate
k_d	dissociation rate
K_D	equilibrium dissociation constant
M	Molar
mM	Millimolar
µM	Micromolar
nM	Nanomolar
°C	degrees Celsius
s	Second

TABLE OF CONTENTS

CHAPTER 1. LITERATURE REVIEW AND INTRODUCTION

1.1. *CANCER*

Cancer is a leading cause of death worldwide, second only to cardiovascular disease. [1] The International Agency for Research on Cancer estimated that in 2018 there were approximately 9.6 million deaths from cancer worldwide and by 2040 this will increase to 16.3 million.[2] There are more than a hundred types of cancer, each distinct in its behaviour and response to treatment. Some cancers result in the formation of solid tumours; however, some, like leukaemia, do not. The factors leading to the development of cancers are as varied as the cancers themselves. Cancer initiation can be triggered by many things, including environmental carcinogens like cigarette smoke, pollution and radiation, biological factors such as viruses and hormones, lifestyle issues such as alcohol consumption, obesity and poor diet, a genetic predisposition or simply age. These initiating factors lead to DNA damage that causes a mutation in the DNA sequence. Mutations that allow abnormal proliferation initiate the cancer and additional mutations lead to rapid growth. Fundamentally, development of cancer occurs when genetic mutations lead to cells that can proliferate in an unregulated manner and invade surrounding normal tissue, eventually invading blood and lymphatic vessels, allowing them to metastasise throughout the body. [3-5]

1.1.1. Cancer therapy

There are a variety of treatment options available and patients will be prescribed treatments based on the type of cancer, how advanced the cancer is and also the patients overall health and age. Some treatments can be harsh and are not well tolerated by the elderly or those with unrelated health issues such as cardiac or respiratory problems. [6] Some patients may only require one treatment, but typically patients undergo a combination of treatments including surgery, chemotherapy and/or radiation therapy. Immunotherapy, hormone therapy and targeted therapy are also used depending on the type and severity of the cancer.

When a tumour is present the first line of treatment is usually surgery. The surgeon may physically remove the entire tumour or debulk the tumour; removing some, but not all of the mass. Debulking is used when removing the entire tumour may cause damage to associated organs and to relieve pain or pressure caused by the tumour. Often the surgeon may also remove lymph nodes or surrounding tissue as a preventative measure to reduce the instance of metastasis. Side-effects and complications arising from surgery can be pain, infection, complications due to anaesthesia, bleeding and blood clots, often leading to increased morbidity. [7] Limitations of surgery include the possibility of undetected residual tumour cells remaining in the patient and nearby tissue and organs can be damaged or removed leading to further complications. Because of these

limitations surgery is often used in combination with radiation and/or chemotherapy, depending on the type, location and size of the tumour. [8]

Radiation therapy uses directed high-dose radiation to kill cancerous cells. Radiation can also be used to shrink tumours prior to surgery or to reduce tumour related symptoms. The side-effects of radiation therapy range from fatigue, burn-like skin irritation and, depending on the site of the radiation therapy, can be severe as gastrointestinal problems, respiratory problems and osteoradionecrosis. [9]

Chemotherapy is perhaps the most well-known cancer therapy and can be used as a sole therapy, or as an adjuvant with surgery and/or radiation therapy. Treatment and efficacy depends on the type and stage of the cancer and chemotherapy may be administered with a curative intent, to prolong life, or for palliative care. A chemotherapy regime consists of one or more cytotoxic drugs that act to inhibit cellular mitosis and induce apoptosis. [10] Mitosis is common to both normal and cancer cells, and normal growth of both cell types is damaged by chemotherapy. The selectivity of chemotherapy arises from the highly proliferative nature of cancer cells, relative to normal cells, and the aim of chemotherapy is to achieve maximum cancer cell death while minimising the degree of toxicity to normal cells. [11] Normal tissues that undergo rapid cell division such as bone marrow, gastrointestinal mucosa, hair follicles and gonads are the most sensitive to chemotherapeutic effects. Accordingly, the most common side-effects of chemotherapy are nausea, vomiting, hair loss, myelosuppression and reduced fertility. Other than fertility complications, most effects are transient and reversible on cessation of treatment. However, an emerging concern is the incidence of non-malignant chronic illness in cancer survivors. While cancer survival rates continue to rise often the quality of life of cancer survivors is significantly diminished, with chronic pain and neuropathy being the most common complaint. An increase in the incidence of pulmonary and cardiac problems, as well as moderately to severely affected mental and general health have been reported. [12, 13] Another drawback of the systemic effects of chemotherapy is the development of drug resistant tumour cells. During treatment cancers can develop resistance to the cytotoxic agents through mechanisms such as DNA mutations and metabolic changes that enable drug inhibition and degradation. [14]

The limitations and side-effects of chemotherapy has led to a rise in the development and use of targeted therapy which specifically targets cancer cells with reduced toxicity to off-target cells. [15]

1.1.2. Targeted cancer therapies

New cancer therapies are becoming more focused on malignant cells with the aim of decreasing off-target effects. As aberrant signalling pathways that occur as a result of mutations that cause the cells to become malignant are identified, studied, and understood they are increasingly being investigated as targets for molecular target-based cancer therapy. [16] These drug targets can be part of the dysregulated pathways, extracellular ligands, cell surface receptors or kinases. The aim is to disrupt an aberrant signalling pathway by inhibiting a protein-protein interaction (PPI) to halt proliferation or metastasis or to deliver a cytotoxic agent directly to the cancer cells. [17, 18] These therapies, which include monoclonal antibodies (mAbs) and small molecule inhibitors, are now a component of therapy for many common cancers and are generally better tolerated than traditional chemotherapy.

The first targeted therapy was the mAb trastuzumab (Herceptin®) directed against the cell surface receptor tyrosine kinase HER2 (ErbB2), which was approved by the FDA in 1998. This antibody (Ab) specifically inhibits the growth of HER2-positive metastatic breast cancer by binding to domain IV of the extracellular portion of the HER2 receptor and arresting cellular mitosis. [19, 20] In 2001 the first rationally designed small-molecule inhibitor, imatinib (sold as Gleevec), was approved for the treatment of chronic myeloid leukaemia (CML) and began a new era in anticancer drug discovery. [21] Imatinib targets a characteristic genetic abnormality of CML which results from a reciprocal translocation between chromosomes 9 and 22. [22] The translocation generates the fusion protein BCR-ABL, a tyrosine kinase; it has been established that the presence of BCR-ABL alone is sufficient to cause CML. [23, 24] Imatinib competitively inhibits ATP binding to BCR-ABL, thus inhibiting tyrosine phosphorylation of proteins involved in BCR-ABL signal transduction; ultimately leading to arrest of growth and apoptosis. Prior to the use of imatinib the five-year survival rate for CML patients was only 30%, the introduction of imatinib increased survival to more than 89%. [25]

HER2 and BCR-ABL are both examples of aberrant, oncogenically activated cell surface receptors that have been targeted to elicit a physiological response that has therapeutic value. Targeted therapeutic strategies have also been shown to sensitise tumour cells to radiation and chemotherapy thereby acting synergistically with traditional therapies. [26, 27]

1.1.3. Receptors

Cell surface receptors are transmembrane (TM) proteins that mediate signals across the plasma membrane. The extracellular segment of the receptor has binding sites for signalling molecules, which, once bound, induce a conformational change and/or a change in oligomerisation state leading to propagation of a signal through to the cytoplasmic portion of the protein which initiates a signalling cascade. Cell surface

receptors can be classified into three broad categories, based on the mechanism they use to effect signalling (Figure 1-1). [28]

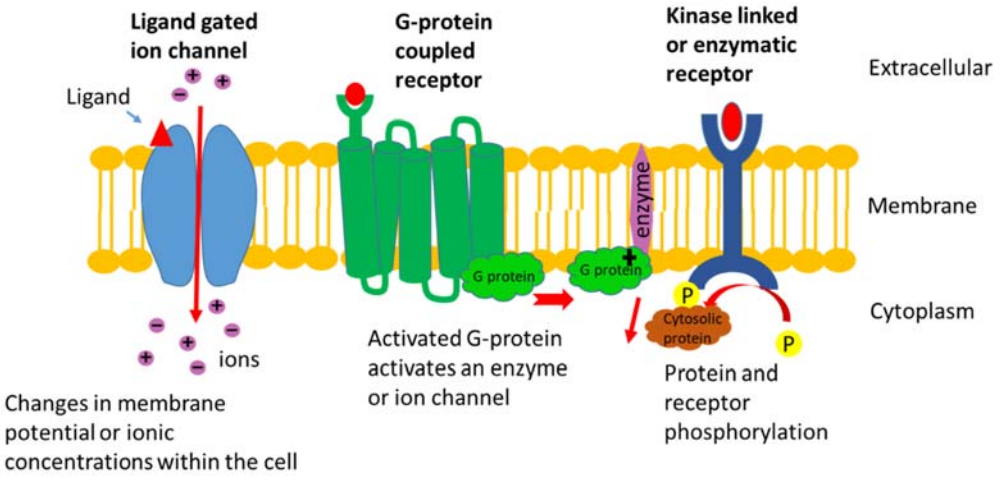


Figure 1-1 Three categories of cell surface receptors. These receptors span the membrane and bind ligands in the extracellular matrix and induce a conformational change or transmit a signal into the cytoplasm. Ligand gated ion channels (left) alter their conformation to open or close a channel to allow a flow of ions across the membrane. G-protein coupled receptors (middle), activate once ligand binding occurs, and the G-protein dissociates from the receptor, carrying a signal to a target, which may be an ion channel or an enzyme. Kinase linked or enzymatic receptors (right) are activated by ligand binding and the cytoplasmic portion of the receptor acts as an enzyme, for example in the case of kinases, to phosphorylate intracellular proteins.

Receptors that can be targeted for therapeutic effect are newly synthesised, overexpressed or abnormally glycosylated on malignant cells. [29] Membrane proteins are increasingly being targeted for their potential therapeutic effect and currently represent more than 60% of the drug targets on the market. [29, 30]

Several factors determine the amenability of targeting cell surface receptors for pharmacological intervention to treat cancer. The receptor should be predominantly expressed on cancer cells compared with normal cells. The target cells should also display a high degree of homogeneity in their receptor expression to minimise non-target effects. [31] The receptors ability to undergo endocytosis may also be an important factor that can be exploited to deliver a toxic agent to induce cell death. [32] The drug itself may consist of Abs, Ab fragments, Ab-drug conjugates (ADCs), proteins, peptides or small molecules.

There have been 6,718 human membrane proteins identified to date, of which 1,352 are receptors. [33] The online Therapeutic Target Database (TTD) is a database managed by the Bioinformatics and Drug Design Group at the National University of Singapore and the Innovative Drug Research and Bioinformatics Group at Zhejiang University. The database provides extensive information, including target validation, of over two thousand therapeutic protein and nucleic acid targets. [34] Databases such as the TTD are an invaluable resource for drug discovery research and clinical medicine in the development of novel therapeutics. Searching the TTD for "receptor" yields 786 results, indicating that 58% of all known cell receptors are currently being used or investigated as a drug target. In 2009, the National Cancer Institute (NCI) issued a prioritisation list of cancer antigens. The list consisted of 75 proteins, 7 of which are receptors. [35] This move towards target-based precision therapies is predicted to become the standard approach for cancer diagnosis and treatment. [36]

A cancer drug target is validated by assessing its specificity for tumour cells and if biological activity can be altered pharmacologically by a ligand to have anti-tumour effects. To exert a modulating effect, the ligand must be able to bind to the target with high affinity; typically, a binding affinity K_D <20 nM is desirable. [37] While this is easily achievable with biological agents such as Abs, for small molecule ligands this level of affinity requires both hydrophobic and polar interactions with the protein target. Such interactions are usually only possible if the ligand can fit into a well-defined pocket or cavity on the target protein's surface. For rationally guided drug discovery, identification of such a binding pocket on the protein target relies on access to detailed structural information. The main methods for resolving the structure of a protein are X-ray crystallography, nuclear magnetic resonance spectroscopy (NMR), Cryo-electron microscopy (CryoEM) and, in the absence of direct structural data, homology modelling. X-ray crystallography is the most common technique and high-resolution crystal structures contain information regarding water molecules, ligand binding sites and the plausible binding mode of ligands. [38] NMR, which is a solution-based technique, can give insights into the dynamics of the target as well as the structure. [39] Recent advances in CryoEM technology have increased the resolving power of the technique to allow the structural analysis of proteins as small as 65 kDa. The advantage of this technique is that proteins can be examined in a near to native state with only small amounts of protein required. Class averages can reveal conformational changes that are not accessible using solid-state crystallography techniques. [40] If structural data is unavailable, it may be possible to construct an homology model of the target protein based on the 3-dimensional (3D) atomic coordinates of a closely related homologue and this model can be used to provide insights into the 3D structure, and possible ligand binding sites of the target protein. Evolutionary related proteins contain structural similarities and small changes in the amino acid sequence of related proteins rarely affect the 3D structure. In practice, where related proteins share over 50% identical

amino acids homology modelling can be used to produce models that are considered accurate enough for drug discovery. [41]

1.1.4. Protein-protein interactions

Protein-protein interactions (PPIs) are defined as physical contacts between proteins that occur in a cell or in a living organism where the contact interfaces have evolved for a specific purpose and are not formed by chance or for generic functions such as protein production or degradation.[42, 43] PPIs form networks of complexes bound by biochemical and/or electrostatic forces and serve a biological role such as protein function, signalling, metabolic pathways and physiological processes. Cellular growth and differentiation, the key drivers of cancer metastasis, are an example of a biological process governed by PPIs.[44]

The formation of PPIs is dependent on the biological context, that is, not all possible PPIs will form in any cell at any time. Rather, the cell type, cell cycle phase, environmental conditions, protein modifications, the presence of cofactors and other binding partners all impact on if and when PPIs form. The physical contact between the proteins that forms the PPIs can be stable and irreversible, as in the case of IL-5 cytokine dimer, or they may be transient and associate and dissociate *in vivo* for example the binding of a ligand to a GPCR (Figure 1-1). [43]

The interactions formed by cytoplasmic or a single transmembrane span proteins are usually oligomerisation of identical subunits, such as homodimerisation, or hetero-complexes that bring different proteins together (Figure 1-2). Homo-complexes are relatively stable and often the hydrophobic surface that forms the interaction is indistinguishable from the protein core and can be considered part of the conformational folding of the protein. [45] Hetero-complexes can be transient, or non-obligate; i.e. the proteins can exist independently and fully folded, and as such the protein interface is typically less hydrophobic and the surface morphology more variable than in homomeric complexes. [46] When the monomers that form non-obligate hetero-complexes are not in their bound state they are in polar environments such as the extracellular matrix. Charged and polar surfaces are arranged around the hydrophobic areas to shield them from the environment. When bound, their hydrophobic contact areas are further shielded from the environment. To achieve this, PPIs may induce conformational changes in one or both binding partners. [47, 48] In spite of this structural and conformational diversity, it is thought that PPIs occur over a large, flat surface area and have been considered generally not viable drug targets because of the lack of suitable binding pockets. [49] More recently, studies have shown that residues present in, or near, protein-protein interfaces that are involved with the interaction, may provide a potential target for pharmacological intervention. [50, 51] These “hot spot” regions are thought to be critical to the binding interaction and

contribute a significant portion of the interaction energy between neighbouring protein subunits (Figure 1-2). [52-54] For drug discovery purposes hot spots ideally cover an area about the same size as a small molecule and have some conformation adaptability. Identifying hot spot regions in protein-protein interfaces provides a starting point for small molecule drug design. It has been shown that small molecules can bind to these regions with higher affinities than the native interaction, thereby inhibiting the PPI. [55-57]

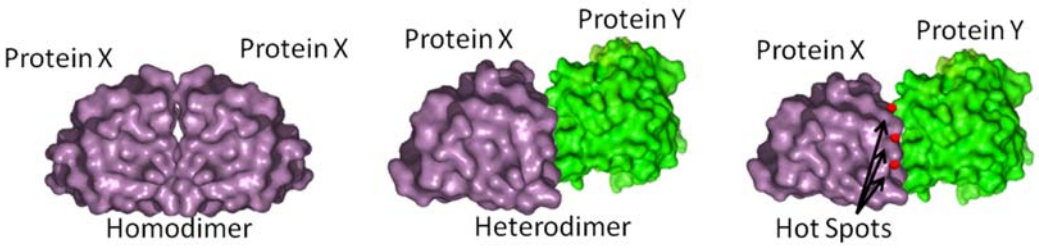


Figure 1-2 Representations of PPIs. A homodimer composed of two monomers of the protein X (left), a heterodimer composed of a monomer of protein X and protein Y (middle) and critical regions of binding known as hot spots (red) (right).

The increased understanding of the molecular mechanisms of cancer biology and advances in computational chemistry have led to a huge increase in protein targets investigated for drug discovery. Current targeted therapies are commonly used as adjuvant therapies along with traditional treatments such as chemotherapy or radiation. The ability to target malignancies from multiple aspects is increasing patient survival and minimising collateral damage to the immune system and other organs. [58]

1.1.5. Therapeutic ligands

Monoclonal antibodies (mAbs) and Ab fragments have been used as therapeutics since the mid 1980's and are an attractive therapeutic tool as they are highly specific to their targets. Abs are used to recruit host immune functions to attack the targeted cell, disrupting cell processes, or as ADCs to deliver a toxin to the cell. [59, 60] The first mAb used to treat patients was muromonab-CD3, approved by the U. S. Food and Drug Administration (FDA) in 1985. [61] Muromonab-CD3 is an immunosuppressant mAb given to supress organ rejection in transplant patients, by targeting the CD3 receptor on the surface of T cells. The limitation of muromonab-CD3 as a therapeutic is that it is a murine mAb produced using mouse-derived cells. Its effectiveness is reduced as the

patient forms anti-mouse Abs that can induce anaphylactic reaction against the foreign mouse protein. [62] The introduction of chimeric Abs containing 70% human sequences in the late 1980's reduced, but did not eradicate, these effects. By the mid 1990's techniques for the selection of fully human Abs from combinatorial approaches or by transgenic animal production, improved efficacy of the Abs and greatly reduced side-effects as they behave as endogenous Abs. As biologicals however, even humanised Abs can elicit an unwanted immune response. [63-66] The biological nature of Ab's and Ab fragments also causes them to be denatured by the gastrointestinal tract, precluding their utility as an orally available drug and limiting their administration route to intravenous. The molecular weight (Mw) of Abs, approximately 150 kDa, prevents them from crossing the blood-brain barrier and tissue penetration is poor (Figure 1-3). As such, some Abs require intra-tumour delivery, which has little benefit for metastatic cancers. [67] So although Abs and Ab fragments have been hugely beneficial in improving cancer therapies, problems remain.

Since Paul Ehrlich, in the 1870's, described the existence of "chemoreceptors" that could be exploited therapeutically, therapeutic medicine has been driven towards the goal of identifying drugs that can bind to biologically important proteins to alter their function to treat human disease. [68] The description and characterisation of the enzyme carbonic anhydrase by Meldrum and Roughton in 1933 led to the discovery that sulfanilamide inhibited carbonic anhydrase, which increases the excretion of sodium and water. [69, 70] This is one of the first examples of a small molecule drug targeting a metabolic pathway. The sequential development of structural variants of sulfanilamide, which were found to be effective antibiotics, hypoglycaemic agents, diuretics and antihypertensive drugs, is illustrative of the structure-based drug development pathway. [68]

Currently identified small molecule ligands are generally less specific than Abs; however, they are chemically defined, less susceptible to degradation and are usually non-immunogenic. In addition, by selectively targeting poorly conserved protein regions, high-affinity small molecule inhibitors have the capability to form the basis of future highly specific drugs that have limited side-effects due to off-target protein binding. Their small size enables them to target cell surface receptors as well as intracellular proteins that are involved with cell growth and metastasis (Figure 1-3). [71] Small molecules make up over 90% of the drugs on the market today. With high throughput screening (HTS) of small molecule libraries, thousands of molecules can be rapidly evaluated against a validated protein target, and positive "hits" can be a starting point for therapeutic development. [72]

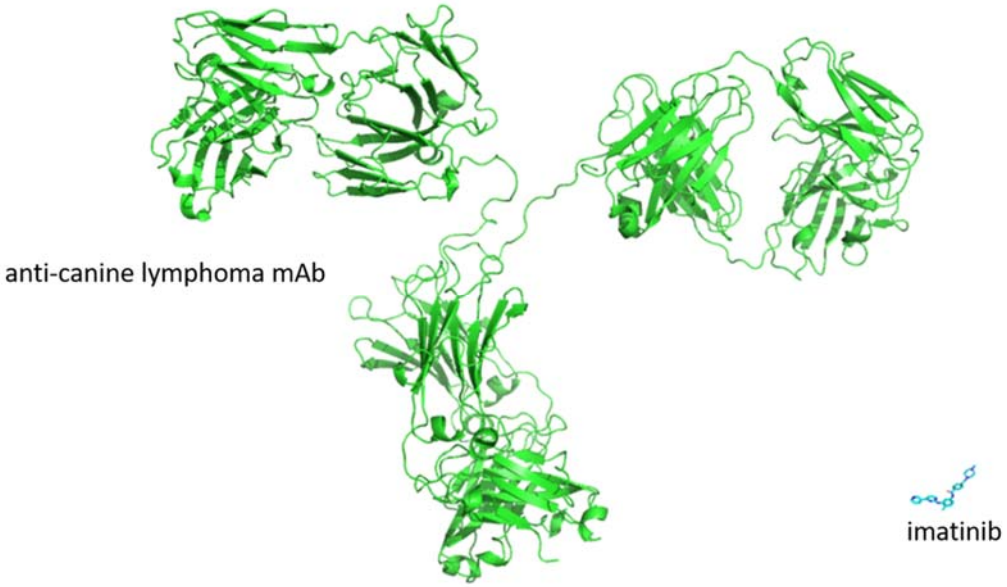


Figure 1-3 Structure of the mAb Mab231 (PDB ID: 1IGT, left) and the small molecule drug imatinib (PubChem CID: 5291, right). This figure is representative of the relative sizes of a typical Ab and small molecule drug. The anti-canine lymphoma mAb is 153 kDa compared to the tyrosine kinase inhibitor imatinib which is 0.5 kDa.

Knowledge of the protein target structure, or of a known ligand allows a rationally designed computational chemistry approach to identify potential binding molecules or ligand mimics. Using a computational, *in silico* screening approach millions of potential drug-like compounds can be rapidly examined for their potential to bind to the target protein, and then representative sets of chemically diverse molecules that possibly bind to the target molecule can be assayed for binding affinity and functional effect on the target protein. [73] An important step in discovering drug-like molecules using this method is developing 'high-throughput' techniques (HTS) to screen hundreds, or thousands, of candidate molecules for binding to the target protein.

1.1.6. High throughput screening

The primary goal of HTS is to identify novel compounds that bind to the target at a low concentration. These "hit" compounds can then be optimised to improve the compound's drug-like properties. [74, 75] Using molecular modelling and property prediction methods to examine the physical properties of the compound, subsequent rounds of *in silico* screening and biophysical assaying can fine-tune the affinity and avidity of drug-like compounds; that can be further refined to improve pharmacokinetics and selectivity and reduce toxicity. The wide range of biological targets that are being investigated has led to the development of many different assay approaches. Generally,

the assays are either biochemical or cell-based and typically use fluorescence-based detection methods. [76] HTS relies heavily on automated liquid handling and detection as well as sophisticated computing software to handle and analyse the data. Automation ranges from simple dilution devices to robotic systems that perform multiple tasks from sample dispensing to data collection, allowing for 24-hour operation and increasing the screening rate. Well-designed screens can provide information regarding compound specificity as well as potency. Analysis of the chemical and physical properties of the hit compounds may reveal recurring key features or functional groups that can form the basis of a structure-activity relationship (SAR). [77] Such features can be critically evaluated for optimisation to increase specificity and affinity. Compounds that act specifically against the target protein are less likely to exhibit off-target toxicity. Multiple positive hit compounds can also help to elucidate important structural features of the molecules involved in selective binding to the target and the development of SARs help further optimise the future compounds. [74]

In 2008 Mayr & Fuerst suggested a “magic triangle of HTS” which describes the fundamental principles of performance management of lead compound discovery (Figure 1-4). [78]

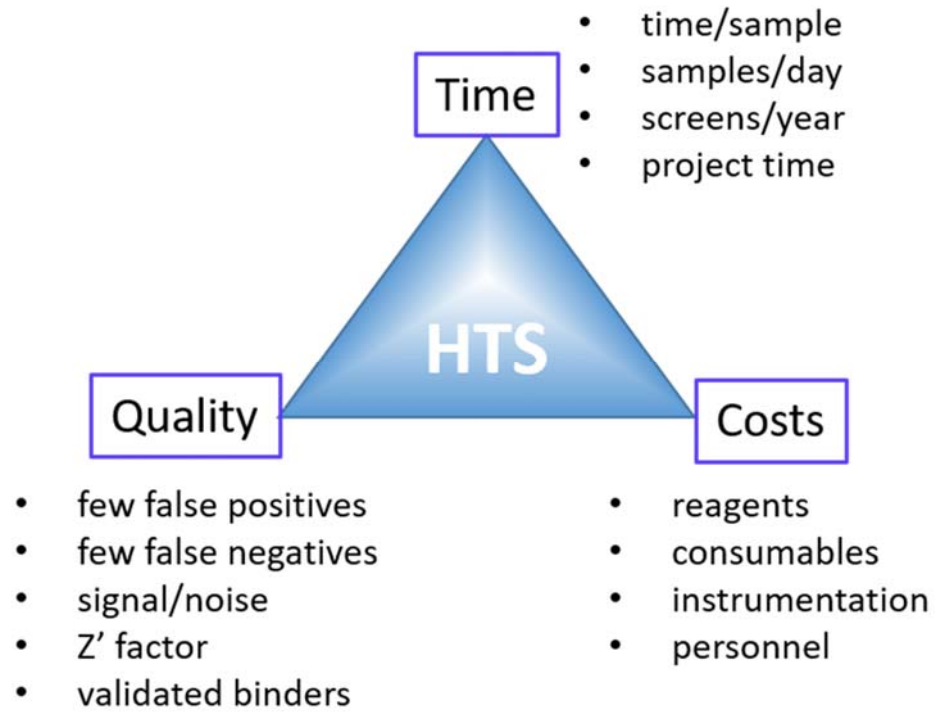


Figure 1-4 The optimisation process for successful HTS. The figure shows the key success factors for lead discovery via HTS, i.e. time, costs and quality. Every change to either factor influences the setup of the other factors. Optimal lead discovery by HTS finds a balance between the three elements. Adapted from Mayr & Fuerst 2009. [78]

Because of the large number of samples analysed in HTS the total handling time for the project needs to be as short as possible. This increases the likelihood of finding novel hit compounds due to the increased number of compounds that can be tested. A faster assay time is also advantageous for sample stability and reproducibility, although due to automation actual screening time represents a minor component of the project's turnaround time. The main time-consuming elements are assay development, data analysis and interpretation, hit list validation and follow-up in secondary techniques. [78] Beyond the initial major costs involved with instrumentation, robotics and computing there are also costs of reagents, consumables and skilled personnel (Figure 1-4). The greatest reagent costs involve the biological test samples, typically proteins or cells, which can be unstable, expensive or difficult to produce. The third factor, quality, is particularly important for large data sets as only assays of high statistical quality can be used for correct data analysis. [79] The generation of false negatives is only problematic when hit rate is very low, while false positives are a drain on time and

resources. Hits also need to be validated by orthogonal techniques to minimise artefacts such as promiscuous binding (Figure 1-4). [80, 81]

The number of techniques available to screen compounds in a high throughput format is almost as varied as the targets. Biophysical and structural methods, previously used for quality control and optimisation, are now increasingly being used to screen compound libraries and validate hit compounds. The most common methods include thermal melt assays, commonly performed using differential scanning fluorimetry (DSF), fluorescence polarisation (FP) assays, surface plasmon resonance (SPR), NMR, isothermal titration calorimetry (ITC), microscale thermophoresis (MST) and protein X-ray crystallography. Each of these techniques has advantages and disadvantages and offer differing information regarding the validation of hit compounds (Table 1). The information ranges from structural binding modes to the energetics of the binding interaction. Application of analogous methods provides the most robust way to differentiate positive binders and to establish SARs.

All of the techniques listed in Table 1 require that the compound be soluble to concentrations above its binding affinity, to varying degrees. For the best quality data, all of the techniques also require that the protein target is homogenous and soluble at relatively high concentrations. X-ray crystallography, for example, typically requires a few milligrams of the protein at more than 5 mg/ml and that the protein crystallises in solid state form, where dynamic fluctuations of protein conformation are inaccessible. NMR requires protein to be soluble at approximately 10 µM and may require isotope labelling. Techniques such as SPR use much less target however require the target to be immobilised on a sensor chip. This immobilisation leads to complications such as steric hindrance of the binding site, disturbances of the target folding and compound binding as well as non-specific binding of the compound to the immobilisation medium. Thermal based techniques such as DSF and ITC are solution based but require milligrams of purified protein and high compound solubility. Fluorescent based techniques such as MST and FP, that are solution based, require much smaller amounts of target and compound; however, the target must be labelled with a fluorophore and the presence of the fluorophore itself may affect binding.

Table 1 Comparison of established biophysical methods for analysis of protein-compound interactions. Adapted from Ciulli et al. [82]

Technique	Advantages	Disadvantages
DSF	• High throughput • Widely applicable to a range of target proteins • Detects direct binding	• High error rate • High material consumption • Plate well variability
FP	• High throughput • Widely applicable to a range of target proteins • Competition binding assay	• Susceptible to false positives and artefact effects • Requires ligand to be labelled
NMR (compound-observed)	• Mid-range throughput • Applicable to a range of target proteins (>10 kDa) • Can be used to assess quality	• Prone to false positives due to compound aggregation or nonspecific effects
NMR (protein-observed)	• Mid-range throughput • Binding site can be identified by peak assignment • K_D measureable from ligand titration	• Limited to small (<30 kDa) and soluble proteins • Requires target proteins to be isotope labelled which can be expensive • High material consumption
ITC	• Direct/competition binding assays • Applicable to a range of target proteins • Low material consumption • Measures several parameters (K_D, ΔG, ΔH, ΔS, n)	• Low throughput • High material consumption • Large heat changes upon binding required for accurate measurements
SPR	• Label-free detection • Readily automated • Applicable to a range of target proteins • Competition and direct binding assays • Low material consumption • Measures several parameters (K_D, *k*a, *k*d and *n*)	• Binding partner is required to be immobilised on a surface • Prone to artefacts due to compound aggregation, immobilisation artefacts or nonspecific effects • Optimisation is time-consuming
X-Ray	• Mid-range throughput • Binding site and ligand binding mode can be identified • Ligand-induced conformational changes can be identified	• Limited to soluble target protein that can be crystallised • Requires access to X-ray sources (in-house, synchrotrons) • Binding site can be occluded by crystal packing • High occupancy of the ligand binding site required
MST	• Applicable to most target proteins • Low material consumption • Competition and direct binding assays	• Ligand requires labelling or needs intrinsic fluorescence • Subject to false positives and artefact effects

Design of an HTS strategy to examine binding to a particular protein target is informed by the physico-chemical properties of the protein and the very nature of possible small

molecule – protein interactions that are predicted to be useful to develop drugs to target protein function. As each protein target is different there is no formulaic approach to examining protein-small molecule interactions, and as such, innovative and creative approaches are required for developing techniques to examine potential drug binding to medically important proteins.

1.2. *LEUKAEMIA*

Leukaemia is a malignant disease of the haematopoietic system which results in the production of increased numbers of abnormal or immature leukocytes. Leukaemia typically begins in either the myeloid or lymphoid progenitor cells in the bone marrow (Figure 1-5).

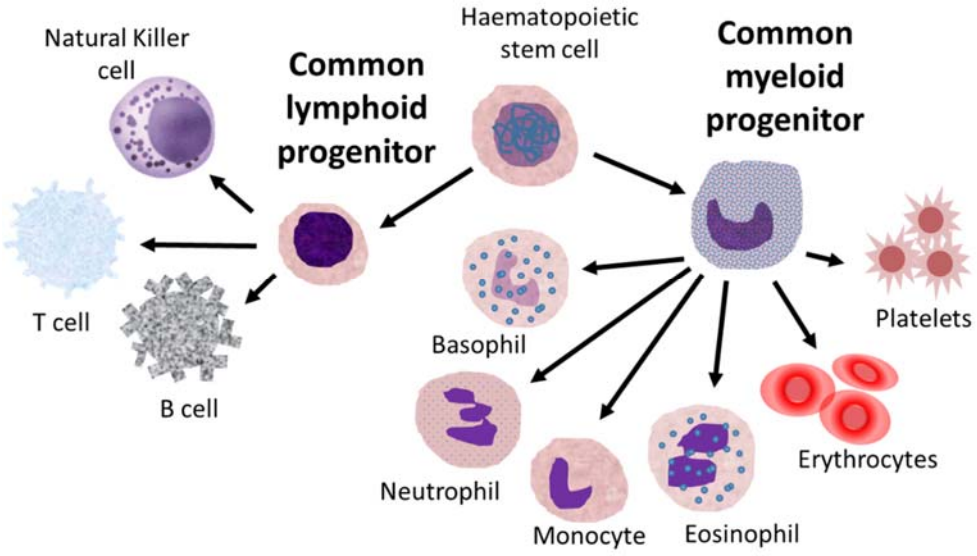


Figure 1-5 Production and differentiation pathways of cells in the haematopoietic system. Leukaemia typically begins in either the lymphoid (left) or the myeloid (right) progenitor cells.

Leukaemia's are classified by the rate at which the cancer progresses: i.e. acute or chronic; and whether they originate in lymphoid or myeloid cells. It is the most common form of blood cancer and typically affects adults over the age of 50 but is also the most common cancer in children under 15. [83]

Chronic and acute lymphocytic leukaemia (CLL and ALL) begins in lymphoid progenitor cells (Figure 1-5). ALL is most common in children under the age of 15 and involves both B and T cells, while the median age at diagnosis for CLL is 72 years and involves only the B cells. Recent advances in supportive care and treatment, due to the greater

understanding of the biology of leukaemic cells, has increased survival rates to over 80% in adults and up to 100% in children. [84-86]

Chronic and acute myeloid leukaemia (CML and AML) begins in the myeloid cells (Figure 1-5). AML is an aggressive malignancy characterised by the impaired differentiation and uncontrolled proliferation of myeloid progenitor cells. The 5-year survival rate for patients under 60 years of age is approximately 24%. For patients over 60 years the survival rate is less than 15% with a median survival of less than 10 months. [87-89] CML is known to be caused by a single genetic mutation, with more than 90% of CML cases resulting from a specific genetic mutation in chromosome 22, known as the Philadelphia chromosome. [90, 91] CML differs from AML in that the myeloid cells do not lose their ability to differentiate. It accounts for 20% of all leukaemia's affecting adults and is most common in middle-aged people. The 5-year survival rate for CML patients treated with the kinase inhibitor imatinib is well over 90%. The success of this drug is largely due to CML having a single aberrant protein to target and the precise specificity of the drug. [92]

1.2.1. Acute myeloid leukaemia

As treatment strategies improve and survival rates increase for most forms of leukaemia, AML retains a high mortality rate. AML is a genetically heterogeneous disease and presents a wide array of clinical symptoms and varied morphology, and as such, is difficult to treat. Generally, first-line treatment is intensive chemotherapy to reduce the number of leukaemic cells to undetectable levels. Complete remission and long-term survival is achievable in less than 45% of patients younger than 60 years and less than 15% for those older. [93] Patients that relapse within 12 months have a less than 20% chance of achieving remission again. For patients who stay in remission longer than 18 months the chance of surviving relapse increases to over 50%. Post-remission strategies include intensive and high dose therapies followed by haematopoietic cell transplantation. AML relapse is typically associated with an increase in molecular complexity and with multiple new subclones and mutations which increase the resistance to cytotoxic chemotherapy. [94, 95] Assessment of various factors evaluates if the patient is a candidate for post-remission intensive therapy and older patients, who often exhibit comorbidities, may only be offered palliative care. Disease recurrence remains the major cause of treatment failure in AML. [87, 88, 96] New strategies to treat this disease are required.

The myeloid differentiation antigen CD33 has been targeted for Ab-based therapies for some time. While the exact physiological function of CD33 is poorly understood, it is expressed on leukaemia blasts in almost all adult and childhood AML patients and has been identified on some adult leukaemic stem cells. [97, 98] The antigen density and proportion of CD33 expression on leukaemic blasts varies from patient to patient but

over 85% of AML patients have CD33 positive blast cells. [99, 100] CD33 is known to undergo endocytosis when bound to an anti-CD33 Ab, this enables direct delivery to the cell of cytotoxic or chemotherapeutic compounds as immunoconjugates. [101] It is this endocytic property, combined with the broad expression pattern in AML, which makes CD33 such an attractive therapeutic target. [102, 103]

Many attempts have been made to target CD33 as a therapy for AML patients. Due to the endocytic properties of CD33 the focus has been on ADCs, but immunotoxins and radionuclides have also been trialled. [104] Lintuzumab, an anti-CD33 mAb, showed promise in early phase trials but failed to improve patient survival. [105]

Unconjugated Ab therapy can engage Natural Killer (NK) cell Ab-dependant cell-mediated cytotoxicity (ADCC) by redirecting NK cells against AML targets. For example, the unconjugated Ab BI36858 is Fc optimised and targets CD33 resulting in improved NK cell-mediated ADCC. [106] Ab B136858, the ADC IMGN779 and the bispecific T cell-engaging Ab AMG 330 are currently in phase I clinical trials. Gemtuzumab ozogamicin (GO), an anti-CD33 ADC, was approved for treatment of AML in 2000. It was subsequently withdrawn in 2010 over toxicity concerns; however, it is currently being reinvestigated as an adjunct therapy as it has been shown to reduce relapse rates. [97]

These promising preliminary studies using CD33 targeted therapies validate CD33 as a therapeutic target for AML; however, while Abs bind with high specificity they may have limited efficacy as a therapeutic. ADCs and bispecific immune-engaging Abs may have on-target but off-leukaemia effects and unwanted immunological reactions. [107] Problems with toxicity may be alleviated with further investigation and the implementation of novel strategies and novel small-molecule compounds targeting CD33.

1.3. *THE SIGLECS AND CD33*

Cluster of Differentiation (CD) are cell surface antigens expressed on cells of the immune system. [108] CD33 is a member of the sialic acid-binding immunoglobulin-like lectins (Siglecs) receptor family. There have been 15 human Siglecs identified to date (Figure 1-6). Siglec-1, 2, 4 and 15 form a distinct evolutionary group and are conserved across all mammals. Siglec-3, also known as CD33, was the first Siglec identified and was originally studied as a myeloid lymphoma marker. CD33 and Siglecs 5 - 11, 14 and 16 share high sequence homology and are collectively referred to as “CD33-related Siglecs”. [109, 110]

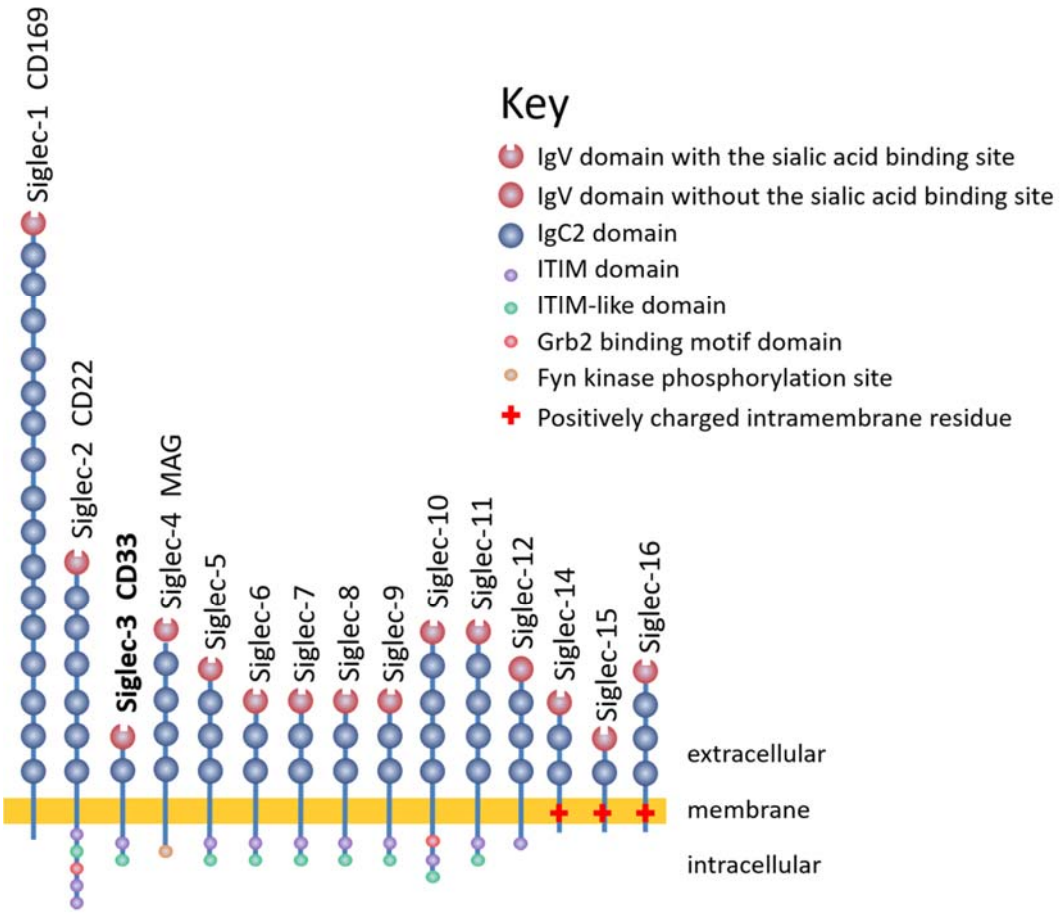


Figure 1-6 Schematic of the structure of human Siglecs and their nomenclature. See key for symbols representing the various domains. The structure of the family members differs by various additions and deletions such as the number of extracellular domains, intracellular immunoreceptor tyrosine-based motifs (ITIMs), a positively charged intramembrane residue or the absence of a sialic acid recognition site (Siglec-12).

While the precise physiological role of CD33 is unclear, the main role of other Siglecs in the immune system appears to be sialic acid recognition, which is required for the differentiation of self and non-self. [111] Sialic acids are negatively charged, nine-carbon monosaccharides, located on the terminating branches of N-glycans, O-glycans and glycosphingolipids. They display wide biological diversity, somewhat due to the differing α linkages to the carbohydrates. They also have an array of natural modifications such as an N-acetyl or hydroxyl group. [112] Sialic acids are highly expressed on outer cell membranes, on secreted glycol-proteins, and on the interior of lysosomal membranes. They play a role in the stabilisation of membranes and transmembrane molecules and

are associated with normal development and immunity, intracellular signalling and host-pathogen interactions. [113]

Siglecs are single-pass type 1 transmembrane proteins and all Siglecs have an extracellular N-terminal immunoglobulin IgV domain. Apart from Siglec-12, all the IgV domains bear the sialic acid-binding site, which contains a conserved arginine. Siglec-12 has an arginine to cysteine mutation in the IgV domain, which removes its sialic acid-binding ability. The conserved arginine forms a critical salt bridge with the carboxylate group of the bound sialic acid. Although sialic acid-binding is dependent on this conserved arginine, structural variability in the binding site confers each Siglec with distinct specificity for differing sialic acid linkages and extended glycan structures which is important when developing isoform specific small molecules. [114-117]

The Siglec extracellular IgV domains are followed by varying numbers of IgC2 domain repeats (Figure 1-6), whose function is unclear, but seem to be effectively spacers that project the sialic acid-binding site away from the cell membrane. [118] The intracellular region of most Siglecs contains a combination of tyrosine motifs composed of one or more immunoreceptor tyrosine-based inhibitory motif (ITIM), an ITIM-like motif, a Grb2-binding or Fyn kinase phosphorylation site (Figure 1-6). The intracellular motifs, when phosphorylated by SRC kinases, recruit and activate SRC homology-2 (SH-2) domains containing tyrosine phosphatases such as SHP-1 and SHP-2. The suppressor of cytokine signalling 3 (SOCS3) competes with SHP-1 and SHP-2 for binding and recruits the ECS (Elongin B/C-Cul2/Cul5-SOCS-box protein) E3 ubiquitin ligase complex which leads to regulation of endocytosis, reduced myeloid cellular activation, proliferation and ultimately proteasomal receptor degradation (Figure 1-7). [97, 104, 119]

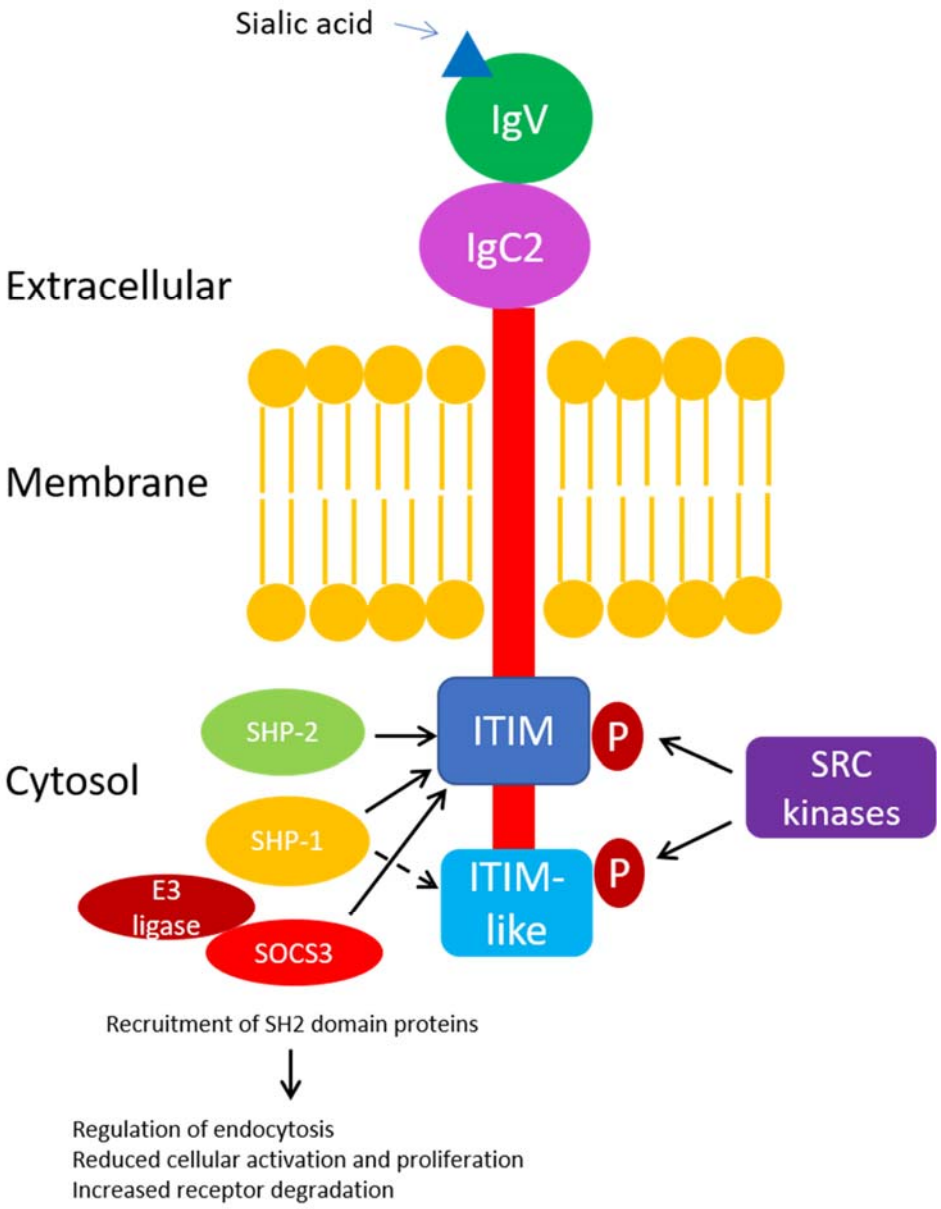


Figure 1-7 Schematic structure of CD33 showing a sialylated ligand binding to the extracellular IgV domain. The protein contains a single transmembrane region, shown in red, and the cytoplasmic portion contains the ITIM and ITIM-like domains which are phosphorylated by SRC kinases. Once phosphorylated the ITIMs act as docking sites for SHP-1 and SHP-2. SOCS3 competes with SHP-1 and SHP-2 and forms an E3 ubiquitin ligase complex.

Most Siglecs are located on a limited set of haematopoietic cells: three exceptions being Siglec-4, which is expressed on oligodendrocytes and Schwann cells, Siglec-6, which is expressed on placental trophoblasts, and Siglec-11, which is present on ovarian stromal fibroblasts. [120-122] Almost all cell types in the immune system express at least one type of Siglec, with some cells expressing multiple types (Table 2).

Table 2 Expression pattern for each Siglec in normal individuals.

Cell type	Siglec
Macrophage	1, 9, 11, 12, 14, 16
B cell	2, 5, 6, 10,
Monocyte	3, 7, 9, 14
Myeloid precursor	3
Oligodendrocyte	4
Neutrophil	5, 9, 14
Trophoblast	6
Natural killer cells	7
Eosinophil	8
Basophil	8
Mast cell	8
Dendritic cell	1, 10
Epithelial cell	12
Osteoclast	15

Most Siglecs are endocytosed when bound by Abs or glycan ligands. This function seems to be important for the clearance of sialylated antigens and in promoting antigen presentation. [123, 124] The endocytic properties of Siglecs, in particular CD33, when targeted with a cytotoxic compound and their restricted expression pattern make them attractive therapeutic targets.

1.3.1. Extracellular structure of CD33

The structure of the human CD33 extracellular domain (i.e. the IgV and IgC2 domains), solved by X-ray crystallography at a resolution of 2.24 Å, was deposited in the PDB in 2016 by Dodd et al. (PDB ID: 5IHB, unpublished). A higher resolution (1.78 Å) structure of the IgV domain in isolation has also been solved by our laboratory (PDB ID: 6D48, Miles et al.) (Figure 1-8). [125]

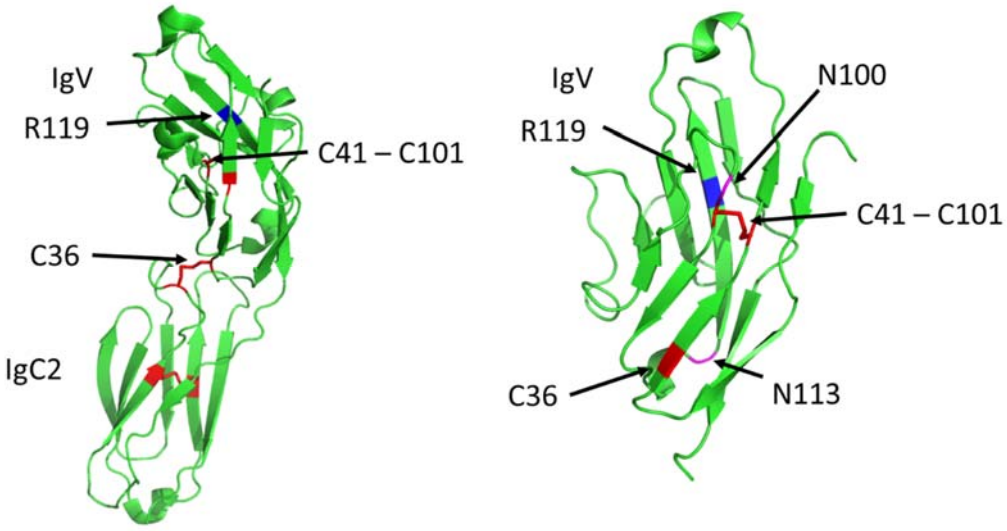


Figure 1-8 The X-ray crystal structure of the extracellular domains of human CD33. Cartoon representations of the extracellular CD33 IgV and IgC2 domains (PDB ID: 5IHB, left) and the higher resolution IgV domain (PDB ID: 6D48, right) showing the location of the disulfide bonded C41-C101 and C36 (red), the R119 sialic acid-binding residue (blue) and two potential N-linked glycosylation sites N100 and N113 (pink).

The unit cell of the extracellular CD33 structure (PDB ID: 5IHB) contains four molecules of CD33 (i.e. Chains A – D) in two different homodimeric arrangements. One dimer has the IgC2 and IgV domain of Chain B interacting with the IgV and IgC2 domains of Chain C, respectively (i.e. in a head-to-tail arrangement). Whereas the second dimer in the unit cell has the IgC2 domain of Chain A interacting with the IgC2 domain of Chain D in a parallel orientation and the IgV domains of both chains are adjacent to each other, although there is little interaction between them (Figure 1-9).

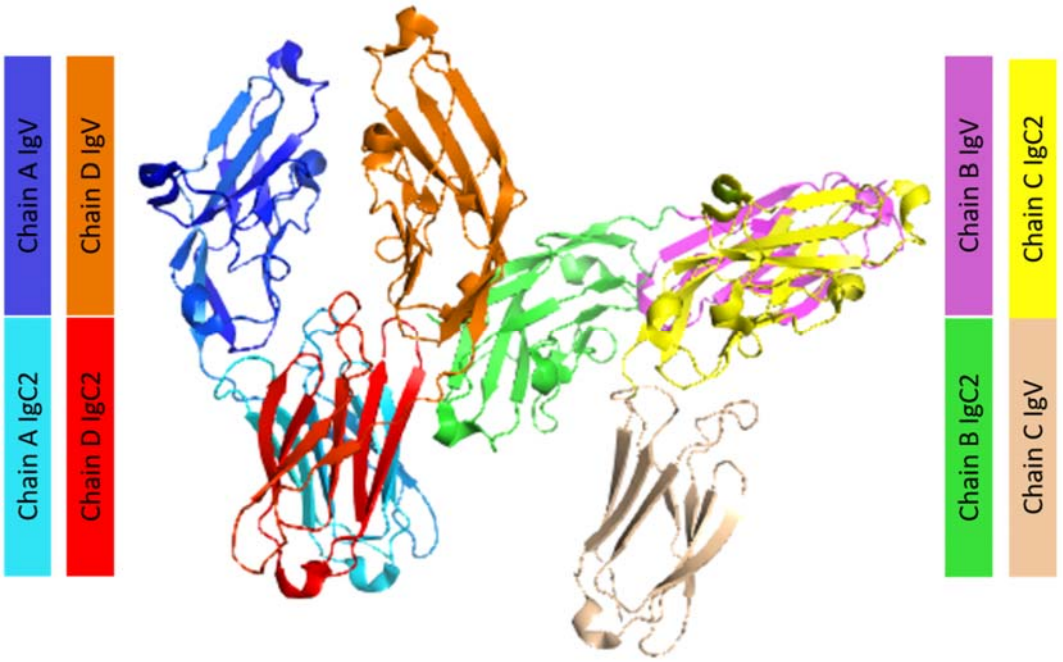


Figure 1-9 The two homodimeric arrangements of CD33. Chains are coloured by subunit showing the parallel arrangement of chains A and D (left, light blue, dark blue, brown and red) and the head-to-tail arrangement of chains B and C (right, green, pink, yellow and beige).

The homodimer formed by Chains A and D (Figure 1-9, Figure 1-10) may have a functional relevance as other Siglecs, e.g. Siglec-4, are known to cluster on the cell surface to bind multi-sialylated ligands and then illicit a functional response. CD33 may also act in the same way.

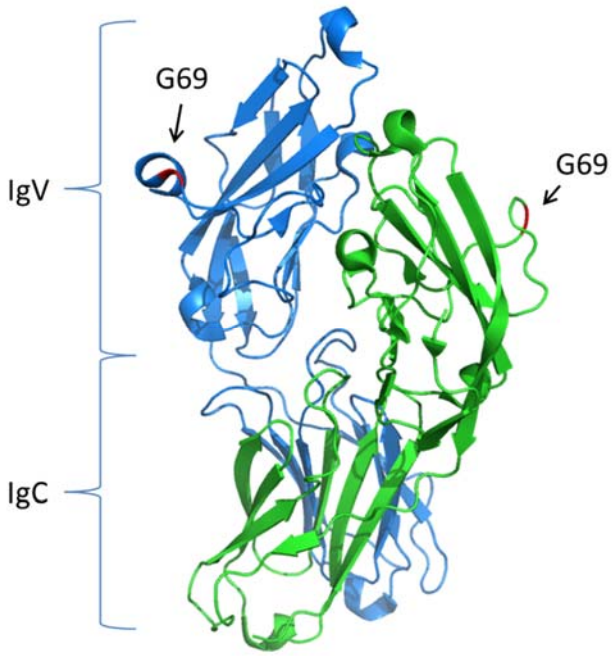


Figure 1-10 Structure of the human CD33 IgV and IgC domains in a potentially biologically relevant homodimer. (PDB ID: 5IHB) In Chain A (blue) residue G69 (red) is part of a helix whereas in Chain D (green) the residue is part of a loop.

The CD33 IgV domain consists of a typical I-type lectin fold formed by 11 β-strands assembling into two β-sheets and contains the critical sialic acid-binding arginine residue, R119 (Figure 1-8). There are three cysteine residues in the IgV domain, two form an intramolecular disulfide bond (C41 – C101) and the third, C36, forms an interdomain disulfide with C169 of the IgC2 domain (Figure 1-8). The intramolecular disulfide bond connects the two β-sheets of the IgV domain and is structurally important for maintaining the correct protein fold. Two N-linked glycosylation sites are present on the CD33 IgV domain. The N-linked glycosylation at N100 is conserved across most of the Siglecs and has been shown to be critical for ligand recognition. Mutagenesis studies have shown that deletion of the equivalent N-linked glycosylation site in CD22 inhibits its sialic acid-binding ability, whereas the same deletion in CD33 unmasks its binding ability. Sialylation of N100 of CD33 has been shown to inhibit ligand binding, potentially by compromising the adjacent disulfide bond (i.e. C41 - C101). [126, 127] The IgC2 domain of CD33 has high sequence and structural fold similarity to the Ig constant region. [123] There are two antiparallel β-sheets, each containing three β-strands (Figure 1-8). Like the IgV domain, the IgC2 domain contains three cysteine residues, all of which are involved in disulfide bonds. An intramolecular disulfide bond is formed by C163 – C212, while C169 forms the interdomain disulfide with C36 of the IgV domain (as

described above). As for the IgV domain, the intramolecular disulfide bond is structurally important and connects the two β-sheets of the IgC2 domain.

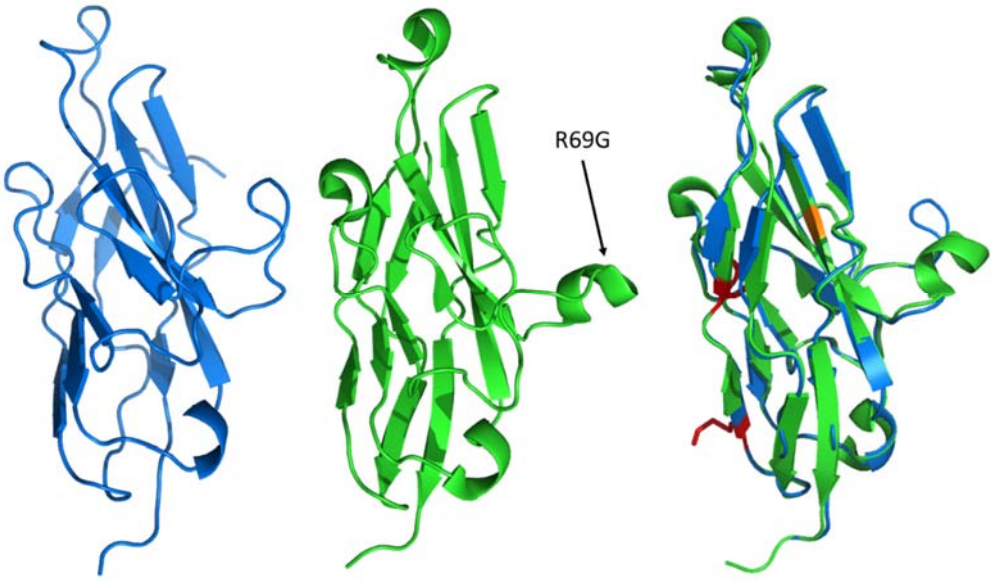


Figure 1-11 Comparison of the wild-type and R69G CD33 IgV domains. The structures of the high resolution wild-type IgV domain structure of Miles et al. (PDB ID: 6D48, blue) and the IgV R69G variant by Dodd et al. (PDB ID: 5IHB, green) were overlaid for comparison. In the overlay (right), the cysteine residues are shown in red and the R119 sialic acid-binding residue in orange.

The extracellular domain structure deposited by Dodd et al. (PDB ID: 5IHB) was expressed in human cell culture, whereas the IgV protein used by Miles et al. [125] was refolded from insoluble material produced recombinantly in *E. coli.* Using a eukaryote expression system would result in glycosylation that may influence the conformation of the protein structure. In addition to the different expression systems used, the Miles et al. IgV domain used the canonical wild-type sequence while Dodd et al. used a natural variant containing the mutation R69G (rs2455069) (Figure 1-10, Figure 1-11). [128] There is a slight difference in the conformation adopted by residues 64 – 70 in the IgV domain between the two crystal structures. Two of the four IgV domains in the Dodd et al. unit cell contain a small helix (Chains A and C), while the other two have a loop (Figure 1-10). The wild-type sequence, with the larger arginine residue at position 69, maintains the loop conformation by forming a salt bridge with D70. Glycine is known as a helix breaker [129], however in this instance it seems to encourage helix formation possibly

by increasing the flexibility of the R69G region during protein folding. Although different sequences and expression systems were used to obtain the two crystal structures, a comparison of their IgV domains demonstrated that they are very similar with a root-mean-square deviation (RMSD) of 0.3 over all Cα atoms. The RMSD is a measure of the average distance between specified atoms of the two structures, in this case the Cα atoms. [130]

The structure of CD33 has been extensively characterised. This structural information combined with its expression in various subtypes of AML and the known clinical benefit demonstrated by GO therapy validates its use as a therapeutic target to treat AML.

1.4. *PROSTATE CANCER*

Prostate cancer (PCa) is the most commonly diagnosed male cancer and is a leading cause of death in most western countries. It is estimated that approximately 16,700 new cases of PCa will be diagnosed in Australia in 2020. [131-135] In spite of high morbidity rates associated with PCa, occurrence is so frequent and onset occurs relatively late in life, such that PCa is a primary factor in morbidity of a small proportion of patients. [136] The 5-year survival rate for localised PCa is more than 95%. Once the cancer metastasises however, the survival rate drops to below 20%. [134] It is estimated that 80% of men over the age of 70 have PCa and this increases to 100% once over the age of 80. [137] The greatest risk factor, therefore, is age. Due to the slow growing nature of PCa the general approach for localised cancer is active surveillance. [138] The most common treatment, if required, is a combination of surgery and radiation. 15 – 35% of PCa patients develop metastasis when tumour cells migrate from the prostate and proliferate in bones. [139] Treatment options for metastatic PCa include androgen deprivation therapy (ADT) and chemotherapy. [140]

Treatments for metastatic disease have a detrimental effect on general health and are hampered by the development of drug resistance. Therefore, diagnosis and estimation of prognosis is important for choosing treatment strategies although there remains no reliable or widely used diagnostic test to distinguish high-risk tumours at an early stage. [141]

Elevated serum levels of prostate-specific antigen (PSA) are used to screen for the presence of PCa, however PSA is produced by both malignant and benign cells and PSA blood tests cannot discriminate clinically important cancers from low-risk tumours. Accurate diagnosis requires transrectal ultrasound-guided biopsy of the prostate, which is an invasive procedure that results in frequent minor complications such as haematuria but also carries the risk of infection and long-term erectile dysfunction. [142]

The investigation of cell surface receptors as diagnostic markers and potential therapeutic targets for PCa has, as for AML, rapidly increased. Prostate membrane specific antigen (PSMA) and the prostate stem cell antigen (PSCA) are present on approximately 90% of human prostate tumours and their expression directly correlates with tumour stage. [143, 144] ProstaScint™ is an ^{111}In-labelled anti-PSMA mAb licensed by the FDA for imaging and staging of PCa. [145] The targeting of PSMA for diagnostic purposes validates its use as a therapeutic target and several clinical trials have looked at the immunogenicity of PSMA peptides for vaccine development. [146]

Human epidermal growth factor receptor 2 (HER2) gene amplification in PCa is directly associated with pathological stage and is a potential diagnostic and therapeutic target. [147] Trastuzumab is currently used to treat HER2 positive metastatic breast cancer and although its use for the treatment of hormone resistant PCa in phase II trials was unsuccessful, its investigation as an adjunct therapy is ongoing. [148, 149]

Proteomic analysis of microvesicles released by metastatic PCa cells showed an enrichment in the cell surface receptor CD151. [150] Elevated levels of CD151 expression are associated with a high tumour grade in a range of cancers. Expression of CD151 was found to be higher in poorly differentiated PCa, which gives it a better prognostic value than the traditional Gleason grading, which is based on the examination of tissue retrieved via biopsy. [151] As such CD151 has clear potential as a prognostic biomarker.

In the transgenic adenocarcinoma of the mouse prostate (TRAMP) PCa model the number of lung metastases was reduced when CD151 was knocked out, whereas the effects of CD151 knock out on the growth of individual metastatic colonies was minimal. [152] This suggests that rather than altering growth rate, CD151 effects tumour dissemination and may be a therapeutic target to prevent cancer metastasis, which is the leading cause of morbidity in PCa patients.

1.5. *CD151 IN PROSTATE CANCER METASTASIS*

PCa progression involves changes in both the extracellular matrix (ECM) underlying prostate epithelial cells and in the cellular receptors for ECM ligands. CD151 is involved with cell adhesion and motility by the formation of PPIs with the ECM ligands laminin-binding integrins. CD151 and integrin expression becomes aberrant in cancer cells. It is likely that these changes are what leads to loss of cellular adhesion and detachment, invasion of the ECM and migration into the vasculature, lymphatics or peritoneal space and eventually metastasis to distant tissues. [153-157]

Integrins are a 24-member family of heterodimeric receptors that play an important role in the regulation of cellular adhesion and migration. Integrin dimers are composed of an α and β subunit, drawn from a pool of 18 different α and 8 different β subunits. [158] Individual subunits are comprised of a large extracellular domain, a transmembrane domain and a cytoplasmic tail. Each of the 24 integrins are tissue specific and are known to regulate cell adhesion, migration and intracellular signalling via recognition of ECM proteins. Several integrin receptors are expressed in normal prostate glands including collagen receptors, laminin receptors and fibronectin receptors. [159-162] The laminin binding α3, α6 and α7 subunits of integrin are the most highly conserved and play a role in normal and pathological conditions. [156]

Laminins are a class of ECM proteins that serve as the major adhesive proteins and mediate cell adhesion to basal membranes. Over 16 laminin isoforms have been identified and are composed of trimers of α, β and γ chains. Laminin nomenclature is based on the trimeric combination e.g. laminin-332 is an α3β3γ2 trimer, also called LM-332. The expression of laminin isoforms differs among tissue types and at differing developmental stages, suggesting that the isoforms are functionally distinct. [163-165] When cells become malignant, the expression and/or function of receptors such as integrin can lead to changes in motility and invasiveness.

Attachment of epithelial cells to the substratum in normal prostate cells occurs via integrins α3β4 and α3β1 to LM-332. Integrin is activated by the binding of the receptor CD151, which induces a conformational change in the integrin allowing it to bind to LM-332 in the ECM. In invasive PCa, LM-332 expression is down regulated or lost [159, 165] and only the laminin binding integrins α6β1 and α3β1 are expressed. [166] This loss of LM-332 may be one of the key events that enables the spread of prostate tumour cells. While α3β1 integrin usually only binds to LM-332, α6β1 integrin binds to a wider array of laminin isoforms, including LM-511 (Figure 1-12). [163] The sustained expression of α6β1 integrins in PCa cells may provide a growth advantage to prostate carcinomas by enabling them to bind LM-511. LM-511 is abundant in the perineurium of the nerves that innervate the prostate gland, which is a route of extra prostatic escape for invasive prostate carcinoma cells. [156, 167] Overall there is accumulating evidence that CD151 activated laminin-binding integrins regulate PCa progression.

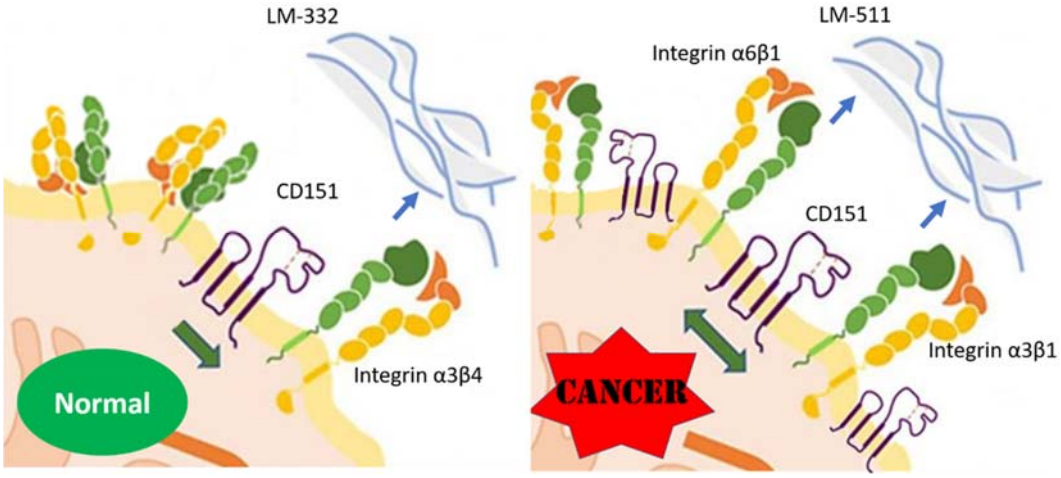


Figure 1-12 Schematic of some of the protein expression changes that occur in prostate cells when the cells become malignant. In normal cells, (left) CD151 binds to integrin α3β4 which induces a conformation change from the integrin's folded, inactive state, to an active upright state where it binds to LM-332. In cancer cells, (right) CD151 expression is upregulated, integrin expression is reduced to α3β1 and α6β1 isoforms only and LM-332 expression is lost. As a consequence, LM-511 becomes the integrin binding partner. Figure adapted from Vences-Catalán and Levy 2018. [168]

Integrins α3β1, α6β1 and α7β1 have been shown to be the predominant laminin binding integrins and form the strongest association with a group of cell surface receptors known as the tetraspanins. [169, 170] CD151 is a member of the tetraspanin receptor family. Residues 186-216 and 195-205 located in the large extracellular loop (LEL) of CD151 were identified by Yauch et al. (2000) and Berditchevski et al. (2001), respectively, as the key integrin interaction sites. [171, 172] Using mutagenesis studies, Kazarov et al. (2002) narrowed the identity of the α3 integrin binding site on the CD151 LEL to residues 194-196, the QRD motif. [173] The CD151 QRD-α3β1 integrin interaction was demonstrated to be high affinity as it was resistant to Triton X-100 disruption. Other CD151 LEL interactions, such as with itself, other integrins or tetraspanins, were Triton X-100 sensitive and shown to be independent of the QRD site. Mutating QRD to INF disrupted α3 and α6 integrin binding.

Further mutagenesis studies by Zevian et al. (2011) [174] closely examined the intricacies of the CD151-integrin interaction. Compared to wild-type CD151, a CD151Palm mutant, in which the six membrane-proximal intracellular cysteine residues that are modified by palmitoylation were mutated to serine residues, binding to α3 integrin was unchanged. The binding of the CD151Palm mutant to the tetraspanin CD9 however, was disrupted. Repeating the QRD to INF mutation of Kazarov et al. (2002) confirmed the lack of integrin binding in the presence of Triton X-100, however binding

still occurred with the weaker detergent Brij 96V. CD151 expressed with the LEL of another tetraspanin, TSPAN7, which is known to have little interaction with α3 integrin, and CD151VR containing a more limited domain swap with TSPAN7 from S158-G207, showed only weak association with α3 integrin. The mutants and CD151 wild-type were all expressed in A431 cells (a human squamous carcinoma cell line) that had been stably silenced for CD151 expression (A431 sh3). The A431 sh3 cells did not bind α3 integrin. CD9 binding to CD151 was not affected by any mutation other than CD151Palm. In addition, CD9 binding to the tetraspanin CD81 was not affected by any of the CD151 mutations or in the A431 Sh3 cells. This study clearly demonstrates that while the CD151 QRD site is important for α3 integrin binding, it is not essential. It also demonstrates that interactions with other tetraspanins are not disrupted by the loss of CD151 or mutations to the CD151 LEL. Yamada et al. (2008) have also shown that the CD151 segments 185-CKTVVALC-192 and 176-GG-177 are involved with the α3 integrin interaction. As mutating these segments of the CD151 LEL removes two disulfide bonds (i.e. C156-C185 and C184-C192), these findings also suggest the possibility that correct folding of CD151 is required for interaction with integrins.

Kazarov et al. (2002) also demonstrated that the CD151 QRD to INF mutant disrupted α3 and α6 integrin dependant cell motility. Targeting the CD151 QRD motif with a mAb or with miRNA-199-3p has been shown to decrease the motility of hepatocellular carcinoma cells, but not reduce cell proliferation. [175, 176] These studies suggest the potential in specifically targeting the CD151 QRD-integrin interaction to disrupt the aberrant behaviour of cancer cells, without disrupting normal function.

The cell surface receptor CD151 therefore presents as an attractive target for cancer therapy. The concept of targeting tetraspanins with mAbs to treat cancer was introduced when the tetraspanins were originally identified. The first anti-tetraspanin mAb with an antiproliferative effect was directed at CD81 on a human lymphoma cell line in 1990. [177] CD151 was first identified in platelets and endothelial cells using a mAb raised against human AML cells. [178] Other mAbs have been generated and used to help understand the function of CD151, with several of them exhibiting the ability to decrease metastasis *in vivo*. [179-181] The majority of the mAbs that target tetraspanins are specifically directed at amino acids located in the variable region of the LEL.

The mechanisms of action for anti-CD151 mAbs are varied. mAb 8C3 dissociates CD151 from integrin α3β1 and attenuates the binding of integrin α3β1 to LM-511. [182] mAb 1A5 shows a strong inhibitory effect on metastasis by preventing tumoural cell detachment and inhibiting invasion of the stroma, possibly through inhibiting integrin activation. [181] Modulation of cell-cell adhesion is induced by mAb 11B1G4 which blocks epithelial-mesenchymal transition, a key step in metastasis. [183] Anti-CD151

mAbs could block cancer at multiple stages, including tumour growth and metastasis by affecting ligand binding such as the CD151-α3β1 integrin interaction. However, CD151 has a widespread distribution on numerous cells and a broad distribution in normal tissue [184] and targeting CD151 may affect the physiological function of these cells.

Patients with a single nucleotide insertion in exon 5 of the CD151 gene, leading to the translation of a truncated protein lacking most of the LEL and hence the integrin binding domain, have been identified. The phenotype arising from this mutation includes sensorineural deafness, fragile skin and anaemia, suggesting that CD151 has functional significance in the skin and inner ear and has a role in erythropoiesis. [185] However; the physiological significance of CD151 is unclear as renal defects are observed in some CD151-null mice [186, 187], although other CD151-null mice appear normal, healthy and fertile. [188]

There are currently many patents for anti-CD151 mAbs or Ab fragments for use in treating cancer, specifically inhibiting metastasis. The targets are varied and include CD151 amino acids 118-180 of the LEL [189], the conformational structure of the LEL [179, 190] and amino acids 113 – 221 of the LEL. [190] However, results of any pre-clinical studies, such as toxicology and pharmacokinetic and pharmacodynamic studies in non-human primates are not yet publicly available.

The specific QRD sequence on the CD151 LEL that is key to the interaction between CD151 and integrins α3β1 and α6β1, is a potential target for the development of small molecules that are selective for CD151 and can inhibit PCa metastasis. Development of therapeutic mAbs requires complex processes and a large financial input. [51] As previously described in section 1.1.4., small molecule development can be much more efficient than developing biological agents to target a specific protein. In addition, small molecules, typically 500 Da or less, can confer more favourable tissue penetration and pharmacokinetics to drug discovery initiatives targeting CD151.

Structure-based assessment of a target, such as the QRD sequence in the CD151 LEL, can provide insights into the druggability of a protein, that is, the binding site must favour interaction with a drug-like compound that can affect biological function. [191, 192] High affinity binding of small molecules to the target protein typically involves intimate interactions with amino acid ligands provided by the 3D fold of the target protein; however, no 3D structures of CD151 are currently available to guide the discovery and development of potential small-molecule ligands. It is therefore necessary to examine existing structures and structure-activity relationships in the tetraspanin family to glean insights into the structure and function of CD151.

1.5.1. The tetraspanin superfamily

Tetraspanins, or the transmembrane 4 superfamily (TM4SF) are small (200 – 350 amino acid) membrane glycoproteins that are expressed in species ranging from marine sponges to mammals. The family is ancient, the first member appearing 570 million years ago and tetraspanins are widespread amongst eukaryotes. [193] There are 33 tetraspanin genes in the human genome, with some members of the family displaying a wide tissue distribution and others being limited to specific tissues; for example, cells in the immune system, brain and tumours. The first protein belonging to this family, ME491/CD63, was characterised in 1988 and hallmark protein motifs were reported in 1990. [194] Overall, proteins in the tetraspanin superfamily are closely related and share 20 – 30% identical protein residues, and contain highly conserved cysteine residues, suggesting that all members share a conserved tertiary structure.

1.5.2. Structure of the tetraspanins

The tetraspanins are characterised by four hydrophobic transmembrane regions (TM1-4) and two extracellular domains: known as the Large and Small extracellular loops (LEL and SEL), a small intracellular loop and cytoplasmic N- and C-termini (Figure 1-13). [195-198] The LEL located between TM3 and TM4 contains four conserved cysteine residues, two in a conserved CCG motif known as the tetraspanin signature. [199] Although several classes of proteins share the same transmembrane topology and conserved residues; they lack the tetraspanin CCG signature so are not considered part of the tetraspanin family.

The crystallographic structure for the tetraspanin CD81 LEL, the only human TM4SF crystal structure reported to date, shows a five-helix bundle stabilised by two disulfide bridges, and sequence homology suggests that the gross structural features of CD81 may be largely conserved among tetraspanins (Figure 1-13). [200, 201] This is supported by the conservation, in all tetraspanins, of the four cysteine residues that form disulfide bonds in the CD81 structure, together with conservation of key structural determinants such as Y127, H151, G158, P176 and I194. [196]

The LEL seems to be organised into two subdomains: the first has a structurally conserved fold, the second has high heterogeneity, is variable in size, secondary structure and fold. The variable subdomain is located within the conserved subdomain and forms a “head” region. Their relative topology is governed by the occurrence of the key disulfide bridges (Figure 1-13). [202] The exposed location and variability of this segment suggests that part of the specific activity of tetraspanins is determined by this LEL region. [171, 203, 204]

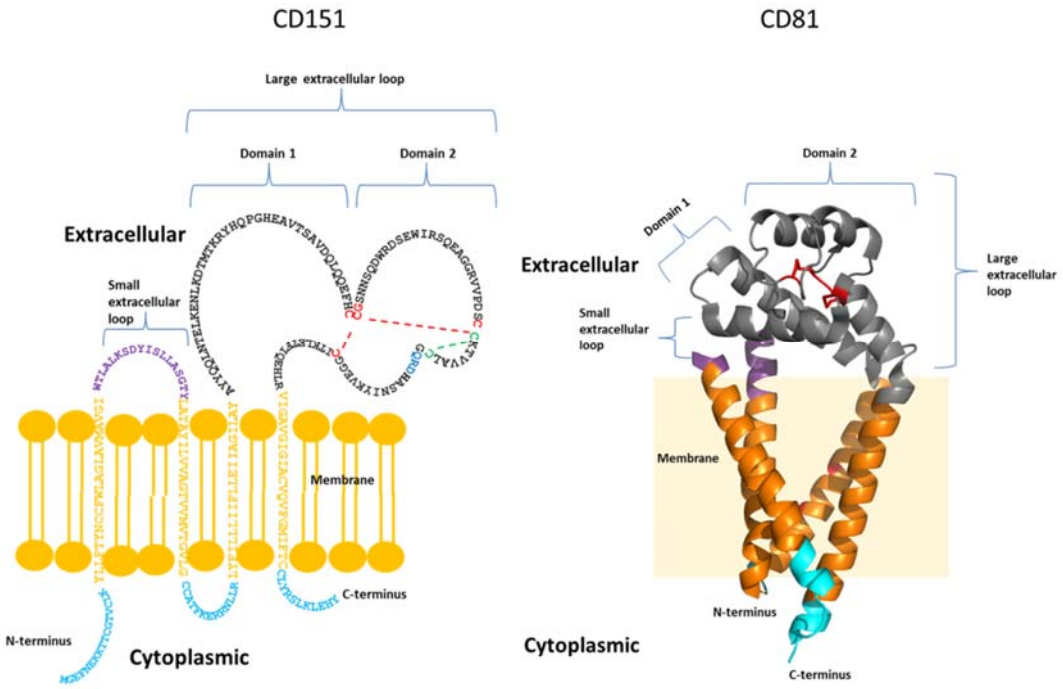


Figure 1-13 Basic schematic of the proposed structure of CD151 and comparison with the known crystal structure of CD81. All tetraspanins are composed of four transmembrane domains (TM1 – 4, orange), a small extracellular loop (SEL, purple) and a large extracellular loop (LEL, black). The conserved CCG motif (red) with the conserved cysteines (red) and the two additional cysteines (green) in the LEL have been highlighted. (left) In addition, the location of the CD151 QRD motif in the LEL is shown in blue. The published structure of CD81 (PDB ID: 5TCX) [201] with the cytoplasmic, transmembrane and extracellular domains colour coded as per the CD151 schematic. (right)

The disulfide bonds between the conserved cysteines in the LEL produce a sub-loop structure (Figure 1-13). [195] Additionally, two to four cysteines within this sub-loop exist in some tetraspanins and possibly participate in disulfide bonding. The remainder of the LEL region shows greater sequence conservation, approximately 20%. [195]

The CD81 crystal structure published by Zimmerman et al. [201] shows the pairing of TM1 with TM2 and TM3 with TM4, the pairs forming a cone shaped conformation with the helices close together at the inner membrane and separated at the outer membrane (Figure 1-13, Figure 1-14). Polar residues in TM1, TM3 and TM4 are predicted to stabilise the transmembrane domain packing by hydrogen bonds, polar interactions and/or interactions with cholesterol. The crystal structure of CD81 shows a cholesterol-binding pocket within the transmembrane domains (Figure 1-14). Membrane cholesterol is involved with the arrangement of tetraspanin microdomains on the cell surface and is required for the formation of tetraspanin-tetraspanin complexes. [205, 206]

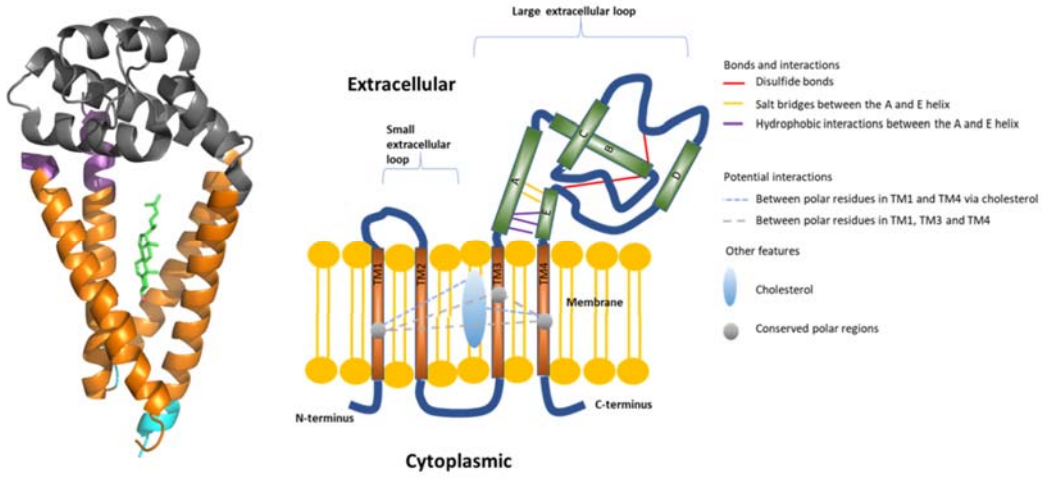


Figure 1-14 The X-ray crystallographic structure of CD81 (PDB ID: 5TCX)[201] showing the cholesterol-binding pocket. Cholesterol is shown as green coloured sticks. (left) A schematic of the proposed general tetraspanin structure, showing potential interactions that stabilise the transmembrane and large extracellular domains. (right) Figure adapted from Deventer et al. 2017. [207]

The LEL is the most widely studied component of the tetraspanins. Its high sequence variability is predicted to contain various functional sites, which mediate PPIs with other membrane proteins and with ligands. [195]

1.5.3. Function

The main role of tetraspanins appears to be organising other proteins into a network of membrane microdomains known as the "tetraspanin web" or tetraspanin enriched microdomains (TEMs). [169, 198, 208] The tetraspanin web plays a role in cell-cell interactions and within this web are complexes in which tetraspanins have specific and high affinity associations with other proteins, including other tetraspanins. Pathologically, but nonetheless informative, human CD81 (hCD81) binds to the E2 envelope protein of Hepatitis C Virus (HCV). Within the variable subdomain of hCD81 an F186L mutation prevented high affinity binding of hCD81 to the HCV E2 protein, whereas a T163A mutation in the same domain increased binding. [203] Aside from pathogenesis, CD151 performs a number of physiologically important roles from which provide important clues about the structure and function of CD151.

Residues 173-SFQ-175 in the murine CD9 variable domain make an important contribution to the sperm fusion activity of oocytes. [209] The SFQ residues are in the

same predicted area of the CD9 LEL as the F186 residue of CD81. High affinity association of CD151 (also known as SFA-1 and PETA3) with integrins is dependent on the 194-QRD-196 site in the CD151 LEL. The CD151 QRD residues are also in the same region of the LEL variable domain as SFQ in CD9 and F186 in CD81. Mutation of the QRD motif in CD151 causes loss of integrin binding and disrupts integrin dependent cell motility. [173]

A secondary interaction network occurs in which proteins associate via palmitoylated tetraspanins acting as linker proteins. The association of lipids, such as gangliosides [210] and cholesterol contribute to the formation of larger tetraspanin complexes. [169, 198] Tetraspanins can associate with integrins and other transmembrane proteins and with signalling enzymes such as phosphatidylinositol-4 kinase protein kinase C. [170, 211, 212] As such, the tetraspanin web provides a foundation for membrane protein signalling (Figure 1-15).

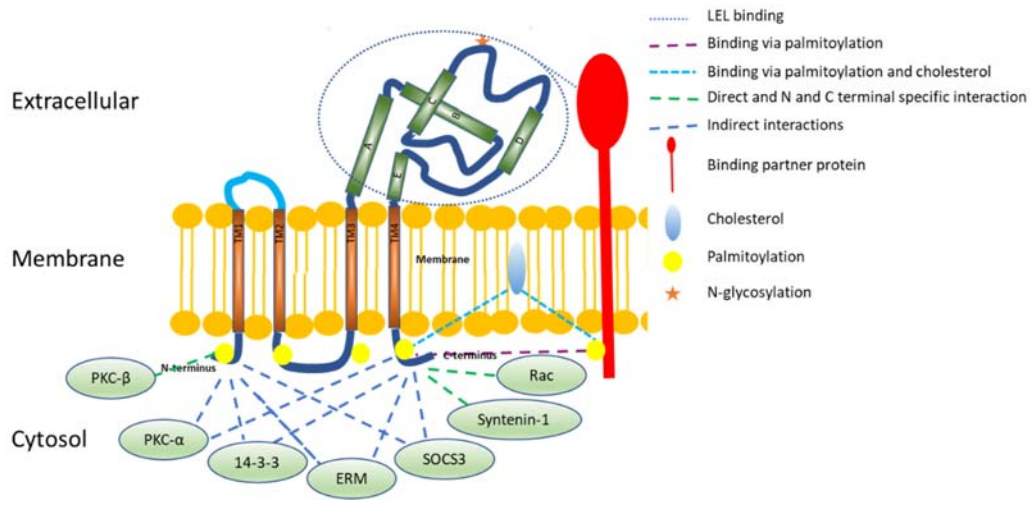


Figure 1-15 Schematic representation of tetraspanin interactions. The binding partner protein could be another tetraspanin or a cell surface protein such as integrin. Tetraspanin interactions are mediated by palmitoylation, cholesterol and direct PPIs.

Many protein signalling networks regulate cell proliferation, motility and survival, and the changes that occur in cancer cells are the result of alterations in cellular drivers of these pathways. Understanding the intricacies of protein signalling networks is crucial

for our understanding of tumour cell behaviour and this knowledge can lead to new strategies for cancer therapy.

1.5.4. The role of CD151 in cancer

CD151 was the first tetraspanin identified as playing a role in cancer development. Its role in the promotion of cellular migration and invasion has been demonstrated in *in vitro* and *in vivo* models. [179, 180, 213-215] CD151 involvement in the initial stages of tumour development has been demonstrated in a diverse range of biological contexts.

The most studied, and seemingly the most important role of CD151 in cancer progression, is the interaction with integrin. [216, 217] The association of CD151 with laminin-binding integrins [218, 219] is crucial in cancer cell migration and invasion. [219-221] Integrins are heterodimeric cell surface receptors made up of non-covalently associated α and β subunits, which link the extracellular matrix to the cytoplasm. [158, 222, 223] Integrin receptors are tissue specific and regulate cell motility and intracellular signalling events by binding to their associated ligands in the ECM while simultaneously binding to intracellular signalling components. As previously stated in section 1.5., the integrin α3, α6 and α7 subunits bind to laminins, which are extracellular proteins important for cell differentiation, migration, adhesion and survival. [224] CD151 has been shown to modulate integrin binding and signalling and regulate cell motility. [182, 225, 226] The highly specific lateral interaction of the CD151 LEL QRD motif with α3, α6 and α7 integrins [169, 172, 219, 227, 228] are high affinity and stoichiometrically coupled. CD151-integrin interactions strengthen the attachment to the ECM. [182] The integrin β1 subunit is associated with the α3, α6 and α7 subunits and is involved with integrin activation, important for anchoring epithelial stem cells to basal membranes and may affect motility of tumour cells. [229]

In cells expressing both CD151 and laminin-binding integrins, CD151 is involved with all integrin-mediated cellular behaviour such as regulation of integrin-ligand interaction, integrin-triggered signalling, direction of integrin intracellular trafficking, recycling and their compartmentalisation on the cell surface. [211] High levels of CD151 are correlated with poor prognosis in a variety of tumours including epithelial malignancies such as carcinomas of the lung, [230] breast, [231-234] colon, [235] pancreas,[236] kidney, [215] liver, [237] oesophagus[238] and prostate. [151]

In the human lung adenocarcinoma cell line A549, a CD151-free pool of integrin α3β1 showed an impaired ability to interact with LM-511. [182] Nishiuchi et al. [182] concluded that an association with CD151 regulates the conformation of α3β1 integrin, which sustains it in its activated state. However, Yang et al. [231] found that removal of CD151 did not diminish the activity of the integrin β1 epitope. Studies in A549 cells have shown that CD151 is involved in the control of two independent integrin α3β1

functions: increased strength of α3β1-mediated cell adhesion and initiation of α3β1-stimulated signalling events involving tyrosine phosphorylation. [226]

CD151 also recruits signalling enzymes into integrin complexes, such as type II phosphatidylinositol-4 kinase [170, 239] or protein kinase C. [212] It is proposed that CD151 is a molecular linker between laminin-binding integrins and growth factor receptors such as epidermal growth factor receptor and c-Met [231, 240, 241] and also may function as a positive regulator of the transforming growth factor β. [232]

By modulating biosynthesis and activation of associated molecules like matrix metalloproteinases (MMPs), CD151 influences tumour invasiveness. [218] CD151-associated integrin signalling induces MMP9 expression [242] and MMP9 binding anchors MMP7 at the cell membrane. [243] In fact, CD151 has been found to participate in nearly all stages of cancer progression associating with numerous proteins involved in tumour progression.

It was recently found that the QRD amino acid motif on the LEL, although crucial for laminin-binding integrin associated cancer progression, is not essential for CD151-integrin association or for the ability of CD151 to promote several different integrin functions. [244] This suggests it may be possible to selectively target the CD151 QRD sequence and inhibit metastatic growth but leave stable cell attachments and normal cell function intact. The specificity of a high affinity small molecule targeting the QRD motif may limit the likelihood of side-effects associated with total CD151 inhibition.

While the intricate details of CD151 interactions with partner proteins are presently poorly understood it appears that modulating the interaction between CD151 and integrins could halt cancer progression in metastatic cancers arising from PCa. Mortality from PCa typically occurs once the cancer has spread outside the prostate. Inhibiting PCa metastasis, and reducing the need for curative therapy, would significantly increase longevity and quality of life for those afflicted. Small molecule-based drugs that target the QRD motif of CD151 are a promising treatment strategy for inhibiting metastasis of this common cancer.

1.6. *AIMS AND OUTLINE OF THIS THESIS*

The membrane proteins CD151 and CD33 are targets for cancer therapy. This thesis aims to use established knowledge of CD33 structure to guide a SPR based screen to identify compounds that bind CD33 IgV domain and are drug development candidates for the treatment of Acute Myeloid Leukaemia. Bioinformatics, recombinant protein production and an array of techniques to characterise protein solubility, conformation and structure will be used to further the understanding of the structure of CD151 LEL to guide a SPR based compound screen. This work will contribute towards structure-based drug design for the inhibition of prostate cancer metastasis.

Chapter 3 describes the development of a SPR assay to identify compounds that bind to the IgV domain of CD33. Various techniques are employed to investigate the optimal construct design, immobilisation methods and assay conditions to establish a reliable assay.

Chapter 4 outlines my findings while undertaking a bioinformatics investigation of CD151 and comparison with similar, structurally defined tetraspanins to guide recombinant protein expression and purification.

Chapter 5 focuses on the production of CD151 LEL recombinant protein, utilising an array of expression and purification tags, expression techniques and various purification methods. The recombinant protein produced was extensively characterised and ultimately shown to be adopting (supposed) non-native folding conformations that reduce stability and solubility.

Chapter 6 describes efforts to crystallise the CD151 LEL produced by the methods described in chapter 5. The limited number of crystals produced were not of sufficient quality for further analysis, however this chapter provides a foundation for future studies.

Chapter 7 expands on the assay development outlined in Chapter 3 and describes the development of a SPR based method for screening compounds that may bind to the QRD motif in the CD151 LEL.

Chapter 8 closes the thesis and is a general discussion addressing the many issues faced during this project including technical challenges, data analysis and critical evaluation of existing literature.

CHAPTER 2. MATERIALS AND METHODS

2.1. *MATERIALS*

2.1.1. Media, buffers and stock solutions

Buffers and stock solutions were prepared with Milli-Q® ultrapure water (Merck Millipore) unless otherwise specified. Sterilisation of large volumes (>100 ml) was by either filtration through a 0.22 µm nylon membrane filter (Merck Millipore) in a reusable polysulfone bottle top filter (Nalgene) or autoclaving at 121°C for > 30 minutes. Small volumes were sterilised by filtration through a 0.22 µm Minisart® syringe filter.

Luria-Bertani Broth (LB) was prepared by dissolving 10 g tryptone, 10 g NaCl and 5 g yeast extract (all Sigma-Aldrich) per litre then sterilised by autoclaving.

Rich Media (RM) was prepared by dissolving 10 g tryptone, 5 g NaCl and 5 g yeast extract per litre then sterilising by autoclaving.

LB agar was prepared as above with the addition of 15 g per litre agar (Sigma-Aldrich).

SOC Outgrowth media (New England Biolabs)

SF-900 II™ (Gibco) Insect cell media

Dulbecco's Modified Eagle's Medium - high glucose (Sigma-Aldrich)

Antibiotics: Ampicillin and *kanamycin* were prepared at 100 mg/ml, filter sterilised and frozen at -20°C in 1 ml aliquots. *Chloramphenicol* was prepared at 100 mg/ml in 100% ethanol and stored at -20°C.

Isopropyl β-D-I-thiogalactopyranoside (IPTG) was prepared at 1 M concentration, filter sterilised and stored at -20°C in 1 ml aliquots.

Gel Loading Dye, Orange (6x) (New England Biolabs) pre-mixed loading buffer with a tracking dye for agarose.

SDS-PAGE sample loading buffer (5x) 250 mM Tris-HCl pH 6.8, 10% SDS, 30% (v/v) glycerol, 0.05% (w/v) bromophenol blue.

SDS-PAGE reducing sample loading buffer as above with the addition of 100 mM 1,4-Dithiothreitol (DTT) or 5% by volume β-mercaptoethanol.

SDS-PAGE running buffer Bolt™ MES SDS running buffer (Life Technologies)

InstantBlue Ultrafast protein stain (Sigma-Aldrich)

Western blot Transfer Buffer Bolt™ Transfer Buffer (Life Technologies)

Western blot Blocking Buffer Odyssey® Blocking Buffer (Millennium Science)

NEB3 buffer (New England Biolabs) 100 mM NaCl, 50 mM Tris-HCl, 10 mM $MgCl_2$,1 mM DTT, pH 7.9 at 25°C.

Bacterial Cell Lysis Buffer 70 mM HEPES pH 7.5, 100 mM NaCl, 0.5 M 3-(1-Pyridinio)-1-propanesulfonate (NDSB), 5% glycerol, 1 mM PMSF, cOmplete EDTA-free Protease Inhibitor Cocktail (1 tablet per 50 ml).

Gel filtration buffer 70 mM HEPES pH 7.5, 100 mM NaCl, 5% glycerol.

IMAC binding Buffer 70 mM HEPES pH 7.5, 100 mM NaCl, 5% glycerol, 20 mM imidazole.

IMAC Elution Buffer 70 mM HEPES pH 7.5, 100 mM NaCl, 5% glycerol, 500 mM imidazole.

Biacore running buffer (HBS-P) 10 mM HEPES pH 7.5, 150 mM NaCl, 0.05% TWEEN®20.

QIAGEN Spin Miniprep kit (Qiagen)

QIAGEN Plasmid Mega kit (Qiagen)

QIAquick Gel Extraction Kit (Qiagen)

NotI-HF (New England Biolabs) High Fidelity (HF) Restriction enzyme

BamHI-HF (New England Biolabs) High Fidelity (HF) Restriction enzyme

2.1.2. Vectors

pUC57 (Genscript) is a double stranded closed circular high copy cloning plasmid isolated from *E. coli.* The vector length is 2710 bp and is isolated from *E. coli* strain DH5α. It contains multiple cloning sites; inserted genes are under the control of the *lac* promoter and the plasmid confers ampicillin resistance.

pET30a(+) (Genscript) is a double stranded closed circular high copy bacterial expression plasmid with a T7 promotor and restriction enzyme cloning. It is 5422 bp in length and confers kanamycin resistance.

pFastbac1 (Genscript) is a double stranded DNA transfer vector (from *E. coli* to expression bacmid) with polyhedrin promoter, restriction enzyme cloning for baculovirus/insect cell expression. It is 4775 bp in length and confers ampicillin resistance and gentamycin selection in insect cells.

pcDNA3.1(+) (Genscript) vector is designed for high-level, constitutive expression in mammalian cell lines. It contains the cytomegalovirus (CMV) enhancer promoter and

SV40 origin for episomal replication. It is 5428 bp in length and confers ampicillin resistance and neomycin selection.

pET-22b(+) (Genscript) vector carries an N-terminal *pelB* signal sequence for periplasmic localisation during bacterial expression. The 5493bp vector carries a T7 promotor and confers ampicillin resistance.

pGEX-6P-2 (Sigma-Aldrich) vector carries an N-terminal GST tag and a PreScission cleavage site. The 4985bp vector contains a tac promoter and confers ampicillin resistance.

2.1.3. Cell strains

2.1.3.1. Bacterial cells

DH5α (ThermoFisher Scientific) Competent cells for subcloning into plasmid vectors. Genotype: F- ϕ80lacZΔM15 Δ(lacZYA-argF)U169 recA1 endA1 hsdR17(rk - , mk +) phoA supE44 thi-1 gyrA96 relA1 λ-

MAX Efficiency® DH10Bac™ (ThermoFisher Scientific) Competent Cells for production of recombinant bacmids used in the Bac-to-Bac® Baculovirus Expression System. The DH10Bac™ *E. coli* strain contains a baculovirus shuttle vector (bacmid) that can recombine with a donor plasmid, pFastBac™.

BL21 (DE3) (New England Biolabs) Chemically competent *E. coli* cells for transformation and protein expression. Contains the T7 RNA polymerase gene under control of the *lacUV5* promoter. The expression strain is deficient in proteases Lon and Dompt and resistant to phage T1 (*fhu*A2) as it is derived from the *E. coli* B strain. Genotype: *fhuA2 [lon] ompT gal (λ DE3) [dcm] ΔhsdS λ DE3 = λ sBamHIo ΔEcoRI-B int::(lacI::PlacUV5::T7 gene1) i21 Δnin5*

Rosetta™ (DE3) pLysS (Novagen) Chemically competent *E. coli* cells for expression are BL21 derivatives designed to enhance the expression of eukaryotic proteins that contain codons rarely used in *E. coli.* pLysS strains express T7 lysozyme, which further suppresses basal expression of T7 RNA polymerase prior to induction. Genotype: F^- *ompT* $hsdS_B(r_B^- m_B^-)$ *gal dcm* (DE3) pLysSRARE (Cam^R)

SHuffle T7 Express (New England Biolabs) Chemically competent *E. coli* B cells engineered to form proteins containing disulfide bonds in the cytoplasm. Constitutively expresses a chromosomal copy of the disulfide bond isomerase DsbC. Genotype: *fhuA2 [lon] ompT ahpC gal* λatt::pNEB3-r1-*cDsbC* (SpecR, $lacI^q$) *ΔtrxB sulA11* R(*mcr-73::miniTn10*--Tet^S)*2 [dcm] R(zgb-210::Tn10* --Tet^S) *endA1 Δgor Δ(mcrC-mrr)114::IS10*

SHuffle T7 (New England Biolabs) Chemically competent *E. coli* K12 cells engineered to form proteins containing disulfide bonds in the cytoplasm. Constitutively expresses a chromosomal copy of the disulfide bond isomerase DsbC. Genotype: F´ *lac*, *pro*, *lacI*^{q} / *Δ(ara-leu)7697 araD139 fhuA2 lacZ::T7 gene1 Δ(phoA)PvuII phoR ahpC* galE (or U) galK λatt*::pNEB3-r1-*cDsbC* (Spec^{R}, *lacI*^{q}) *ΔtrxB rpsL150*(Str^{R}) *Δgor Δ(malF)3*

Lemo21(DE3) (New England Biolabs) Chemically competent *E. coli* BL21(DE3) cells containing the Lemo System™. This strain is deficient in proteases Lon and OmpT. Genotype: *fhuA2 [lon] ompT gal (λ DE3) [dcm] ΔhsdS/ pLemo*(Cam^{R}) *λ DE3 = λ sBamHIo ΔEcoRI-B int::(lacI::PlacUV5::T7 gene1) i21 Δnin5* pLemo = pACYC184-*PrhaBAD-lysY*

2.1.3.2. Insect cells

Sf21 (ThermoFisher) is an insect cell line used to isolate and propagate recombinant baculoviral stocks and to produce recombinant proteins. The cells originated at the USDA Insect Pathology Laboratory, where they were derived from the pupal ovarian tissue of the fall army worm, *Spodoptera frugiperda*.

Sf9 (ThermoFisher) is an insect cell line used to isolate and propagate recombinant baculoviral stocks and to produce recombinant proteins. The cells originated at the USDA Insect Pathology Laboratory, from the parental IPLBSF-21 (Sf21) cell line, which was derived from the pupal ovarian tissue of the fall army worm, *Spodoptera frugiperda*.

2.1.3.3. Mammalian cells

HEK 293T (ATCC) are Human Embryonic Kidney cells that stably express SV40 large T antigen; that acts as a helicase to replicate plasmids, such as pCDNA3.1, that carry an SV40 origin of replication, in transfected cells.

2.2. *METHODS*

2.2.1. CD151 DNA and plasmid constructs

The protein sequence of the topological LEL domain of human CD151 (TSPAN24) was obtained from the UniProtKB database (UniProt ID P48509 [113-221]). Plasmid constructs to be expressed in *E. coli* and insect cells were codon optimised, synthesised and cloned into commercial vectors by Genscript or in-house.

2.2.2. CD151 protein expression in bacteria

2.2.2.1. Transformation of chemically competent bacterial cells

A tube containing 50 µl of competent cells was thawed on ice for 10 minutes. 1 – 5 µl containing 1 pg – 100 ng of plasmid DNA was added to the tube and mixed. The mixture was incubated on ice for 30 minutes, then heat shocked at 42°C for 30 seconds then placed on ice for 5 minutes without mixing. 950 µl of room temperature (RT) SOC Outgrowth media was added to the cells and the mixture was incubated at either 37°C, or 30°C for SHuffle cells, with shaking for 60 minutes. 50, 100 and 500 µl of cells were plated onto LB agar with appropriate selection antibiotics and incubated at either 37°C or 30°C overnight.

2.2.2.2. Glycerol stocks

10 ml LB supplemented with antibiotics was incubated with a transformed colony at 30°C with shaking overnight. 500 µl of the overnight culture was mixed with 500 µl of 50% glycerol (filter sterilised) and stored at -80°C.

2.2.2.3. Purification of plasmid DNA and subcloning

pUC57 gene constructs to be expressed in *E. coli* were cloned into DH5α cells as above. 10 ml overnight cultures grown in LB with appropriate selection antibiotics were divided into 5 x 1 ml glycerol stocks, prepared as above, or for DNA isolation. 5 ml of the overnight culture was processed using the Qiagen miniprep kit as per manufacturer instructions to isolate the plasmid DNA.

The gene insert was cut from pUC57 plasmid using 10 units of restriction enzymes Not-1 and BamH1 in a 35 µl reaction mix containing DNA, NEB3 buffer, BSA made up to volume with dH_2O. The mixture was incubated at 37°C for 90 minutes then mixed 1:6 with Orange (6x) gel loading dye and assessed on a 1% agarose gel run for 1 hour at 100 V. Gel was soaked in ethidium bromide for 10 minutes then washed and visualised under UV light. The visible band at the appropriate size was excised from the gel and the DNA was extracted using a Qiagen QIAquick Gel Extraction Kit as per manufacturer's instructions.

DNA was ligated into pET30a(+) expression vector using T4 DNA ligase (New England BioLabs) in a 10 µl reaction mix with ligase, T4 DNA ligase buffer (10x) (New England BioLabs) and DNA at either 15 or 30 ng, made up to final volume with nuclease-free dH_2O. The mixture was incubated at 16°C for 4 hours then transformed into an expression cell line.

2.2.2.4. CD151 Protein expression in bacterial cells

LB (or RM for SHuffle cells) with appropriate antibiotics was inoculated with a single colony from either a fresh transformation growing on LB agar or from a frozen glycerol stock and grown overnight at 30°C with shaking. This overnight starter culture was then diluted to an optical density of 0.1 when measured at 600 nm (OD600), referenced against sterile media. Cultures were grown at 37°C (or 30°C for SHuffle cells) with shaking appropriate to the vessel size for adequate aeration until cell growth reached mid exponential phase i.e. OD600 between 0.6 and 0.8. Protein expression was induced with 0.5 mM IPTG (unless stated otherwise) and the cultures were either maintained at their current temperature for 3 - 4 hours or chilled to 16°C for overnight expression. Harvested cultures were centrifuged at 4000 *g* for 20 minutes to sediment the bacteria and the media was decanted. The cell pellets were resuspended in lysis buffer, approximately 5 ml of lysis buffer per litre of original culture, and frozen at -20°C overnight and then either processed or moved to -80°C for storage.

2.2.2.5. Cell lysis

Frozen cell pellets were thawed on ice. Small volumes (<20 ml) were lysed by sonication of the cell suspension in several 5 - 20 second bursts with cooling on ice between. Larger volumes were lysed by three passes through a chilled EmulsiFlex-C5 Homogenizer (AVESTIN, Canada) with a homogenising pressure > 1500 psi. The lysed cells were centrifuged at 20,000 *g* for 30 minutes at 4°C to sediment cell debris and insoluble protein. The supernatant, containing the soluble protein, was decanted and the pellet was resuspended in 8 M urea, 70 mM HEPES, 100 mM NaCl pH 7. Soluble lysate and the urea solubilised pellet were both filtered through a 0.45 µm filter. Samples of the expression cultures pre- and post-induction, during expression, and the soluble and insoluble lysate fractions were analysed by SDS-PAGE to monitor protein production.

2.2.3. CD151 protein expression in insect cells

2.2.3.1. Cell counts

Cells were counted using a Neubauer type haemocytometer. Cells were diluted with Trypan blue at a ratio of 1:1 then diluted 1:4 with PBS. A coverslip was placed onto the haemocytometer and 10 µl of the cell mixture was pipetted into each of the two counting chambers. The haemocytometer was then viewed with a light microscope at 100x magnification and unstained (viable) cells in the outer four and middle squares were counted. The total cell count was then divided by the number of squares counted and multiplied by eight to account for the dilution of the cells with Trypan blue and PBS. The resulting number was then multiplied by 10,000 to determine the number of cells per ml of the suspension culture.

2.2.3.2. Plasmid transformation

pFastbac-CD151 113-220 was transformed into DH5α cells as above and plated onto LB agar + ampicillin and grown overnight. Colonies were selected and grown overnight in 10 ml LB + ampicillin, and plasmid DNA was isolated using a Qiagen Miniprep kit as per manufacturer's instruction.

2.2.3.3. Generating recombinant bacmid

Purified pFastbac-CD151 113-220 plasmid was transformed into MAX Efficiency® DH10Bac™ cells as above and plated onto LB agar containing 50 µg/ml kanamycin, 7 µg/ml gentamicin, 10 µg/ml tetracycline, 100 µg/ml X-gal and 40 µg/ml IPTG. Plates were incubated for 48 hours at 37°C. 10 white colonies were picked from the plate, dipped into PCR mix then transferred to LB with 50 µg/ml kanamycin, 7 µg/ml gentamicin, and 10 µg/ml tetracycline. Presence of the CD151 gene in the recombinant bacmid was verified by PCR using Universal forward and reverse primers:

For PCR, a 25 µl reaction mix per colony was prepared (and one extra for luck):

Component	**25 µl reaction**	**Final concentration**
10x standard *Phusion* reaction buffer	2.5 µl	1x
10 mM dNTPs	0.5 µl	200 µM
10 µM Forward primer	0.5 µl	0.2 µM
10 µM Reverse primer	0.5 µl	0.2 µM
Template DNA (from colony)		
Phusion DNA polymerase	0.125 µl	1.25 units /50 µl PCR
Nuclease free water	20.87 µl	to 25 µl total volume

PCR tubes were transferred from ice to the PCR machine with the block preheated to 95˚C and thermocycling performed:

Stage	Temperature °C	Duration
Initial denaturation	95	30 seconds
30 cycles of:	95	15-30 seconds
	45 - 68	1-60 seconds
	68	1 minute per kb
Final extension		
	68	5 minutes
Hold	4 – 10	

Positive colonies were identified by agarose gel electrophoresis.

Bacmid was purified from positive colonies in culture using Qaigen Miniprep kit as per manufacturer's instructions.

2.2.3.4. Generating virus P1 stock

In a 6-well plate:

Seed 8 x 10^5cells per well in 2 ml media without antibiotics, allow cells to attach for 15 minutes at RT.

For each well prepare:

- 8 µl Cellfectin II in 100 µl media (no antibiotics), mix well.
- 3 µl (1 -2 µg) baculovirus DNA in 100 µl media (no antibiotics), mix gently.
- Combine DNA with Cellfectin, mix gently and incubate at RT 15 – 30 minutes.
- Remove media from cells, replace with 800 µl media without antibiotics.
- Add 210 µl DNA-Cellfectin mix dropwise onto the cells, incubate at 27˚C for 3 - 5 hours.
- Remove transfection mix and replace with 2 ml media, with antibiotics.

Incubate cells at 27˚C for 72 hours.

Harvest virus, centrifuge 500 *g* 2 minutes to remove cell debris, aliquot into sterile 2 ml tubes, store at 4°C.

2.2.3.5. Amplifying baculovirus stock

P1 viral stock was amplified using a multiplicity of infection (MOI) of 0.1 plaque forming units (pfu) per cell. The amount of inoculum required was calculated using the formula:

$$\boldsymbol{inoculum\ required} = \frac{\boldsymbol{MOI} \times \boldsymbol{number\ of\ cells}}{\boldsymbol{titre\ of\ viral\ stock} *} \quad (\boldsymbol{ml})$$

Equation 2-1

*Viral titre was not experimentally determined and assumed to be 5 x 10^6 pfu/ml.

*Sf*9 or *Sf*21 cells were plated into a 6-well plate at 2 x 10^6 cells/well. Cells were allowed to attach for 1 hour then the appropriate amount (as calculated above) of P1 viral stock was added to each well and the cells were incubated for 48 hours in a 27°C humidified incubator. Virus was harvested, media centrifuged at 500 *g* for 2 minutes then stored in sterile 2 ml tubes at 4°C. This is P2 viral stock.

Amplification was repeated to produce high-titre P3 viral stock.

2.2.3.6. Expression of recombinant protein

SF9 or SF21 cells were grown in Sf-900 II SFM™ to a density of 1 x 10^6 cells/ml in 450 ml media and infected with 50 ml P3 virus. Cells were grown in a 27°C humidified incubator for 72 hours then media was harvested, centrifuged and stored at -20°C.

2.2.4. CD151 protein expression in mammalian cells

2.2.4.1. Plasmid amplification

Synthetic DNA cloned into pcDNA 3.1 (GenScript) were first transformed into *E. coli* DH5α cells for amplification under ampicillin selection. A 200 ml culture of DH5α transformed with pcDNA 3.1 was grown overnight at 37°C with shaking in LB supplemented with 100 µg/ml ampicillin. Cells were harvested by centrifugation at 4000 *g* for 20 minutes at 4°C in a Heraeus Cryofuge™ 6000i low-speed centrifuge (Thermo Electron Corporation). Plasmids were extracted from harvested cells with the QIAGEN Plasmid Mega Kit (Qiagen) according to the manufacturer's instructions.

2.2.4.2. Transfection, protein expression and harvest

HEK 293F cells (ThermoFisher) were cultured in 2 x 100 ml Opti-MEM® expression medium (Invitrogen), maintained at 37°C with an atmosphere of 8% CO_2 until cell density reached 1 x 10^6 cells/ml. Transient transfections were performed using 50 µl of 293fectin™ transfection reagent (Invitrogen) and 50 µg of plasmid DNA diluted to a total volume of 3 ml in Opti-MEM® (Life Technologies). The lipid-DNA was incubated for 20 - 30 minutes at RT to allow the DNA – 293fectin™ complexes to form. The formed complexes were then added to 100 ml of cultured HEK 293F cells for transfection. Samples were taken at day 2 and 3 then cells were harvested by centrifugation at 800 *g*

for 5 minutes at 4°C. The supernatants (culture media) containing the secreted proteins were filtered through a 0.22 µm membrane filter prior to purification.

2.2.5. CD151 protein purification

2.2.5.1. Affinity purification chromatography

Purifications were carried out using an ÄKTA™ purifier Chromatography system at 4°C, monitoring sample UV absorbance at 280 nm and conductivity, system pressure limit set at 0.3 MPa, and a flow rate of 5 ml/minute, unless specified. A HisTrap™ HP column (GE) was equilibrated with 5 column volumes (CV) of dH_2O, followed by 5 CV of binding buffer (70 mM HEPES, 100 mM NaCl, 5% glycerol, 20 mM imidazole, pI, pH 7.5). The soluble lysate was applied to the column using the in-line sample pump and the column washed with >10 CV binding buffer until a stable baseline was achieved. Protein was eluted with a 10 CV gradient of 0 - 100% elution buffer (binding buffer + 500 mM imidazole) followed by 5 CV at 100% elution buffer. Eluted protein was collected in 1 – 3 ml fractions. Fractions containing protein were analysed by SDS-PAGE.

2.2.5.2. Affinity purification under denaturing conditions

Purification was carried out as above using 8 M urea, 70 mM HEPES, 100 mM, NaCl pH 7 as binding buffer, washing with 6 M urea, 70 mM HEPES, 100 mM NaCl, pH 7 and eluting with 6 M urea, 70 mM HEPES, 100 mM NaCl, pH 7, 500 mM imidazole.

2.2.5.3. On-column refold

Urea solubilised protein was applied to a 5 ml HisTrap™ HP column, as above and washed with binding buffer containing 6 M urea until a stable baseline was achieved. The column was then washed with a 0 - 100% gradient of binding buffer without urea over 20 CV at 0.5 ml/minute and washed for a further 5 CV or until a stable baseline. Refolded protein was eluted with binding buffer + 500 mM imidazole as above.

2.2.5.4. Glutathione sepharose columns

Purifications were carried out using an ÄKTA™ purifier Chromatography system at 4°C, monitoring sample UV absorbance at 280 nm and conductivity, system pressure limit set at 0.3 MPa, and a flow rate of 5 ml/minute, unless otherwise specified.

A GSTrap™ FF 5 ml column (GE) was equilibrated with 5 CV of binding buffer PBS, 5 mM EDTA pH 7. The sample was applied to the column using the in-line sample pump with a flow rate of 1 ml/minute, then washed with >5 CV binding buffer, until the baseline stabilised. Bound protein was eluted with 50 mM Tris-HCl, 10 mM reduced glutathione, pH 8, and 1 – 3 ml fractions were collected. Fractions containing protein were analysed by SDS-PAGE.

2.2.5.5. Dextrin sepharose columns

Purifications were carried out using an ÄKTA™ purifier Chromatography system at 4°C, monitoring sample UV absorbance at 280 nm and conductivity, system pressure limit set at 0.3 MPa, and a flow rate of 5 ml/minute, unless specified.

An MBPTrap™ HP column (GE) was equilibrated with 5 CV of dH_2O, followed by 5 CV of binding buffer (70 mM HEPES, 100 mM NaCl, 5% glycerol, pI, pH 7). The soluble lysate was applied to the column using the in-line sample pump and the column washed with >10 CV binding buffer or until a stable baseline was achieved. Protein was eluted with a 10 CV gradient of 0 - 100% elution buffer (binding buffer + 10 mM maltose) followed by 5 CV at 100% elution buffer. Eluted protein was collected in 1 – 3 ml fractions. Fractions containing protein were analysed by SDS-PAGE.

2.2.5.6. Desalting and buffer exchange

Desalting and buffer exchange for volumes <15 ml was carried out with a HiPrep 26/10 Desalting column (GE) using an ÄKTA™ purifier Chromatography system at RT, monitoring sample UV absorbance at 280 nm and conductivity, system pressure limit set at 0.3 MPa, and a flow rate of 10 ml/minute. Sample was applied to the column using an in-line sample loop and eluted under constant buffer conditions into 5 ml fractions.

Samples >15 ml were dialysed against the relevant buffer, >200 times the sample volume, using SnakeSkin™ Dialysis Tubing 22 mm with a 3.5 kDa molecular weight (Mw) cut-off (Thermo Scientific), at 4°C overnight.

2.2.5.7. Ion exchange chromatography

Purifications were carried out using an ÄKTA™ purifier Chromatography system at 4°C, monitoring sample UV absorbance at 280 nm and conductivity, system pressure limit set at 0.3 MPa, and a flow rate of 5 ml/minute, unless otherwise specified.

A HiTrap™ Q HP 5 ml column (GE) was equilibrated with 5 CV of dH_2O, followed by 10 CV of binding buffer without NaCl. The sample was applied to the column using the in-line sample pump and the column washed with >10 CV binding buffer or until a stable baseline was achieved. Protein was eluted with a 10 CV gradient of 0 - 100% elution buffer (binding buffer with 1 M NaCl) followed by 5 CV at 100% elution buffer. Eluted protein was collected in 1 – 3 ml fractions. Fractions containing protein were analysed by SDS-PAGE.

2.2.5.8. Size exclusion chromatography

Purifications were carried out using an ÄKTA™ purifier Chromatography system at RT, monitoring sample UV absorbance at 280 nm and conductivity.

Small scale purifications, (<0.5 ml protein solution) were carried out using either a Superdex 75 or Superdex 200 10/300GL column (GE). Larger volumes of protein (1 - 5 ml) were purified using either a HiLoad 16/600 Superdex 75 pg or HiLoad 16/600 Superdex 200 pg column (GE). System pressure limit was set at 1.5 MPa and flow rate at 0.5 ml/minute for 10/300 columns and 0.3 MPa and 1 ml/minute for 16/600 columns, unless specified.

Columns were equilibrated in binding buffer, and sample was applied to the column using the in-line sample loop. Samples were eluted isocratically with 1.5 CV of binding buffer. Eluted protein was collected in 0.5 ml or 2 ml fractions and fractions containing protein were analysed by SDS-PAGE and Western blot.

2.2.6. Protein concentration measurement

2.2.6.1. Molecular weight and extinction coefficient

The amino acid sequence of the protein construct, including any relevant purification tags, was entered into the online ExPASy ProtParam tool http://web.expasy.org/protparam/ which calculates various physical and chemical parameters such as the Mw, theoretical Isoelectric point (pI) , amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY).

Protein concentration was routinely measured after each purification step using a NanoDrop 2000c Spectrophotometer (Thermo Scientific) by measuring UV absorbance at 280 nm as per manufacturer's instructions. The protein concentration was calculated using the Mw and extinction coefficient calculated as above.

2.2.6.2. Bradford assay

Protein concentration was measured periodically using the Pierce™ Coomassie Plus (Bradford) Assay kit (ThermoFisher Scientific). 10 µl of each standard or the unknown sample was added to 300 µl of the Coomasie Plus Reagent in a 96-well plate. The plate was mixed using a plate shaker and incubated for 10 minutes at RT. The absorbance at 595 nm was measured using a POLARstar microplate reader. Standards were prepared from bovine serum albumin at 2 mg/ml, diluted from 1500 – 125 µg/ml. All standards and samples were measured with a minimum of two replicates and each plate included several blank samples for reference.

2.2.6.3. Protein volume concentration

When required, protein samples were concentrated using Amicon Ultra Centrifugal Filter Units (Merk) with a Mw cut-off less than one third of the Mw of the protein.

2.3. *PROTEIN CHARACTERISATION*

2.3.1. Polyacrylamide gel electrophoresis

SDS-PAGE was carried out using precast Bolt™12% Bis-Tris Plus Gels (ThermoFisher Scientific). Gels were assembled into a Bolt™ Mini Gel Tank and the tank filled with Bolt™ MES SDS running buffer prepared from 20x concentrate. Samples were normalised to approximately 0.1 mg/ml by diluting with water and 6x SDS-PAGE loading buffer, vortexed, briefly centrifuged, then 10 – 15 µl was loaded into the gel wells. SeeBlue™ Plus2 Prestained Protein Standard was loaded into the far-left lane to provide Mw markers. Electrophoretic protein separation was achieved at a constant voltage of 165 V constant for 35 minutes. Gels were removed from the plastic casing and soaked in InstantBlue™ Protein Stain for >15 minutes, excess stain was removed with water and gels were scanned using the LI-COR Odyssey® CLx Imaging system at 700 and 800 nm wavelengths.

2.3.2. Western blotting

SDS-PAGE was carried out, as above, and once the gel was removed from the casing the gel was sandwiched with Immobilon®FL PVDF membrane and Whatman® cellulose chromatography paper, assembled into a Mini Blot Module (ThermoFisher Scientific) soaked and filled with Bolt™ Transfer buffer. Proteins were transferred at a constant field strength of 10 V for 60 minutes. After transfer the membrane was dried at 37°C for 1 hour then blocked in Odyssey® Blocking Buffer (TBS) for 1 hour at RT or overnight at 4°C. The membrane was then incubated for 1 hour at RT, shaking with the relevant primary Ab, diluted in blocking buffer including 0.1% Tween-20. The membrane was washed extensively with TBS 0.1% Tween-20 then incubated with the relevant IR (near infrared) labelled secondary Ab diluted in blocking buffer including 0.1% Tween-20 and 0.01% SDS. The membrane was washed extensively with TBS 0.1% Tween-20 with a final wash in TBS to remove the Tween-20 before being scanned using LI-COR Odyssey® CLx Imaging system at 700 and 800 nm wavelengths.

2.3.3. Dynamic light scattering

The size distribution profile of protein samples was assessed for oligomeric and aggregation state using dynamic light scattering (DLS). Purified samples were centrifuged at 15,000 *g* for 5 minutes then transferred to a 40 µl disposable plastic micro cuvette. DLS measurements were carried out using a Malvern Zetasizer Nano Z.

2.3.4. Circular dichroism spectroscopy

Secondary structure of purified protein was analysed using circular dichroism (CD). Protein samples were buffer exchanged using a Zeba™ Spin Desalting Column into 10 mM sodium fluoride, pH 7, and diluted to 0.2 mg/ml. Samples were transferred to a 1

mm quartz cuvette and measurements were performed using a Jasco-815 CD Spectrophotometer. 730 data points were obtained in the spectrum from 180 to 260 nm with 0.1 nm intervals, at 20°C. Data pitch of 1 nm, response time of 2 seconds and a scanning speed of 20 nm/minute was used, and readings were averaged over 4 accumulations. The measurements were referenced against a buffer blank sample. CD spectra were deconvoluted to estimate the helix, sheet and disordered content of the sample using the program Contin-LL reference set 3 on the DichroWeb Server.[245, 246] http://dichroweb.cryst.bbk.ac.uk/html/home.shtml.

2.3.5. Mass spectrometry

Molecular mass of the purified samples was measured using liquid chromatography-time-of-flight (LC-TOF). 10 µl samples were separated by size using a Sepax Bic-C4, 5 µm, 300 Å, 2.1 x 50 mm column on an Agilent HPLC, running buffer 40% acetonitrile, 0.1% TFA and analysed using an Agilent 6220 LC/ESI-TOF mass spectrometer. Data were deconvoluted using Agilent MassHunter Qualitative Analysis Software V6.

2.3.6. Differential scanning fluorimetry

Protein thermal stability was measured using differential scanning fluorimetry (DSF) to perform a fluorescence based thermal shift assay. Purified protein samples at >0.5 mg/ml were mixed 1:1 with a 20x dilution of SYPRO® Orange Protein Stain, then 1 µl of the protein/dye mix was added to 9 µl of each buffer condition to be tested in a 96-well 0.2 ml skirted PCR White plate (Life Technologies). The plate was sealed with a Microseal® B PCR plate sealing film, optical grade (Bio-Rad) and samples were analysed using a Bio-Rad C1000 TouchTM Thermal Cycler set to cycle from 20 to 80°C in 0.2°C increments for 1.8 seconds. Melting temperature (*T*m) was calculated using the Bio-Rad CFX Manager 3.1 software.

2.3.7. Surface plasmon resonance

Protein interactions were analysed by surface plasmon resonance (SPR) using a Biacore T200 (GE) and the Biacore T200 Control Software Version 2.0.

2.3.7.1. Consumables

Series S CM5 sensor chip (GE) carboxymethylated dextran covalently attached to a gold surface

Series S NTA sensor chip (GE) carboxymethylated dextran pre-immobilised with nitrilotriacetic acid

NiHC 1500 Sensor chip (Xantec) Poly-nitrilotriacetic acid (NTA) derivatised linear polycarbonate hydrogel

SAHC 1500M Sensor chip (Xantec) Streptavidin, immobilised in a linear polycarboxylate hydrogel

CMD 500M Sensor chip (XanTec) Carboxymethyldextran hydrogel

2.3.7.2. Immobilisation of protein on an SPR sensor chip

2.3.7.2.1. Amine coupling to chip surface

Protein was immobilised on either a GE CM5, CM7 or XanTec CMD 500M sensor chip via amine coupling. Chips were preconditioned as per manufacturer's recommendation. The instrument was primed with 10 mM HEPES, 150 mM NaCl (pH 7.4), 0.005% TWEEN®20 (HBS-P) to allow surface equilibration. Immobilisation was carried out at a flow rate of 10 µl/minute. The chip surface was first activated with a mixture of 0.4 M 1-ethyl-3-(3-dimethylaminopropyl)-carbodiimide (EDC) and 0.1 M N-hydroxysuccinimide (NHS) in water (NHS/EDC) to give reactive succinimide esters in the carboxymethylated dextran matrix. When the protein (i.e. SPR ligand) is passed over the chip surface the esters react with primary amines on the protein surface to covalently link the protein to the dextran matrix. Once the desired immobilisation level was reached any uncoupled succinimide esters are blocked with 1 M ethanolamine-HCl pH 8.5.

2.3.7.2.2. Protein thiol coupling to chip surface

For coupling to a sensor chip via a protein thiol group, a standard carboxymethylated dextran (CM) sensor chip (as above) is used. Chips were preconditioned as per manufacturer's recommendation. The instrument was primed with HBS-P to allow surface equilibration. Immobilisation was carried out at a flow rate of 10 µl/minute. The chip surface was activated with NHS/EDC and the formed esters were modified with 20 mM 2-(2pyridinyldithio)ethanolamine (PDEA) in 0.1 M sodium acetate, 1.0 M sodium chloride to reactive disulfide groups. The protein was then coupled by the reaction of a free -SH group on the protein with the disulfide group on the sensor chip. The chip surface is blocked with 50 mM cysteine-NaCl in 0.1 M sodium acetate, 1.0 M sodium chloride to deactivate excessive reactive groups.

2.3.7.2.3. Capture via protein His tag

His-tagged protein was captured on either a Series S NTA or NiHC 1500 sensor chip via Ni^{2+}/NTA chelation. Chips were preconditioned as per manufacturer's recommendation. The instrument was primed with HBS-P to allow surface equilibration. Immobilisation was carried out at a flow rate of 10 µl/minute. The chip surface was saturated with nickel by injecting 0.5 mM $NiCl_2$ into the running buffer for 60 seconds. The protein was injected across the chip surface and captured on the Ni^{2+} via the His tag.

2.3.7.2.4. Capture via biotin label

Chemically biotinylated or biotinylated avidin (AviTag™)-tagged CD33 was captured on a SAHC 1500M sensor chip via streptavidin capture. Chips were preconditioned as per manufacturer's recommendation. The instrument was primed with HBS-P to allow

surface equilibration. Immobilisation was carried out at a flow rate of 10 µl/minute. The chip surface was saturated with nickel by injecting 0.5 mM $NiCl_2$ diluted in running buffer for 60 seconds. The protein was injected across the chip and captured on the streptavidin surface via biotin.

2.3.7.2.5. Combined capture/couple method

Protein was immobilised on either a Series S NTA or NiHC 1500 sensor chip. His-tagged protein was captured via Ni^{2+}/NTA chelation, and then covalently immobilised using amine coupling chemistry (GE Healthcare Laboratory Guidelines 29-0057-17 AB). Chips were preconditioned as per manufacturer's recommendation. The instrument was primed with HBS-P to allow surface equilibration. Immobilisation was carried out at a flow rate of 10 µl/minute. The chip surface was saturated with nickel by injecting 0.5 mM $NiCl_2$ for 60 seconds. The surface was activated with NHS/EDC (Amine Coupling Kit, GE Life Sciences) for 420 seconds at a flow rate of 10 µl/minute to derivatise the carboxymethylated dextran surface with reactive NHS ester groups. Protein was diluted in HBS-P and injected across the chip surface. Remaining NHS esters were blocked with a 420 second injection of ethanolamine.

2.3.8. Calculations

To calculate the approximate amount of ligand to immobilise (R_{ligand}) on the sensor chip surface to achieve a desired maximum binding response (R_{max}), the following equations were used:

$$\boldsymbol{R_{max} = (R_l Mw_a n_l) Mw_l^{-1}}$$

Equation 2-2

where R_{max} is the maximum binding response, R_l is the amount of ligand immobilised, Mw_a is the molecular weight of the analyte and Mw_l is the molecular weight of the ligand and n_l is the valency of the ligand.

The R_{max} calculation is theoretical and assumes all immobilised ligand is accessible and functional. To calculate the actual percentage of functional ligand, using the R_{max} obtained from a binding experiment the following equation is applied:

$$\boldsymbol{Functional\ ligand = \left(\frac{R_{max}}{R_l}\right)\left(\frac{Mw_l}{Mw_a}\right)100\ (\%)}$$

Equation 2-3

In this thesis, the “ligand” refers to the protein and the “analyte” to the small molecule compound.

2.3.9. Compound screening

Compound screens were carried out in 70 mM HEPES pH 7, 100 mM NaCl, 0.05% T20, 2% DMSO (HBS-P+). A 10 mM DMSO solution of each compound was diluted 1/20 directly into 96-well plates containing 1.02x running buffer without DMSO to give a final concentration of 500 µM with 2% DMSO. Plates were sealed immediately, mixed, then centrifuged at 4000 rpm for 5 minutes. Any wells containing visible precipitate were excluded from analysis.

Binding experiments were performed at 25^{o}C by injecting the compounds over all flow cells at 30 µl/minute for 30 seconds with a dissociation time of 420 seconds. Solvent correction and control cycles were included every 20 cycles. Data were solvent corrected, reference subtracted, adjusted for Mw, quality controlled and evaluated using the Biacore T200 Evaluation Software V. 2.0. Any compounds binding more than the negative control were further analysed for kinetic and affinity analysis.

Kinetic and affinity analysis was performed as above by injecting compound solutions in two-fold dilutions and at a minimum of five concentrations from 5 mM or lower depending on compound solubility, in triplicate. Kinetic and affinity data were solvent corrected, reference subtracted and blank subtracted using the Biacore T200 evaluation software V.2.0. Kinetic constants were determined by curve fitting using a 1:1 binding model. Association and dissociation curves were individually fit to experimental data points from discrete experiments.

2.3.10. Protein crystallisation

0.2 µl of purified protein at >2 mg/ml was dispensed into subwell 1 of each well of a UV+ Low Profile Microplate (Rigaku) with 50 µl of precipitant in the well reservoir using the Crystal Gryphon liquid handling robot (Art Robbins Instruments). Plates were sealed with UV friendly Clearview Sheets (Molecular Dimensions) and housed at either room temperature (22°C) or 4°C. Plates were regularly inspected using a Minstrel HT UV crystal drop imager (Rigaku) with both UV and visible light for crystalline precipitate.

2.3.11. BLAST analysis of protein sequences

BLAST analysis of protein sequences were conducted via the NCBI webserver using Protein BLAST, blastp suite, selecting the required database and organism and blastp (protein-protein BLAST) algorithm https://blast.ncbi.nlm.nih.gov

2.3.12. Sequence alignments

Multiple sequence alignments were carried out using either CLC Sequence Viewer Version 8.0 via QIAGEN Aarhus A/S www.qiagenbioinformatics.com, or T-Coffee via the Centre for Genomic Regulation of Barcelona http://tcoffee.crg.cat/, Clustal Omega as implemented on the UniProt web site (http://uniprot.org)or manually using Microsoft Excel version 1810.

2.3.13. Post-translational modifications

Post-translational modifications were predicted by literature search and sequence analysis by ExPASy Bioinformatics Resource Portal.
https://www.expasy.org/resources/search/keywords:post-translational%20modification

2.3.14. CD151 LEL homology modelling

Homology models of the CD151 LEL were constructed, based on the results of multiple sequence alignments, using the Protein Homology/analogy Recognition Engine V2.0 (Phyre^2) http://www.sbg.bio.ic.ac.uk/phyre2 and examined for veracity using SYBYL-X 2.1.1 (Certara LP., Princeton, NJ, USA) http://www.certara.com. Homology models were inspected and figures generated using PyMol (PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC).

CHAPTER 3. DISCOVERY OF CD33 INHIBITORS

3.1. *INTRODUCTION*

AML is considered to be a disease of older people and its occurrence in those younger than 45 years is uncommon, with the average age at diagnosis being 68. [247] Older patients are generally unable to tolerate intensive chemotherapy, which limits treatment options, and the median survival rate is only 5 to 10 months.[87] There is an urgent need for new treatments. Therapies targeting CD33 have been explored since the 1990's, focussing on Abs and ADCs. Although there has been some success with this approach, the development of these treatments has been hampered by specificity, low target expression and slow internalisation kinetics. [104] Because of their size Abs cannot pass through the cell membrane and are less efficient for tissue penetration and blood clearance than small molecule compounds. Any protein therapeutic is potentially immunogenic, while small molecule agents are mostly non-immunogenic. [248] There are clear advantages to finding a small molecule agent for the treatment of AML, particularly for overcoming the difficulties associated with Ab-based therapeutics.

The crystal structures of both the IgV domain alone and the full extracellular domain (i.e. IgV and IgC2 domains) of hCD33 have been solved with and without ligands bound to the sialic acid-binding domain (Figure 1-8, Figure 1-10)(PDB IDs: 6D48, 6D49, 6D4A, 5IHB, 5J06 and 5J0B). [125] Little is known about the structure of the transmembrane domain and the cytoplasmic tail. CD33 ligands all contain a critical sialic acid moiety whereby the carboxylate of the sialic acid is engaged by a salt-bridge interaction with CD33 residue R119. The IgV domain contains amino acids 19 – 135, with three cysteine residues that form both inter (C36 with C169 in the adjacent IgC2 domain) and intra (C41 - C101) domain disulfide bonds. Structural modelling by Dr Tracy Nero, in our laboratory, has revealed a druggable pocket that may be targeted to develop drugs to treat diseases such as AML. (unpublished data)

Using computational chemistry techniques our research group identified distinct classes of small molecules that were predicted to bind to CD33 and influence function. By optimising the construct design for the CD33 IgV domain, immobilisation technique and assay conditions, my work has led to development of an SPR-based binding assay to screen and characterise potential CD33 inhibitors. These small-molecules have the potential to lead to the discovery of drugs that can be therapeutically beneficial in treating cancers, such as AML.

3.2. *MATERIALS AND METHODS*

General methods utilised in this chapter have been described in Chapter 2. Method development and assay optimisation are detailed in Results and Discussion (section 3.3).

3.2.1. Wild-type and mutant CD33 IgV domain constructs

The initial CD33 IgV domain protein expression constructs and purification strategies in our research group were designed by Dr. Luke Miles; however, after exhaustive exploration this methodology proved to be irreproducible. Subsequent expression construct design and purification strategies were established by myself, and Jasmina Markulić; another graduate researcher in our research team.

Codon optimised (*E. coli*) cDNA encoding residues D18 to H143 of human CD33 was synthesised and sub-cloned into a pET-30a+ vector by Genscript with the following variations (Table 3):

Table 3 List of the CD33 fusion constructs used in this project. C-terminal tags are non-cleavable.

CONSTRUCT	N-TERMINAL TAG	CLEAVAGE SITE	MUTATION	C-TERMINAL TAG
His-CD33 wild-type	Octa-His	TEV		
His-CD33 R119A	Octa-His	TEV	R119A	
His-CD33 C36S	Octa-His	TEV	C36S	
His-CD33 C36S, R119A	Octa-His	TEV	C36S, R119A	
Avi-CD33 C36S	Octa-His	TEV	C36S	AviTag™
Avi-CD33 C36S, R119A	Octa-His	TEV	C36S, R119A	AviTag™

An octa-His tag was chosen to simplify purification by increasing the affinity of the heterologous protein to the immobilised Ni^{2+} on the IMAC resin. [249] TEV cleavage site allows for the removal of the His tag which may impact conformational heterogeneity required for protein crystallisation. [250, 251] A non-cleavable, C-terminal AviTag™ was included for capturing the protein on a streptavidin chip for SPR (see 2.3.7.2.4).

Protein purification and quality assurance of the CD33 wild-type IgV was performed by Jasmina Markulić for a different project focused on Alzheimer's disease and is not shown here. The protein was thoroughly assessed for conformation and purity and found to be suitable for use in SPR assays.

3.3. *RESULTS AND DISCUSSION*

3.3.1. Surface plasmon resonance assay development

Data quality is dictated by assay conditions. Variables such as SPR ligand stability, analyte solubility, non-specific binding, mass transfer effects and aggregation are

dependent on buffer choice, flow rate, SPR ligand surface density and analyte concentration. Careful optimisation and control of these variables is necessary to minimise binding artefacts and to produce robust, reproducible data. In these studies, the SPR “ligand” refers to the CD33 protein and the “analyte” is the small molecule compound.

3.3.1.1. pH optimisation of CD33 immobilisation

For covalent coupling to dextran-carboxy based sensor chips the ligand is concentrated to the sensor chip surface by electrostatic attraction. Efficient ligand attraction requires that the pH of the ligand lies between the pk_a of the surface and the isoelectric point (pI) of the ligand. At pH >3.5 the dextran matrix carries a net negative charge and at pH < pI the ligand will have a net positive charge. Low ionic buffer strength also favours the electrostatic interaction and buffers with 10 - 20 mM total cation concentration are optimal. [252]

Wild-type CD33 IgV domain was purified by gel filtration into 10 mM HEPES, 100 mM NaCl, 0.005% TWEEN®20, pH 8 (HBS-P) with a final concentration of 127 µg/ml. CD33 has a pI of 7 so a range of immobilisation buffers from pH 6.5 to 5.5 were tested to optimise the surface pre-concentration on a standard carboxymethylated dextran CM5 sensor chip.

A pH optimisation against a blank, inactivated CM5 sensor chip surface was performed with wild-type CD33 diluted in 5 mM maleic acid pH 6.5, 6.0 and 10 mM sodium acetate pH 5.5 to 40 µg/ml. No discernible pre-concentration was observed (Figure 3-1) suggesting that these buffers were unable to bring down the sample pH to an appropriate level. The pH screen was repeated at a lower pH range using 10 mM sodium acetate at pH 5, 4.5 and 4. (Figure 3-2)

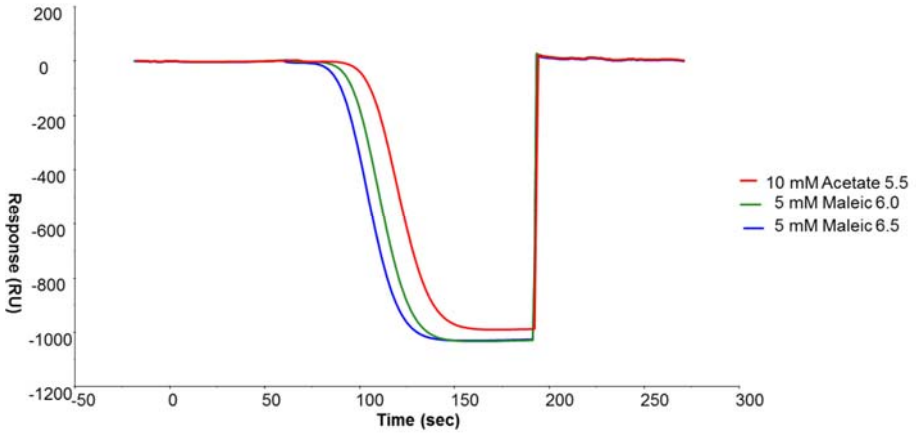


Figure 3-1 Comparison of the electrostatic binding response of wild-type CD33 IgV domain. CD33 40 µg/ml in 5 mM maleic acid buffer pH 6.5 (blue) and 6 (green) and 10 mM sodium acetate buffer pH 5.5 (red). CD33 was injected for 2 minutes at a flow rate of 10 µl/minute across a blank CM5 sensor chip. The chip surface was regenerated with pulses of 50 mM NaOH (not shown).

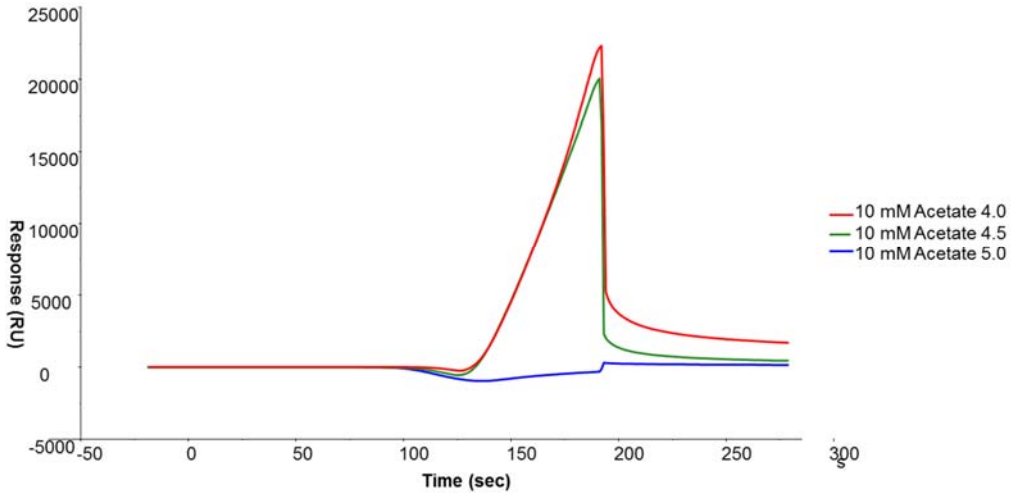


Figure 3-2 Further comparison of the electrostatic binding response of wild-type CD33 IgV domain. CD33 40 µg/ml in 10 mM sodium acetate buffer pH 4 (red), pH 4.5 (green) and pH 5 (blue). CD33 was injected for 2 minutes at a flow rate of 10 µl/minute across a blank CM5 sensor chip. The chip surface was regenerated with pulses of 50 mM NaOH (not shown).

The optimal buffer for immobilisation not only gives good pre-concentration but has a pH that preserves the stability of the protein. As shown in Figure 3-2, dilution into a buffer at pH 4 produced a fast pre-concentration with a high capture level, pH 4.5 gave a fast pre-concentration with a lower capture level and pH 5 gave a very slow, almost negligible pre-concentration. Covalent cross-linking efficiency can be reduced at lower pH and the increased capture rate may reflect a tendency for non-specific interaction. [253] The stability of CD33 under various pH conditions and buffers was examined by Jasmina Markulić and CD33 was found to be stable in pH 4.5 (data not shown). To minimise cross-linking and to preserve protein integrity, pH 4.5 was chosen for the immobilisation buffer. It is likely that due to the buffering capacity of the pH 8 sample buffer the actual sample pH following dilution is much higher than 4.5, however due to the small amounts of protein available and the low volume of the reaction solution the pH was not physically tested.

3.3.1.2. Thiol coupling of CD33 protein to sensor chip surface

For coupling to a sensor chip via a ligand thiol group, a standard CM5 sensor chip is activated with NHS/EDC and the formed esters are modified with 2-(2pyridinyldithio) ethanolamine (PDEA) to reactive disulfide groups. The ligand is then coupled by the reaction of a free thiol (i.e. –SH) group on the ligand with the disulfide group on the chip surface. This immobilisation technique exploits the free C36 residue on the wild-type CD33 IgV domain (Figure 1-8) and allows homogenous orientation of the immobilised protein, with the binding site clearly accessible to any binding partners. Using the pre-concentration pH 4.5 (Figure 3-2), CD33 was coupled to a CM7 chip via thiol coupling. (Figure 3-3) The concentration of ligand required varies depending on the activation and availability of the binding site on the ligand. A general rule-of-thumb is that the ligand should be between 5 and 50 μg/ml to give an adequate chip surface density of the immobilised ligand. As there is only one available thiol group on CD33, the concentration needed to be in the higher range.

CM7 chips have the same carboxymethylated dextran chemistry as the standard CM5 chips but with three-fold higher capacity. A higher protein immobilisation density is preferable when dealing with small molecule analytes.

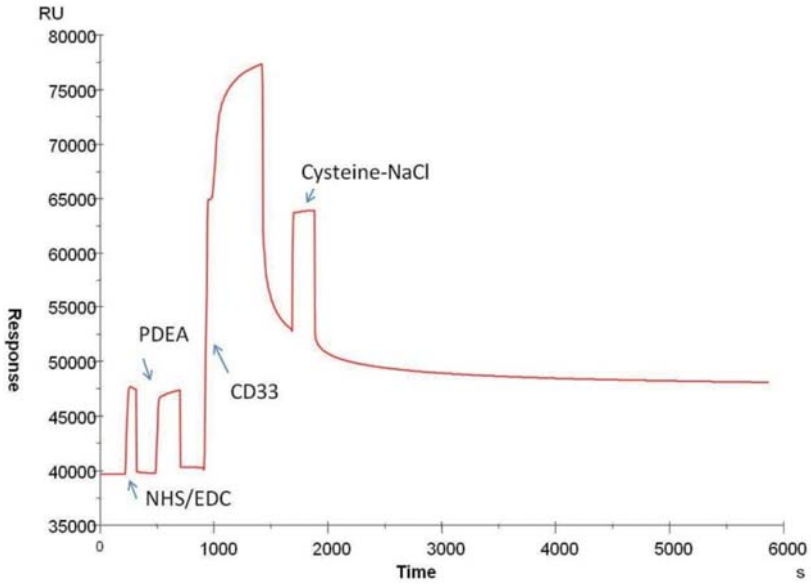


Figure 3-3 Thiol coupling of wild-type CD33 IgV domain to a CM7 chip. A 1:1 mixture of 0.4 M EDC and 0.1 M NHS was injected for 2 minutes to activate the chip surface (NHS/EDC). 80 mM PDEA in 50 mM sodium borate pH 8.5 was then injected for 4 minutes to introduce disulfide groups. CD33 50 µg/ml in 10 mM sodium acetate buffer pH 4.5 was injected for 7 minutes followed by a 4-minute injection of 50 mM L-cysteine-NaCl pH 4 to deactivate excessive reactive groups. Flow rate was 10 µl/minute. 7792 RU of CD33 was immobilised.

3.3.2. Positive control compounds 3'- and 6'-sialyllactose

CD33 is known to bind sialyl acids [254, 255], therefore 3'-sialyllactose and 6'-sialyllactose (Figure 3-24) were used as positive control compounds to demonstrate the binding activity of the immobilised wild-type CD33 IgV domain. [256] The reported affinities of 3'- and 6'-sialyllactose for wild-type CD33 are 8.7 and 8 mM, respectively, determined using an SPR assay. [256] While these affinities are extremely low and close to the detection limits of the Biacore T200 used for SPR, the two compounds were readily available and the only positive controls available at the time. A CD33 specific Ab could have been used but is not suitable for comparison with small molecules. A much higher affinity small molecule positive control, P22, became available later in the project (see 3.3.11.4). 3'-Sialyllactose and 6'-sialyllactose in 0.5, 1, 10% solution diluted in running buffer were injected across the immobilised wild-type CD33 IgV domain to test binding. (Figure 3-4)

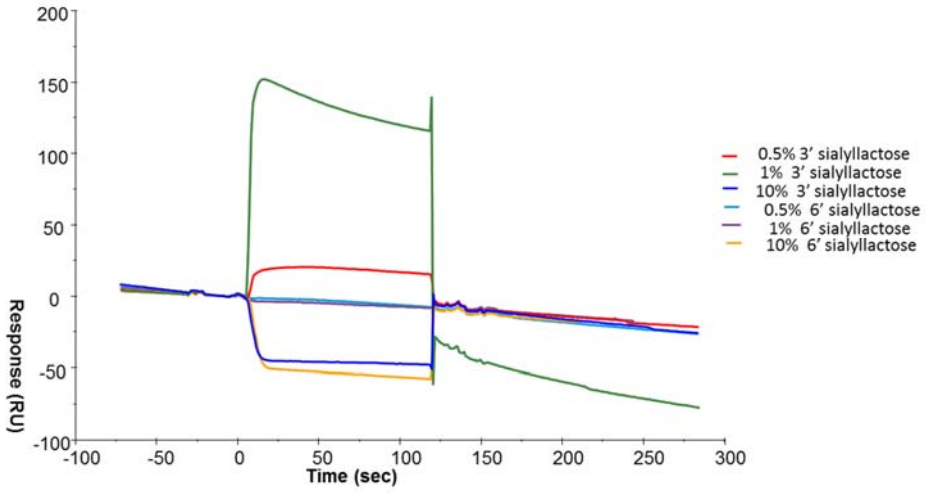


Figure 3-4 Sensorgram showing the binding response of 3'- and 6'-sialyllactose to wild-type CD33. 3'-Sialyllactose at 0.5 (red), 1 (green) and 10% (dark blue) and 6'-sialyllactose at 0.5 (light blue), 1 (brown) and 10% (yellow) solutions were injected across the immobilised wild-type CD33 IgV domain. The 3'-and 6'-sialyllactoses were injected for 2 minutes at 30 µl/minute.

3'-Sialyllactose at 0.5% solution appeared to bind to CD33, however the binding did not appear to be concentration dependant and the bulk shift from buffer mismatch was masking any binding at the higher concentrations. In Figure 3-4 the green curve, 1% 3'-sialyllactose, shows a huge bulk shift with no apparent binding. The blue and yellow curves representing 10% solutions appear "upside-down" indicating there was more response to the reference surface which results in negative data upon subtraction. (Figure 3-4) The 6'-sialyllactose did not seem to bind to the wild-type CD33 at all. As noted above, the reported affinities of 3'- and 6'-sialyllactose for CD33 are 8.7 and 8 mM, respectively, close to the detection limits for an SPR assay. As such, the lack of apparent binding may be simply due to the very weak real binding signal being swamped by effects of buffer mismatching. It is also possible that the CD33 IgV domain was unfolding or, possibly immobilising, via one of the structural cysteine residues (C41 and C101, refer to section 1.3.1). Unfolded protein would lose specific binding for sialylated sugars. Given their reported low affinity for CD33, 3'- and 6'-sialyllactoses are not ideal positive controls; the lack of apparent binding does not provide definitive evidence of the activity or otherwise of the immobilised wild-type CD33 IgV domain; however, the lack of a high-affinity positive control hampers the quantitative evaluation of these findings.

3.3.2.1. Sensor chip surface stability

Once the wild-type CD33 protein was immobilised, the stability of the CM7 sensor chip surface was assessed by observing the behaviour of the baseline signal following

repeated injections of running buffer (100 mM Tris, 200 mM NaCl pH 8). Some drift was apparent during surface activity checks with both sialyllactoses (Figure 3-4), raising concerns that stability of the CM7 sensor chip surface would become an issue during the lengthy compound screening experiments. Further stability tests of repeated injection of running buffer over cycles of more than 20 minutes were performed to ascertain long term stability of the CD33 protein. (Figure 3-5)

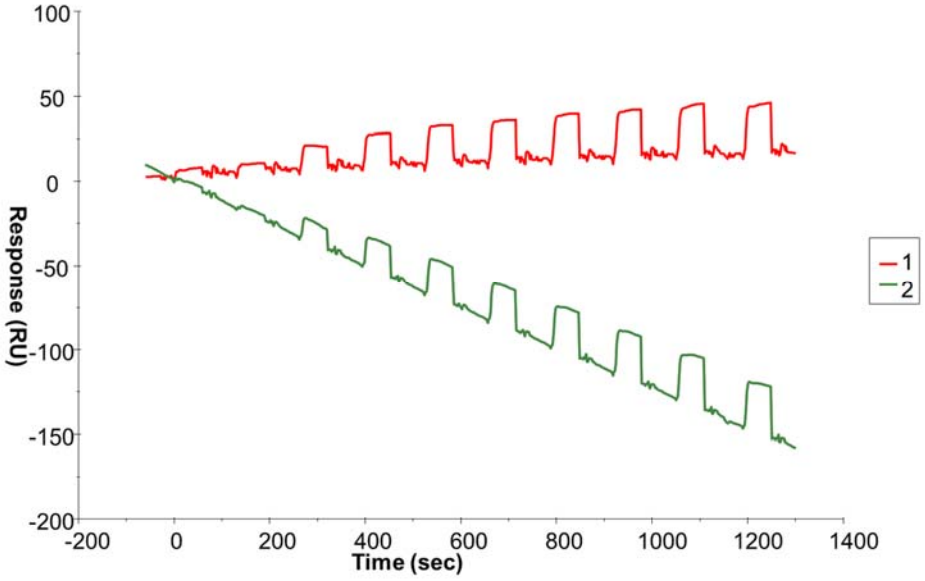


Figure 3-5 Sensorgram showing repeated injections of running buffer across the immobilised wild-type CD33 IgV domain in flow cell 2 (green) and the blank CM7 chip surface (red).

The CM7 sensor chip surface showed no sign of stabilising after several cycles of running buffer injections (Figure 3-5). In situations like this, treatment of the surface with a short injection (5 - 30 seconds) of NHS/EDC after protein immobilisation can help to stabilise the chip surface by cross-linking the proteins to each other and to the dextran matrix. [257] The cross-linking step is followed with a pulse injection of ethanolamine to inactivate any remaining succinimide esters. In an attempt to stabilise the wild-type CD33 surface, cross-linking was employed (Figure 3-6). Following the NHS/EDC cross-linking stabilisation step, further buffer injections were carried out to ascertain the effectiveness of this stabilisation method. (Figure 3-7)

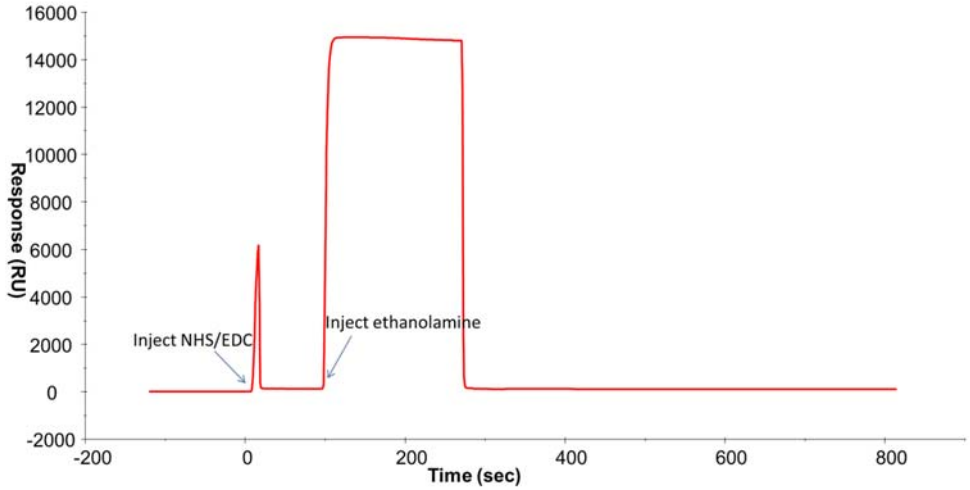


Figure 3-6 Sensorgram showing injection of NHS/EDC for 18 seconds to stabilise the surface, followed by injection of ethanolamine for 180 seconds to deactivate remaining succinimide esters.

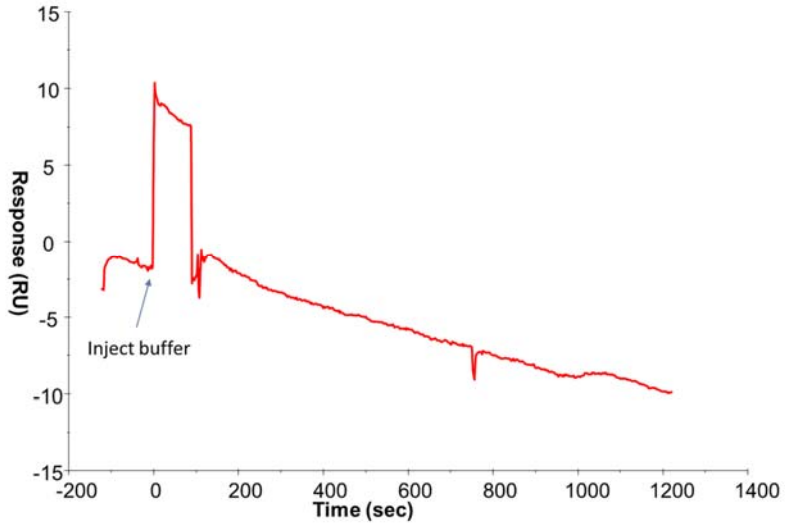


Figure 3-7 Sensorgram of the injection of buffer over the NHS/EDC cross-linked surface. Running buffer was injected at 30 µl/minute for 2 minutes, then the cycle continued for 20 minutes to observe the stability of the baseline.

Unfortunately, there was still considerable baseline drift after the cross-linking step (Figure 3-7), so the thiol-coupling method of immobilisation was abandoned.

3.3.2.2. Capture of His-tagged CD33 to NTA chip surface

Nitrilotriacetic acid (NTA) sensor chips capture His-tagged proteins using the chelation of Ni^{2+} by NTA on the chip surface and histidine residues in the protein poly-His tag. Wild-type CD33 IgV domain was expressed with an N-terminus His_8 tag and a TEV cleavage site to facilitate purification. The His_8 tag was chosen over the traditional His_6 to increase the tag's affinity to the immobilised Ni^{2+} on the IMAC resin used for purification, but also to increase the capture stability of the protein to the NTA chip. Capture via a His tag is advantageous as it orients the protein in a homogenous manner and the capture does not require the electrostatic pre-concentration step needed for covalent coupling, allowing the protein to remain in conformationally favourable buffer conditions. It is possible that lowering the pH for pre-concentration could cause the protein to lose activity.

For His capture the NTA chip was conditioned with a one-minute injection of 350 mM EDTA in running buffer and then the chip surface was charged with a one-minute injection of 0.5 mM $NiCl_2$. The NTA chip surface was then washed with a one-minute injection of 3 mM EDTA to remove excess Ni^{2+} before the protein was captured. Uncleaved CD33 with a His_8 tag was injected at 10 µl/minute for one minute at 200 nM in running buffer, resulting in the capture of approximately 2000 RU of protein onto the chip surface. (Figure 3-8) Typically the affinity of histidines to nickel is in the range of 1 µM and under ideal conditions would be sufficient to allow analysis of subsequent analyte binding. However, the microenvironment created by the moieties in proximity to the His-tag and the buffer environment can reduce this affinity and increase the dissociation of the His-tagged protein from the surface. [258, 259] Side-chains on the surface of the protein such as cysteine, lysine, tyrosine and tryptophan may bind to a chelated metal and high ligand concentrations increase the availability of these low affinity binding sites. [260] As such, the high ligand densities required to generate a binding response from a small molecule may produce complex binding curves and less stable ligand capture. These effects are likely responsible for the unstable baseline observed using this ligand capturing technique.

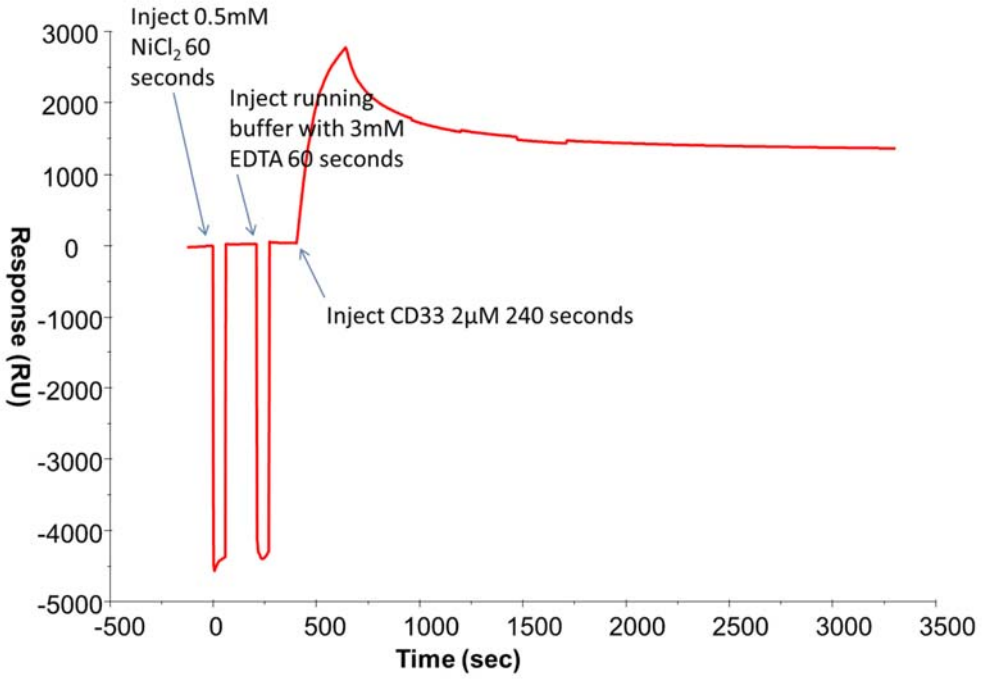


Figure 3-8 His-tagged wild-type CD33 IgV domain captured on an NTA chip. The chip surface was activated with a 60 second injection of $NiCl_2$ and washed with running buffer containing 3 mM EDTA prior to the protein being injected. This method captured approximated 2000 RU of CD33 protein. However, the baseline was not stable as evidenced by the slow decrease in response following CD33 injection.

As previously, 3'-sialyllactose and 6'-sialyllactose 0.5% solutions were injected separately across the captured wild-type CD33 protein surface but did not appear to bind. (Figure 3-9) The immobilised protein was also unstable, and the baseline continued to drift as the CD33 protein dissociated from the chip surface. This level of baseline drift would mask any sialyllactose binding. Repeated injections of buffer did not stabilise the surface and eventually all the captured protein was lost. A small amount of baseline drift is correctable with reference subtraction but the dissociation of the CD33 protein from the capture was too great for this surface to be useful for small molecule binding studies and an alternative method was sought.

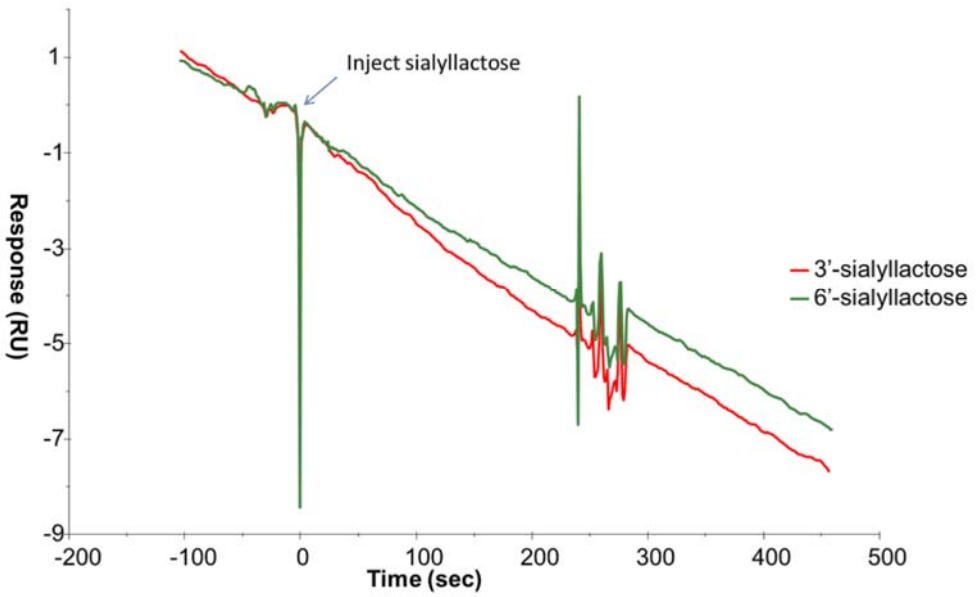


Figure 3-9 Sensorgram of the binding response of 3'-sialyllactose (red) and 6'sialyllactose (green) to His captured wild-type CD33 IgV domain. Samples were injected for 4 minutes at a rate of 30 µl/minute.

3.3.2.3. Capture/couple of His-tagged CD33 to an NTA chip

NTA sensor chips carry the same unmodified carboxymethyl groups as the chips used for covalent coupling, such as the CM series. After charging the NTA sensor chip with Ni^{2+} (as above in section 3.3.2.2.), the carboxy groups can be activated with EDC/NHS in a similar manner to covalent coupling. The affinity of the His-tagged protein for the nickel will concentrate the protein on the chip surface in the same way the electrostatic pre-concentration does for covalent coupling, with the added advantage of homogenous orientation of the protein and without the need to lower the pH or ionic strength of the buffer. Once the protein is captured via the His tag, any available primary amines from a lysine residue or the N-terminus of the protein can react to form a covalent bond with the succinamide esters in the dextran matrix. This method produces a permanently coupled, stable, non-drifting surface. [261] Using this method over 5500 RU of wild-type CD33 protein was immobilised on an NTA chip (GE) and a stable baseline was established. (Figure 3-10)

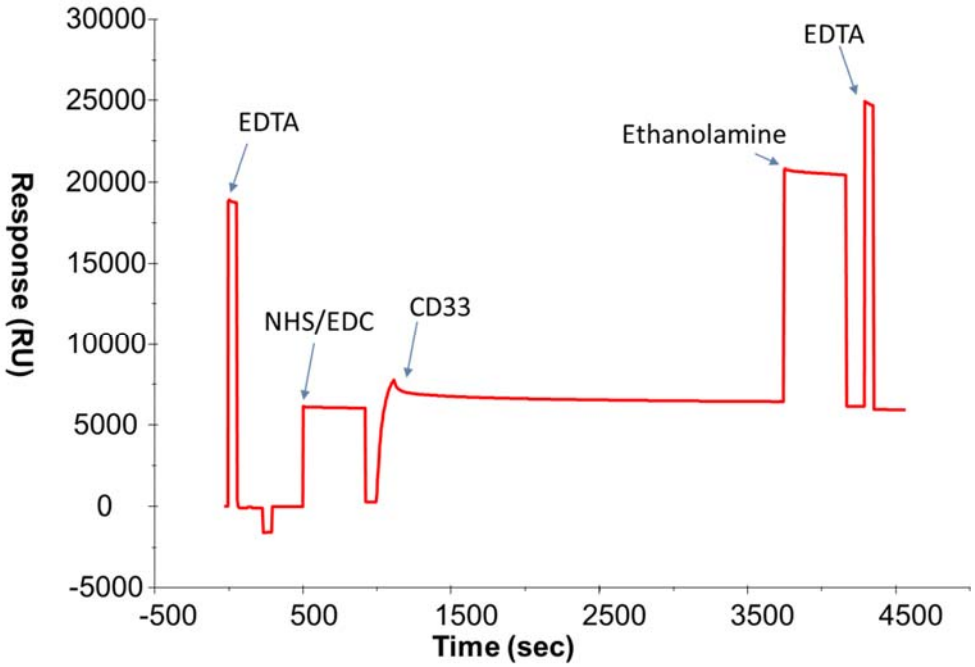


Figure 3-10 Capture/coupling of wild-type CD33 IgV domain on an NTA sensor chip. After charging the chip surface with Ni^{2+} and activating the carboxy groups with EDC/NHS, a 200 nM solution of CD33 was injected for 120 seconds at 10 µl/minute capturing over 5500 RU of protein on the chip. The remaining activated carboxy groups were blocked with ethanolamine and residual Ni^{2+}removed with EDTA.

3.3.3. Estimation of CD33 protein surface density on the sensor chip surface

Using the R_{max} formula (Equation 2-2) and based on an average analyte Mw of 300 Da for a small molecule screen, Mw of 15270 Da for wild-type CD33 IgV domain, a 1:1 binding interaction and an R_{max} of 100 RU, then 5090 RU of CD33 protein should be immobilised on the NTA sensor chip surface. This sensor chip surface density was achieved with CD33 thiol covalent coupling to CM7 chips but resulted in an unstable surface. The His tag capture technique was unsuitable as CD33 could not be captured to a high enough level onto the NTA chips and the resulting surface was unstable. The capture/couple immobilisation technique gave a suitable stable surface density of wild-type CD33 protein on the NTA chip.

3.3.4. Negative binding control protein

As a further assay control, CD33 R119 was mutated to an alanine (R119A). As previously described, R119 is highly conserved across the Siglec family and is involved in the critical interaction with the carboxyl group of the sialylated ligand. [125] This amino acid

substitution should disrupt ligand binding to the sialic acid-binding pocket of CD33. The R119A mutant was capture/coupled to the same density as the wild-type CD33 protein on a parallel flow cell on the same NTA chip (not shown). In all SPR experiments it is crucial to include surface controls to correct for the effects of signal drift and non-specific binding. Using the R119A mutant as the control surface would have simplified examination of small molecule binding specific to the wild-type protein, however as we were also interested in examining the effect of the mutation on binding kinetics the CD33 R119A mutant was treated as an additional target protein and not simply a negative control.

3.3.5. Buffer optimisation

Optimal buffer conditions for wild-type and R119A CD33 purification were examined previously in our laboratory by Jasmina Markulic using thermal melt analysis. CD33 was found to be most stable in 100 mM Tris-HCl 200 mM NaCl pH 8 and both the wild-type and R119A mutant melted at the same temperature (data not shown). [125] In the interest of preserving protein stocks and for time management only CD33 wild-type was used for buffer optimisation.

Screening of small molecule compounds requires the addition of DMSO to the running buffer to facilitate compound solubility. As some proteins can be sensitive to DMSO, a thermal melt assay was performed on CD33 wild-type across a range of DMSO concentrations from 0 to 11.25% in both TBS and PBS (Figure 3-11).

For the wild-type CD33 IgV domain a thermal melt temperature (Tm) of 46.3°C and 46.9°C was observed at 0% DMSO in TBS and PBS, respectively. In both buffers thermal melt analysis indicated that wild-type CD33 was stable in DMSO with a small decrease in stability between 5% and 11.25% (Figure 3-11). Based on these results, a maximum of 5% DMSO was used in all subsequent CD33 biophysical experiments. Concurrent small molecule-CD33 IgV domain NMR binding experiments conducted by Dr. Luke Miles (our laboratory) were run using PBS, so it was important to establish whether this buffer could affect the binding kinetics of the CD33 protein. From thermal melt data shown in Figure 3-11, it appears that wild-type CD33 behaves similarly in both PBS and TBS.

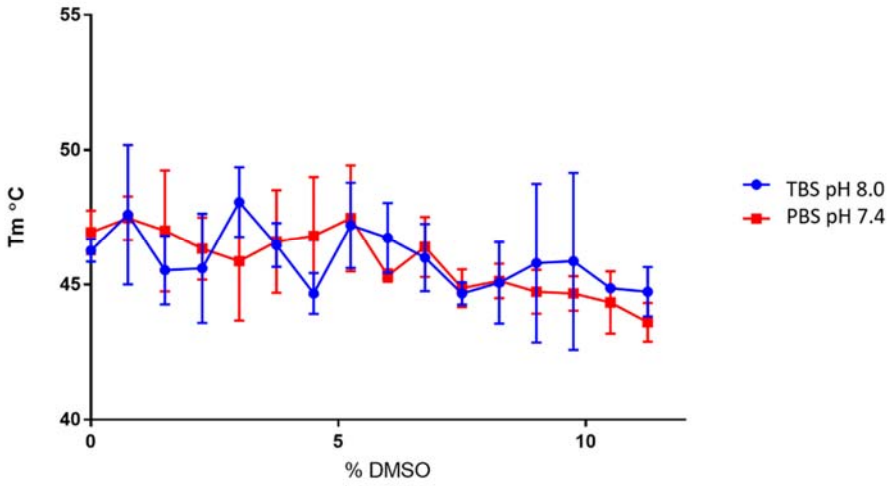


Figure 3-11 Plot of the thermal melt temperature (Tm) of wild-type CD33 IgV domain in TBS pH 8 (blue) and PBS (red) with increasing amounts of DMSO. Samples were run in triplicate; the mean and standard deviation are shown.

3.3.6. Compound selection process

As described in section 1.3.1, the crystal structure of the CD33 extracellular region was deposited in the PDB (PDB ID: 5IHB, 5J06 and 5J0B) in 2017 (unpublished). The three structures deposited by Dodd et al. are the apo CD33 R69G mutant (PDB ID:5IHB), and complexes with 3'-sialyllactose (PDB ID: 5J06) and 6'-sialyllactose (PDB ID: 5J0B) arising from ligand soaks. A higher resolution apo wild-type CD33 IgV domain structure has also been solved by our laboratory (PDB ID: 6D48). [125] The sialic acid-binding region of CD33 was identified by interaction of the sialyllactose ligands with the conserved arginine residue (R119) and compared to that of all published ligand-Siglec IgV domain crystal structures. The region where sialylated ligands bind to the Siglec IgV domain is quite flat and would be considered undruggable. The GT1b analogue from the hSiglec-7 complex (PDB ID:2HRL)[262] contains a 2-(trimethylsilyl)ethyl substituted glucose (Glc) moiety that occupies a crevice near the conserved arginine residue (i.e. R119 in CD33) (Figure 3-12). A similar crevice is observed in the crystal structures of all published Siglecs. The 2-(trimethylsilyl)ethyl substituted Glc moiety is a similar Mw (approximately 280 Da) and physical size as a small molecule and suggested that the crevice may be a hot-spot suitable for small molecule binding (further rationale is outlined in section 1.1.4).

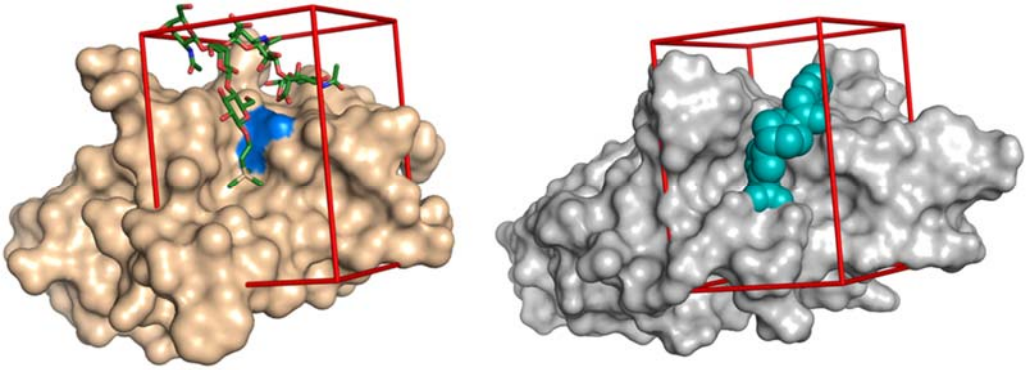


Figure 3-12 Structure of the hSiglec-7 IgV domain in complex with the GT1b analogue and comparison with CD33. hSiglec-7 is depicted as a light brown molecular surface and the GT1b analogue as green sticks. The location of the conserved arginine residue in hSiglec-7 is highlighted in blue. (PDB ID: 2HRL, left) [262]. The 2-(trimethylsilyl)ethyl sidechain on the glucose sugar hangs down into the crevice near the conserved critical arginine residue. The structure of wild-type CD33 IgV (PDB ID: 6D48, unpublished) is shown in a similar orientation to that of hSiglec-7 and depicted with a grey molecular surface (right). A small molecule compound identified from the virtual screen (cyan spheres) is shown bound to the equivalent crevice near the critical R119 residue of CD33. The red box denotes the virtual screen search area.

Dr. Tracy Nero (our laboratory) used the high-resolution apo wild-type CD33 IgV domain structure for virtual screening of our in-house library of commercially available drug-like compounds to identify small molecules that could bind to the conserved crevice and inhibit endogenous ligand binding to CD33. The virtual screen was carried out using FRED V3.2.0.2. (OpenEye, Santa Fe, NM USA) on our dedicated cluster of Linux computers. Over 10 million commercially available compounds, which include databases from the National Cancer institute, Sigma-Aldrich, Maybridge, Interbioscreen, Chembridge, Enamine and Asinex, were screened. The virtual screening results were sorted on ChemGauss4 docking scores and the top 1000 ranked compounds inspected. A set of 75 chemically and structurally diverse compounds was selected and purchased. The 75 compounds were then screened for solubility and non-specific binding to the SPR sensor chip surface.

3.3.7. Compound solubility assessment

The 75 compounds, and subsequent analogues of compounds 2952 and 2971 that were later purchased (refer to section 3.3.6.), were found to have varying levels of solubility in 100% DMSO or in aqueous buffer containing 5% DMSO. Solubility was assessed by visual examination for particulate matter using a Leica zoom 2000 illuminated stereo microscope. Stock solutions of compounds were prepared at 10 mM in 100% DMSO. The highest concentration used in aqueous buffer + 5% DMSO was 500 µM. To dilute the

compounds, the buffer was prepared at 1.05 x concentration and then a 1/20 dilution of the 100% DMSO dissolved compound was made. This results in a final concentration of compound at 500 µM in 1 x buffer with 5% DMSO. When compounds were insoluble using this method of dilution, two different approaches were taken:

1. To 5 µl of compound in 100% DMSO, 5 µl of aqueous buffer without DMSO was added. Once the compound was dissolved in 50% DMSO, 10 µl of aqueous buffer was then added. The compound was incrementally diluted in this manner until DMSO was diluted to 6.25% then the final dilution was into buffer + 5% DMSO. The discrepancy in DMSO concentration is accounted for with a solvent correction curve.
2. The compound was diluted in 100% DMSO to 200 µM then diluted in 1.05 x buffer to 10 µM (1/20).

If neither of these methods solubilised the compound it was excluded from analysis. Of the original 75 compounds, all but 2 were soluble at 500 µM in TBS containing 5% DMSO. The same solubility analysis was carried out for the analogues of compounds 2952 and 2971. The 27 analogues of 2952 were all soluble at 500 µM in TBS containing 5% DMSO, whereas only 10 of the 60 2971 analogues were soluble.

3.3.8. Clean screen

All the soluble compounds were screened against a blank NTA sensor chip to identify non-specific binding prior to testing against the CD33 protein. Eight of the initial compounds and one 2952 analogue displayed binding to the chip surface and were excluded from further analysis.

3.3.9. Compound screening

CD33 IgV domain was capture/coupled on an NTA sensor chip to 3163 RU. The Rmax formula (Equation 2-2) calculates that for a Rmax of 20 RU, which is an appropriate response for a small molecule, 1000 RU of CD33 protein should be immobilised. This calculation assumes that the protein is 100% homogenous and biologically active. To compensate for any misfolded or denatured protein in the sample this number was tripled. Compounds were screened, in duplicate at 100 µM, against the immobilised wild-type and R119A mutant CD33 IgV domain (Figure 3-13). A regeneration step with a 30 second pulse of 100 mM HCl was included to ensure all compound was dissociated from the chip surface at the end of each cycle.

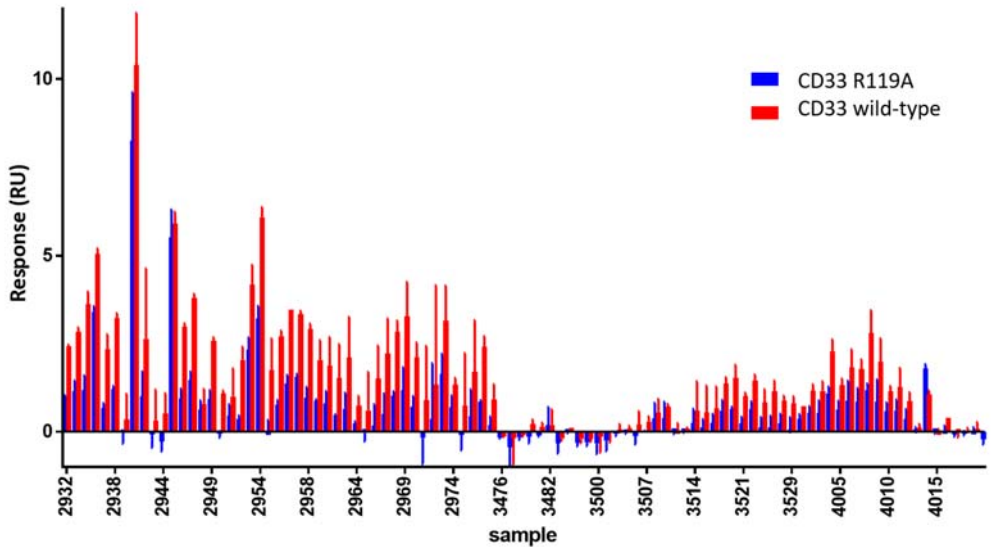


Figure 3-13 Compounds binding to wild-type and R119A CD33 IgV domain. Binding to wild-type CD33 is shown in red and binding to the R119A mutant in blue. Compounds that bound preferentially to the wild-type protein were selected for further analysis. Data are means and standard deviation calculated from *n* = 2 experiments.

The binding curves were careful examined. Compounds that generated binding curves displaying a typical association and dissociation phase, the amount of binding was within the calculated Rmax (<60 RU) and were reproduced for each experiment were considered positive binders. These positive binders were then compared to assess the differences between the binding to wild-type CD33 IgG domain and to the negative control arginine mutant. Compounds that bound preferentially to the wild-type over the arginine mutant were considered to be potentially specific to the sialic acid binding site. This assessment led to the identification of 17 compounds showing specific binding to the wild-type CD33 IgV domain (Figure 3-14).

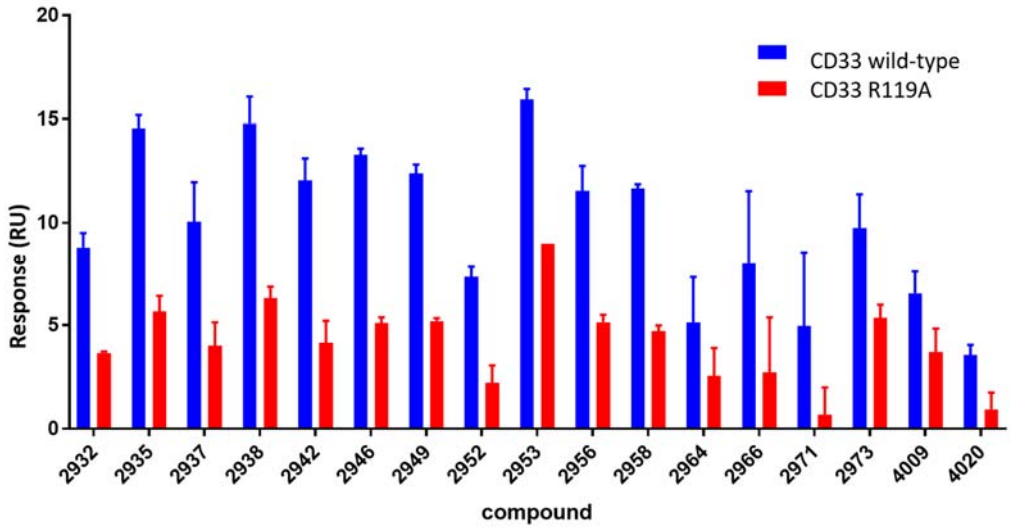


Figure 3-14 Plot of the positively binding compounds selected for further analysis. Binding to CD33 wild-type shown in blue and binding to the R119A mutant in red. Data are means and standard deviation calculated from *n* = 2 experiments.

The 17 compounds identified as potential hits were tested at five concentrations between 7.8 and 500 µM; 6 compounds were confirmed with affinities for CD33 ranging from K_D = 70 µM – 473 µM (Table 4, Figure 3-15).

Table 4 Kinetic evaluation of compounds binding to CD33 wild-type and R119A mutant. Samples were tested in duplicate and values given are the average.

	CD33 WILD-TYPE			**CD33 R119A**		
	k_a (1/Ms)	k_d (1/s)	K_D (M)	k_a (1/Ms)	k_d (1/s)	K_D (M)
2932	233	0.07	0.000287	No binding		
2943	205	0.04	0.000188	No binding		
2952	133.9	0.01	0.00007	No binding		
2953	96.84	0.04	0.000434	No binding		
2966	167.2	0.08	0.000473	No binding		
2971	372.4	0.05	0.000143	66.41	0.07	0.001

The two best binders, based on affinity and specificity, were deemed to be compounds 2952 and 2971. (Table 4, Figure 3-15) Although 2971 did appear to bind to the R119A

mutant, the affinity was low and it is possible that this compound has an alternative binding mode (or binding location) to the other compounds.

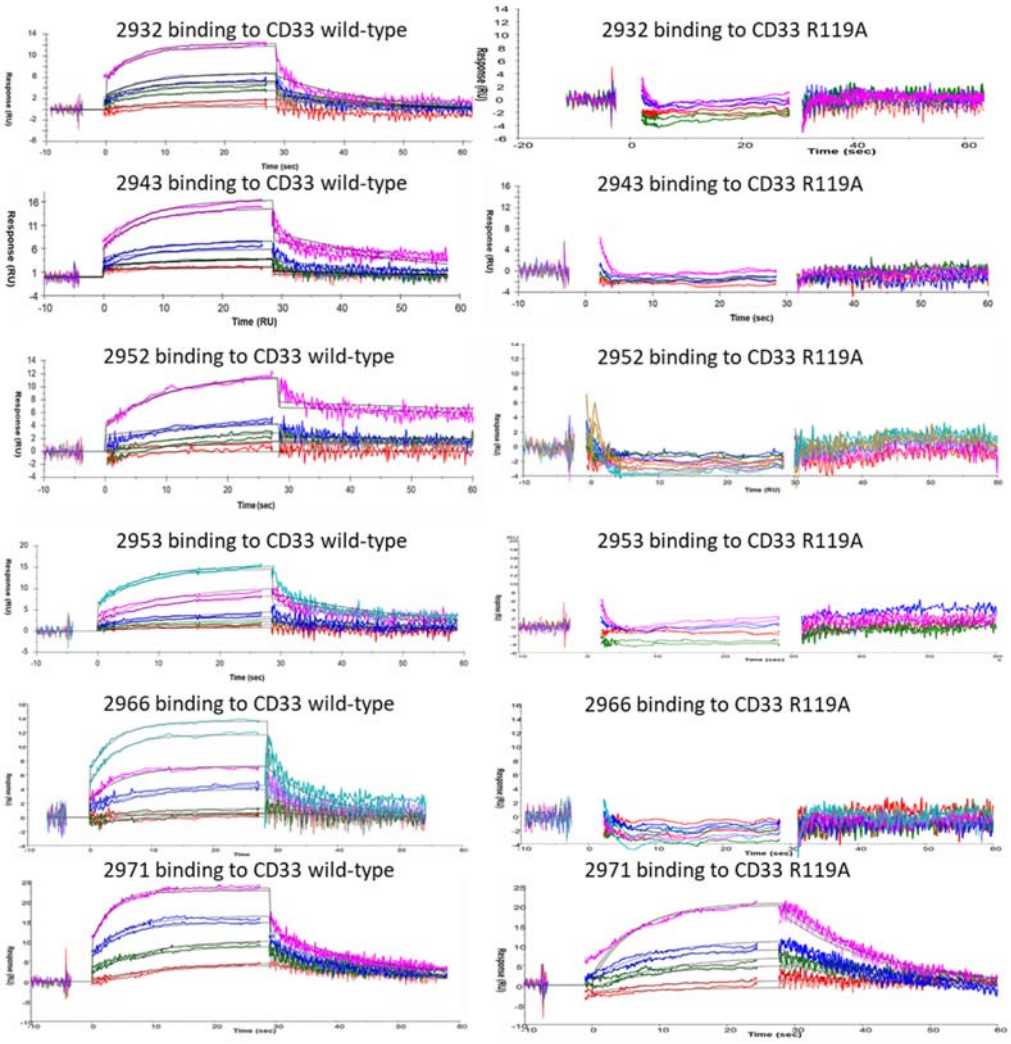


Figure 3-15 Binding of compounds 2932, 2943, 2952, 2953, 2966 and 2971 to CD33 wild-type (left) and the R119A mutant (right). Compounds were run at the following concentrations: 250, 125, 62.5, 31.25, 15.6, 7.8 µM. Black lines show the fit to a 1:1 binding model. Compounds were injected for 30 seconds at 60 µl/minute, in duplicate.

Although compound binding was observed and 6 compounds with binding affinities below 500 µM for wild-type CD33 were identified, there were two main problems encountered during these SPR studies:

1. The purification of the CD33 IgV domain involved refolding denatured insoluble protein extracted from inclusion bodies. The free IgV domain cysteine residue (C36) seemed to cause protein instability during purification, possibly due to

misfolding resulting from incorrect disulfide bond formation within the CD33 IgV domain and/or disulfide bond formation between CD33 molecules. Misfolded protein was potentially a cause for the low binding activity observed for the positive control ligands, 3'- and 6'-sialyllactose, with the immobilised CD33 protein.

2. The GE NTA sensor chips are designed for the repeated stripping and recapturing of the His-tagged protein and can be reused many times. When used for the permanent capture/couple method, however, they become single use and are not cost effective.

To address the problems encountered during CD33 IgV domain purification, the C36 residue was mutated to serine (C36S). Mutating out the free cysteine residue increased protein stability and streamlined the refold/purification process (not shown). Biotinylation of the C36S CD33 IgV domain was investigated to provide an alternative chemistry for immobilisation on the surface of SPR chips, and the use of more cost-effective sensor chips was explored.

3.3.10. Quality assessment of CD33 IgV domain C36S mutant and C36S, R119A double mutant

Protein production and purification was carried out by Dr. Stefan Herman. In order to confirm that the protein was of correct size the samples were analysed by LC-TOF (section 3.3.10.1) and tertiary structure the protein samples were analysed using CD (section 3.3.10.2). Protein that was evaluated as correct size and structure by these methods was considered suitable for use in SPR.

3.3.10.1. LC-MS analysis

Samples of the CD33 C36S and C36S, R119A mutant were analysed by mass spectrometry.

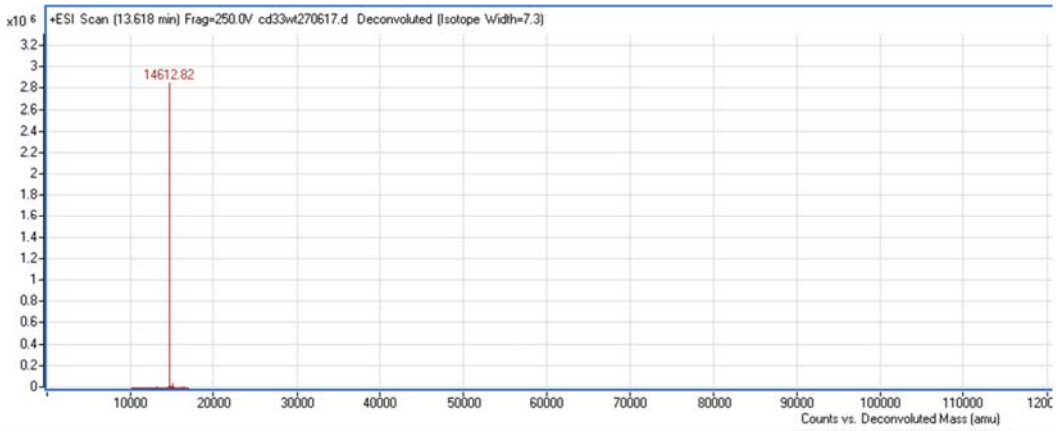


Figure 3-16 LC-TOF analysis of the purified CD33 C36S IgV.

The predicted monoisotopic mass of CD33 C36S IgV is 14614.3 Da. LC-TOF analysis shows a single protein peak with a predicted protein mass of 14612.82 Da. The difference in mass is likely due to the disulfide bond forming between C41 and C101 resulting in the loss of two H^+ (Figure 1-8). This result confirms that the protein is the correct size and the disulfide bond has formed.

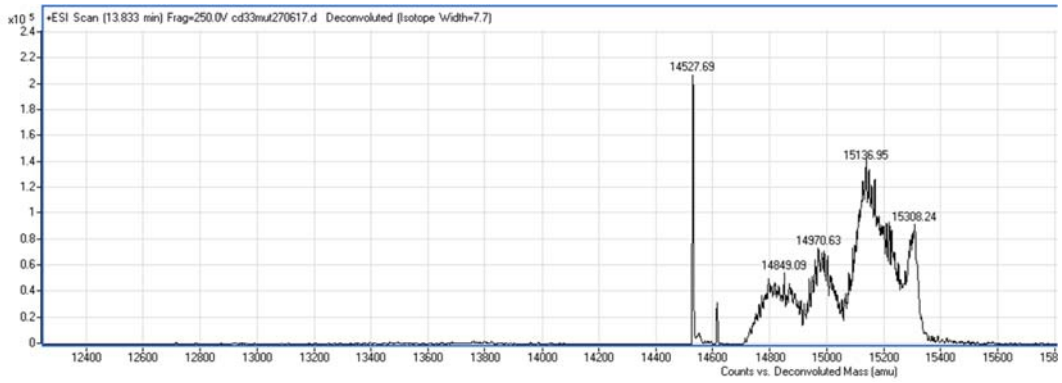


Figure 3-17 LC-TOF analysis of the purified CD33 C36S R119A IgV.

The predicted monoisotopic mass of CD33 C36 S R119A is 14545.2 Da. LC-TOF analysis shows the main protein peak with a predicted protein mass of 14527.39 Da. (Figure 3-17). The difference in mass of 18 Da is too large to be accounted for by the predicted disulfide, and the reason for the discrepancy is unknown. It may be due to a demethylation or dehydration event. The spectra showed some contaminating protein peaks which may be contributing to the anomalous result. This protein was further purified using SEC to remove any contaminating proteins (data not shown).

3.3.10.2. CD analysis of the CD33 IgV domain C36S mutant and C36S, R119A double mutant

A sample of the CD33 C36S and C36S, R119A mutant was buffer exchanged into 50 mM NaF and analysed by CD using the Contin-LL method and reference data set 4. [263] The spectra for both proteins concurs with the crystallographic structure (Figure 1-8) that the protein is mostly strand, indicating the proteins likely have correct tertiary structure. (Figure 3-18, Table 5, Figure 3-19, Table 6)

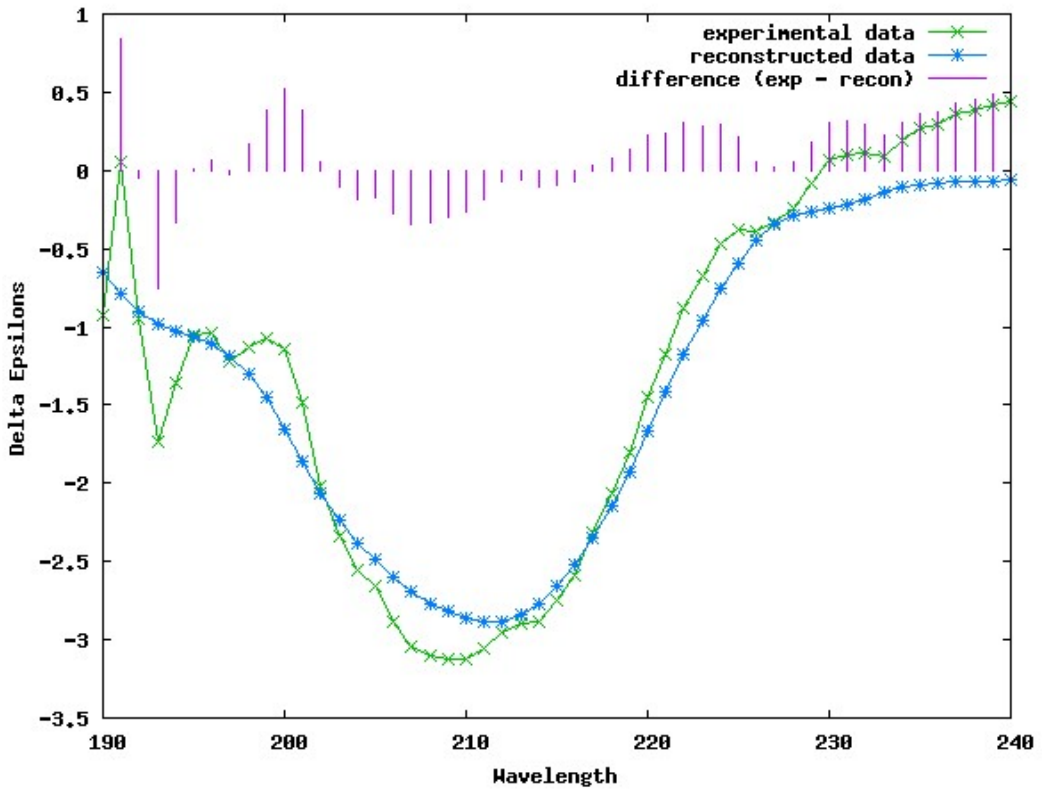


Figure 3-18 CD spectra of the CD33 C36S IgV. Experimental data are plotted in green; the calculated spectrum derived from the calculated output secondary structure is plotted in blue and the difference spectra is depicted in vertical lines in pink. The plot shows a reasonable fit to the calculated curve with some deviations in the lower wavelengths.

Table 5 The content of secondary structure predicted from the CD spectra of CD33 C36S IgV.

Result	Helix1	Helix2	Strand1	Strand2	Turns	Unordered	Total
1	0.044	0.036	0.247	0.112	0.210	0.351	1
2	0.017	0.031	0.237	0.115	0.232	0.367	0.999

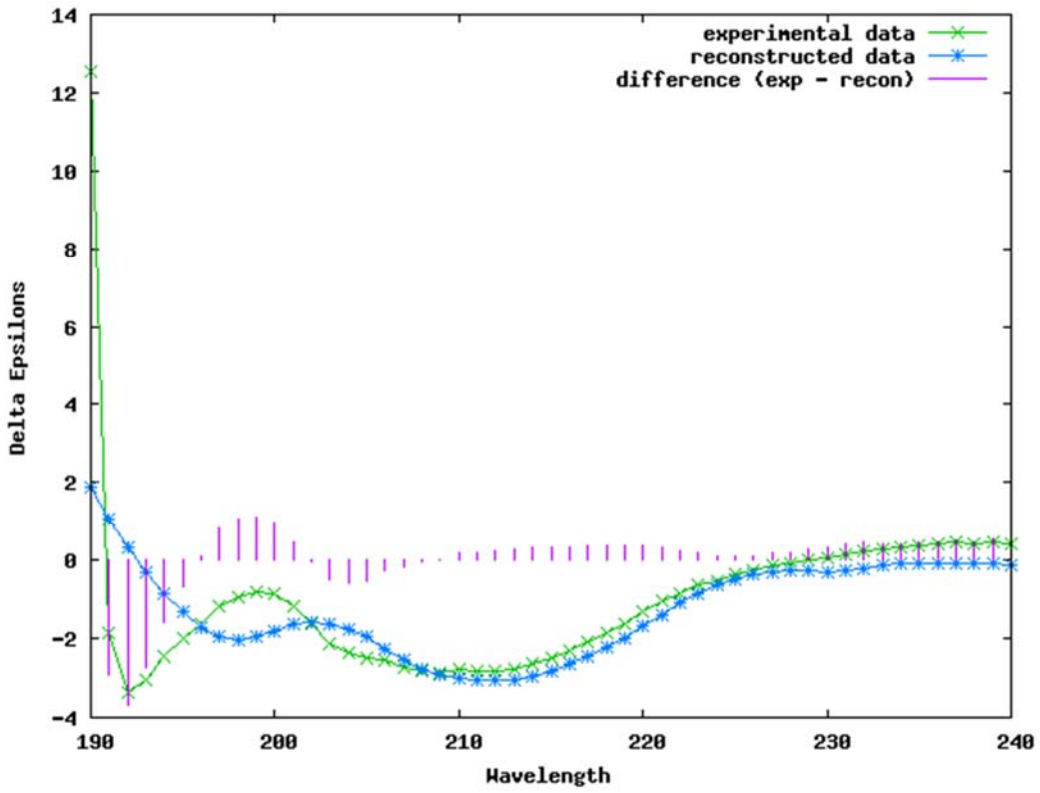


Figure 3-19 CD spectra of the CD33 C36S R119A IgV. Experimental data are plotted in green; the calculated spectrum derived from the calculated output secondary structure is plotted in blue and the difference spectra is depicted in vertical lines in pink. The experimental data shows a good fit to the calculated curve.

Table 6 The content of secondary structure predicted from the CD spectra of CD33 C36S R119A IgV.

Result	Helix1	Helix2	Strand1	Strand2	Turns	Unordered	Total
1	0.000	0.000	0.279	0.000	0.000	0.721	1
2	0.000	0.004	0.390	0.065	0.365	0.176	1

3.3.11. Immobilising the CD33 IgV domain C36S mutant and C36S, R119A double mutant

3.3.11.1. Immobilising CD33 by biotin-streptavidin capture

Biotin is a small (244 Da), water-soluble vitamin that binds with a high affinity and specificity to avidin and streptavidin. SPR sensor chips pre-coated with streptavidin allow for the capture of biotinylated protein resulting in a highly stable surface. The avidin-biotin complex is the highest affinity non-covalent interaction known, with a K_D of 10^{-15} M, and is resistant to extremes of pH, temperature and proteolysis. Two different methods of biotinylation, of both the C36S CD33 IgV domain protein and the C36S, R119A double mutant were trialled: chemical biotinylation via an amine residue and enzymatic biotinylation via an AviTag™.

3.3.11.2. Chemical biotinylation of CD33 protein

CD33 C36S and CD33 C36S, R119A were chemically biotinylated using EZ-Link™ Sulfo-NHS-LC-LC-Biotin reagent (ThermoFisher Scientific # 21338). The N-hydroxysulfosuccinimide (NHS) esters of the reagent reacts with primary amines (-NH_2) on the side-chain of lysine residues to form a permanent amide bond biotinylating the protein. The LC-LC component of the reagent is a pegylated spacer arm which extends to a distance of 30.5 Å, this flexible arm helps to minimise any steric hindrance of the protein's ligand binding site caused by the biotin. [264]

The CD33 proteins were buffer exchanged into PBS to remove primary amines present in the Tris buffer. EZ-Link reagent was added to the protein at a 0.8:1 molar ratio. This low ratio is to ensure the protein is "minimally biotinylated" i.e. an average of one biotin molecule or less per protein molecule. The reaction mixtures were incubated on ice for one hour, then any excess or unreacted reagent was removed by size exclusion chromatography. This procedure ensured the homogeneity of the protein was retained after buffer exchange and biotinylation.

CD33 C36S-biotin and CD33 C36S, R119A-biotin were captured to an average of 3000 RU on separate flow channels of a SAHC 1500M sensor chip pre-coated with streptavidin in a linear polycarboxylate hydrogel (XanTec). This was the saturation level of capture i.e. injecting more CD33 protein did not increase the capture level (Figure 3-20).

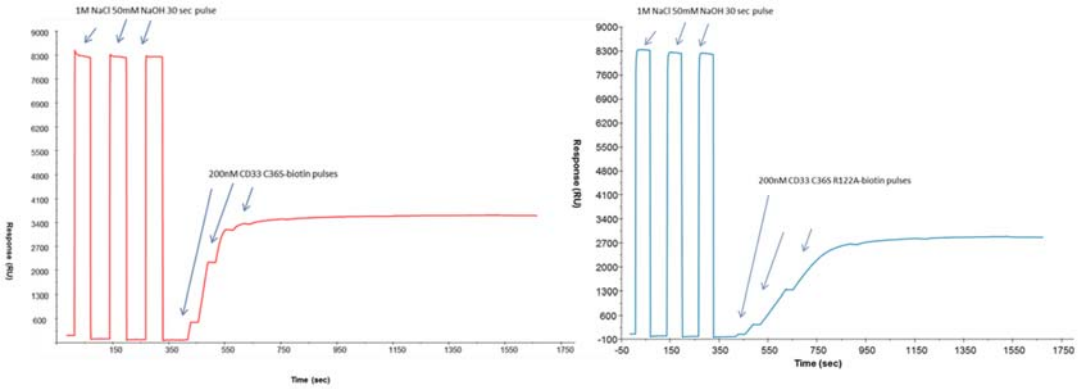


Figure 3-20 Capture of chemically biotinylated CD33 C36S on flow cell 2 and CD33 C36S, R119A on flow cell 4 of an SAHC sensor chip preimmobilised with streptavidin. The chip surface was conditioned with three 30 second pulses of 1 M NaCl in 50 mM NaOH prior to injecting the protein. CD33 protein was captured to a density of approximately 3000 RU.

3.3.11.3. AviTag™-CD33 protein

A non-cleavable AviTag™ was included on the C-terminal end of the CD33 constructs with a TEV cleavable octa-His tag on the N-terminus to facilitate purification. Once CD33 was purified, the His tag was cleaved and the protein was enzymatically biotinylated using a Beira Biotin-protein ligase reaction kit (cat # BirA500 Avidity LLC), which covalently links a single biotin to the 15 amino acid peptide tag. The control protein CD33 C36S, R119A was constructed in the same manner.

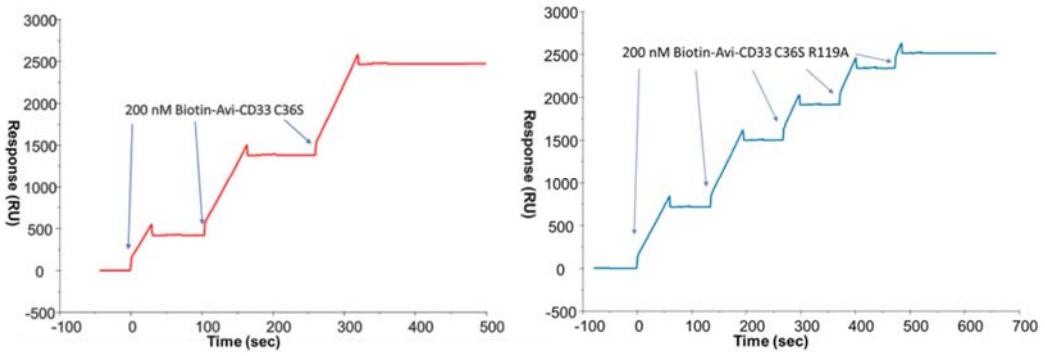


Figure 3-21 Capture of biotinylated AviTag™-CD33 C36S and CD33 C36S, R119A on flow cells 2 and 4, respectively, of an SAHC 1500M sensor chip. Chip conditioning was performed separately. CD33 protein was injected in short bursts at 10 µl/minute until the capture level was approximately 2500 RU.

CD33 C36S and CD33 C36S, R119A were captured to approximately 2500 RU on an SAHC 1500M sensor chip (XanTec) as above (section 3.3.11.2)(Figure 3-21).

3.3.11.4. Stability of CD33 C36S proteins

A comparison of the stability of the CD33 proteins biotinylated by different methods was performed using thermal melt analysis (DSF). Although the AviTag™-biotin construct generally showed a lower Tm than the chemically biotinylated protein, the response in different buffer conditions was consistent. (Figure 3-22) These findings showed that the protein constructs exhibited similar dependence of protein stability on buffering conditions.

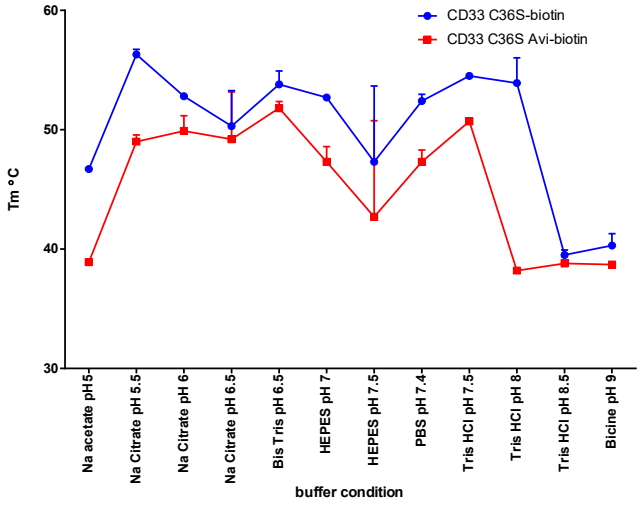


Figure 3-22 A comparison of the effect of the two different biotinylation methods on CD33 protein stability. The chemically biotinylated protein (blue) generally has a higher Tm than the enzymatically biotinylated protein. However, the response in each condition was consistent. Samples were tested in duplicate and means and standard deviation are shown.

3.3.12. Mass transfer control

Before the analyte can bind to the ligand it needs to move from the bulk solution towards the sensor chip surface. This initial movement is known as mass transfer and arises due to convection and diffusion of the analyte. [265] The rate of mass transfer is dependent on the diffusion coefficient of the analyte and the flow rate of the solution. Mass transfer limitations arise when diffusion of the analyte in bulk solution is slower than the binding rate of the analyte to the ligand, resulting in an artificial limitation of analyte association kinetics due to reduced local concentration of analyte. Kinetic constants calculated under these conditions reflect the mass transfer rate rather than true binding kinetics. [253] A mass transfer control experiment establishes if the observed binding rate varies with flow rate. The kinetic models used in the Biacore T200 Evaluation software include a term for mass transfer; however, it is prudent to empirically examine the effects of fluid dynamics on binding to negate mass transfer effects. To this end binding of CDC33 C36S to immobilised P22-biotin on a SAHC sensor chip was examined at three different flow rates (Figure 3-23).

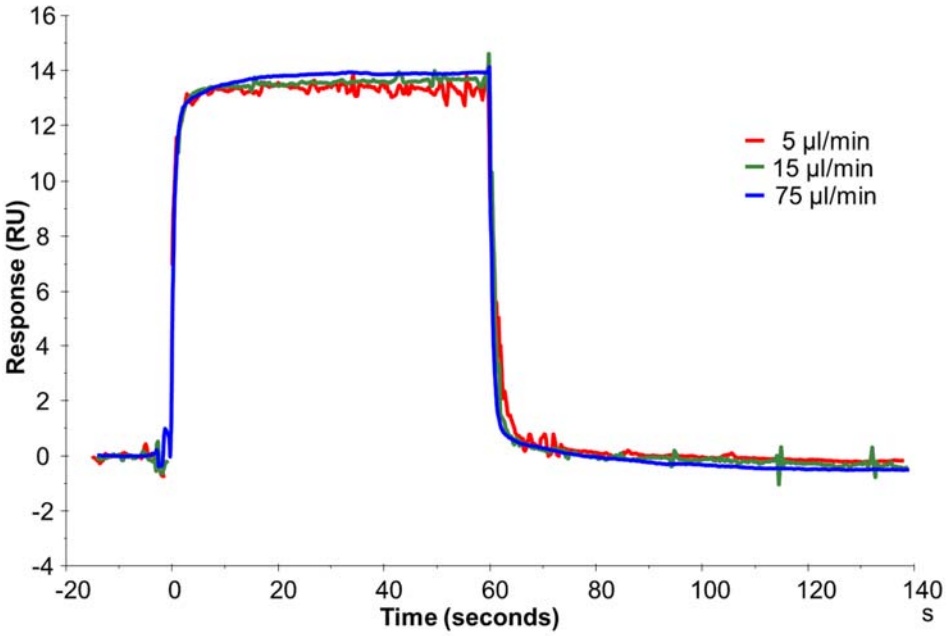


Figure 3-23 Binding of CD33 C36S at 0.1 µM to immobilised P22 at three different flow rates.

The mass transfer control experiment (Figure 3-23) showed that there was little difference in binding of the CD33 C36S to P22 at different flow rates. The amount of binding was consistent at different flow rates; however, at the higher flow rate of 75 µl/minute the curve displayed a higher signal to noise ratio, and dissociation was slightly

faster. The Biacore T200 specifications indicate that the instrument baseline noise should be < 0.03 RU. [266] As the noise associated with the 5 and 15 µl/min binding curves was greater than 0.03 RU there is a chance the noise could mask binding signal. As such, all future experiments were run at >60 µl/minute flow rate.

3.3.13. Validation of C36S mutants

Several high affinity sialic acid analogues selective for CD33 have been identified by Paulson et al. [267] Compound 22 (P22), a 2,5,9-trisubstituted sialic acid mimetic, was reported to have an IC_{50} of 11 µM using a flow cytometry assay. Its relatively small size and high binding affinity for the CD33 IgV domain made it an ideal positive control for SPR binding assays. The compound became available to use for this project in my latter experiments although its availability and high cost of synthesis only allowed limited use. (Figure 3-24).

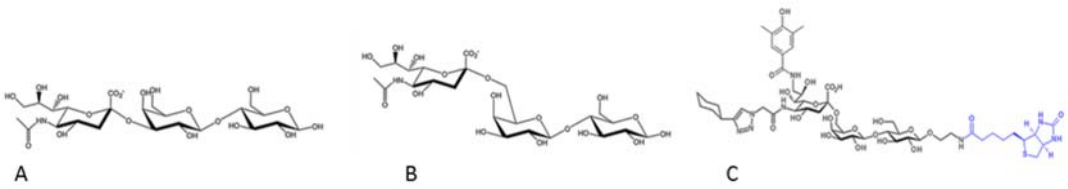


Figure 3-24 Structure of (A) 3'-sialyllactose, (B) 6'-sialyllactose and (C) P22 with the additional biotin shown in blue.

The binding of the P22 compound to CD33 IgV was assessed to the wild-type (Figure 3-26), the C36S mutant (Figure 3-25), and the R119A negative control mutant (Figure 3-27). His-tagged CD33 wild-type IgV domain was captured on a Series S NTA chip (GE) via Ni^{2+}/NTA chelation then covalently immobilised using amine coupling chemistry (GE Healthcare Laboratory Guidelines 29-0057-17 AB). Biotinylated AviTag™- CD33 C36S was immobilised on a SAHC 1500M sensor chip (XanTec) via streptavidin capture, to an average density of 6500 RU. P22 binding was assessed over a concentration range from 3 to 243 µM.

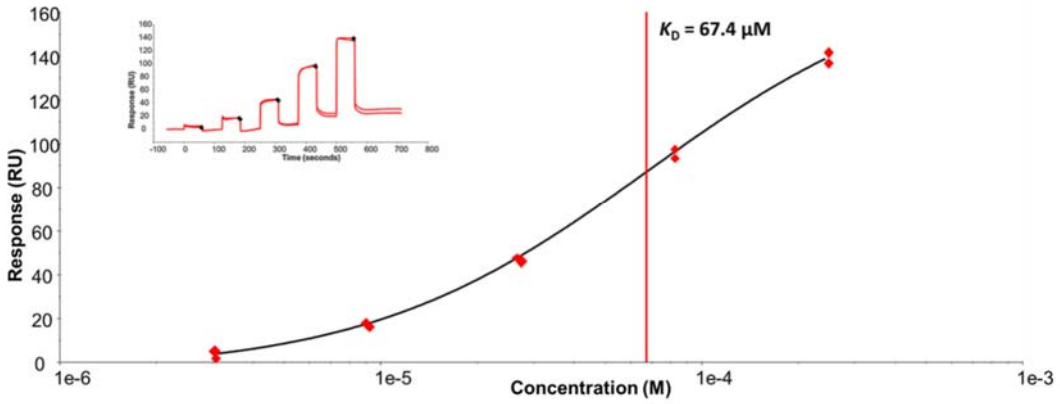


Figure 3-25 Equilibrium analysis of P22 binding to immobilised CD33 C36S. Response at equilibrium (t = 45 – 55 seconds, indicated by the black dot, insert) plotted against P22 concentration. Solid black line represents the fit of a Langmuir isotherm to the experimental data points yielding K_D = 67.4 µM, with 50% R_{max} indicated by the vertical red line. P22 was tested at concentrations of 3, 9, 27, 81, 243 µM. Binding measurements were performed in duplicate and all data sets are overlayed (insert).

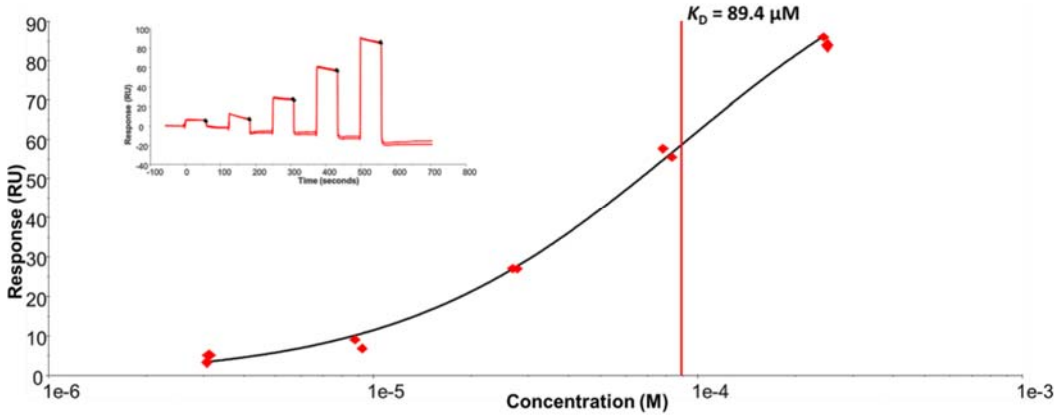


Figure 3-26 Equilibrium analysis of P22 binding to immobilised CD33 wild-type. Response at equilibrium (t = 45 – 55 seconds, indicated by the black dot, insert) plotted against P22 concentration. Solid black line represents the fit of a Langmuir isotherm to experimental data points, yielding K_D = 89.4 µM, at 50% R_{max} indicated by the vertical red line. P22 was tested at concentrations of 3, 9, 27, 81, 243 µM. Binding measurements were performed in duplicate and all data sets are overlayed (insert).

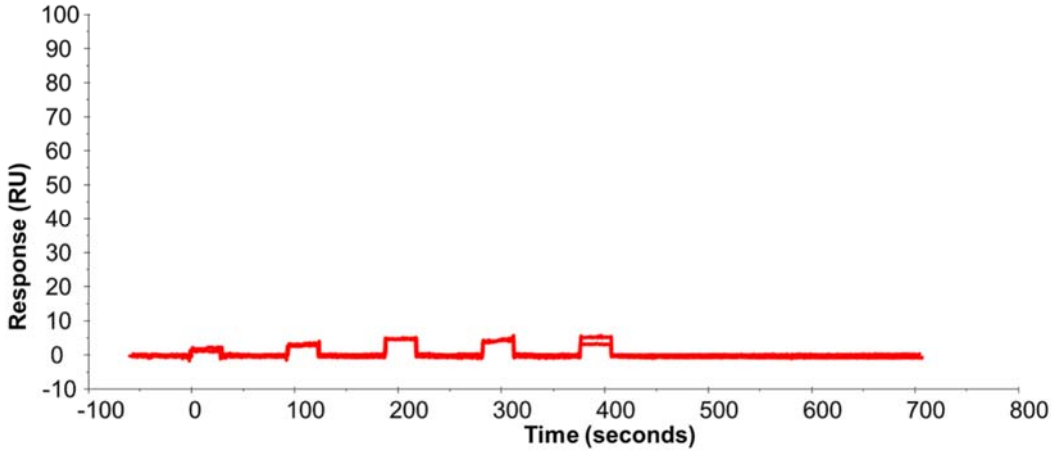


Figure 3-27 P22 binding to immobilised negative control protein CD33 C36S, R119A. P22 was tested at concentrations of 3, 9, 27, 81, 243 µM. Binding measurements were performed in duplicate and all data sets are overlayed.

Binding analyses gave similar results for P22 binding to the the wild-type CD33 IgV domain (K_D = 89.4 µM, Figure 3-25) and C36S mutant (K_D = 67.4 µM, Figure 3-26), indicating that the C36S does not disrupt binding, as anticipated. P22 did not bind to the R119A mutant, which supports the premise that P22 binds specifically to the sialic acid binding site. The binding data, was however, problematic. Non-specific binding of the P22 to the C36S R119A mutant (Figure 3-27) is evident as there is a slight increase in response at higher concentrations of P22. There is also baseline drift evident with P22 binding to the wild-type (inset, Figure 3-26). To address this issue the analysis was repeated in the reverse orientation with the biotinylated-P22 compound immobilised via streptavidin capture on a Xantec SAHC chip (Figure 3-28, Figure 3-29, Figure 3-30) to an average density of 570 RU.

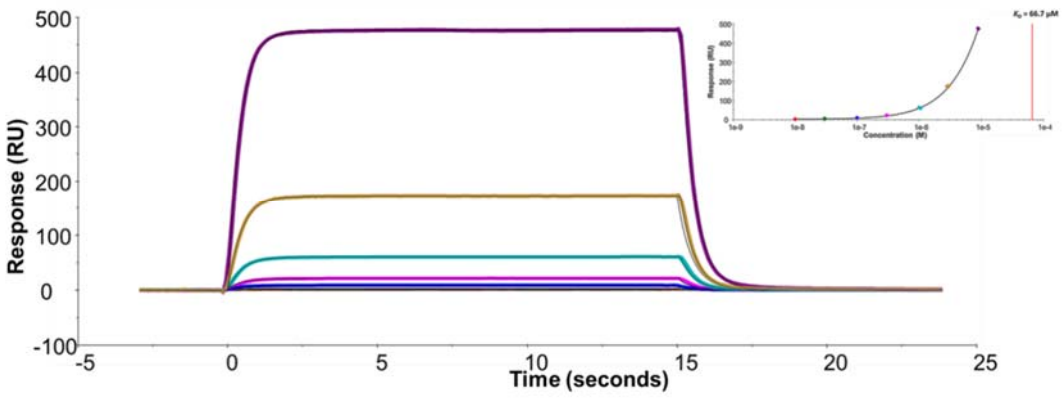


Figure 3-28 Kinetic analysis of CD33 C36S binding to immobilised P22-biotin. K_D = 57.4 µM was calculated using a global fit 1:1 binding model (black lines). Inset shows the equilibrium analysis of the same data (K_D = 66.7 µM). CD33 C36S was tested at concentrations of 0.03, 0.1, 0.3, 1, 3, 9 µM. Binding measurements were performed in duplicate and all data sets are overlayed.

CD33 C36S bound to immobilised P22 with K_D = 57.4 µM (Figure 3-28), very close to the K_D from the reverse orientation (K_D = 67.4 µM, Figure 3-25). Equilibrium analysis of the same data gave a much closer K_D = 66.7 µM to that from the reverse orientation (K_D = 67.4 µM, Figure 3-25), however the data indicated that the concentration range was insufficient as the isotherm does not reach an inflection point (inset, Figure 3-28). The K_D derived from equilibrium analysis parallels that obtained by kinetic analysis (K_D = 57.4 µM); however, the lack of binding data at higher concentration values reduces confidence in the equilibrium derived value, suggesting that the K_D from kinetic analysis is more accurate.

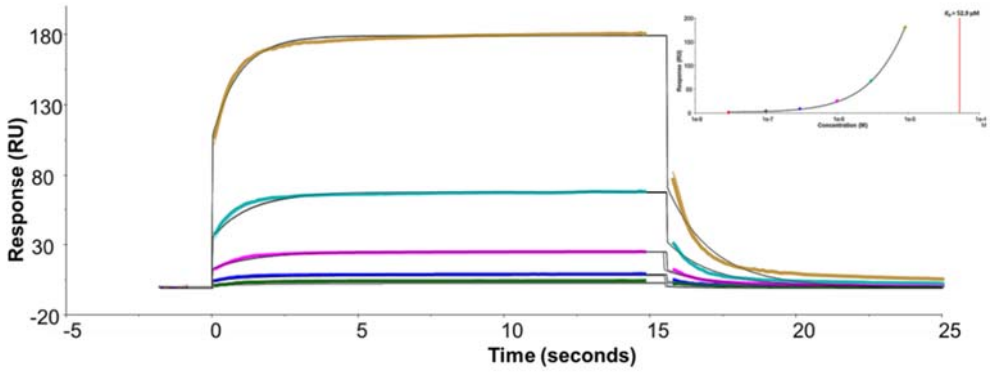


Figure 3-29 Kinetic analysis of CD33 wild-type binding to immobilised P22-biotin. A K_D = 18 µM was calculated using a global fit 1:1 binding model (black line). Inset shows the equilibrium analysis of the same data (K_D = 52.9 µM). CD33 wild-type was tested at concentrations of 0.03, 0.1, 0.3, 1, 3, 9 µM. Binding measurements were performed in duplicate and data sets are overlayed.

Wild-type CD33 IgV bound to immobilised P22 with K_D = 18 µM (Figure 3-29). The K_D derived from equilibrium analysis of 52.9 µM (insert, Figure 3-29) is three-fold lower than that obtained by kinetic analysis (K_D = 18 µM). As was observed with P22 binding to the C36S mutant, the concentration range was insufficient as the isotherm does not reach an inflection point and reduces confidence in the equilibrium analysis derived value. The three fold lower K_D does indicate that there is a difference in binding in this orientation.

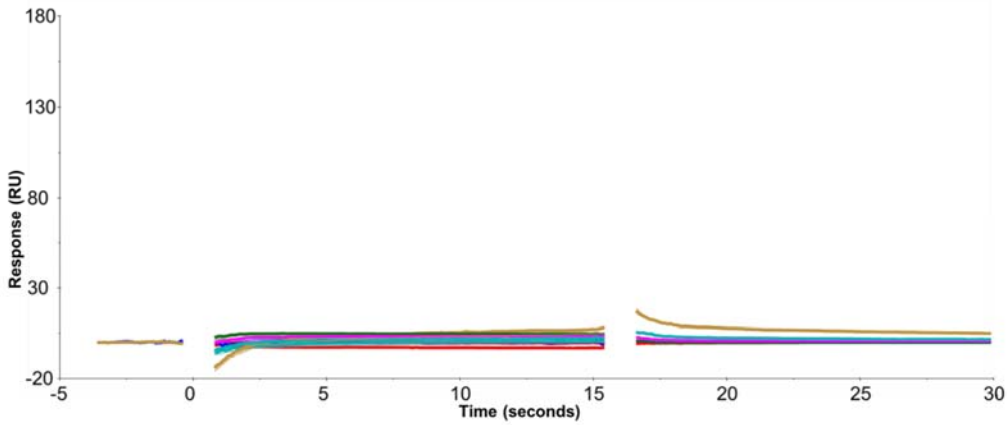


Figure 3-30 Kinetic analysis of CD33 C36S, R119A binding to immobilised P22-biotin. CD33 C36S, R119A was tested at concentrations of 0.03, 0.1, 0.3, 1, 3, 9 µM. Binding measurements were performed in duplicate and data sets are overlayed.

The negative control CD33 C36S R119A mutant did not appear to bind to the P22 (Figure 3-30). Results of the binding experiments are summarised below (Table 7, Table 8).

Table 7 Summary of the equilibrium analysis of P22 binding to immobilised CD33 C36S and CD33 wild-type.

	K_D µM	R_{max} (RU)	Expected R_{max} (RU)	%Functional ligand
CD33 C36S	67.4	183	254	72
CD33 wild-type	89.4	118	195	60

Table 8 Summary of the kinetic evaluation of CD33 C36S and CD33 wild-type binding to immobilised P22.

	k_a (1/Ms)	k_d (1/s)	K_D µM	R_{max} (RU)	Expected R_{max} (RU)	%Functional ligand
CD33 C36S	5.44E+04	3.118	57.4	3341	13389	21
CD33 wild-type	6.07E+04	1.092	18.0	249	4118	6

Binding of the CD33 wild-type and C36S mutant to P22 compound as either ligand or analyte suggests that binding to the sialic acid-binding region is not compromised by the C36S mutation. Immobilised CD33 C36S and CD33 wild-type show similar binding affinities to P22 (Table 7). With P22 immobilised and the CD33 constructs in solution, binding analysis suggests that wild-type CD33 is acting as a bivalent analyte, that is, has two identical binding sites. Two binding sites will give rise to a stabilisation of the ligand-analyte complex without extra response but shifts the equilibrium constant to a more stable interaction. This is apparent with the CD33 wild-type showing an approximately three-fold higher affinity than the CD33 C36S mutant. This effect appears to arise from a decreased rate of dissociation of P22 leading to a decreased dissociation rate, as K_D is calculated by *kd*/*ka* (Table 8), and strengthens the hypothesis that in solution wild-type CD33 IgV domain forms a homodimer via an intermolecular disulfide bond involving the C36 residue. This validated the decision to proceed with the C36S mutation for further SPR studies. The equilibrium analyses performed on the CD33 C36S and CD33 wild-type indicated that the concentration range used in these experiments was inadequate for accurate analysis. Langmuir's isotherm takes the form of a rectangular hyperbola, or when logarithmically transformed (inset, Figure 3-28, Figure 3-29) a sigmoidal curve with an half-saturation value representing, for the present purpose, the K_D (Figure 3-25 and Figure 3-26). Fits of the Langmuir isotherm to the experimental data points for both CD33 constructs do not reach this half-saturation values and hence the apparent K_D is extrapolated from an artificially forced symmetry. A higher concentration range, to at least 0.5 M (10 x the apparent K_D) would have been preferable; however, amounts of protein and P22 compound were limited and this was not possible. The low % Functional ligand value for the CD33 constructs in solution suggests that steric hindrance may inhibit protein binding to immobilised P22 (Table 8). This calculation was performed using Equation 2-2 based on the Mw for the CD33 wild-type monomer; however, If the Mw for the dimer is used the % Functional ligand is reduced two-fold.

Irrespective of the CD33 wild-type in solution data, the average K_D for the P22 compound binding to CD33 C36S in solution was approximately 70 µM. Rillahan et al. suggest that CD33 interacts with P22 with an IC_{50} of 11 µM by flow cytometry. [267] The IC_{50} represents the concentration of an inhibitor that is required for 50% inhibition *in vitro,* and may not faithfully reflect the chemistry of ligand binding, owing to indeterminate factors that affect the avidity of ligand binding in evoking a measurable inhibitory response. [268] Alternatively, the dissociation equilibrium constant, $K_{D,}$ describes the equilibrium between forwards and backwards rates of ligand binding, such that, at the equilibrium concentration the probabilities of ligand binding and unbinding are equal. Typically, the IC_{50} value is used where observation of the effect of ligand binding is inferred to approximate the value of K_D, as in the case of Rillahan et al. who used flow cytometry techniques to establish the IC_{50} of compounds targeted to Siglecs conjugated to liposomal nanoparticles. [267] IC_{50} and K_D values are comparable when used to rank the relative affinities of ligands but are only comparable, in terms of ligand binding affinity, where the capacity of ligand binding to provoke a measurable biological response approaches plurality. [269] In the current instance, it is informative to note that the IC_{50} and K_D values are both of µM magnitude.

3.3.14. Selection of 2952 and 2971 analogues

Based on affinity and specificity selection compounds 2952 and 2971 were identified as the best binders from the 73 compounds screened (section 3.3.9).

A Unity 2D structural analogue and chemical similarity search of our in-house drug-like small molecule database, consisting of ~4 million commercially available compounds, was conducted by Dr. Tracy Nero (our laboratory) using the 2D structures of compounds 2952 and 2971 as the search queries within SYBYL-X 2.1.1 (Certara, L.P., http:// certara. com/http://certara.com). The Tanimoto similarity score is a metric which evaluates how similar two molecules are to each other based on the intersections and unions of the molecular fingerprints. That standard cut-off for biologically similar molecules is 70%. [270] There were 27 available analogues of 2952 (i.e. structurally and/or chemically similar compounds, Tanimoto similarity score >75%), all of which were purchased for screening against CD33. For 2971 there were 248 analogues (Tanimoto similarity score >80%), of which 60 were purchased for screening against CD33. [271] Of the total 87 analogues purchased, only 27 analogues of 2952 and 10 of 2971 were soluble in PBS + 5% DMSO to 500 µM. PBS was chosen as the running buffer for the SPR screen to conform to analogous control experiments that were simultaneously conducted using NMR techniques by Dr. Luke Miles. as shown above (Figure 3-22), a consistent and similar response was obtained for SPR experiments conducted in both PBS and TBS (refer to section 3.3.5).

3.3.14.1. Screen of 2952 and 2971 analogues

The 37 analogue compounds were initially screened at 100 µM against chemically biotinylated proteins, however no binding was observed (not shown). Screening experiments were therefore repeated at 500 µM compound. (Figure 3-31)

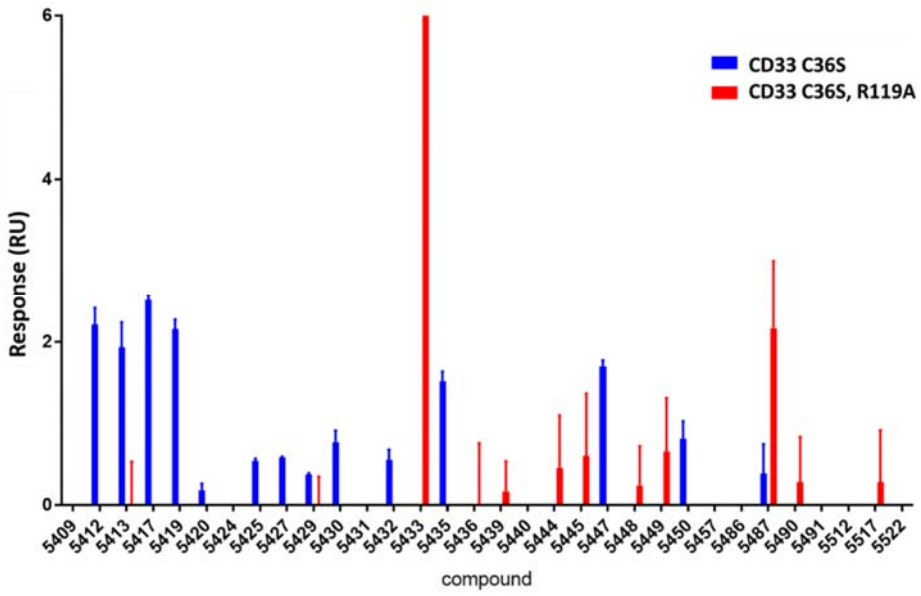


Figure 3-31 Plot of analogue binding response at 500 µM to chemically biotinylated CD33 C36S (blue) and the negative control protein CD33 C36S, R119A (red). Data are means and standard deviation calculated from *n* = 2 experiments.

The compounds were rescreened against the enzymatically biotinylated CD33 C36S and CD33 C36S, R119A at 100 µM (Figure 3-32).

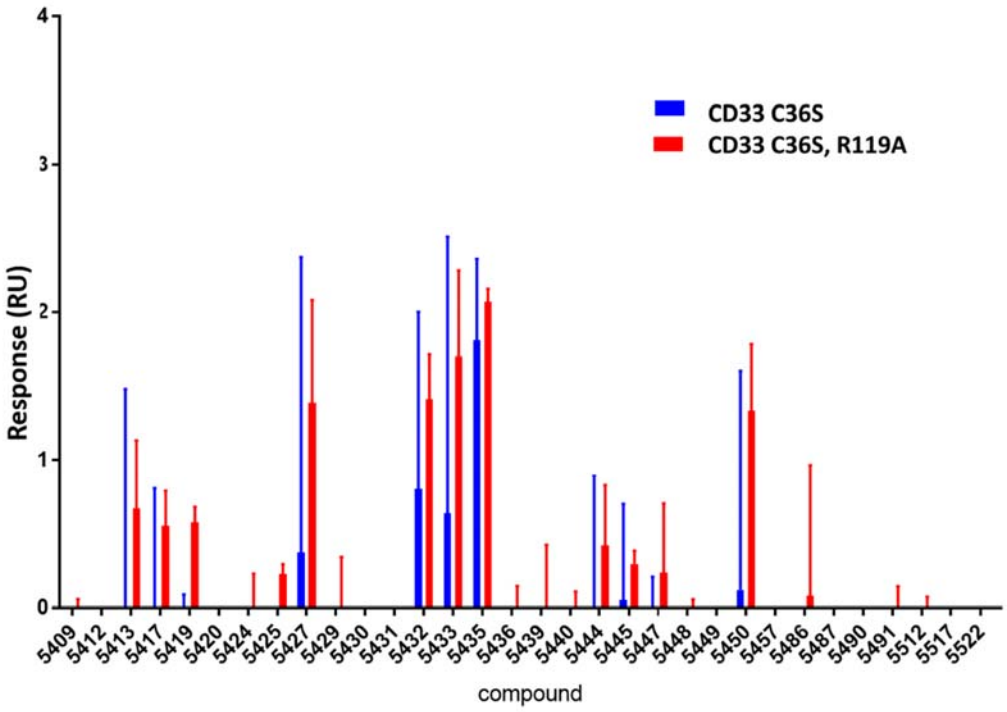


Figure 3-32 Plot of analogue binding response at 100 µM to enzymatically biotinylated AviTag™- CD33 C36S (blue) and the negative control protein CD33 C36S, R119A (red). Data are means and standard deviation calculated from *n* = 2 experiments.

Eight compounds exhibited specific binding above the background level to the chemically biotinylated CD33 C36S protein (Figure 3-31), while no compounds showed specificity for the AviTag™CD33 C36S protein. (Figure 3-32) A number of compounds exhibited greater binding and/or specificity to the CD33 C36S, R119A negative control protein, presumably binding to a different site on the CD33 surface. This finding could be of interest if the compounds allosterically disrupted sialic acid binding, however this was outside the scope of this project and not investigated. The positive binding compounds from the chemically biotinylated protein screen were then compared with their responses from the AviTag™ protein screen (Figure 3-33).

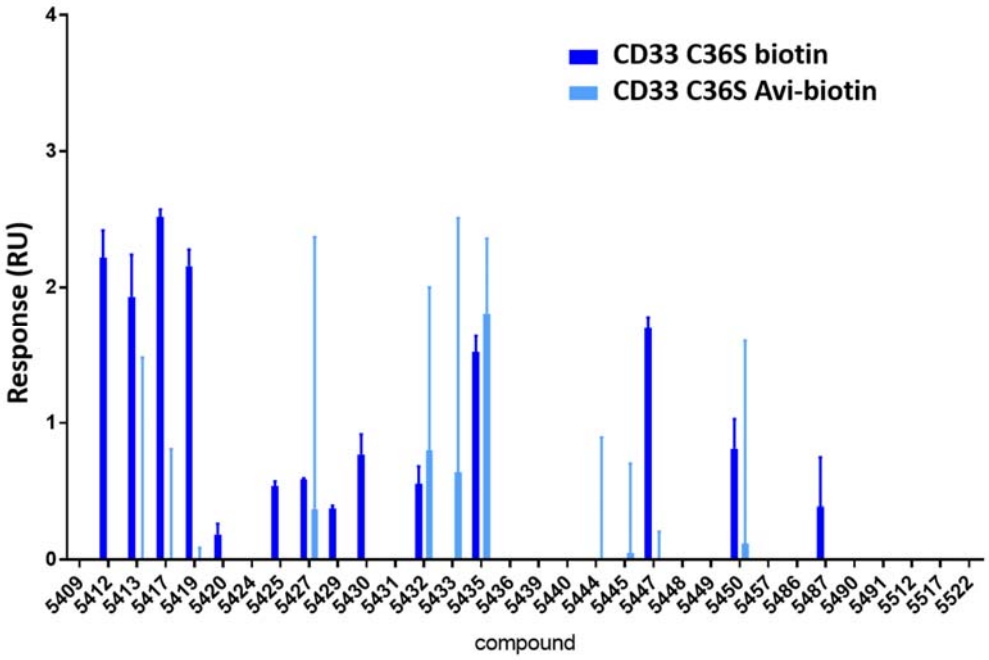


Figure 3-33 Binding response of the compounds to the chemically biotinylated CD33 C36S (dark blue) and the AviTag™- CD33 C36S (light blue). Data are means and standard deviation calculated from *n* = 2 experiments.

Four compounds, 5427, 5432 and 5435 and 5450 bound to both the chemically and enzymatically biotinylated CD33 C36S protein (Figure 3-33), however they also bound to the enzymatically biotinylated R119A mutant negative control protein (Figure 3-32). The data suggests that these compounds do not bind to the targeted R119 pocket and also that the biotinylation method or the different protein constructs may be causing binding anomalies. Binding studies for the chemically biotinylated proteins were carried out at a higher compound concentration (500 µM) than for the enzymatically biotinylated proteins and it is possible that some non-specific binding is occurring.

Chemical biotinylation through the primary amines of the protein (i.e. lysine residues) has the potential to obstruct the compound binding site. The long spacer arm of the biotin linker should allow enough flexibility to overcome this, however, as binding was only observed at a higher concentration this is still potentially a problem.

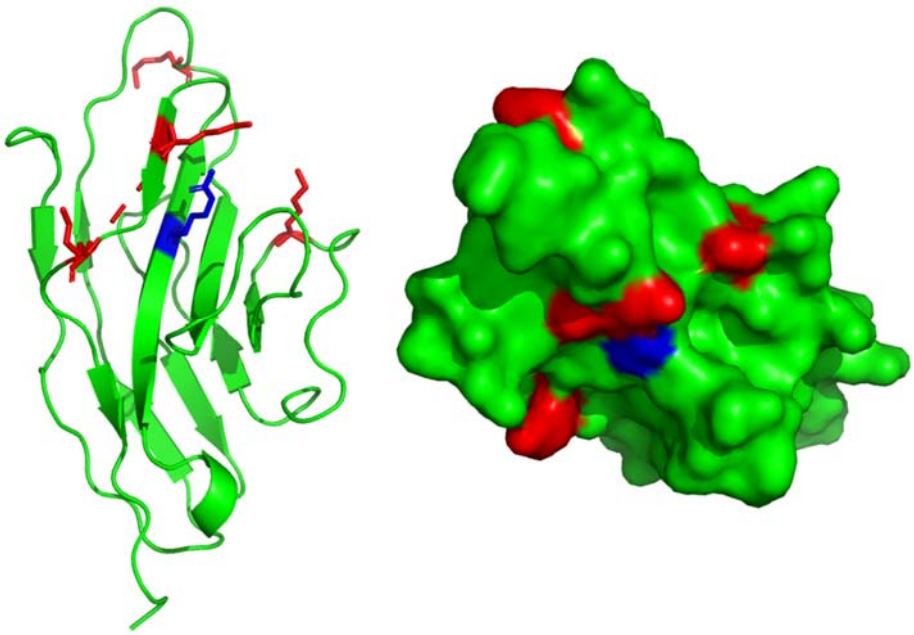


Figure 3-34 Structure of the IgV domain of CD33. The left image shows a cartoon representation of the CD33 IgV domain with the sialic acid-binding residue R119 (blue) and all of the lysine residues (red) displayed as sticks. The image on the right shows the molecular surface of the CD33 IgV domain in a view rotated 90^0 about the Y-axis from that shown in the cartoon image, revealing that R119 is closely flanked by three lysine residues.

The structural model of the CD33 IgV domain (Figure 3-34) reveals that the targeted R119 residue is surrounded by primary amine containing lysine residues, all of which are potential binding sites for the NHS esters of the Sulfo-NHS-LC-LC-Biotin reagent. It is possible, even with the length of the LC-LC linker region on the biotin label, that the sialic acid-binding pocket and/or the crevice near R119 could be obscured by chemical biotinylation.

3.4. *SUMMARY AND FUTURE DIRECTIONS*

Optimisation of the assay conditions and construct design has enabled the development of a reliable SPR assay, which will be invaluable for future small molecule screening campaigns to identify compounds with high affinity for the CD33 sialic acid-binding site. It was found that the best approach for recombinant bacterial expression of CD33 IgV was to mutate the C36 residue to S to avoid dimerization of the protein in solution, and inclusion of an Avi tag in the recombinant construct to facilitate capture on a streptavidin surface in the biosensor. Running the biosensor assays at a higher flow rate

of > 60 µl/min in PBS + 0.05% Tween 20 improved the signal quality. The assay has been validated by the positive control compound P22, which binds with µM affinity to R119 in the CD33 sialic acid-binding site and fails, as predicted, to bind to the R119A mutant (Figure 3-29, Figure 3-30).

Several compounds that bind to the sialic acid-binding site were identified. Analogues of two of these compounds were found to bind independently of R119 and specifically to the R119A mutant. Mutating the arginine to an alanine would increase the size of the crevice in the IgV domain of CD33, as arginine has a guanidino side chain and is far larger than the methyl sidechain of alanine (Figure 3-35). This may be evidence that the compounds are binding to other residues in the crevice targeted by the virtual screen or to another site entirely, (refer to Figure 3-12) rather than the sialic acid-binding site residue R119.

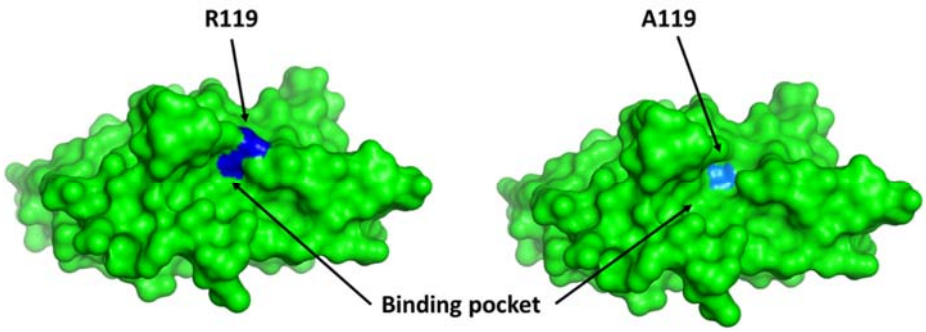


Figure 3-35 Comparison of the wild-type CD33 IgV domain and an *in-silico* R119A mutant. The molecular surface of wild-type CD33 (PDB ID: 6D48, left) [125] showing the arginine residue at 119 (dark blue) and a potential binding pocket/crevice beneath the R119. A model of the R119A CD33 mutant was constructed (right), with the introduced alanine residue coloured light blue. The R119A mutation creates a larger crevice for small molecules to bind.

The 2952 and 2971 analogues only bound to the chemically biotinylated CD33 IgV domain C36S protein at a concentration of 500 µM, indicating that despite the long spacer arm, the biotinylation process may be obstructing the compound binding site on CD33. Repeating the screen at a lower concentration (i.e. 100 µM) against the AviTag™-CD33 protein showed greater binding of the 2952 and 2971 analogues to the negative control protein (C36S, R119A double mutant) than to the target CD33 C36S protein, suggesting that the analogues are binding to a site on the CD33 IgV domain which is

independent of R119. The data indicates that the structural changes introduced into the parent 2952 and 2971 compounds during the analogue search process have not improved affinity to the targeted sialic acid-binding site and that the compounds are not binding directly to R119. As suggested above, it is possible that the binding site of the analogue compounds may even be optimised by the R119A mutation. Alternatively, these compounds may be in fact binding to a completely different site on the IgV domain of CD33.

Mutating the free C36 residue in the CD33 IgV domain improved protein stability and purification yield. However, introducing a new immobilisation technique at the same time as introducing the C36S mutation compounded the complexity of validating the SPR assay. Introducing an AviTag™ to the protein enabled a higher immobilisation level and a more stable surface than was previously possible with the His tag capture/couple technique. More rigorous assay validation would require the original compound screen to be repeated against AviTag™ wild-type CD33 and a capture/coupled His-tagged C36S mutant. Unfortunately, the positive control P22 compound only became available later during my PhD project but binding studies clearly demonstrated the C36S mutation did not affect the sialic acid-binding site. The initial use of 3'- and 6'-sialyllactose as positive controls was difficult as their low affinity (approximately 8 mM [256]) was close to the detection limit of the Biacore T200 instrument and most likely produced false-negative results. An obvious choice for a positive control ligand would be the GT1b ganglioside analogue that was used to identify the binding pocket targeted in the virtual screening (see paragraph 3.3.6). The crystallographic structure of the GT1b analogue was taken from the hSiglec-7 complex published by Attrill et al (PDB ID:2HRL). The GT1b analogue was custom made, in-house, and is not available commercially so could not be obtained for this project. [262]

When screening for novel inhibitors to a putative binding site, a positive control for assay validation is often unavailable. Secondary biophysical techniques to validate the data are crucial in these situations. Positive binding compounds identified from the SPR studies described in this chapter are now being investigated using crystallographic studies, cell-based assays and NMR-based assays for further validation. Findings arising from the experiments shown in this chapter have formed the basis for an ongoing industry collaboration to identify and characterise novel small molecule inhibitors of CD33 function.

CHAPTER 4. BIOINFORMATICS ANALYSIS OF CD151

4.1. *INTRODUCTION*

Computational analyses of protein sequences can provide insights into the structure and function of uncharacterised or poorly characterised proteins and provide strategies for structural and functional studies. Post-translational modifications (PTMs) such as glycosylation and phosphorylation can affect protein folding, stability and biological activity. Other PTMs such as the cleavage of peptide bonds or formation of disulfide bonds are important for protein tertiary and quaternary structure. Many databases have been created to collect information regarding specific motifs, consensus patterns and sites within protein sequences that can be used to predict PTMs. [272] Predicting PTMs becomes salient when choosing an expression system for heterologous protein as prokaryote and eukaryote systems have varying abilities to facilitate requirements. For example, *E. coli* is a widely used expression system which can be grown easily and is cost effective, yet cannot process common PTMs such as glycosylation and phosphorylation which may impact protein structure and activity. [273]

As described previously (section 1.5.1), tetraspanins are found throughout the animal kingdom as well as in lower eukaryotes, fungi, amoebas and plants. There are 33 human tetraspanins and the lack of prokaryotic homologues suggests that tetraspanins and eukaryotes co-evolved, suggesting a role for tetraspanins in the evolution of intercellular interactions. [274] The tetraspanins range from 200 - 300 amino acids in length and are characterised by four transmembrane (TM) domains, a highly conserved small extracellular loop (SEL) and a large extracellular loop (LEL), which is composed of a constant region and a variable region. The LEL constant region is formed by three helices and the variable region contains two to four disulfide bonds. [275] This rigid and compartmentalised LEL structure prevents domain shuffling and duplication, so the tetraspanin's structural evolution is limited to amino acid substitutions, insertions and deletions.[276]

Using a bioinformatics approach to compare structures and protein sequences of members of the tetraspanin family, including CD151, I was able to gain insights into salient features of the CD151 structure. This data could be used to inform protein expression and purification experiments in order to further characterise CD151 using crystallographic and other biophysical techniques.

4.2. *RESULTS AND DISCUSSION*

4.2.1. Blast search

The Basic Local Alignment Search Tool (BLAST) [277] can be used to predict evolutionary relationships, identify gene families and predict the function of novel proteins. A search was performed to identify 3D structures of proteins related to CD151 (UniProt ID: P48509) which may be suitable to use as templates for modelling the CD151 structure. Searching the NCBI Protein Data Bank protein (PDB) database using Blastp (protein-protein BLAST) and the full length 253 amino acid sequence of CD151 as the search query returned a single result, Chain A of human tetraspanin CD81 (PDB ID: 5TCX_A). [278]

Altering the BLAST parameters and repeating the search for human proteins related to CD151 returned 65 proteins with an E-value of less than 1×10^{-4}, indicating that they are similar enough to be considered closely related proteins: all 65 proteins were identified as tetraspanins. Disregarding duplicate proteins arising from multiple isoforms reduced the sample of closely related human proteins to 26 tetraspanins from the overall family of 33. The sequence identity amongst the 26 selected tetraspanin proteins ranged between 21% and 56%.

4.2.2. Multiple sequence alignment

Aligning the 26 human tetraspanins identified by the BLAST search, and colour coding into regions of highest similarity, reveals that the LEL has very low sequence homology between the tetraspanins (Figure 4-1). The sequence alignment shows that, apart from the tetraspanin signature (i.e. the CCG motif) and the conserved third cysteine residue, there are large segments of sequence variability in the LEL across the human tetraspanins. The integrin binding site on the CD151 LEL, i.e. the QRD motif, is unique to CD151 (refer to 1.5.2). The Align function utilised within UniProt (https://www.uniprot.org) uses the Clustal Omega program to carry out multiple sequence alignments to a reference sequence. [279] Using this alignment algorithm, the fourth conserved cysteine fails to align across the 26 selected sequences (Figure 4-1); however, aligning the sequences of all 33 human tetraspanins using CLC sequence viewer successfully aligns all four conserved cysteine residues (Figure 4-2).

There are 9 regions of CD151 listed in the UniProt record (P48509) as either a topological domain or TM domain. The topological domains are described as either cytoplasmic or extracellular. When reviewing tetraspanin sequences extracted from the UniProt database it became apparent that shortcomings in domain prediction algorithms led to inaccurate and misleading representations for several of the identified CD151-like tetraspanins. [280] In some cases not all of the nine domains were assigned and some assignations seemed unlikely, as discussed in detail below (Table 9).

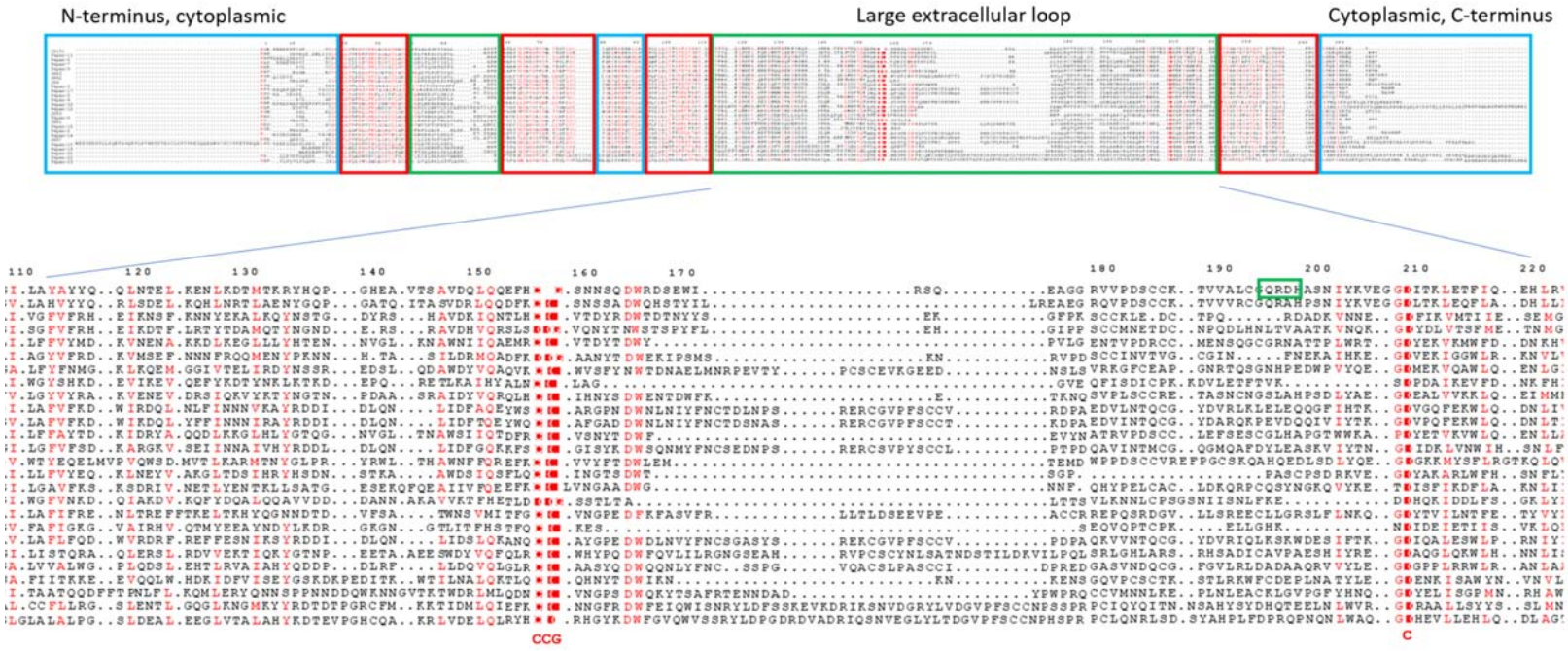


Figure 4-1 Similarity map of the multiple sequence alignment of 26 human tetraspanins.

The alignment of 25 human tetraspanins with CD151 (top sequence) performed using Clustal Omega shows regions of similarity (red text), with boxes indicating the approximate location of the cytoplasmic (light blue), transmembrane (red) and extracellular (green) regions. Extracting the segment of sequences that roughly corresponds to the large extracellular loop (LEL) reveals the strictly conserved CCG motif and a third cysteine residue (red).

Figure 4-2 Multiple sequence alignment of the LEL of the 33 known human tetraspanins.

This alignment was constructed using CLC Sequence viewer Version 8.0. Borders of the LEL are clipped at the start of the transmembrane regions. Cysteine residues are coloured yellow.

Table 9 Topology of the 33 human tetraspanins extracted from the UniProt database. Cyt = cytoplasmic, TM = transmembrane, SEL = small extracellular loop, LEL = large extracellular loop. Numbers refer to amino acid positions. Tetraspanin topologies highlighted in yellow have been identified as containing anomalies or omissions.

	Cyt	TM1	SEL	TM2	Cyt	TM3	LEL	TM4	Cyt
CD9	2 - 12	13 - 33	34 - 55	56 - 76	77 - 87	88 - 111	112 - 195	196 - 221	222 - 228
TSN2	1 - 13	14 - 34	35 - 54	55 - 75	76 - 90	91 - 111	112 - 188	189 - 209	210 - 221
CD81	1 - 12	13 - 33	34 - 63	64 - 84	85 - 89	90 - 112	113 - 201	202 - 224	225 - 236
TSN8	1 - 9	10 - 33	34 - 57	58 - 72	73 - 83	84 - 109	110 - 205	206 - 230	231 - 237
CD151	1 - 18	19 - 39	40 - 57	58 - 78	79 - 91	92 - 112	113 - 221	222 - 242	243 - 253
TSN11		19 - 39		63 - 83		93 - 113		220 - 240	
TSN18	1 - 13	14 - 34	35 - 49	50 - 70	71 - 83	84 - 104	105 - 223	224 - 244	245 - 248
TSN1	1 - 11	12 - 32	33 -52	53 - 73	74 - 88	89 - 109	110 - 211	212 - 232	233 - 241
TSN9	1 - 13	14 - 34	35 - 55	56 - 76	77 - 85	86 - 106	107 - 203	204 - 224	225 - 239
TSN4	1 - 13	14 - 34	35 - 55	56 - 76	77 - 85	86 - 106	107 - 201	202 - 222	223 - 238
CD53	1 - 11	12 - 32	33 - 54	55 - 69	70 - 80	81 - 106	107 - 181	182 - 206	207 - 219
TSN6	1 - 19	20 - 40	41 - 59	60 - 80	81 - 93	94 - 114	115 - 208	209 - 229	230 - 245
TSN7	1 - 16	17 - 40	41 - 56	57 - 75	76 - 86	87 - 112	113 - 213	214 - 234	235 - 249
TSN3	1 - 11	12 - 32	33 - 50	51 - 71	72 - 85	86 - 106	107 - 212	213 - 233	234 - 253
CD63	2 - 11	12 - 32	33 - 51	52 - 72	73 - 81	82 - 102	103 - 203	204 - 224	225 - 238
TSN12	1 -12	13 - 33	34 - 59	60 - 80	81 - 89	90 - 110	111 - 224	225 - 245	246 - 305
TSN17	1 - 19	20 - 40	41 - 63	64 - 84	85 - 94	94 - 115	116 - 234	235 - 255	256 - 270
TSN5	1 - 17	18 - 38	39 - 61	62 - 82	83 - 92	93 - 113	114 - 232	233 - 253	254 - 268
TSN14	1 - 17	18 - 38	39 - 61	62 - 82	83 - 92	93 - 113	114 - 232	233 - 253	254 - 270
TSN33	1 - 24	25 - 45	46 - 64	65 - 85	86 - 96	97 - 117	118 - 235	236 - 256	257 - 283
TSN15	1 - 23	24 - 44	45 - 62	63 - 83	84 - 93	94 - 114	115 - 235	236 - 256	257 - 294
TSN10	1 - 78	79 - 99	100 - 120	121 - 141	142 - 154	155 - 175	176 - 355		
CD37	1 - 17	18 - 38	39 - 59	60 - 74	75 - 85	86 - 111	112 - 241	242 - 266	267 - 281
CD82	1 - 11	12 - 32	33 - 53	54 - 72	73 - 83	84 - 110	111 - 228	229 - 250	251 - 267
UPK1B	1 - 15	16 - 36	37 - 60	61 - 81	82 - 86	87 - 107	108 - 229	230 - 250	251 - 260
UPK1A	1 - 14	15 - 35	36 - 59	60 - 86	87 - 91	92 - 112	113 - 230	231 - 252	253 - 258
TSN19		16 - 36		59 - 79		87 - 107			
TSN16	1 - 13	14 - 34	35 - 37	38 - 58	59	60 - 80	81 - 94	95 - 115	116 - 245
ROM1	1 - 19	20 - 44	45 - 64	65 - 84	85 - 102	103 - 125	126 - 263	264 - 286	287 - 351
PRPH2	1 - 24	25 - 43	44 - 61	62 - 80	81 - 99	100 - 123	124 - 264	265 - 290	291 - 346
TSN13	1 - 19	20 - 40	41 - 44	45 - 65	66 - 72	73 - 93	94 - 167	168 - 188	189 - 204
TSN31	1 - 12	13 - 33	34 - 44	45 - 65	66 - 72	73 - 93	94 - 173	174 - 194	195 - 210
TSN32		14 - 34		60 - 80		90 - 110		203 - 223	

The UniProt database annotates TM domains when they have either been determined experimentally and there is evidence for the location, or the structure of the protein and its family have been reviewed and annotations are assigned by sequence similarity or models. TM domains are also predicted using software from external providers. [281]

TSN11 (UniProtKB ID: A1L157, TSN11_HUMAN) and TSN32 (UniProt ID: Q96QS1, TSN32_HUMAN) do not have the topological domains identified, but the TM domains are numbered so it is a simple matter to extrapolate the missing information. The TSN19, (UniProtKB ID: P0C672, TSN19_HUMAN) entry lists only three TM regions (Table 9). By comparing tetraspanin 19 from the European domestic ferret (UniProtKB ID: M3Y8M8, M3Y8M8_MUSPF) with the human TSN19 sequence reveals 70% homology. M3Y8M8 (European domestic ferret) has four TM domains assigned. TM 1 – 3 for both tetraspanins have similar alignments, so it would be reasonable to assume that the

assignment of M3Y8M8 TM4 (residues 228 – 253) would be applicable to the human TSN19 tetraspanin.

The topological assignments for TSN16 seem unlikely as the SEL has only three amino acids, the cytoplasmic loop has one amino acid and the LEL has 13 amino acids, which places the CCG signature motif and the conserved four cysteines on the cytoplasmic C-terminal tail. The molecular characteristics of TSN16, also known as TM4-B, were published in 1999 by Puls et al. Using multiple alignments with 20 other tetraspanins, they predicted the hydrophobic TM regions to be 14 – 39, 59 – 80, 85 – 111 and 218 – 241. [282] This arrangement makes more sense spatially, in terms of the size of the extracellular regions and the predicted LEL (residues 112 – 217) includes the CCG motif and conserved cysteines.

TSN10 (UniProtKB ID: Q9H1Z9, TSN10_HUMAN) has three TM regions assigned, leaving a lengthy 179 amino acid extracellular C-terminal region which includes the CCG motif and conserved cysteines. TSN10, also known as oculospanin, is expressed in the eye and was first identified by Wistow et al. in 2002 during their preparation of a cDNA library from human post-mortem eye tissues. [283] Wistow et al. predicted the TM regions of TSN10, including the fourth TM segment missing from the UniProt database annotations, as spanning amino acids 80 – 102, 122 – 144, 156 – 178 and 291 – 313.

UniProt is a constantly expanding database that is an invaluable resource for research; however, it is important to note that the automated systems used to curate protein families are fallible and, at best, provide a loose framework for further investigation. The combined sequence and hydrophobicity alignment of all 33 known human tetraspanins illustrates the conserved nature of this family of proteins (Figure 4-3).

The only human tetraspanin whose 3D atomic structure has been solved is CD81 (section 1.5.2, Figure 1-8, Figure 1-10). Comparing the sequence alignment and relative hydrophobicity of CD151 and CD81 in light of the 3D structure, gives a clearer view of the conserved nature of the TM domains and the variability of the LEL (Figure 4-4).

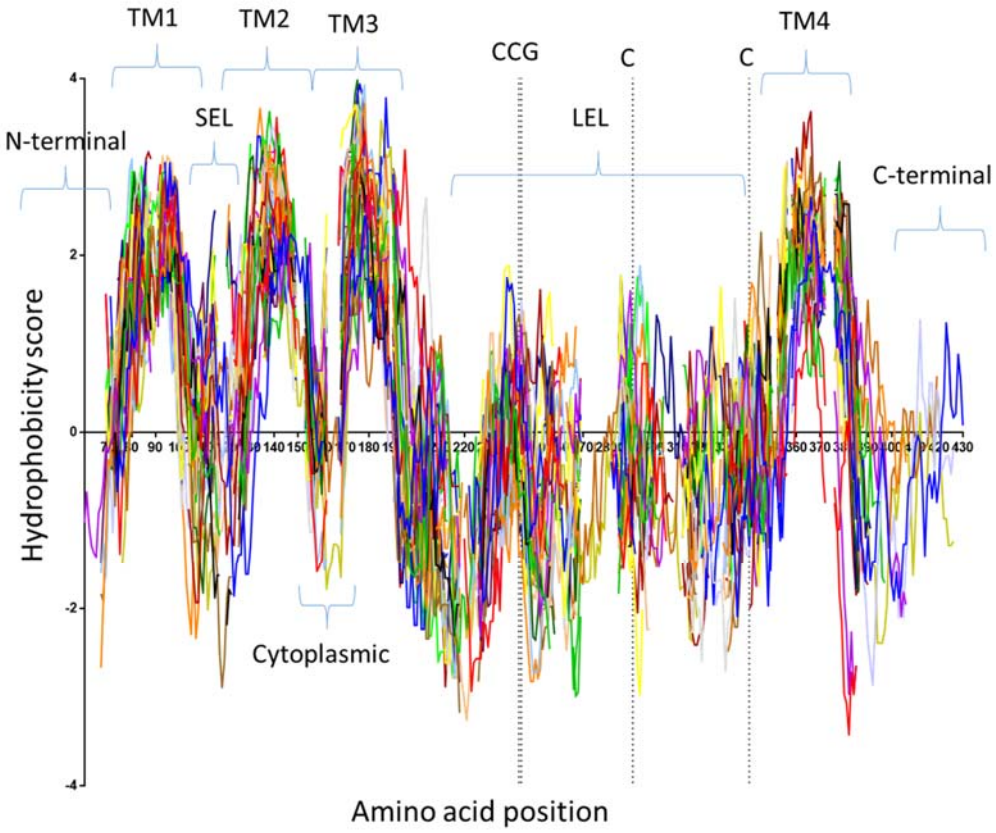


Figure 4-3 Alignment of the 33 human tetraspanins by sequence and hydrophobicity. The conserved CCG motif and conserved cysteines are indicated by vertical dotted lines. Hydrophobicity was determined using the Kyte & Doolittle scale [284] and the sequence alignment was constructed using CLC Sequence Viewer Version 8. TM = transmembrane, SEL = small extracellular loop, LEL = large extracellular loop.

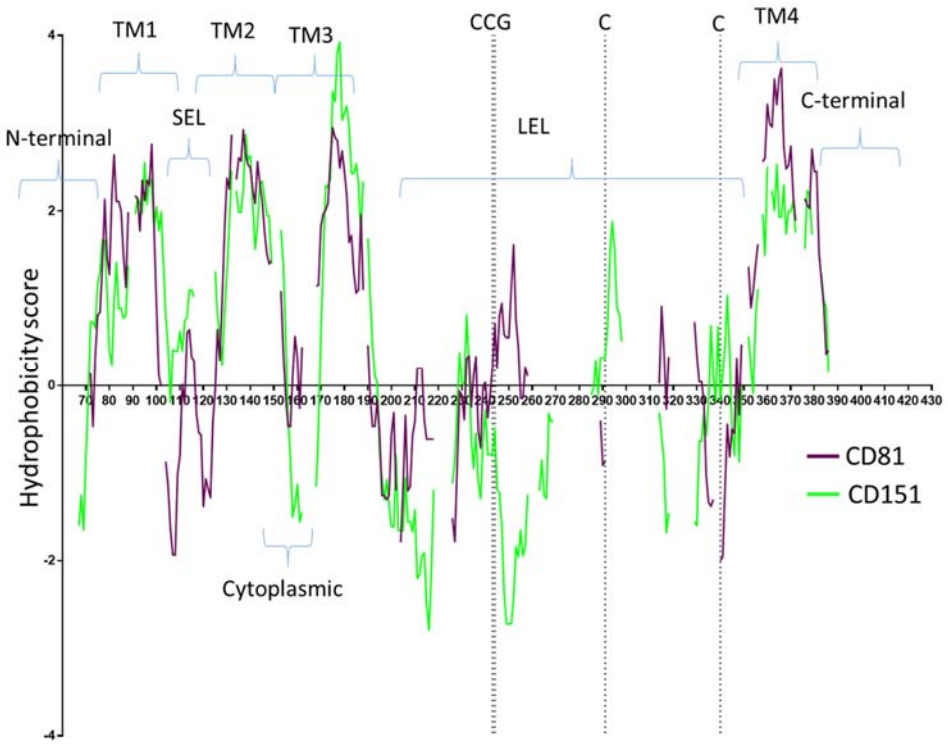


Figure 4-4 Alignment of CD81 (purple) and CD151 (green) by sequence and hydrophobicity. The conserved CCG motif and conserved cysteines are indicated by vertical dotted lines. Hydrophobicity was determined using the Kyte & Doolittle scale [203] and sequence alignment constructed using CLC Sequence Viewer Version 8. TM = transmembrane, SEL = small extracellular loop, LEL = large extracellular loop.

4.2.3. Predicted post translational modifications in the LEL

4.2.3.1. Glycosylation sites

N-linked glycosylation is a common PTM of the extracellular regions of human transmembrane proteins and is known to play a role in the regulation of protein stability and folding in the ER. [285] N-linked glycosylation occurs when a glycan attaches to the amide nitrogen of an asparagine. This can be predicted to occur by the presence of the protein sequence Asn-Xaa-Ser/Thr, where Xaa is any amino acid. A single CD151 N-linked glycosylation site is predicted by sequence analysis at N159 in the LEL. [286] It has been demonstrated that glycosylation of CD151 modulates the glycosylation of integrin α3β1, which is crucial for the CD151-α3β1 dependent migratory function. [287] This could be an important factor when designing inhibitors of the CD151-α3β1 interaction. However, the binding interaction between CD151 and integrin α3β1 itself is not inhibited by the mutation of N159 in CD151, although the migratory activity of the complex is. [287] CD151 produced in a bacterial expression system would lack glycosylation of N159 but

would still be useful for identifying inhibitors of CD151 function since the CD151 LEL still binds integrin α3β1 in the absence of glycosylation. This is an important point as variable glycosylation increases protein heterogeneity and surface entropy; therefore, being able to produce non-glycosylated protein in bacteria is advantageous for crystallographic studies that require homogenous protein samples. [288]

4.2.3.2. Disulfide bonds and oligomeric status

The CD151 LEL contains six cysteine residues. Four of these are conserved across all tetraspanins and are known to form disulfide bonds. [289] Five tetraspanins contain only the four conserved cysteines, eight contain eight cysteines, two contain seven and the remainder contain six. The two tetraspanins containing seven cysteines are the retinal proteins PRPH2 and ROM1, which form covalently coupled heterodimers with each other via the seventh cysteine. [290] To date there is no evidence that the additional two cysteine residues in CD151 form either an inter- or intramolecular disulfide bond. There are some reports that tetraspanins, including CD151, can form homodimers *in vivo.* [174] There is evidence, however, that this is through the membrane associated cysteine residues and not through the extracellular loop. [291-293]

The originally published crystal structure of the CD81 LEL (in 2004) identified a hydrophobic region in the “head domain” and stated that the LEL forms a dimer. [294] The more recent (2016) full length structure of CD81 revealed a monomeric form and the authors stated that the previously observed dimer was likely a non-native crystallisation effect due to the absence of the TM region. [278] Studies of CD151 produced recombinantly are few and biochemical analysis of the purified protein is limited. Generally, recombinantly expressed CD151 is reported as "correctly folded" as confirmed by Western blot analysis with reactivity to a conformation recognising Ab, but apparent Mw or oligomeric status is not reported. [295, 296] Other studies report recombinant CD151 having large Mw’s due to oligomerisation, but these studies fail to offer any analysis or insight in to the molecular mechanism by which oligomers are formed. [297, 298]

My studies indicate that all six cysteine residues in the LEL of CD151 form intramolecular disulfide bonds and that the CD151 protein, when correctly folded, is monomeric. Evidence supporting this notion is detailed in section 5.3; and moreover, my experimental findings suggest that oligomers of the CD151 LEL arise from promiscuous disulfide bonding between incorrectly folded CD151 monomers.

4.2.3.3. Palmitoylation

Protein lipidation enhances hydrophobicity and acts to anchor proteins in the cell membrane, which is important for localisation, partitioning into domains and the physical interaction of proteins within the cell. [299, 300] The most common

mechanisms of lipidation are via myristoylation, farnesylation (also known as prenylation), and palmitoylation.

Myristoylation occurs in cytosolic proteins that begin with the sequence 1-MG-2, the initiating methionine is cleaved co-translationally and the fatty acid myristate is linked to G2 via an amide bond. Farnesylation is the process by which a cytoplasmic cysteine residue in a CaaX motif at the C-terminal end of the protein is post-translationally modified by the thioester linkage of an isoprenoid lipid, which leads to the proteolytic removal of the three terminal amino acids and methylation of the carboxyl group of the farnesylated cysteine. [301, 302] Both of these modifications are irreversible.

Palmitoylation is the post-translational addition of the fatty acid palmitic acid to integral and peripheral membrane cysteine residues, and less frequently, threonine and serine. The majority of palmitoylation occurs in the form of the reversible S-palmitoylation via a thioester linkage to a membrane integral cysteine. N-Palmitoylation occurs when the cysteine residue is located on the N-terminus of the protein, the palmitate temporarily binds to the cysteine as for S-palmitoylation but this is then rearranged to an amide bond. [303] O-Palmitoylation is the less common linkage of palmitate to the hydroxyl group of serine or threonine residues. [304] The dynamic nature of S-palmitoylation plays an important role in protein localisation, accumulation, secretion, stability and function by altering conformation, ligand binding and membrane affinity. [300]

S-Palmitoylation generally falls into four broad groups: transmembrane proteins palmitoylated on cytoplasmic cysteine residues located near the TM domain, proteins that are first myristoylated at an N-terminal glycine residue, proteins that are first farnesylated at a C-terminal CaaX box and peripheral membrane proteins that are modified with palmitate only. [299, 305, 306] Tetraspanins can potentially undergo all of these palmitoylation events, as all have cysteines at the cytoplasmic side of the TM domains, six of the 33 human tetraspanins have an N-terminal glycine G2, and one has the CaaX motif. CD151 has a G2 residue that could potentially be myristoylated and six cytoplasmic cysteine residues that are all potential palmitoylation sites.

Using [^{3}H]palmitate labelling, Yang et al. (2002) demonstrated that CD151 cytoplasmic cysteine residues C11, C15, C241, C242 are palmitoylated. [298] Mutation of the four cytoplasmic cysteines to serine, i.e. removing palmitoylation, reduced the association of CD151 with CD9, but did not affect binding to the α3 subunit of integrin. This was an important detail to note when considering the expression of the extracellular component of CD151 in isolation. Palmitoylation has been shown to be involved with receptor function and, in the case of GPCRs, binding of agonists/antagonists. [307] The ability of CD151 to bind α3 integrin without the palmitoylation PTMs confirms this is not the case for CD151 and that the LEL alone is potentially still functionally active. Yang et al. found that mutating the four cytoplasmic cysteines reduced [^{3}H]palmitate labelling

by more than 90%. The residual signal was proposed to come from the additional two cytoplasmic cysteine residues at C79 and C80; however, mutation of these cysteines resulted in total loss of CD151 expression. This intriguing result was not further commented on in the publication, so it is unclear if (1) the loss of CD151 expression is due to the cysteines forming a critical structural element of the full length protein, (2) they were part of a stop-transfer signal sequence or (3) the result was due to experimental error. The authors also noted that CD151 formed dimers and large multimers, raising the possibility that at least some of their protein was incorrectly folded.

4.2.4. Secondary and tertiary structure prediction and homology modelling

The only published experimental structures of any tetraspanins are the LEL and full length hCD81 [308, 309] and the LEL of the tetraspanin TSP-2 from the human blood fluke *Schistosoma mansoni.* [310] Both of these tetraspanins contain only four conserved cysteines. Alignment of the hCD151, hCD81 and sm-TSP-2 LELs indicates that these three tetraspanins share only 5% sequence identity in this region, whereas random chance typically leads to a similarity of 6% (Figure 4-5). [311]

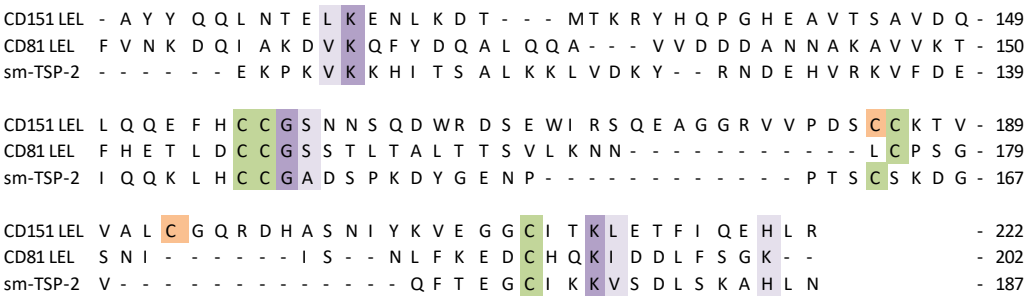


Figure 4-5 Sequence alignment of the LEL of hCD151, hCD81 and sm-TSP-2. Regions of sequence identity are highlighted in purple with conservative substitutions in light purple, the conserved cysteine residues in green, and the additional cysteine residues of the CD151 LEL are highlighted in orange. Sequence alignment performed using T-Coffee (section 2.3.12).

The Protein Homology/analogY Recognition Engine V 2.0 (Phyre^2) was used to construct homology models for the LEL of hCD151 (refer to section 2.3.14). [312] Two similar models were obtained, one used the LEL of hCD81 as the structural template (PDB ID: 5TCX)[308] and the second model used the EC2 domain of the *Shistosoma mansoni* tetraspanin sm-TSP-2 (PDB ID: 2M7Z). [310] The hCD151 LEL is 20 amino acids longer than the LEL in hCD81 and the two sequences share 12.7% sequence identity. (Figure

4-5) The more recently published structure of the sm-TSP-2 EC2 domain has 19.8% sequence identity with the CD151 LEL, in spite of the CD151 LEL sequence containing 29 more amino acids. (Figure 4-5) However, both CD151 LEL models contained only two of the three experimentally determined disulfide bonds (C155-C208 and C156-C185), the third disulfide bond (C184 – C192) was not modelled (Figure 4-6, left). Two antiparallel β-strands were predicted in the regions W169 – R178 and T187 – A198 in both CD151 LEL models, even though neither template protein structure contained any β-strands. Empirical analysis of purified protein using circular dichroism based techniques also does not support that CD151 LEL contains any β-sheet (Figure 5-18). Based on homology modelling predictions the two antiparallel β-strands located in the region T187-A198, placed C184 ~25 Å from C192 (Figure 4-6, left) rendering the formation of a disulfide bond between these two residues as physically impossible. Thus, the conformation of the hCD151 LEL model constructed using the sm-TSP-2 EC2 domain as the template (chosen as it has a higher sequence identity with CD151 compared to CD81) was manually modified in the region K186 – G207 to enable formation of the, experimentally verified, disulfide bond between C184 – C192.

Modelling the C184 – C192 disulfide was performed using SYBYL-X 2.1.1 (refer to section 2.3.14). The loop replacement tool within the Biopolymer module of SYBYL-X 2.1.1. was used to alter the conformation of CD151 LEL residues K186 – G207, to bring C184 and C192 within 5 Å of each other. A covalent bond was then defined between these two cysteine residues and the model subjected to optimisation to determine a global energy minima permissive of the forced covalent interaction between C184-C192. The AMBER 7 FF02 molecular mechanics force field and AMBER partial atomic charges, along with the conjugate gradient convergence method, were used for the geometry optimisation step; termination of the optimisation was achieved when the gradient difference of successive steps was <0.05 kCal/mol.Å or 2000 iterations was reached. [313] All other geometry optimisation parameters were left at default values. There is no experimental data to indicate the conformation(s) adopted by the loop regions in between the three disulfide bonds (i.e. residues 157 – 183, 186 – 191 and 193 – 207); hence these three loop regions were modelled in random low energy conformations. The three disulfide CD151 LEL model was deemed to be a good quality model by Procheck, with 91.8% of all residues in allowed regions of the Ramachandran plot. [314] The final hCD151 LEL model is shown in Figure 4-6 (right). Further, subsequent analysis of purified hCD151 LEL by CD predicted a high α-helix content and failed to support the presence of β-sheet secondary structure, which is consistent with the modified homology model shown in the right panel of Figure 4-6 (Figure 5-18).

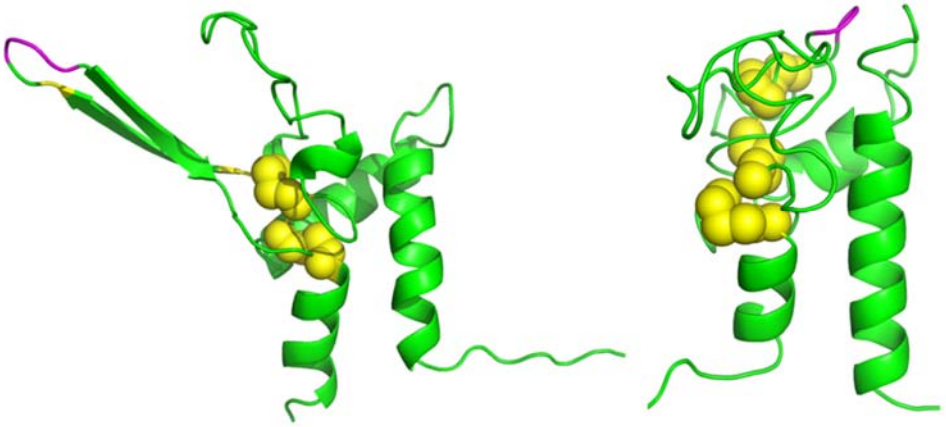


Figure 4-6 Human CD151 LEL homology models. The original hCD151 LEL model was constructed using the LEL of the related human tetraspanin, CD81, as the template. Only two of the three experimentally defined disulfide bonds were modelled, whereas the third cysteine (yellow cartoon) was located approx. 25 Å from possible disulfide linkages. (left) The final hCD151 LEL model, constructed using the LEL of sm-TSP-2 as the template and manual modification to the conformation adopted by residues K186 – G207, permits all three disulfide bonds. Disulfide linked cysteine residues are shown as yellow spheres and the critical QRD motif (i.e. the integrin binding site) is shown in pink.

4.3. *CONCLUSION*

The bioinformatical characterisation of CD151 provided insights into relevant features of the human CD151 protein. Sequence alignments and modelling of the LEL suggested appropriate domain boundaries for protein expression. PTMs in the CD151 LEL were found to be limited to a single N-linked glycosylation site and two conserved disulfide bonds. It has been established that the glycosylation in the CD151 LEL it is not required for protein inhibitor studies, thus raising the potential for protein expression using *E. coli.* This expression method is faster, has higher yields and is much more cost effective than mammalian or insect cell expression. Homology modelling of the CD151 LEL was refined to include a third disulfide bond which resulted in an overall conformation that favourably orientates the QRD integrin binding site in an accessible location. The presence of the disulfide bonds and the potential for them to be structurally significant guides the protein purification strategies to exclude reducing agents, a common inclusion in purification buffers. The model predicts a high helical content indicates that CD analysis may be a useful tool for quality assessment.

CHAPTER 5. CD151 LEL PROTEIN PRODUCTION

5.1. *INTRODUCTION*

Structural and biochemical analysis of a protein is greatly facilitated by access to substantial quantities of protein in a natively folded, highly purified form. The production of recombinant proteins is a common technique for obtaining a protein of interest that is amenable to laboratory-based experimentation. The choice of expression system is usually guided by the post-translational complexity of the protein, but also by cost and availability.

E. coli as a host organism for recombinant protein expression is a widely used system which offers fast growth, low cost and high yields. If the protein requires eukaryotic post translational modification (PTM) such as glycosylation or disulfide bonding however, a prokaryotic expression system may not be adequate. More complex protein expression strategies such as insect or mammalian cell-based systems may address these issues but the cost of specialised culture media, access to specialised equipment, time to establish and propagate cells and, often, poor protein yields may outweigh the potential shortcomings of prokaryotic protein expression systems. Expression, and purification from *E. coli* remains the preferred system for recombinant protein expression and is usually the initial choice.

Strains of *E. coli* have been developed to mitigate some of the problems associated with prokaryotic expression of eukaryotic proteins. For example, Rosetta and CodonPlus host strains are derivatives of BL21 DE3 *E. coli* that are designed to enhance the expression of eukaryotic proteins that contain codons rarely used in *prokaryotes.* SHuffle and Origami cells, also derivatives of BL21 DE3, have a modified cytoplasm to facilitate disulfide bond formation. Commercially available *E. coli* expression vectors that promote soluble expression by incorporating fusion tags such as SUMO and Maltose Binding Protein (MBP) work well for some proteins and can circumvent the need to resort to the more expensive expression systems.

It is difficult to predict which expression system will work best for a particular protein and decisions should be informed by knowledge of the proteins structural intricacies and the end use of the recombinant protein. Reviewing the relevant literature and performing bioinformatic analysis can inform and prioritise the best strategies, as discussed in chapters 1 and 4. Published methods for recombinant expression of the protein or family members are an invaluable resource; however, in many instances these methods are, at best, a starting point for exploring protein expression and purification strategies.

5.2. *DESIGN OF CD151 LEL CONSTRUCTS*

The amino acid sequence for the LEL of CD151 was taken from UniProt entry P48509. CD151 is a 253 amino acid protein that contains four transmembrane (TM) regions, three cytoplasmic regions and two extracellular regions: the small extracellular loop (SEL) and the large extracellular loop (LEL). As discussed in Chapter 1, the LEL is a compelling target for drug discovery and was the region of interest for this project. The LEL spans 108 amino acids from residues 113 – 221 and contains the integrin binding site. Bioinformatic analysis revealed that the LEL contains two conserved disulfide bonds with the potential for a third, and a single N-linked glycosylation site. These PTMs influence choices made during protein expression and purification. All expression constructs were based on the 108 amino acid sequence of the CD151 LEL.

Various constructs were designed for use during the project and are summarised in Table 10. Plasmid constructs to be expressed in *E. coli,* insect or mammalian cells were codon optimised, synthesised and cloned into commercial vectors by Genscript or in-house.

Table 10 Summary of the constructs used for this project.

Name	N-terminal Tag(s)and protease site	Cloning sites	Expression Vector	Mw (Da)
GST-CD151 LEL	GST-PreScission	BamH1-Not1	pGEX-6p2	39290
His-Avi-CD151 LEL	His_8-TEV-AviTag™	BamH1-Not1	pET30a(+)	17550.2
His-CD151 LEL	His_8-TEV	BamH1-Not1	pET30a(+)	15176.6
His-CD151 LEL C184S C192S	His_8-TEV	BamH1-Not1	pET30a(+)	15144.5
His-GFP-CD151 LEL	His_6-thrombin-GFP	Ndel-BamH1	pET28a	40267
TrxA-His-CD151	TrxA-His_6-TEV	Msc1-Not1	pET32a+	27080
PelB-His-CD151	PelB-His_6-TEV	Ndel1-Not1	pET-22b(+)	17687.6
His-MBP-CD151 LEL	His_8-MBP-TEV	BamH1-Not1	pET30a(+)	55127
MBP(A-D)-CD151 LEL	MBP(A-D)	BamH1-Not1	pMALX(A-D)	52797.6 – 53013.9
His-SUMO-CD151 LEL	His_6-SUMO-ULP1	Nde1-BamH1	pET30a(+)	24782.6
His-SUMO-CD151 LEL Pross	His_6-SUMO-ULP1	Nde1-BamH1	pET30a(+)	24932.8
IgK-CD151 LEL-His	Ig Kappa (C-terminal His_6)	Nhe1-Not1	pcDNA3.1(+)	13692
HBM-His-CD151 LEL	HBM-His_6-TEV	Not1-Xbal	pFastBac1	14252.6

5.3. *BACTERIAL EXPRESSION*

5.3.1. GST-CD151 LEL

Published methods for recombinant expression of CD151 LEL in bacteria fuse the N-terminus of the LEL to a glutathione S-transferase (GST) tag. [297, 315] GST rapidly folds into a stable and soluble protein upon translation and has been shown to act as a chaperone to facilitate protein folding. [316, 317] The GST-CD151 expression vector described in Table 10 was transformed into BL21 DE3 cells for expression.

Purification of GST-tagged proteins is based on the affinity of GST to the glutathione ligand coupled to a matrix such as the GSTrap™ columns (GE) (refer to section 2.2.5.4). Protein is eluted from the affinity column by reduced glutathione, which is a mild, non-denaturing condition that does not affect the protein's native structure and function. Inclusion of a PreScission Protease cleavage site for removal of the GST tag allows for purification of native protein. PreScission Protease is a fusion protein of GST and human rhinovirus type 14 3C protease. [318] The protease specifically recognises the amino acid sequence Leu-Glu-Val-Leu-Phe-Gln↓Gly-Pro, cleaving between the Gln and Gly residues (as indicated by the arrow). PreScission Protease is maximally active at 4°C so cleavage can be performed at low temperatures, potentially retaining the stability of the target protein. The GST tag allows easy removal from the cleavage reaction and facilitates on-column cleavage which can simplify purification and tag cleavage into a single step process.

5.3.1.1. Expression in BL21 DE3 cells

5.3.1.1.1. Small-scale expression

Small scale 100 ml expressions were used to optimise for media, temperature and time by comparing expression levels of cultures in LB and TB, at 24°C and 37°C, over intervals ranging from 30 minutes to overnight (15 hours). Samples of each culture were taken every half hour for 6.5 hours then overnight and analysed by SDS-PAGE.

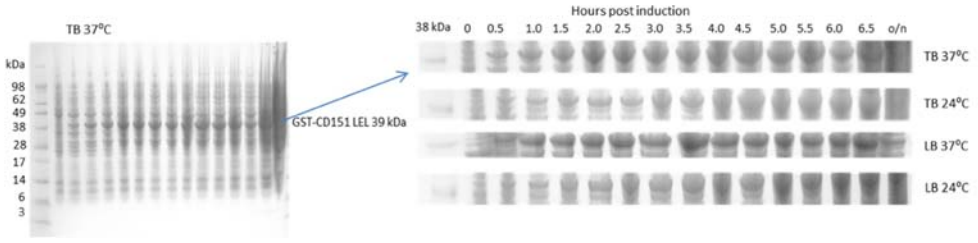


Figure 5-1 SDS-PAGE analysis of a time-course of the expression of GST-CD151 LEL in either TB or LB media The whole SDS-PAGE gel for TB at 37°C is given as an example (left) and the bands corresponding to a 39 kDa protein on all gels cropped for easier comparison. o/n represents overnight growth.

LB at 37^{o}C seemed to perform best as a growth media and protein levels did not appear to increase after 3.5 hours (Figure 5-1). LB media at 37^{o}C and harvesting 3.5 hours post-induction was chosen for all future expressions of GST-CD151 LEL.

A trial 2 L expression using LB media at 37°C and 3.5-hour induction was performed to assess protein production and purification. Samples were taken hourly post-induction and analysed by SDS-PAGE and Western blot (Figure 5-2).

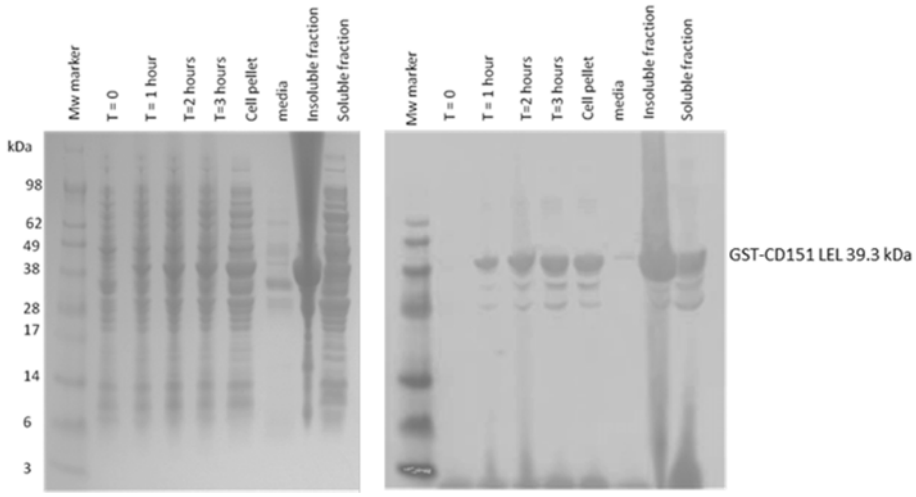


Figure 5-2 SDS-PAGE analysis of GST-CD151 expression. 1 ml samples were taken at hourly time points and the cell pellet was lysed and the soluble and insoluble fractions were separated. Samples were analysed by SDS-PAGE (left image) and Western blot probed with anti-GST Ab (right image).

While there was a large amount of GST fusion protein in the insoluble fraction, there was also GST fusion protein present in the soluble fraction. The soluble fraction of the cell lysate was purified over a GSTrap™ 5 ml column and fractions of the flow through, wash and eluted protein were analysed using SDS-PAGE (Figure 5-3).

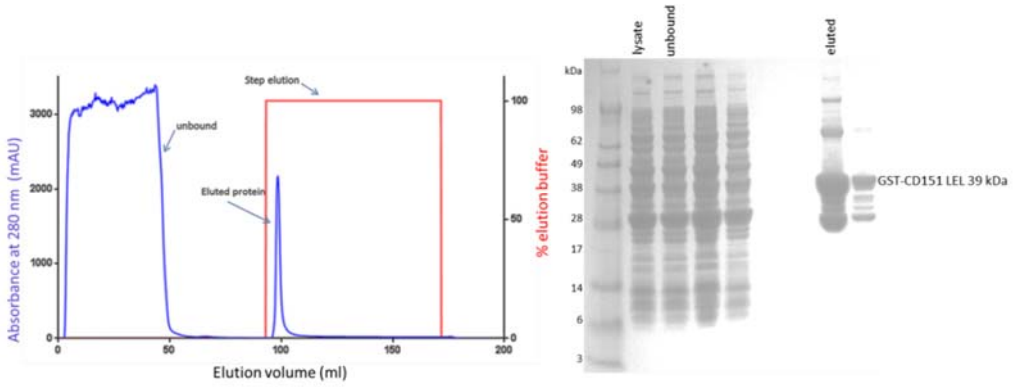


Figure 5-3 Elution profile of GST-CD151 purification over a GSTrap 5ml column in PBS. The blue line represents the absorbance at A280 nm (mAU) of flow through and eluted material and the red line indicates the concentration of elution buffer. Column eluted with 10 mM reduced glutathione in PBS (left). Cell lysate, unbound fractions and eluted protein analysed by SDS-PAGE (right).

SDS-PAGE (Figure 5-3) showed that although protein appeared in the unbound column flow through, most of the protein eluted in a single peak. The unbound material may be misfolded, or the column may have reached its binding limit. GST is a 26 kDa protein which often degrades upon denaturation and reduction for protein gel electrophoresis. [319] Degradation of the GST fusion protein is most likely responsible for the band observed at approximately 28 kDa.

A 500 µl sample of the protein eluted from the GSTrap (Figure 5-3) was further purified by size exclusion chromatography using a Superdex 75™ column (Figure 5-4).

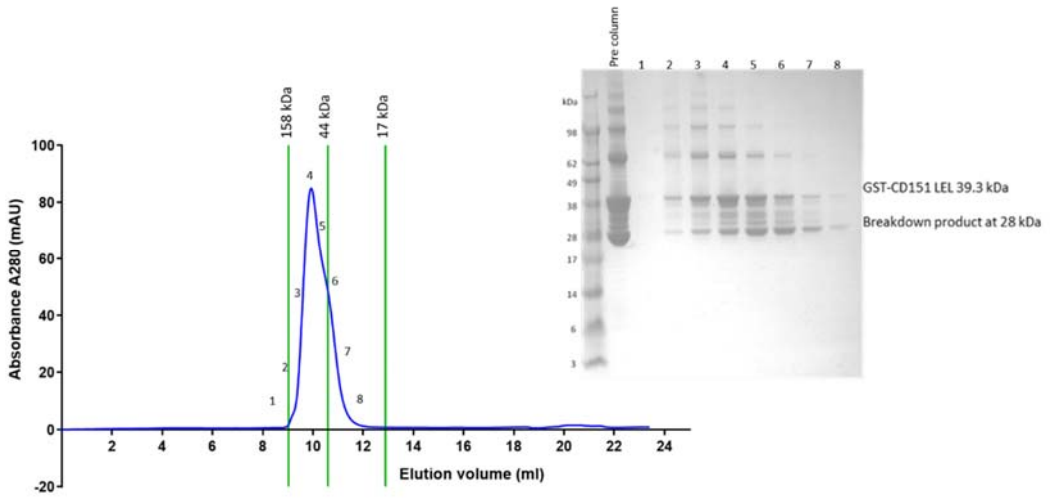


Figure 5-4 Size exclusion purification of affinity purified GST-CD151 using a Superdex 75 10 300 (GE) column (left). The blue line represents the eluted protein absorbance at

A280 nm (mAU). The column was pre-calibrated with Mw standards, their elution times and Mws indicated by the green lines. Eluted fractions were analysed by SDS-PAGE (right). Numbered fractions on the chromatogram correspond to the numbered lanes on SDS-PAGE.

A broad protein peak eluted across several fractions, suggesting a heterogeneous sample and that the protein was breaking down in solution and not simply an artefact of SDS-PAGE analysis. Fractions of the eluted protein were analysed using SDS-PAGE (Figure 5-4), confirming that there was breakdown product not adequately separated from the intact protein during size exclusion purification.

5.3.1.1. Buffer optimisation by thermal melt assay

To find a buffer condition that may increase protein stability during purification a sample of the size exclusion fraction containing the greatest amount of intact GST-CD151 was used for a thermal stability assay by Differential Scanning fluorimetry (DSF). Fluorescence based thermal shift assays quantify the change in thermal stability of a protein under various buffering conditions. The fluorescent dye SYPRO® Orange binds non-specifically to the hydrophobic core of proteins, which is exposed during thermal denaturation. As the protein further denatures and aggregates it begins to incorporate water which quenches the fluorescence. The midpoint of the unfolding transition from no fluorescence to maximum fluorescence is taken as the melting temperature, T_m. SYPRO orange has an excitation maxima of 300/472 nm and the emission maxima is 570 nm. The assay is performed using a real time-PCR machine which incrementally ramps the temperature while monitoring fluorescence. Using 96-well plates up to 48 different conditions can be analysed in duplicate.

GST-CD151 LEL protein from fraction 3 from the size exclusion purification seemed to have the most correct size protein than other fractions with the least amount of the 28 kDa contaminant (Figure 5-4). A sample of fraction 3 was assayed across a range of pH and salt concentrations. (Figure 5-5)

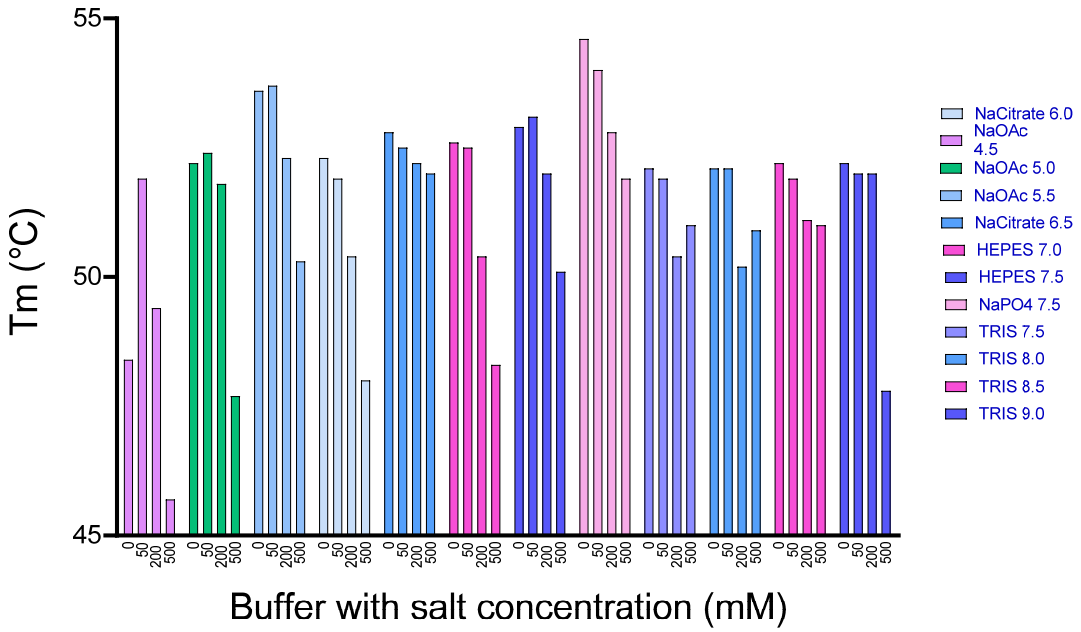


Figure 5-5 Melting temperature of GST-CD151 as determined by DSF thermal melt assay. The different buffers were tested with a salt concentration of 0, 50, 200 or 500 mM. Each well contained 3 µg of protein.

Examining the raw data revealed some anomalies with the assay. Fifteen of the conditions gave very high fluorescence compared with the rest of the conditions and all gave a T_m of exactly 52°C. (Figure 5-6) Fourteen of these conditions were at the edge of the 96-well plate (Figure 5-7). An "edge effect" is a widely reported phenomenon that plagues microtiter plate assays, generally due to increased evaporation or thermal gradients at the edge of the plate. [320, 321]

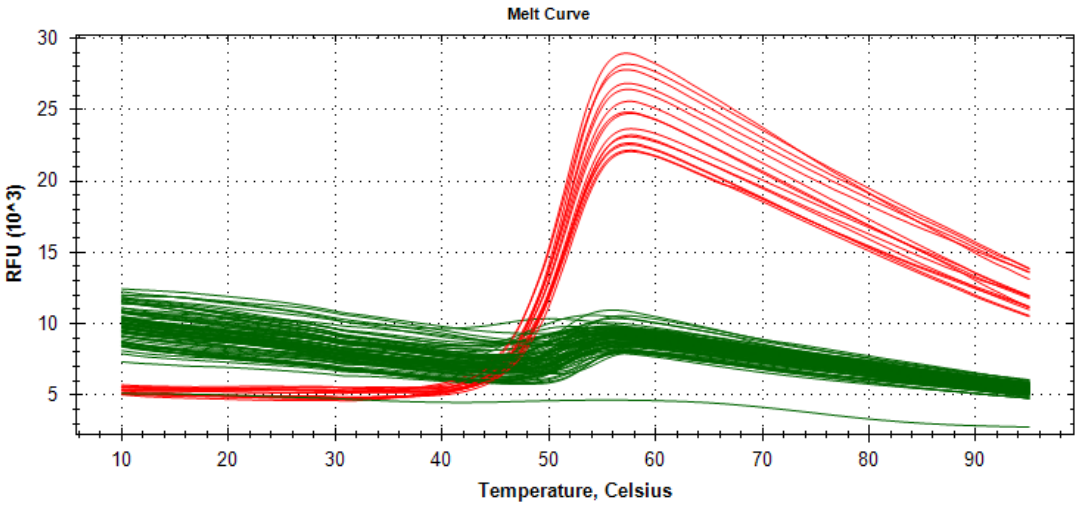


Figure 5-6 Relative fluorescence units (RFU) of the thermal melting curves of GST-CD151 LEL plotted against temperature. The curves in red indicate anomalous data.

	1	2	3	4	5	6	7	8	9	10	11	12
A	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk
B	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk
C	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk
D	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk
E	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk
F	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk
G	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk
H	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk

Figure 5-7 Wells generating an unusually high fluorescence signal highlighted in blue.

The location of the wells and the consistency of the T_m suggest that the increased fluorescence may be due to an edge effect rather than an accurate T_m.

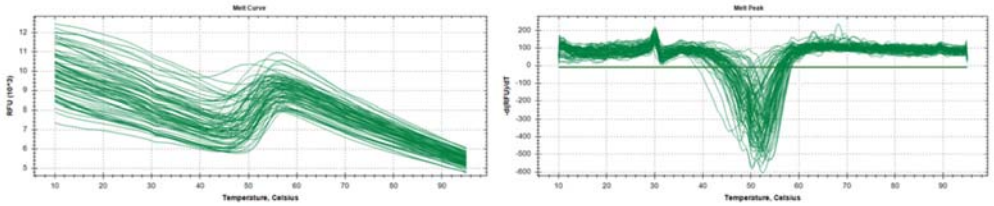


Figure 5-8 Relative fluorescence units (RFU) vs temperature (left) and the first derivative of RFU (-d(RFU)/dT) vs temperature (right) of the thermal melt assay of GST-CD151 under various pH and salt concentrations. The T_m is calculated as the temperature at the minimum of the first derivative.

Disregarding the anomalous readings reveals that most of the fluorescence curves start with a high reading, suggesting that that protein is already partially unfolded or contains solvent accessible hydrophobic regions that are binding the dye (Figure 5-8). Plotting the T_m v. condition shows a clear trend that the least stable conditions were low pH with high salt and the most stable condition was neutral pH with low salt (Figure 5-9). Based on these results, future purifications of GST-CD151 LEL were carried out in 100 mM sodium phosphate buffer pH 7.5 with 50 mM NaCl. Although the condition with no salt gave a higher T_m (over half a degree) than with 50 mM NaCl, it is preferable to include some salt for purification to prevent non-specific ionic interactions between the protein and the chromatographic resin. Sometime after this experiment was performed, I discovered that the brand of microtitre plate used (Bio-Rad Laboratories Pty Ltd. HSP9655 Hard shell PCR plate 96 well) was prone to auto-fluorescence at 56°C (data not shown); changing to a different plate manufacturer (Life Technologies Australia Pty Ltd AB0800W 0.2 ml skirted 96 well plate low profile) solved the anomalous readings (not shown).

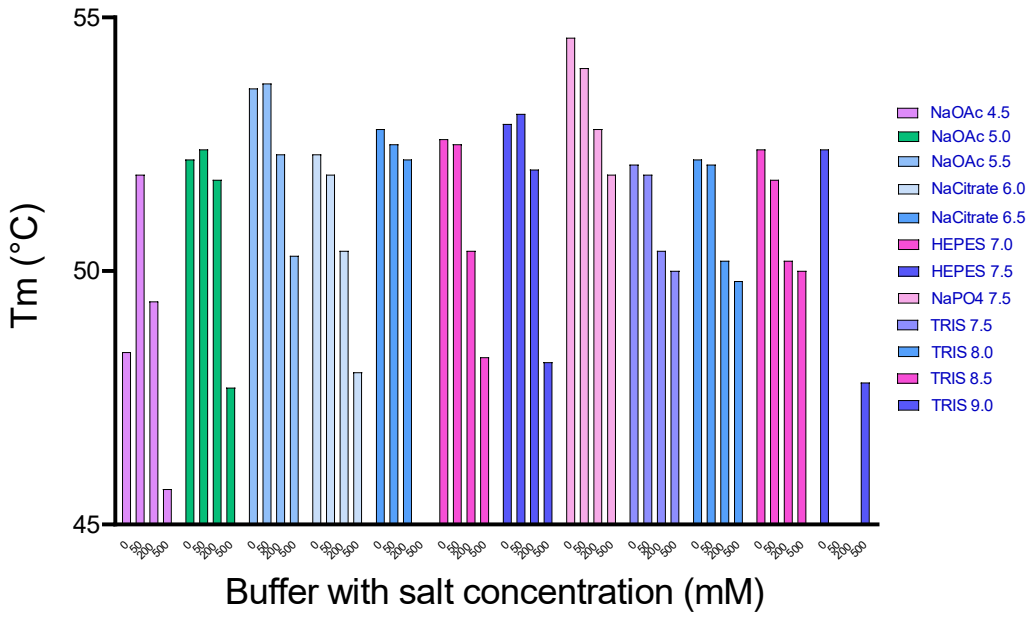


Figure 5-9 Plot of the thermal melt (T_m) of GST-CD151 vs buffer and salt condition. T_m was typically higher in low or no salt and highest at neutral pH in sodium phosphate buffer.

5.3.1.1. Large-scale expression

Protein was again expressed in a 3.5 L culture, reducing the temperature at induction to 16°C and limiting growth post-induction to 2 hours only to minimise protein degradation. 8.8 mg of protein was affinity purified and 4 x 100 µg aliquots of protein were incubated with 15, 4, 2 and 0 µg of PreScission protease at 4°C for 4 hours. Cleavage efficiency was analysed using SDS-PAGE (Figure 5-10).

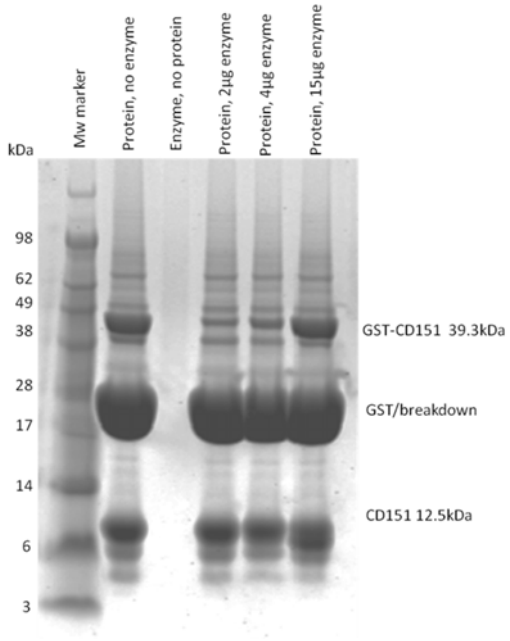


Figure 5-10 SDS-PAGE analysis of GST tag cleavage by PreScission protease. The lane with protein incubated with 2 µg of protease, labelled as "enzyme", has the greatest reduction in whole protein at 39.3 kDa.

Affinity purified protein, no protease added, in lane 2 of the SDS-PAGE gel in Figure 5-10 shows a strong band running above the 6 kDa marker. This band has been present in previous purifications but not in such quantities. It is unlikely that the low temperature expression increased protein degradation, but may have increased the stability of a breakdown intermediate product. The band of GST-CD151 fusion protein was visibly reduced with the addition of protease, however there did not seem to be a corresponding increase in the size of the GST or CD151 bands, suggesting that the protease was active, but the resulting protein products do not contain the full length CD151 LEL fusion protein. Surprisingly, increasing the amount of protease seemed to reduce cleavage. The Mw of PreScission protease is 46 kDa and the increase in band size as the concentration of protease increases may be due to the presence of the protease itself. However, this is unlikely as the lane without protease gives an almost identical banding pattern as the lane with the most protease. The lane with protease only has no visible bands despite the 15 µg of protease being added. The most likely explanation is that the low ratio of protease to protein is optimal for protease activity.

The SDS-PAGE gel in Figure 5-2 shows the ≈ 28 kDa contaminating protein was present in equal amounts to the GST-CD151 LEL fusion protein in the crude, soluble fraction of the cell lysate. However, the corresponding Western blot probed with anti-GST Ab was much more reactive to the fusion protein than the contaminant. This suggests that

although there may be a component of the 26 kDa GST present in that fraction most of the protein may be unrelated. Further purification methods were trialled to remove the contaminating protein.

5.3.1.1.1. Anion exchange chromatography

Cell lysate was purified over an anion exchange column HiTrap™ Q HP 5ml (GE) column and eluted with a 20-column volume gradient of 0 – 1 M NaCl (Figure 5-11).

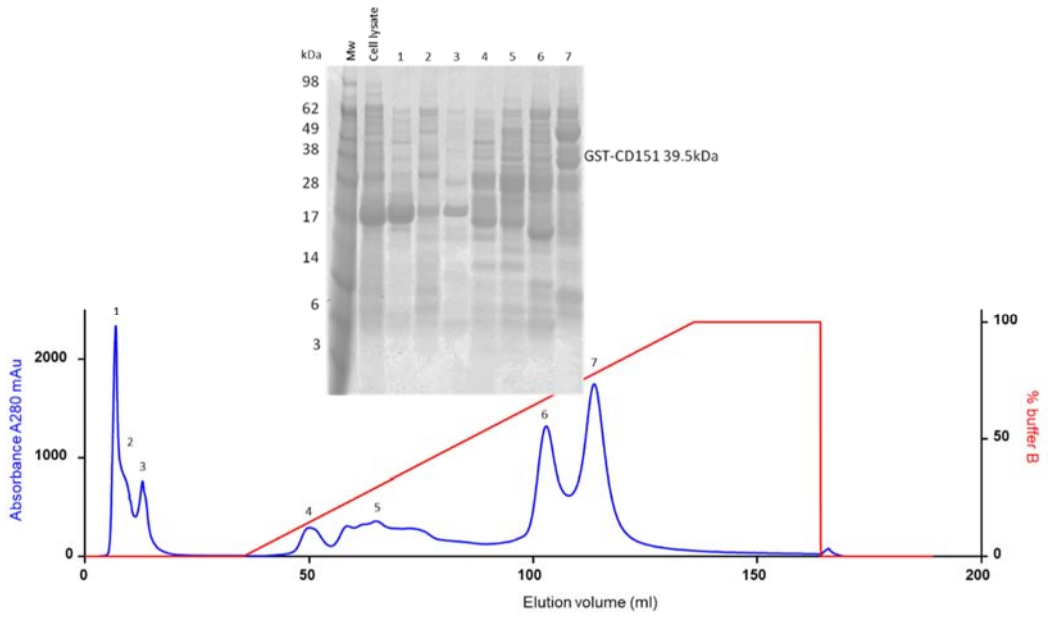


Figure 5-11 Chromatogram of anion exchange chromatography of GST-CD151 LEL cell lysate with a 20 CV gradient elution 0 – 1 M NaCl. The blue line represents the flow through and eluted protein absorbance at A280 nm (mAU) and the red line represents the % of buffer B. Fractions from the unbound flow through and the 4 eluted peaks were analysed using SDS-PAGE, numbered peaks correspond to numbered lanes on the gel. (insert)

Although the anion exchange fractionated the lysate into 4 distinct protein peaks (Figure 5-11), analysis by SDS-PAGE revealed the peaks contained a heterogeneous mixture of protein with no separation of the fusion protein from the contaminants. As separating the contaminant with chromatographic techniques was unsuccessful an alternative method was investigated.

5.3.1.2. Ammonium sulfate precipitation

Ammonium sulfate precipitation is a commonly used method for fractionating complex protein mixtures. Protein solubility varies proportionally to the percentage of saturation of the ammonium sulfate. By increasing the concentration of ammonium sulfate in a step wise manner it is possible to salt out the different proteins at each step.

A larger scale 10 L expression of the GST-CD151 LEL was carried out with reduced temperature and induction time as previously. Cells were lysed using the EmulsiFlex-C5 Homogenizer and the soluble fraction clarified by centrifugation. A small-scale trial ammonium sulfate cut was performed at 1.5, 2, 2.5 and 3 M ammonium sulfate and the fractions analysed by SDS-PAGE. (Figure 5-12)

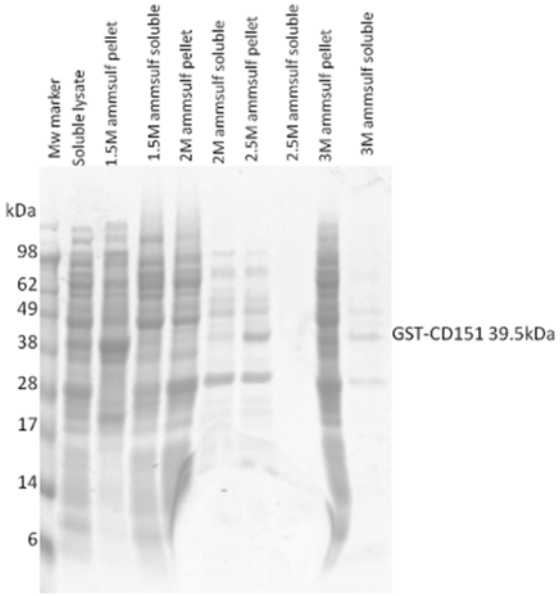


Figure 5-12 SDS-PAGE analysis of ammonium sulfate precipitation of GST-CD151 LEL cell lysate. Lanes show the result of 1.5, 2, 2.5 or 3 M ammonium sulfate.

The best separation of the fusion protein and the 28 kDa contaminant seemed to be at 1.5 M ammonium sulfate; although, there did seem to be GST-CD151 LEL present in other fractions. Unfortunately, the gel distorted due to the high ammonium sulfate content so the banding on the gel was difficult to interpret. Further analysis confirmed that 1.5 M ammonium sulfate was sufficient to precipitate GST-CD151 LEL protein (Figure 5-13).

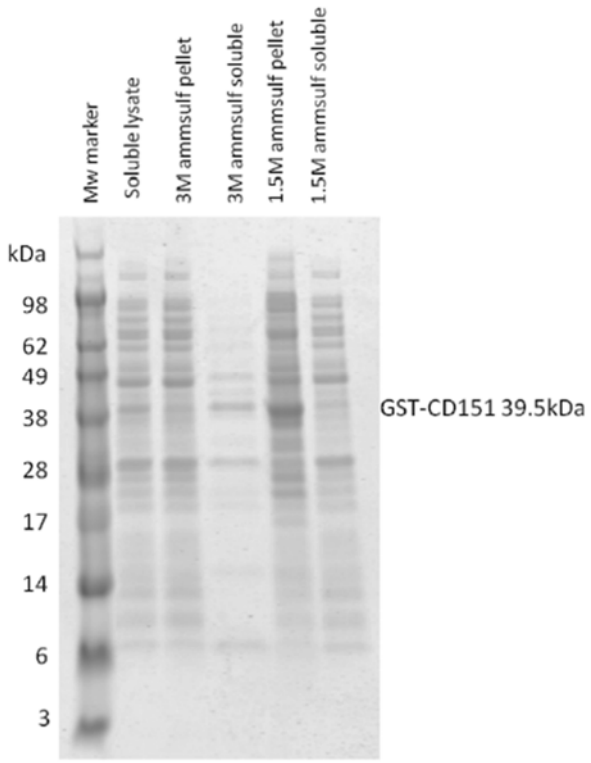


Figure 5-13 SDS-PAGE analysis of ammonium sulfate precipitation of GST-CD151 LEL cell lysate at 1.5 and 3 M.

As shown in Figure 5-13 most of the fusion protein is present in the 1.5 M ammonium sulfate pellet. There are other contaminating proteins present in the pellet but the contaminating 28 kDa protein stays soluble at 1.5 M ammonium sulfate.

The remaining 50 ml of cell lysate from the 10 L expression was precipitated with 1.5 M ammonium sulfate and the precipitated protein was pelleted by centrifugation. The pellet was dissolved in 20 ml PBS with protease inhibitors and purified over a GSTrap™ column as previously. Approximately 22 mg of affinity purified protein was collected. The protein was desalted into PreScission Protease cleavage buffer, 400 µg of protease added and the protein was incubated on ice for 4.5 hours. Cleaved protein was separated from intact protein and protease using a GSTrap™ column, such that the unbound flow through fraction from the column should contain the cleaved CD151 LEL while the cleaved GST tag and the GST tagged PreScission protease are bound to the column and eluted with 10 mM reduced glutathione. Fractions from the column were analysed using SDS-PAGE (Figure 5-14).

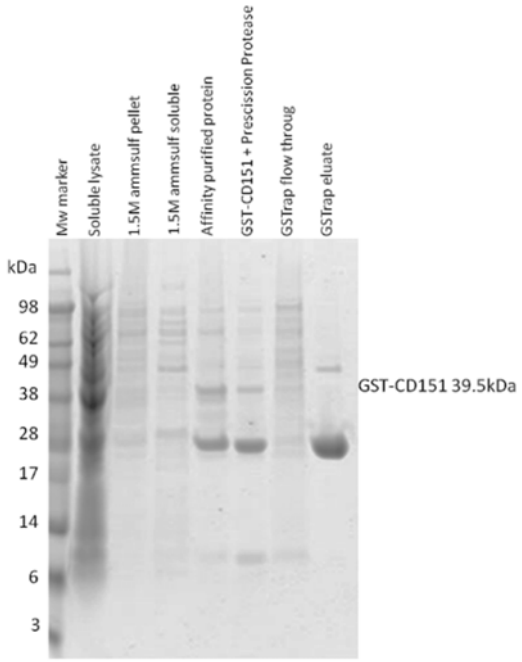


Figure 5-14 SDS-PAGE analysis of the stages of purification of GST-CD151 LEL fusion protein from a 10 L expression.

As shown in Figure 5-14 the ammonium sulfate precipitation did not completely remove the 28 kDa contaminating protein. The PreScission Protease cleavage reaction does decrease the band of protein at 39 kDa, and produce a band at approximately 8 kDa, as would be expected upon removal of the GST. However, this protein is too small to be intact CD151 LEL and is likely a breakdown product of the LEL. This notion is supported by the presence of the band in the cell lysate, prior to purification or cleavage, suggesting that the fusion protein is inherently unstable.

5.3.1.3. Purification and refolding of insoluble GST-CD151 LEL

Bacterial expressions of GST-CD151 LEL produced a large amount of insoluble protein at the expected Mw (Figure 5-2). The insoluble fractions from all previous expressions were pooled and solubilised in 8 M urea. The material was clarified by centrifugation at 20,000 *g* for 40 minutes and dialysed against PBS overnight at 4°C. The dialysed material was centrifuged to separate precipitated protein and samples from each step analysed by SDS-PAGE (Figure 5-15).

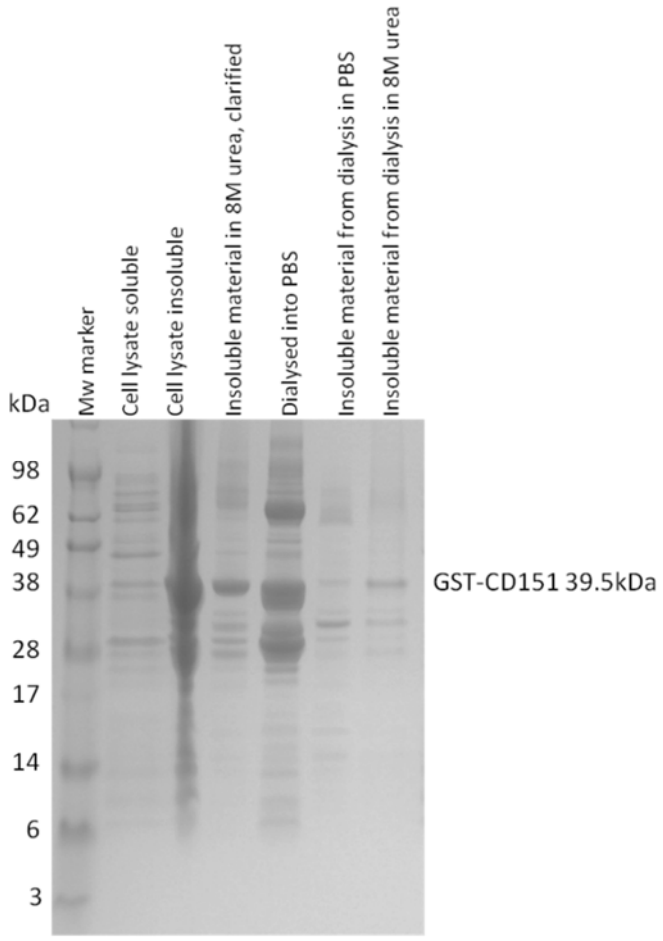


Figure 5-15 SDS-PAGE analysis of protein from the insoluble fraction of the cell lysate solubilised in 8 M urea, dialysed into PBS and the precipitated material from dialysis.

The urea solubilised material predominantly contained the fusion protein. This remained soluble after dialysis into PBS suggesting the protein has refolded, however the 28 kDa breakdown product was present in equal amounts. A sample of the solubilised material in PBS was further purified by size exclusion chromatography using a HiLoad Superdex 75 26 60 Prep Grade column and fractions were analysed using SDS-PAGE. (Figure 5-16).

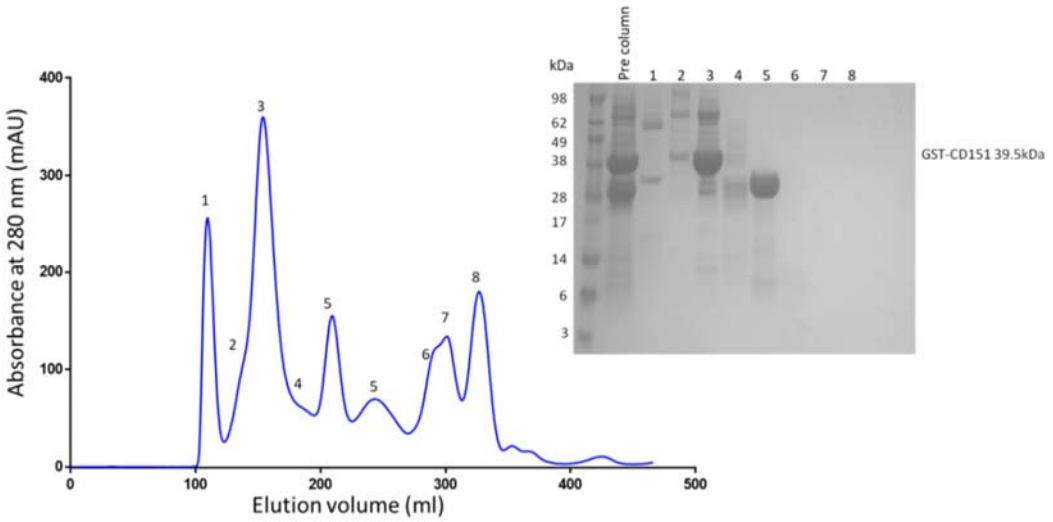


Figure 5-16 Size exclusion purification of the solubilised material in PBS. The blue line represents the eluted protein absorbance at A280 nm (mAU). Analysis of the fractions by SDS-PAGE (inset) reveals that the peak labelled 3 appears to contain the 39.5 kDa GST-CD151 LEL. Numbers on the eluted peaks correspond to the lanes on the gel.

Peak 3 from the Superdex 75 purified protein (Figure 5-16) was dialysed into PreScission Protease buffer and protease added at 1:50 ratio and incubated overnight at 4°C. The cleaved protein was purified using size exclusion chromatography and fractions from the size exclusion purification were analysed by SDS-PAGE (Figure 5-17).

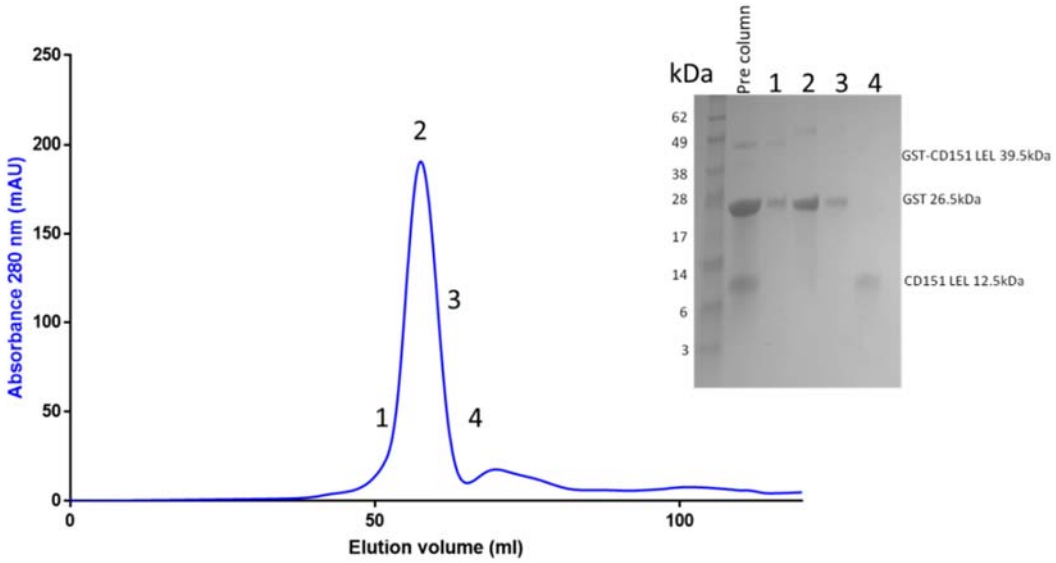


Figure 5-17 Size exclusion purification of the GST-CD151 LEL cleaved by PreScission Protease. The blue line represents the protein absorbance at 280 nm. Numbers on the eluted peak (left) correspond to the lanes on the SDS-PAGE gel (right).

SDS-PAGE analysis of the size exclusion fractions of the GST-CD151 LEL fusion protein cleaved with PreScission Protease shows only a very faint band of uncleaved protein at 39.5 kDa, a large band of the cleaved GST tag at 26.4 kDa and a diffuse band between 6 and 14 kDa which may be the 12.5 kDa CD151 LEL. Samples of uncleaved and cleaved CD151 were examined using circular dichroism spectroscopy (CD) (Figure 5-18).

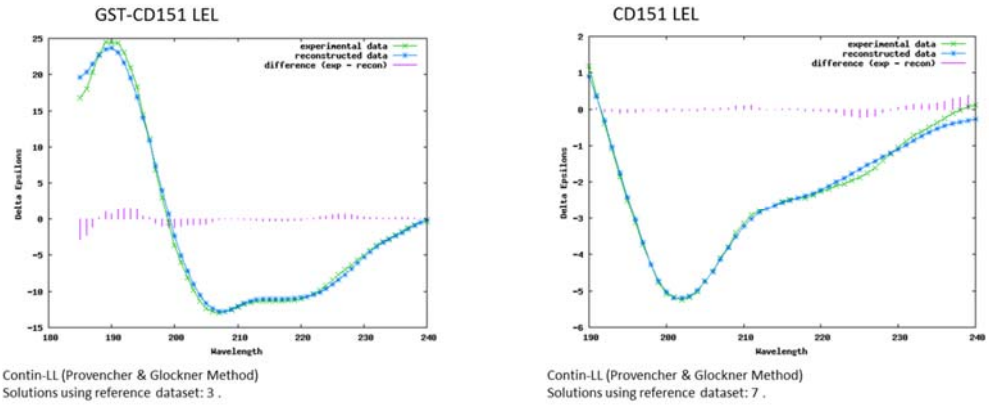


Figure 5-18 Graphical output of DICHROWEB protein secondary structure analysis from CD spectroscopic data. Experimental data are plotted in green; the calculated spectrum derived from the calculated output secondary structure is plotted in blue and the difference spectra is depicted in vertical lines in pink.

The plot for GST-CD151 LEL (Figure 5-18, left) shows a reasonable fit to the calculated data, with some deviation at lower wavelengths, the shape of the curve corresponding with that expected of a helical protein. The plot for CD151 LEL (Figure 5-18, right) shows a reasonable fit to the calculated data, with some deviation in the upper wavelength, the shape suggesting mostly disordered protein.

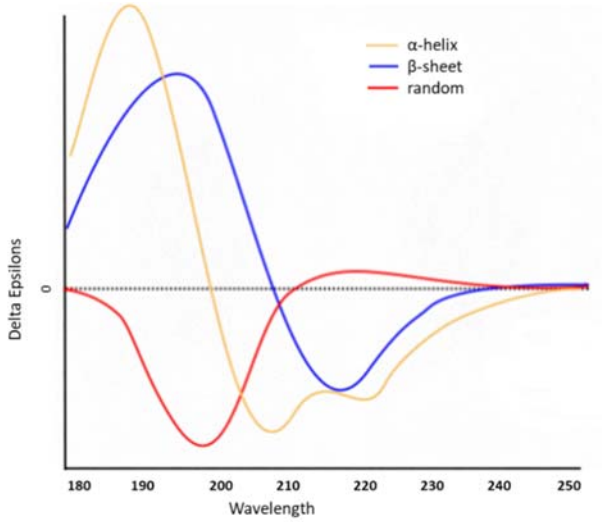


Figure 5-19 Characteristic far-UV CD spectra of pure secondary structures. The spectrum for an all-helix protein (yellow), all-sheet (blue) and a random (disorderly) protein. (red)

The CD spectra for both the GST-CD151 LEL and the CD151 LEL were reproduced reasonably well by the theoretical projections (Figure 5-18). Comparison to representative curves for protein with pure secondary structures suggested that the GST-CD151 LEL contains mostly helix and that the cleaved CD151 LEL is disordered. Examination of the deconvoluted secondary structure composition and total content summary shows that about 20% of the GST-CD151 LEL is unordered (Table 11) and up to 80% of the CD151 LEL is unordered (Table 12). As the CD151 LEL makes up approximately 30% of the total weight of the GST-CD151 LEL construct, these findings suggest that the GST tag was correctly folded but the CD151 LEL was disordered.

Table 11 Summary of the deconvoluted CD data listing the secondary structure composition and total content of GST-CD151 LEL. The analysis shows mostly helical protein with a small amount of unordered content.

Result	Helix1	Helix2	Strand1	Strand2	Turns	Unordered	Total
1	0.507	0.193	0.000	0.079	0.027	0.194	1
2	0.511	0.178	0.000	0.087	0.002	0.222	1

Table 12 Summary of the deconvoluted CD data listing the secondary structure composition and total content of CD151 LEL. The analysis shows a mixture of helix and strand with a large amount of disordered protein.

Result	Helix1	Helix2	Strand1	Strand2	Turns	Unordered	Total
1	0.064	0.063	0.087	0.051	0.123	0.612	1
2	0.030	0.054	0.100	0.000	0.012	0.805	1.001

These data show that the CD151 LEL produced by this method is not correctly folded. The diffuse band of CD151 LEL observed in SDS-PAGE analysis suggests a heterogeneous mix of protein species that may indicate varying disulfide bonding (Figure 5-17). The CD spectra data indicate that the CD151 LEL, after cleavage from the GST tag, is mostly disordered. The low yield, difficult purification and poor quality of the end product make this expression system unviable. As such, an alternative expression system was required with the hypothesis that standard bacterial expression was not allowing correct formation of disulfide bonds.

5.3.1.4. Expression of GST-CD151 LEL in SHuffle™ cells

5.3.1.4.1. Small-scale expression

An alternative expression host, *E. coli* strain SHuffle, which is genetically engineered to promote the cytoplasmic production of disulfide-bonded proteins, was trialled. The pGEX-6p2 vector containing the CD151 LEL gene was transformed into chemically competent SHuffle T7 express cells. A small-scale expression to test for protein production was carried out using Rich Media (RM) for 4.5 hours at 30°C, inducing expression at OD 600 nm = 0.6 with 0.5 mM IPTG. Samples were taken at induction and at two time points before and at harvest to monitor protein production. Analysis of

samples by SDS-PAGE showed expression of recombinant GST-CD151 LEL at > 2.5 h post-induction (Figure 5-20).

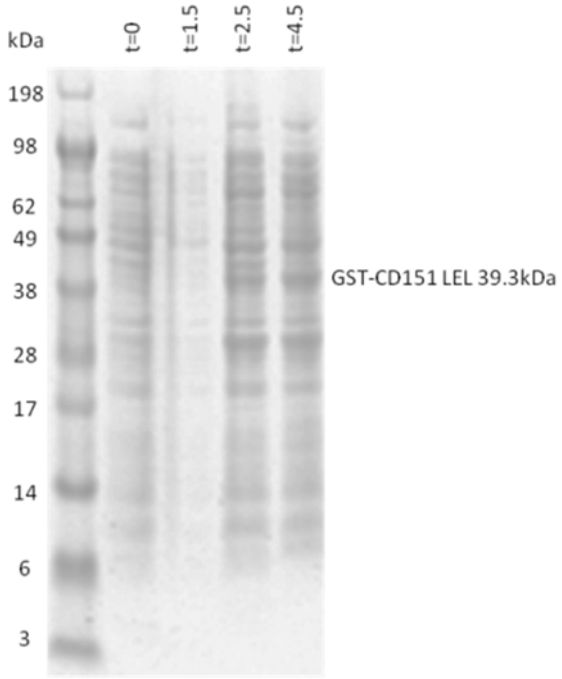


Figure 5-20 SDS-PAGE analysis of GST-CD151 LEL production in SHuffle cells.

5.3.1.4.2. Large-scale expression

An 8 L culture of pGEX-6p2/GST-CD151 LEL transformed SHuffle cells was grown and protein expression carried out as for the small scale trial described above. Cells were harvested, resuspended in lysis buffer and lysed using an EmulsiFlex-C5 Homogenizer. The lysate was centrifuged at 20,000 x g for 30 minutes at 4°C. The supernatant was decanted, diluted to 80 ml in 50 mM HEPES 100 mM NaCl 0.5 mM $MgCl_2$ pH 7.5, purified over a 5 ml GSTrap™ column, and eluted with the same buffer supplemented with 10 mM reduced glutathione. Samples were analysed using SDS-PAGE (Figure 5-21).

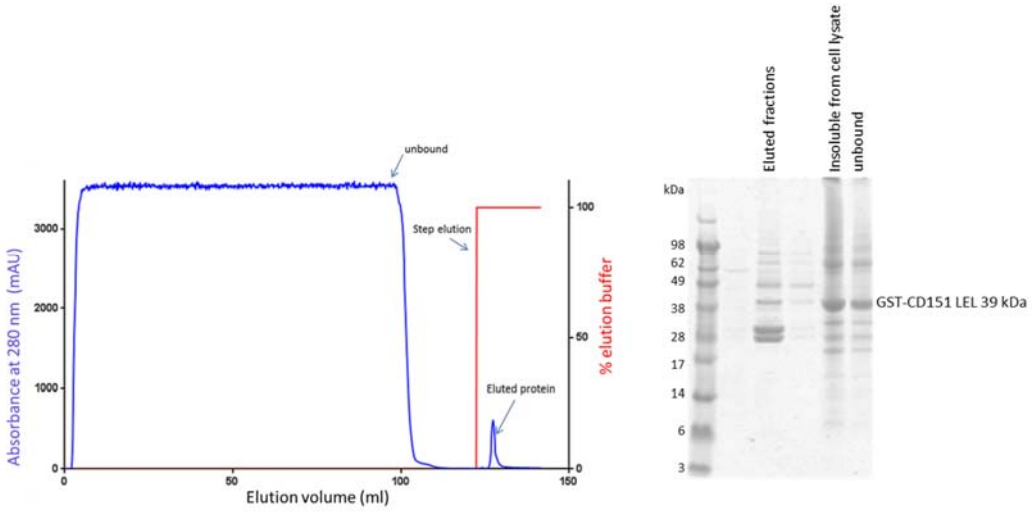


Figure 5-21 Chromatogram of the purification of the soluble fraction of cell lysate by GSTrap (left). The blue line represents the flow through and eluted protein absorbance at A280 mAU, the red line represents the % of elution buffer. Unbound and eluted fractions were analysed by SDS-PAGE (right).

The eluted fractions contained a small amount of GST-CD151 LEL but most of the protein was in the insoluble fraction or the column flow through. (Figure 5-21) GST binding to GSTrap resin is conformation dependant; therefore, failed retention of the soluble fusion protein on the GSTrap column indicates that the protein was incorrectly folded. The eluted protein contained 1.8 mg total protein; however, the GST-CD151 LEL band represents only a fraction of the total protein, rendering this expression method unsuitable for production of quantities that would be commensurate with biophysical and crystallographic analysis.

5.3.2. His-Avi-CD151 LEL

Facilitation of an SPR assay to screen for potential inhibitors of CD151 function requires recombinant protein that can be coupled to an SPR chip. Many methods for coupling proteins are available, the most common method being covalent coupling via a free amine group on the protein surface. This method, however, is not selective, resulting in a heterogeneous mix of orientations of the protein on the chip surface; a population of which will, by virtue of the coupling chemistry, likely obscure the targeted binding site. Alternative methods utilising tags added to the protein C- or N-terminus during expression will ensure homogenous orientation on the chip and a solvent accessible, unperturbed, binding site. In order to examine ligand binding to CD151 LEL, I designed a construct with a cleavable C-terminal His_6 tag, for affinity purification, preceded by a non-cleavable AviTag™. Once biotinylated through the AviTag™, the protein can be

captured on a streptavidin chip. The affinity for biotin to streptavidin is in the pM range which makes the coupling highly stable. Additionally, the high-affinity capture chemistry acts as a further purification step to remove contaminating protein from the sample, and results in a stable, homogenous surface that is optimal for examining ligand binding kinetics.

5.3.2.1. Small-scale expression in BL21 DE3 cells

His_6-TEV-Avi-CD151 LEL gene in pET30a(+) vector was transformed into BL21 DE3 cells and a small scale 100 ml culture grown for a trial protein expression as previously described. Fractions from the expression were analysed using SDS-PAGE (Figure 5-22).

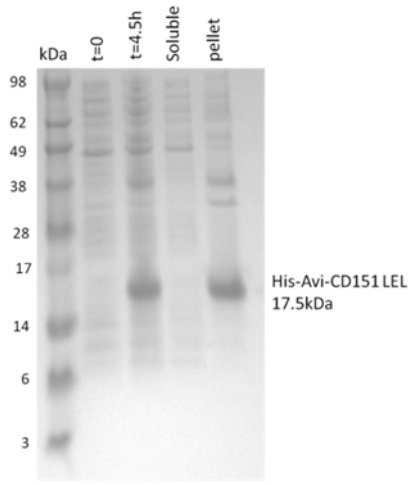


Figure 5-22 SDS-PAGE analysis of His_6-Avi-CD151 expression in BL21 DE3 cells. t represents time post-induction.

These experiments showed that, post-induction, protein expression was confined to the insoluble fraction, and the expressed protein appeared to migrate in the SDS-gel matrix at less than the predicted MW of 17.5 kDa. These experiments suggested that the protein was unfolded and therefore, the previous strategy (page 129) of using SHuffle cells, to promote disulfide bond formation, was employed to promote correct folding of the recombinant protein.

5.3.2.2. Small-scale expression in SHuffle cells

The pET30a(+) vector containing the His_6-TEV-Avi-CD151 LEL gene was transformed into SHuffle T7 Express cells. A small-scale expression to test for protein production was carried out using Rich Media (RM) for 4 hours at 30°C, inducing expression at OD600 = 0.6 with 0.5 mM IPTG, and samples were analysed by SDS-PAGE (Figure 5-23).

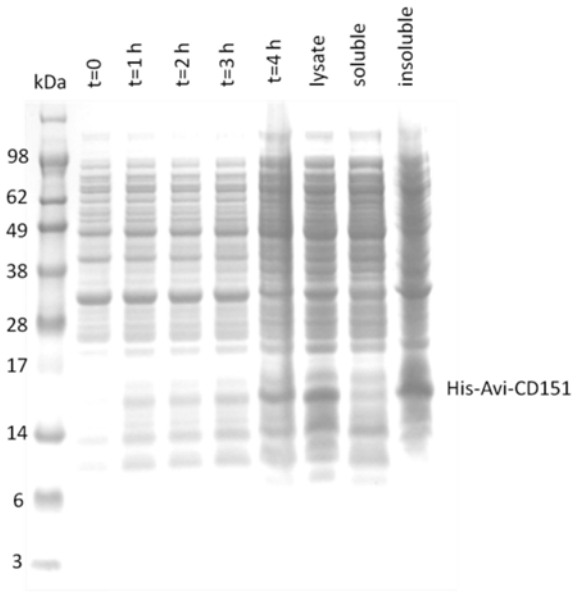


Figure 5-23 SDS-PAGE analysis of samples from His_6-Avi-CD151 expression in SHuffle cells. Samples were taken at induction (t=0) and for the next four hours. The cell lysate was then fractionated into soluble and insoluble protein by centrifugation.

Although most of the fusion protein was consigned to the insoluble fraction, these experiments suggested that a small proportion of His_6-Avi-CD151 was present in the soluble fraction. This material was diluted in PBS and purified over a HisTrap™ column (Figure 5-24).

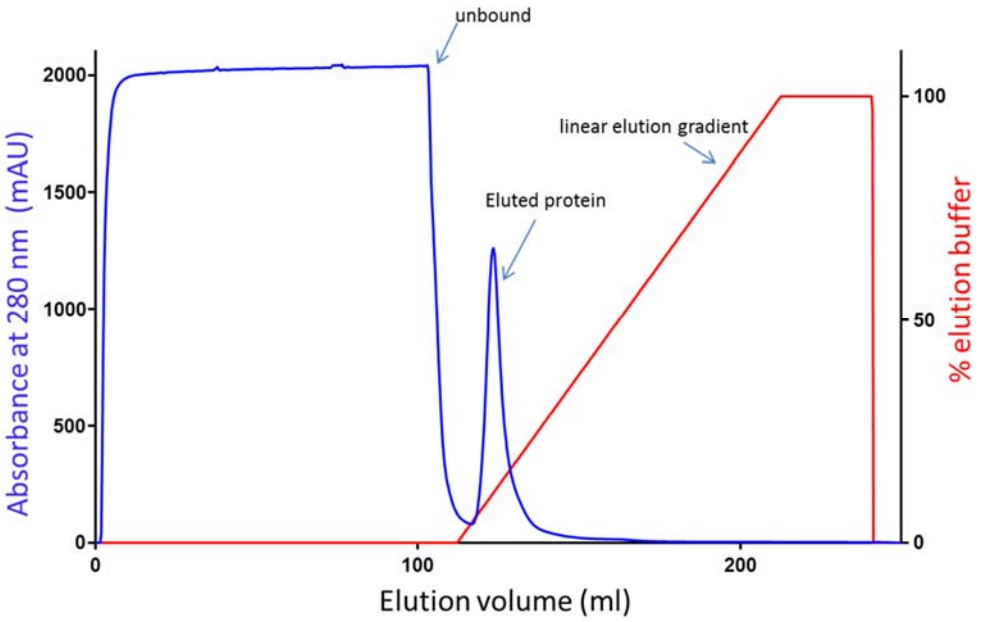


Figure 5-24 Chromatogram of HisTrap purification of His_6-Avi-CD151. The blue line represents the flow through and eluted protein absorbance at 280 nm, the red line indicates % of elution buffer.

Eluted fractions were concentrated to 0.5 ml and purified over a Superdex 75 gel filtration column (Figure 5-25).

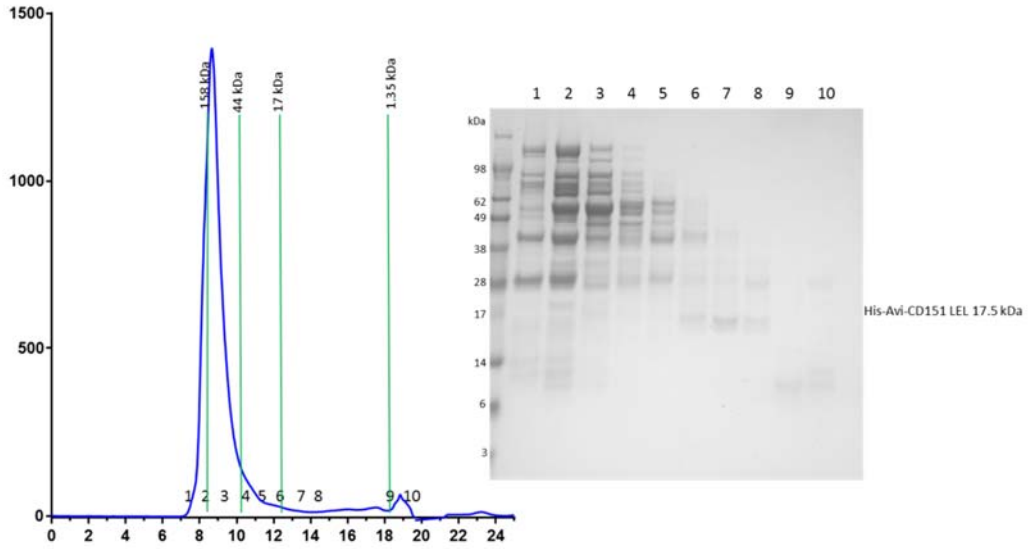


Figure 5-25 Chromatogram of elution profile of the His_6-Avi-CD151 LEL HisTrap fractions purified over Superdex 75 10 300 gel filtration column. (left) The column was pre-calibrated with Mw standards, their elution times and Mws indicated by the green lines. Eluted fractions were analysed by SDS-PAGE (right). Numbered fractions on the chromatogram correspond to the numbered lanes on SDS-PAGE.

Gel filtration standard 17 kDa myoglobin elutes at 12.9 ml, so the 17.5 kDa His-Avi-CD151 LEL should elute immediately preceding. Analysis of the fractions using SDS-PAGE showed faint bands in the fractions eluting at 11.35 ml (fractions 6, 7, 8) which possibly correspond to His_6-Avi-CD151. (Figure 5-25)

Western blot analysis was performed, using an Ab reactive to the histidine tag, to confirm the presence of the CD151 fusion protein (Figure 5-26). The anti-His Ab was reactive only to protein in the insoluble fractions or at a much higher Mw than predicted for the His_6-Avi-CD151 protein. The anti-CD151 Ab, which binds only to correctly folded CD151, was reactive only to the insoluble material. These results suggest that the CD151 fusion protein produced from prokaryotic SHuffle cell expression was only present in the insoluble fraction or as incorrectly folded high order oligomers.

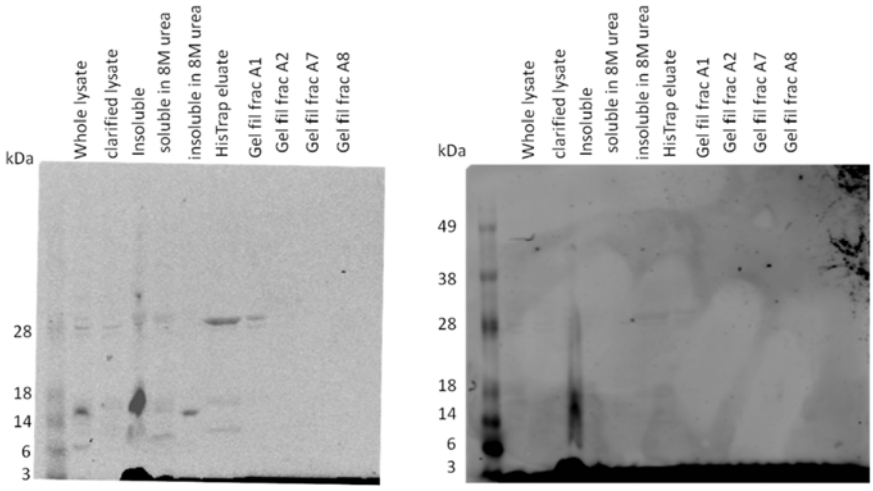


Figure 5-26 Western blot analysis of the stages of purification of the CD151 fusion protein. The blot was performed twice, probing first with anti-His Ab (left) then the second blot probed with anti-CD151 Ab (right).

For the anti-CD151 Ab to react to the insoluble material suggests that the CD151 LEL adopts a conformation that is recognised by the conformational Ab but that the protein is not soluble. One possibility is that insolubility of the protein may be due to inter-subunit interaction of hydrophobic areas on the surface of the protein, and it is possible that buffer conditions could be modified to increase solubility.

Using fraction 1 from the gel filtration purification, which appeared to contain dimers of the CD151 fusion protein (Figure 5-25), several buffer additives were used to examine their effect on the aggregation of the protein in solution using dynamic light scattering (DLS) based analysis (Figure 5-27).

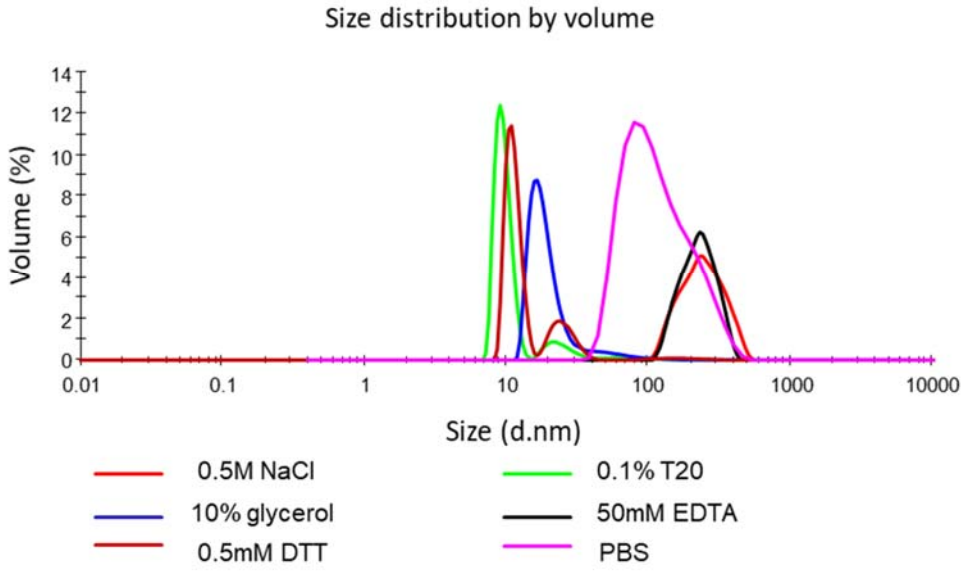


Figure 5-27 The effect of different buffer additives to oligomeric His_6-Avi-CD151 analysed by DLS. Pink = PBS, orange = PBS+0.5 M NaCl, grey = PBS+50 mM EDTA, blue = PBS+10% glycerol, red = PBS+0.5 mM DTT, green = PBS + 0.1% TWEEN®20.

The purified protein in PBS buffer, shown in pink, shows a broad size distribution averaging at 100 nm. PBS buffer contains 150 mM NaCl. Increasing the ionic strength of the buffer by increasing the salt content may reduce aggregation of protein by shielding electrostatic interactions between dipolar and charged functional groups. In this instance, increasing the salt concentration to 0.5 M, shown in orange, appears to have destabilised the protein further and caused an increase in aggregation. Addition of 50 mM EDTA, shown in grey, also increased the aggregation of the protein. As EDTA chelates divalent metal ions, this result suggests ions like Ca^{2+} and Mg^{2+} may help stabilise the protein.

Stabilising osmolytes such as glycerol are thought to interact with the exposed amide backbone of proteins. [322] Therefore, addition of a stabilising osmolyte possibly favours the native state by preventing promiscuous inter-subunit interactions involving the main-chain and protects against aggregation. Addition of 10% glycerol to the protein, shown in blue, reduced the hydrodynamic radius of soluble protein suggesting the glycerol reduced aggregation.

Dithiothreitol (DTT) is a redox reagent used to reduce protein disulfide bonds. DTT is typically used at concentration of 1 – 100 mM to reduce solvent accessible disulfide bonds. Addition of DTT at 0.5 mM (Figure 5-27), shown in red, reduced the protein size to a defined peak at 10 nm, strongly suggesting that the soluble protein forms

intermolecular disulfide bonded oligomers. At this low concentration, some intramolecular disulfides might remain intact, although non-native disulfides may be less stable and more readily reduced. The second, smaller peak in the DTT treated sample indicates the presence of larger protein complexes suggesting that the complexes are not due solely to intermolecular disulfide bonds. The most interesting and dramatic effect on protein size was from the addition of 0.1% TWEEN®20, shown in green (Figure 5-27), which decreased protein size slightly more than reduction with DTT. This suggests that the proteins are self-associating through hydrophobic patches that are disrupted by the amphiphilic nature of the detergent. A second peak of larger protein was again present in the sample, suggesting an association not disrupted by detergent and most likely due to disulfide bonded oligomers. These results suggest that buffer conditions may be varied to improve solubility of correctly disulfide bonded but misfolded protein that is present in the insoluble fraction of the cell lysate.

5.3.2.3. Protein refolding buffer selection

The DSF thermal melt assay is the most common method for monitoring the unfolding of a protein under various buffering conditions, see page 115. One of the limitations of this technique however, is that the Sypro Orange dye will react to native hydrophobic regions and will also fluoresce in hydrophobic environments created by the presence of detergents. The signal produced under these circumstances will mask the signal associated with protein unfolding. [323] The assay also requires approximately 250 µg of soluble protein per 96-well plate assay [324]. Inability to produce soluble, correctly folded, CD151 LEL in significant quantities reduced the amenability of this technique, and necessitated innovation of a novel method to examine protein stability in different solvent conditions.

Because there are infinitely variable permutations of electrolytes and additives that could be tested for correct folding of CD151 LEL, a high throughput method was required to screen for suitable solvent conditions. Using the CD151 Ab 11G5α (Abcam ab33315), which is recognises a 3D epitope of intact CD151 antigen, and a tetra-His Ab (Qiagen 34670) amine coupled to a CM5 (GE) sensor chip, I developed a method to screen buffer conditions that favoured refolding of denatured protein.

5.3.2.3.1. Method

Insoluble His_6-Avi-CD151 protein was solubilised in 8 M urea and purified using a 5 ml HisTrap™, eluting with 6 M urea and 500 mM imidazole. 20 µl of the solubilised material was transferred into each well of a 96-well deep well block containing 1 ml of various buffers and additives. The block was incubated overnight at 4°C with gentle rocking to allow the protein to refold. The block was centrifuged at 4000 x g for 5 minutes to sediment any insoluble material and 200 µl of the soluble fractions were transferred to a 96-well plate for analysis using the Biacore T200.

By injecting the different fractions over the immobilised Abs, the total amount of protein could be measured by the amount of binding to the tetra-His Ab and the amount of correctly folded protein could be measured by the amount of binding to the CD151 Ab. The change in refractive index caused by the different buffers was compensated by taking the measurement after the injection had ended but before the protein had dissociated from the chip surface. This is referred to as the "stability" report point of the curve (Figure 5-28).

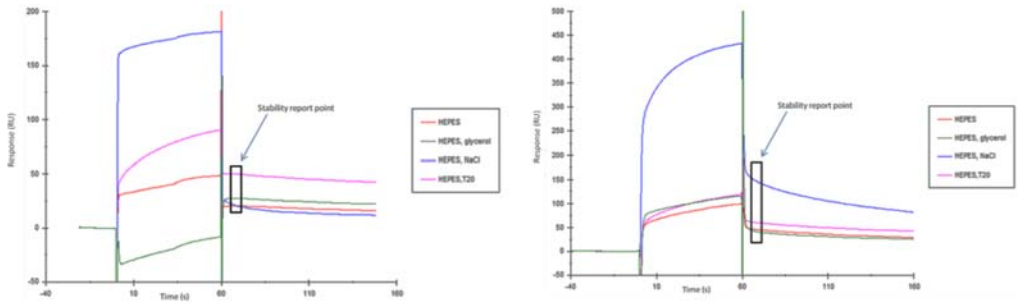


Figure 5-28 SPR sensorgrams showing the binding response of refolded protein in 50 mM HEPES pH 7.5, 50 mM HEPES 10% glycerol, 50 mM HEPES 150 mM NaCl, 50 mM HEPES 0.1% TWEEN®20 to CD151 Ab (left) and tetra-His Ab (right).

From the representative example above, it was apparent that the addition of 0.1% TWEEN®20 to the buffer improved binding to the CD151 Ab and suggested that detergent is required for the correct folding of the CD151 fusion protein. Binding to the tetra-His Ab however, is improved with the addition of NaCl suggesting that salt is required for increased solubility, although it does not favour correct folding as the increased salt did not increase binding to the CD151 Ab. This result indicates that solubility does not necessarily increase in parallel with the prevalence of correctly folded protein, and that the protein may be capable of forming non-native, soluble conformations. As the CD151 and His Abs were immobilised on the SPR chip to the same level the amount of protein binding to the Abs would normally be considered comparable; however, owing to the unknown affinities of the Abs in this instance the amount of binding cannot be considered quantitative.

5.3.2.3.2. Results

By plotting the binding response, at the stability report point, as a function of buffer condition it is possible to develop a buffer profile that favours solubility and correct conformation (Figure 5-29, Figure 5-30).

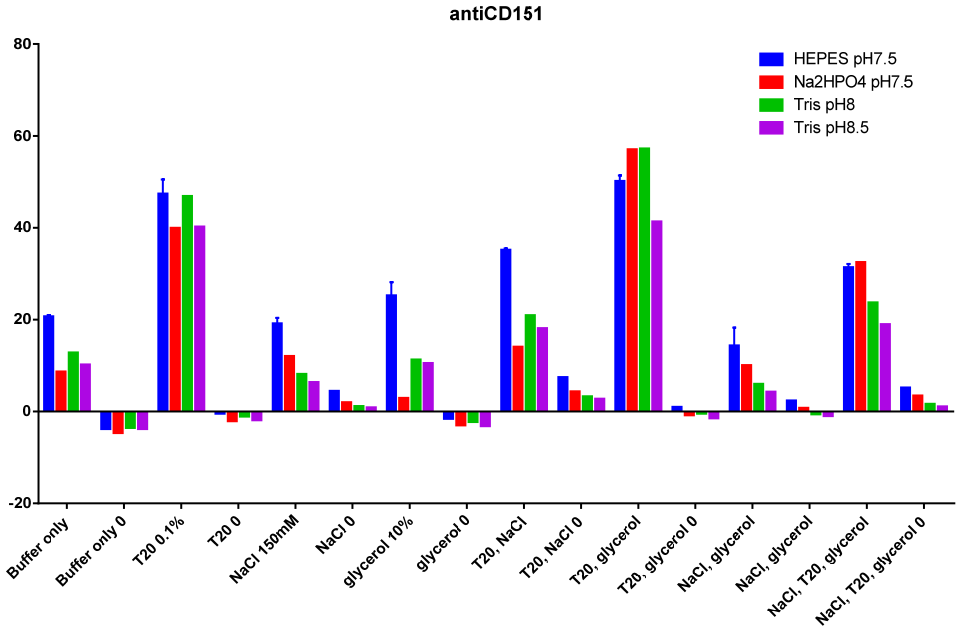


Figure 5-29 Binding response of refolded His_6-Avi-CD151 to a CD151 Ab under various buffer conditions. Measurements were performed in triplicate and results are shown as average ± standard deviation.

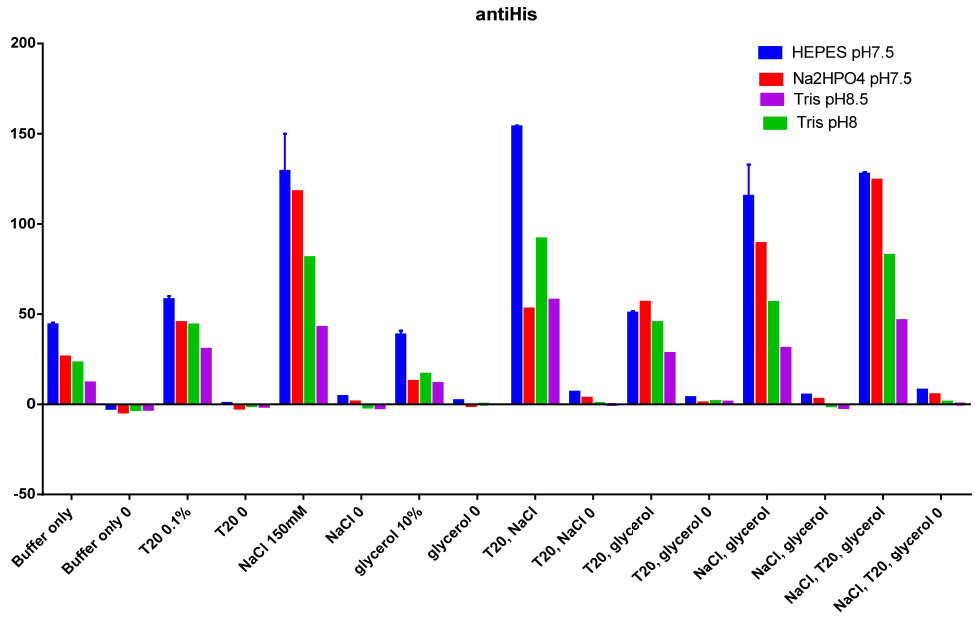


Figure 5-30 Binding response of refolded His_6-Avi-CD151 to a tetra-His Ab under various buffer conditions. Measurements were performed in triplicate and results are shown as average ± standard deviation.

The theoretical pI of the His_6-Avi-CD151 fusion protein is 5.92, as calculated from the protein sequence by ExPASy. [325] Thiol-disulfide exchange is inhibited at low pH and a pH >7 is typically required for disulfide bond formation. [326, 327] As such, I chose to screen buffers above the pI as to give the protein a net negative charge to increase solubility [328] and above pH 7 to promote disulfide formation and correct folding. Three standard buffers used routinely in protein purification and biochemical assays in this pH range are HEPES, phosphate and Tris-HCl. Adequate buffering capacity is usually reached at concentrations higher than 25 mM; however, concentrations up to 50 mM may be required after the addition of protein and other additives. [329] To ensure adequate buffering under all conditions all buffers were tested at 50 mM. Standard concentrations of NaCl at 150 mM, glycerol at 10% and TWEEN®20 at 0.1% were chosen as a starting point for additives (Figure 5-29, Figure 5-30).

Consistent with the observed reduction in protein hydrodynamic radius in the previous experiment using DLS (Figure 5-27) the greatest increase in binding to the CD151 Ab was with the addition of 0.1% TWEEN®20 (Figure 5-29). Binding response was usually better in HEPES buffer. The addition of NaCl and glycerol did not seem to improve binding to the CD151 Ab suggesting that they are not required for correct folding. However, the greatest increase in response to the tetra-His Ab was with the addition of NaCl (Figure 5-30). At 500 mM NaCl, as tested in DLS (page 137) protein aggregation increased and this may be what is leading to the increased response to the His Ab; however, no corresponding decrease in binding to the CD151 Ab was observed so it is unlikely that NaCl is affecting protein conformation. Again, HEPES buffer usually gave the best binding response to the His Ab. These findings suggest a HEPES buffer with low NaCl and TWEEN®20 favoured solubility and folding during protein purification.

TWEEN®20 at 0.1% is over 13 times the critical micelle concentration (CMC) of 0.007% [330] and at this concentration is non-dialysable and should be avoided in structural studies. [331] As such, it is not an ideal detergent for protein purification. (3-((3-cholamidopropyl) dimethylammonio)-1-propanesulfonate) or CHAPS, a zwitterionic detergent, is known to solubilise protein by attenuating protein-protein aggregates without the denaturing effects that are often attributed to harsher detergents such as Tween. [332] It is also readily dialysable if removal from the buffer is required for structural and biochemical studies. CHAPS is typically used at 0.1% (1.6 mM), which is lower than the CMC range of 6 – 10 mM, so I decided to test a concentration range of 1 – 5 mM.

To determine if the inclusion of NaCl was increasing solubility or simply increasing protein size (which also gives rise to an increased signal in SPR), I titrated NaCl, KCl, NaF and Na_2SO_4 from 50 mM to 400 mM. As the protein showed increased aggregation by DLS in the presence of EDTA, which chelates divalent metal ions, (Figure 5-27) it is possible that metal ions, such as Ca^{2+} and Mg^{2+}, are structural cofactors required for correct protein conformation. As such, I also tested the effect of low concentrations of $CaCl_2$ and $MgCl_2$ as calcium and magnesium ions are often observed in protein structures. [333] The refolding method was repeated, as above (page 139), for further buffer refinement with the above additives (Figure 5-31, Figure 5-32).

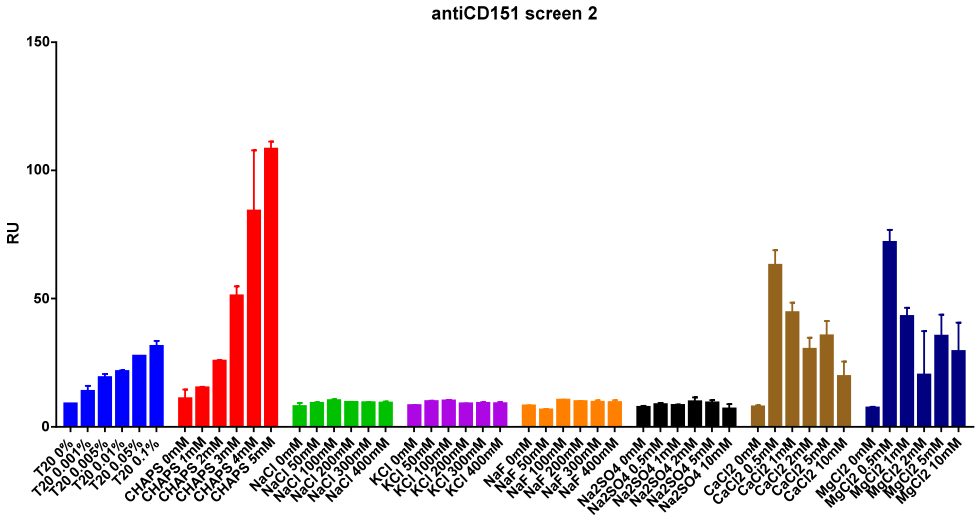


Figure 5-31 Binding response of refolded His_6-Avi-CD151 to a CD151 Ab in 50 mM HEPES pH 7.5 with various buffer additives. Measurements were performed in triplicate and results are shown as average ± standard deviation.

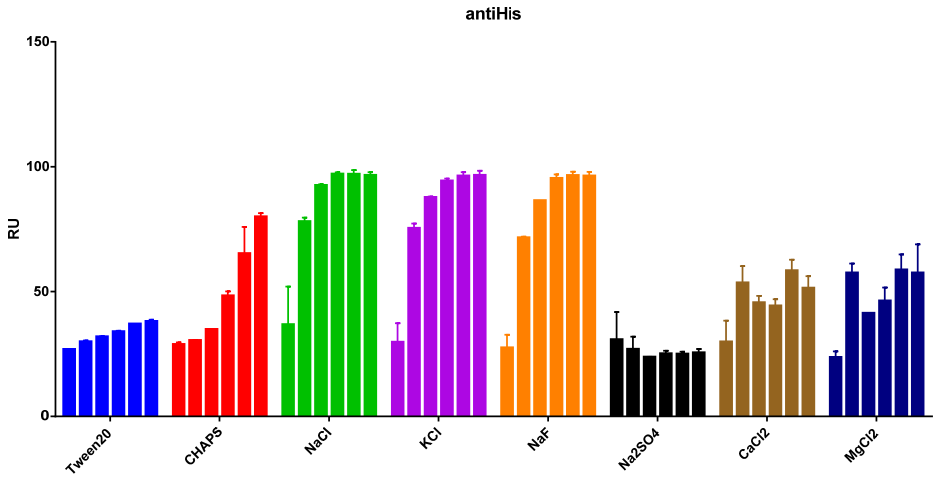


Figure 5-32 Binding response of refolded His_6-Avi-CD151 to a His Ab in 50 mM HEPES pH 7.5 with various buffer additives. Abscissa labels are as shown in Figure 5-31. Measurements were performed in triplicate and results are shown as average ± standard deviation.

Concentrations below 3 mM CHAPS did not improve protein folding when compared to TWEEN®20. However, at 4 mM and 5 mM the binding response increased more than two-fold above that with TWEEN®20. There was no discernible difference in binding with the addition of up to 400 mM NaCl, KCl, NaF or Na_2SO_4. Both $CaCl_2$ and $MgCl_2$ improved binding to the CD151 Ab at 0.5 mM (Figure 5-31). Protein solubility, as measured by the amount of binding to the His Ab, was also improved with the addition of 4 or 5 mM CHAPS. NaCl, KCl and NaF all behaved the same, with an increase in binding response at 50, 100 and 200 mM but no increase above 200 mM. Na_2SO_4 did not affect protein solubility and $CaCl_2$ and $MgCl_2$ doubled the binding response at 0.5 mM (Figure 5-32). These results suggest that it is the negative chloride and fluoride ions that are interacting with the protein. Analysis by CD requires the sample to be free of chloride ions and this result confirms that substitution with fluoride will not change the protein conformation or solubility, which is important information for downstream characterisation.

The CMC of CHAPS decreases with increasing salt concentration [334] so a further refolding screen was carried out, as above (page 139), to refine NaCl, CHAPS, $CaCl_2$ and $MgCl_2$ concentrations.

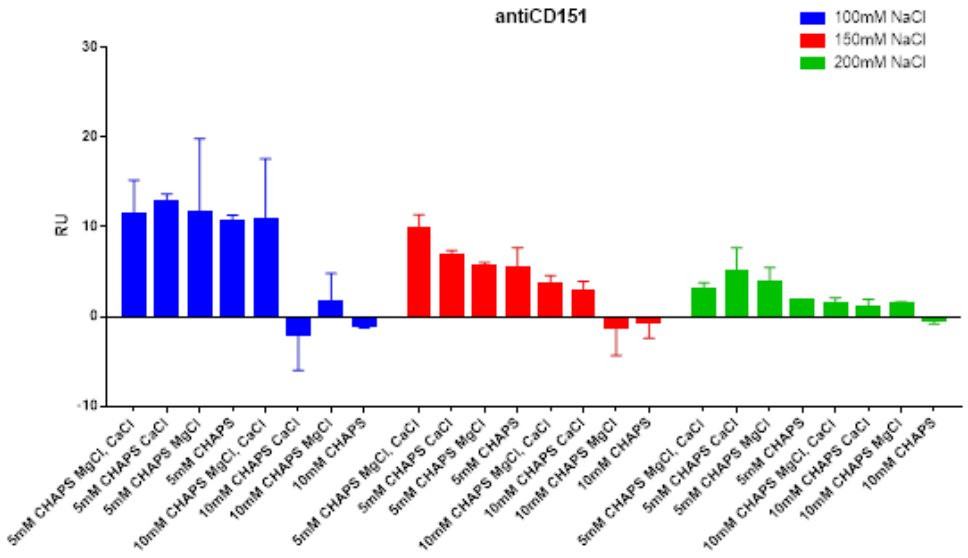


Figure 5-33 Binding response of refolded His_6-Avi-CD151 to a CD151 Ab in 50 mM HEPES pH 7.5 with varying CHAPS, NaCl, $MgCl_2$ and $CaCl_2$ concentrations. Measurements were performed in triplicate and results are shown as average ± standard deviation.

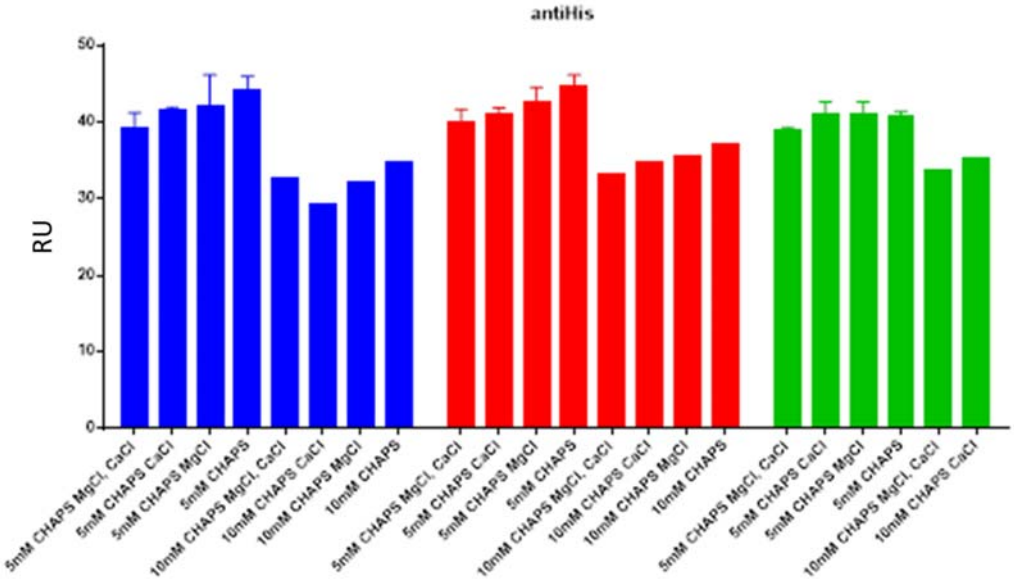


Figure 5-34 Binding response of refolded His_6-Avi-CD151 to a His Ab in 50 mM HEPES pH 7.5 with varying CHAPS, NaCl, MgCl2 and CaCl2 concentrations. Measurements were performed in triplicate and results are shown as average ± standard deviation.

Correct protein folding, as measured by binding to the CD151 Ab was best in 100 mM NaCl. The inclusion of 10 mM CHAPS was detrimental to protein folding with no binding being measured at this concentration in any salt concentration. The inclusion of $MgCl_2$ and $CaCl_2$ with the 10 mM CHAPS improved binding but only as much as with 5 mM CHAPS, so the increase in CHAPS to above CMC is not required for protein folding. $MgCl_2$ was slightly better at refolding the protein than $CaCl_2$ (Figure 5-33).

Relative binding to the anti-His antibody indicated that 5 mM CHAPS improved protein solubility more than 10 mM CHAPS and $MgCl_2$ was again slightly advantageous. No difference in solubility with varying NaCl concentration was observed, as previously (Figure 5-34).

From these results I chose the following buffer composition: 50 mM HEPES pH 7.5, 100 mM NaCl, 0.5 mM $MgCl_2$, 5 mM CHAPS (CHAPS buffer).

5.3.2.4. Protein refolding method selection

5.3.2.4.1. Dilution

The dilution method for refolding proteins is widely used. Denatured proteins are diluted by >50 fold into a non-denaturing buffer, usually containing redox reagents such as glutathione to allow native disulfide bonds to form. [335] Given that the CD151 Ab was reactive to the insoluble and the refolded protein it can be assumed that the disulfide bonds are already formed in the cytoplasm of SHuffle cells. As such, I did not include any redox reagents in the refold buffer screening. For the refold buffer screen I used a 50-fold dilution, which should be adequate to induce protein refolding; however, there was evidence of aggregation as some wells contained precipitate. Aggregation can be a function of concentration so the dilution of the denatured protein into the refold buffer needs to be large enough to reduce protein concentration sufficiently to avoid intermolecular aggregation. Due to the diffusion coefficient of the denaturant being much larger than that of the protein, the denaturant diffuses faster than the protein and the denatured protein can aggregate at the point of dilution. [336] Also due to the expense of CHAPS and the lack of means to concentrate large volumes, I wanted to find a manageable refold dilution that discouraged aggregation.

Four different dilution volumes were tested: 5, 10, 25 and 50 times at both room temperature (RT) and at 4^{o}C. 20, 40, 100 and 200 µl of denatured protein in 8 M urea in CHAPS buffer was slowly dropped into 1 ml of CHAPS buffer and incubated overnight at either RT or 4^{o}C with rocking. After overnight incubation the samples were centrifuged at 13,000 x g for 5 minutes and the soluble fraction transferred to a new tube. The protein content of the soluble material was estimated by A280 nm absorbance and all samples were diluted to 150 µg/ml for testing by SPR. (Figure 5-36)

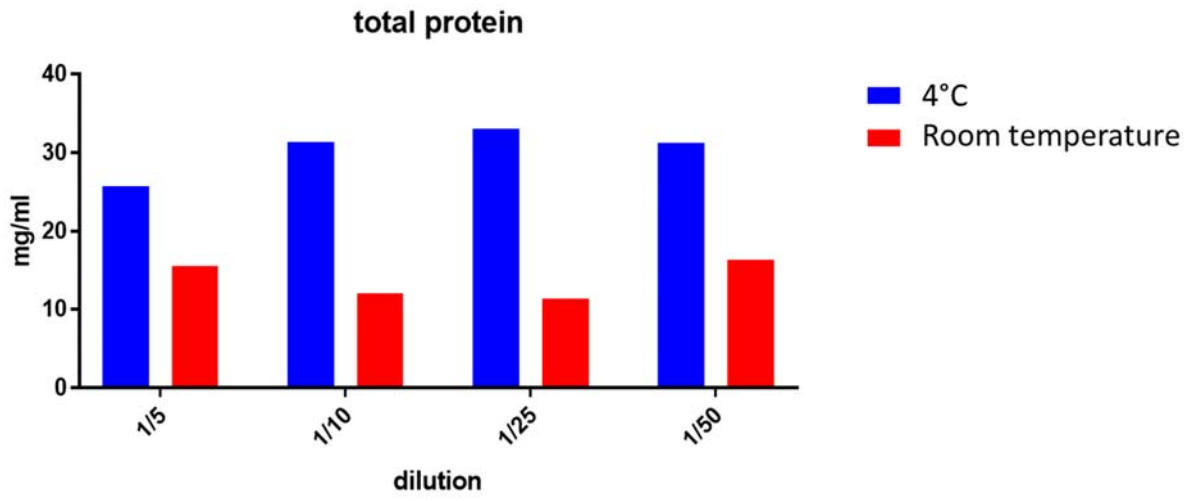


Figure 5-35 Total amount of soluble protein in the different refold conditions, as estimated by A280 nm. Protein concentration is corrected for dilution.

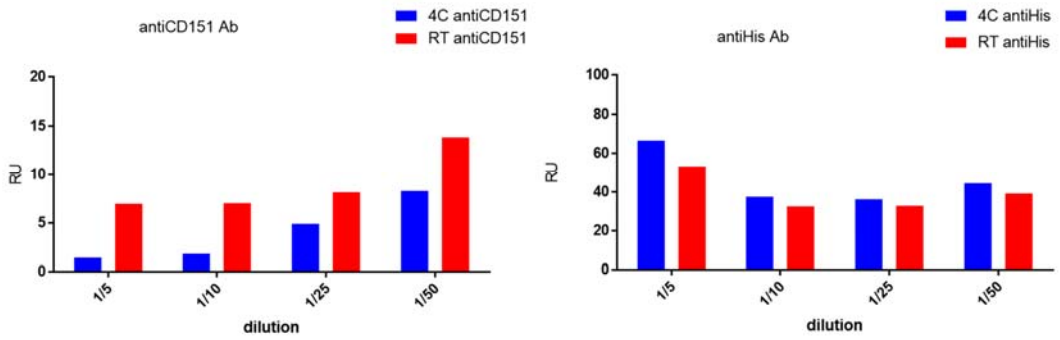


Figure 5-36 Comparison of different dilutions and temperature for refolding denatured His_6-Avi-CD151 protein. Protein concentrations were adjusted to 150 µg/ml for the binding experiments. Binding to the anti-CD151 Ab represents amount of correctly folded protein, binding to the anti-His Ab represents total soluble protein.

The refold at 4^{o}C resulted in an average of two-fold greater amounts of total protein in solution, with little difference between dilutions (Figure 5-35). The difference due to temperature may be a result of slower kinetics of the diffusion at a lower temperature discouraging aggregation or may simply be due to protein degradation in the warmer conditions. Conversely, the amount of protein binding to the anti-CD151 Ab was greater for the RT samples, with close to a two-fold increase in the amount of correctly folded

protein in the 1/50 dilution sample (Figure 5-36). Although the amount of protein binding to the anti-His Ab which represents total soluble His-tagged protein is greater in the 4°C samples, as per the total protein amounts, the amount of binding of the 1/5 dilution sample is highest. This may be due to protein aggregation at the lower dilution, resulting in oligomers binding to the Ab which would generate a greater binding response than monomer.

The soluble fractions of the refolds were normalised to 0.3 mg/ml and the urea solubilised precipitate was diluted 25-fold and samples analysed by SDS-PAGE (Figure 5-37).

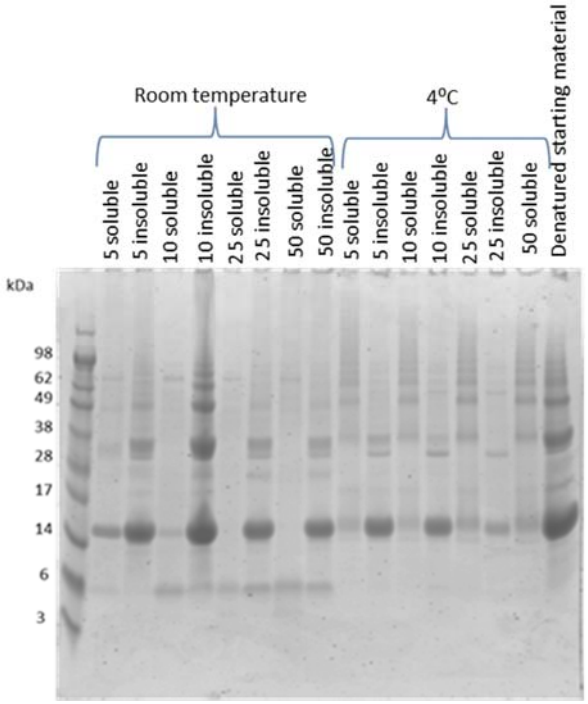


Figure 5-37 SDS-PAGE analysis of the soluble and insoluble fractions of refolded His_6-Avi-CD151 at different dilutions and temperatures.

As anticipated, the protein refolded at RT contained more aggregate and breakdown product than the protein refolded at 4°C, and for the 1/25 and 1/50 dilutions did not contain any protein at the expected size. This is somewhat disconcerting as the samples bound to the anti-CD151 Ab (Figure 5-36). This result raises the possibility that the Ab may not be entirely specific for correctly folded CD151 or may be able to recognise a 3D epitope that is retained in non-natively folded CD151. The sample of protein refolded at 50-fold dilution did not contain any visible precipitate (sample not included on gel) at 4°C. All samples contained higher Mw protein suggesting that some of the protein

aggregated to form oligomers. Examining the gel for the sample with the least oligomeric protein and breakdown products indicates that the 1/5 dilution at RT was best for refolding, in terms of the correct size of protein generated. However, the proportion of insoluble protein in this condition was also very high.

Overall, it appeared that refolding at 1/5 dilution at RT was the best option. This is not an ideal situation due to the large amounts of starting material required to generate correctly folded protein, and thus this method proved too inefficient to be practical for producing substantial quantities of high-quality protein. To develop a refolding technique more suitable to producing correctly folded protein in sufficient quantities for structural biology and biophysical experiments I developed a method for refolding denatured protein while it was immobilised on an affinity chromatography column.

5.3.2.4.2. On-column refold

The purpose of these experiments was to develop a technique for promoting the correct folding of CD151 LEL that was more amenable to laboratory handling than the dilution refolding approach detailed above. To this end I decided to streamline the refolding process by attempting to refold denatured protein *in situ* during the cation-affinity purification step. [337]

The insoluble fraction of the cell lysate from a 6 L His_6-Avi-CD151 expression in SHuffle cells was solubilised in 100 ml of 8 M urea in CHAPS buffer and loaded onto a 5 ml HisTrap™FF column. The column was washed with 5 column volumes (CV) of 6 M urea in CHAPS buffer (page 139) then slowly washed with a gradient of 0 – 100% CHAPS buffer over 20 CV. The column was then washed with 5 CV of CHAPS buffer before the refolded protein was eluted with 5 CV of 0 – 100% CHAPS buffer + 500 mM imidazole (Figure 5-38).

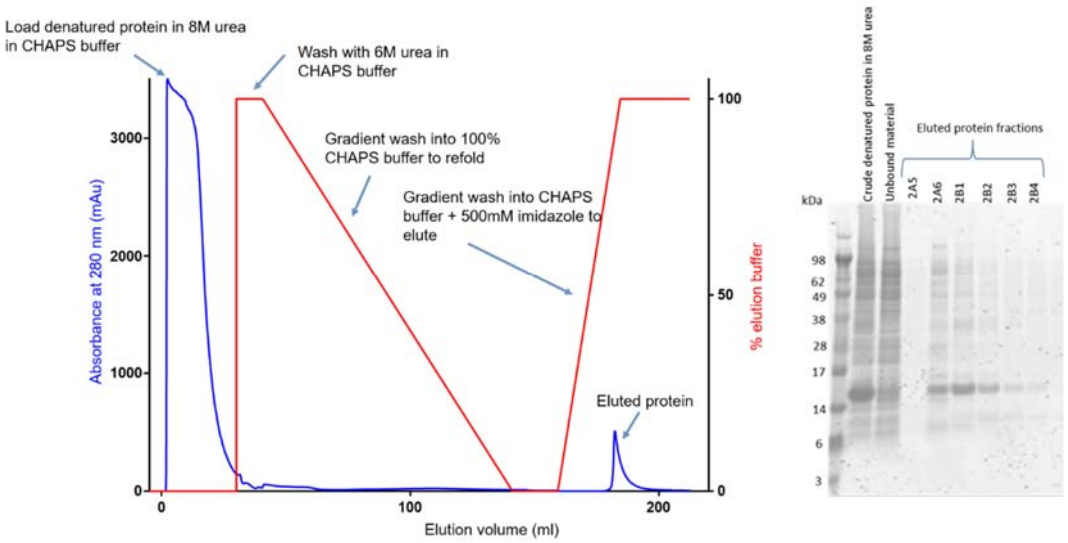


Figure 5-38 On-column refold of denatured His_6-Avi-CD151 into CHAPS buffer. Protein solubilised in 8 M urea in CHAPS buffer was loaded onto a 5 ml HisTrap column and unbound material eluted with 6 M urea in CHAPS buffer. A gradient from 100% 6 M urea in CHAPS buffer to 100% CHAPS buffer over 20 CV was washed over the protein to refold. Refolded protein was eluted with a 10 CV gradient of CHAPS buffer + 500 mM imidazole (left) The blue line represents the flow through and eluted protein absorbance at A280 nm (mAU), the red line represents the % buffer B. Samples from the unbound and eluted material were analysed using SDS-PAGE (right).

Analysis by SDS-PAGE shows the refolded protein to be the correct size (Figure 5-38). The fractions were pooled, desalted into CHAPS buffer to remove the imidazole and stored overnight at 4^{o}C. During refrigeration a large amount of precipitate formed, suggesting that although the protein was solubilising it was still not folding correctly and was unstable. While this method was superior to the dilution method of refolding, as it produced cleaner protein while using much less buffer and with higher yields, it was still problematic in that the protein produced is unstable on storage.

5.3.3. His-CD151

In the previous section I aimed to produce His_6-Avi-CD151 for SPR studies. In this section I aimed to produce His_6-TEV-CD151 protein for structural studies. Quite often His tags are found to inhibit protein crystallisation due to their flexibility so a cleavable His_6 tag was designed to overcome this potential problem.

5.3.3.1. Small-scale expression

His_6-TEV-CD151 LEL in pET30a(+) vector was transformed into SHuffle C3029H Express and SHuffle C3026H cells and a small scale 20 ml culture of three clones of each

transformation grown for a trial protein expression as previously. Fractions from the expression were analysed using SDS-PAGE (Figure 5-39).

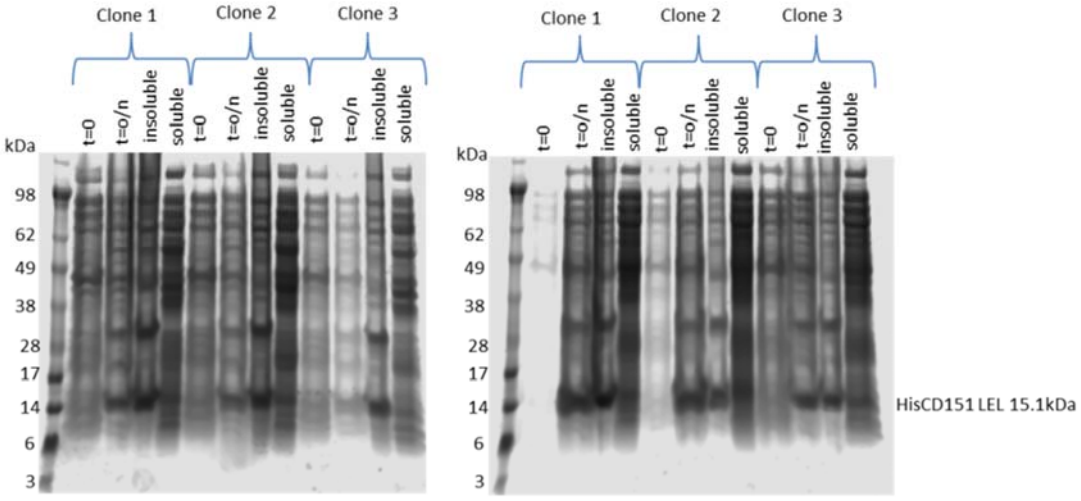


Figure 5-39 SDS-PAGE analysis of His_6-CD151 LEL expression in C3026H express SHuffle cells (left) and in C3029H SHuffle cells (right).

There was no discernible difference in the expression in either cell type or between clones and little material corresponding to heterologously expressed protein was visible in the soluble fraction.

5.3.3.2. Large-scale expression

A larger 2 L culture was grown of C3026H clone 1 and C3029H clone 2. Cells were grown to OD600 nm = 0.7, induced with 0.5 mM IPTG then grown for 4 hours at 30°C before harvesting the cells, resuspending in CHAPS buffer and processing as previously described.

The soluble fraction of cell lysates were purified over a HisTrap™, eluted fractions containing protein were pooled and an analytical fractionation performed by gel filtration using a Superdex 200 10 300 column (Figure 5-40).

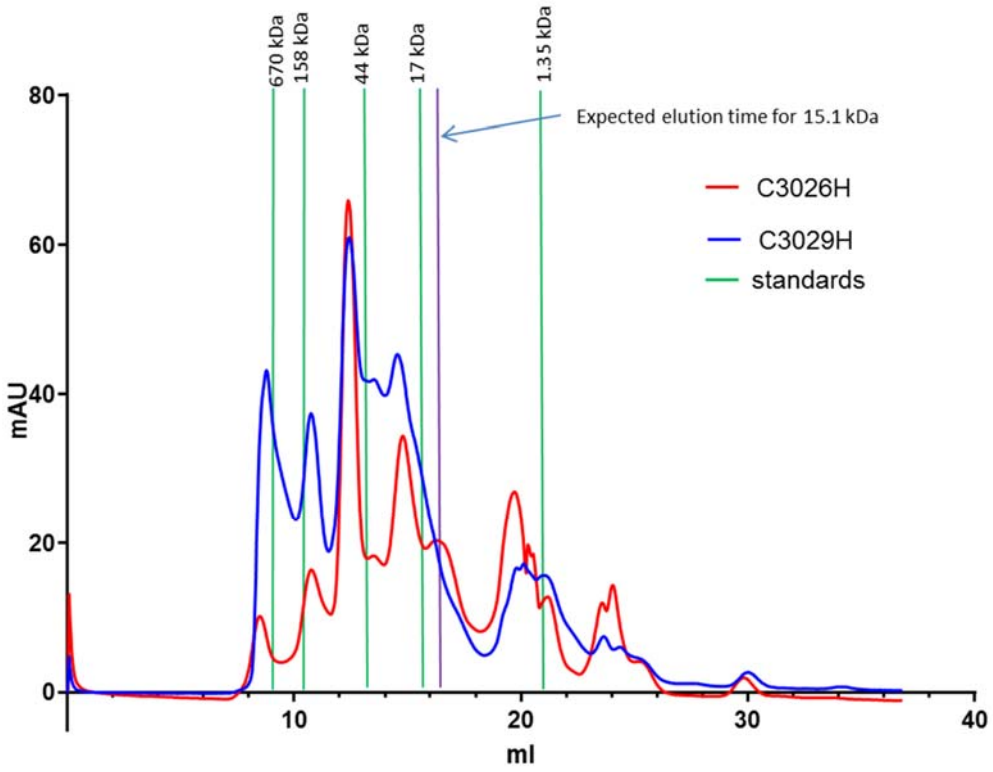


Figure 5-40 Gel filtration profile of His_6-CD151 LEL expressed in SHuffle C3026H (red) and C3029H (blue) cells compared to the elution times of Mw standards (green). Samples were analysed using a Superdex 200 10 300 increase column.

Gel filtration standards myoglobin (Mw 17 kDa) and vitamin B12 (Mw 1350) eluted at 15.9 and 21.2 ml, respectively. Extrapolating from the His_6-CD151 LEL, at 15.1 kDa, should elute at 16.7 ml. There is a protein peak at the expected elution time for a 15.1 kDa protein in the C3026H material but it is not the most abundant species and the peak was not prominent. (Figure 5-40) Pooled fractions from HisTrap purification of C3026H were further purified by anion exchange and fractions examined by SDS-PAGE (Figure 5-41).

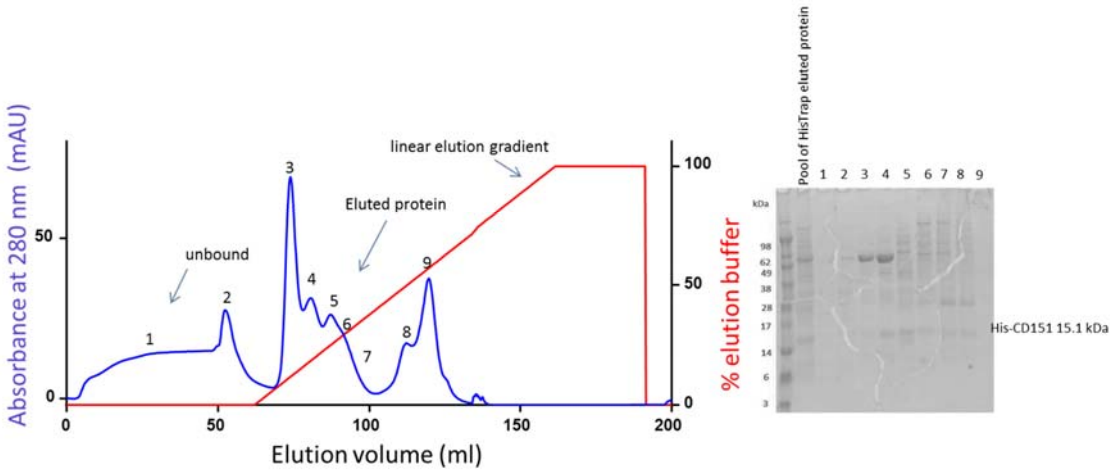


Figure 5-41 Anion exchange of HisTrap purified cell lysate from SHuffle cells C3026H. The blue line represents the eluted protein absorbance at A280 nm (mAU), corresponding to the left ordinate axis, the red line corresponds to the right ordinate axis. SDS-PAGE gel showing corresponding fractions from the purification. PAGE analysis showed that all eluted fractions appeared to contain His_6-CD151 LEL (right panel).

Purification by anion exchange was unable to separate the His_6-CD151 LEL from the contaminating proteins (Figure 5-41).

The insoluble material from the expression using SHuffle C3029H cells was solubilised in 8 M urea in CHAPS buffer, refolded on the HisTrap column, as previously described, and eluted with CHAPS buffer + 500 mM imidazole. Eluted fractions containing protein were analysed by SDS-PAGE (Figure 5-42).

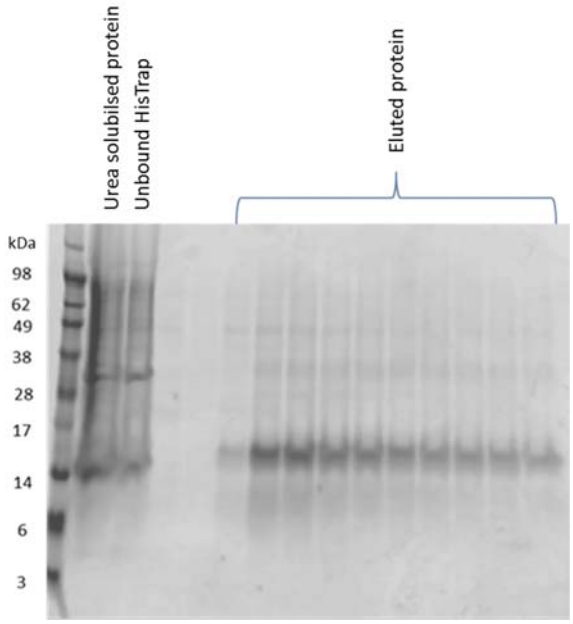


Figure 5-42 SDS-PAGE analysis of the fractions eluted from an on-column refold of His_6-CD151 LEL.

The eluted fractions were pooled and contained 3.7 mg of protein, by A280 nm estimation, in 27 ml total volume. The protein was concentrated to 4 ml using a 6 kDa cut-off spin filter, and during concentration a visible precipitate formed. After centrifugation the final soluble protein concentration was 0.5 mg/ml in 4ml. A sample of the protein was analysed by CD (Figure 5-43). The experimental plot shows a good fit to the theoretical data, with some deviation at lower wavelengths.

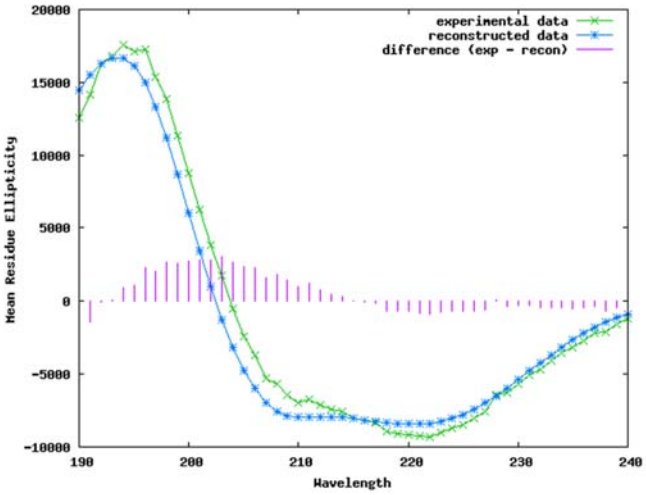


Figure 5-43 Secondary structure analysis from the CD spectra of His_6-CD151 LEL refolded from the insoluble fraction of SHuffle C3029H expression. Experimental data are plotted in green; the calculated spectrum derived from the predicted secondary structure is plotted in blue and the difference spectra is depicted by vertical lines in pink.

Table 13 Summary of deconvoluted CD data listing the secondary structure composition of His_6-CD151 LEL showing total helix content.

Result	Helix1	Helix2	Strand1	Strand2	Turns	Unordered	Total
1	0.000	1.000	0.000	0.000	0.000	0.000	1
2	0.564	0.436	0.000	0.000	0.000	0.000	1

The CD spectra of the refolded His_6-CD151 LEL showed defined secondary structure, predominantly consisting of α-helical elements, which was more consistent with the modelled structure of CD151 LEL (Figure 4-6) than previous results (Figure 5-18,Table 12).

5.3.4. His_6-CD151 C184S C192S

Of the 34 tetraspanins found in mammals, the only published structural data is for CD81 – the crystallographic extracellular domain structure was published in 2001 [200] and the full-length structure in 2016. [201] The extracellular domain of sm-TSP-2, a tetraspanin found in the blood fluke *Schistosoma mansoni* was also published in 2014. [310] Both CD81 and sm-TSP-2 contain four cysteine residues in the LEL, which form two disulfide bonds and are conserved across all tetraspanins. [195, 291] CD151 contains an extra two cysteines which are hypothesised to form a third disulfide bond (Figure 5-44).

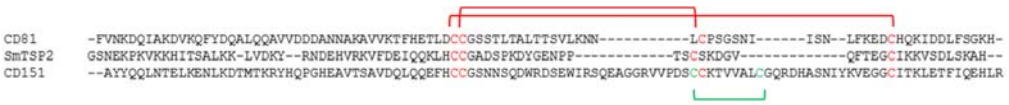


Figure 5-44 Sequence alignment of the large extracellular loop of CD81, sm-TSP-2 and CD151 showing the four conserved cysteines and their respective disulfide bonds in red and the additional two CD151 cysteines and disulfide bond in green.

Published structural studies of CD81 and sm-TSP-2 LELs used protein produced in *E. coli* [308, 310, 338] without any need for refolding or the use of bacteria with a modified cytosol. This suggests that it may be the additional cysteines in the CD151 LEL that are misfolding and causing the oligomerisation observed during protein purification. To test this hypothesis, I designed a construct with the cysteines at residue 184 and 192 of CD151 LEL mutated to serine.

5.3.4.1. Small-scale expression

His_6-TEV-CD151 LEL C184S C192S cDNA in the pET30a(+) vector was transformed into SHuffle T7 Express cells alongside the His_6-TEV-CD151 LEL wild-type and a small scale 100 ml culture of each transformation grown for a trial protein expression as described previously. The cells were resuspended in lysis buffer and lysed by sonication. Samples were analysed by SDS-PAGE (Figure 5-45).

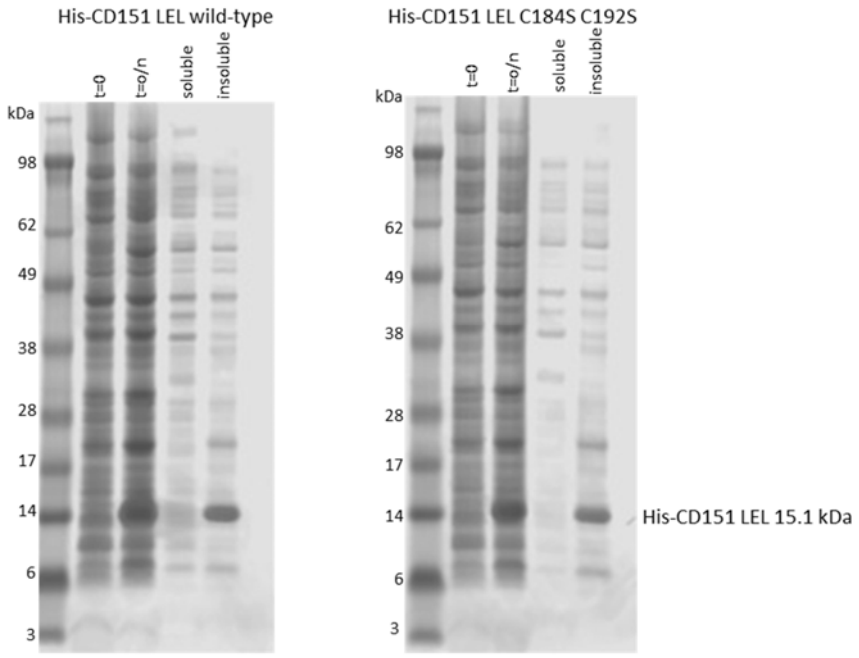


Figure 5-45 SDS-PAGE analysis of the expression and purification of His_6-CD151 wild-type (left) and C184S C192S mutant in SHuffle T7 Express cells. Cells were induced at OD600 = 0.6 and grown at 16^{o}C overnight.

There was no discernible difference in the expression levels of protein and no visible band of protein in the soluble fraction of the C184S C192S mutant. This suggests that, even with only two disulfide bonds, the protein is still not able to fold. Alternatively, it may also demonstrate the structural importance of the third disulfide. The LEL of CD151 is 20 amino acids longer than that of CD81 and 29 amino acids longer than the LEL of sm-TSP-2. Without the third disulfide in the CD151 LEL the protein may be too conformationally flexible to successfully form the correctly folded species. Kazarov et al. (2002) and Yauch et al. (2000), while investigating the interaction site of CD151 – α3 integrin binding, mutated C192 to a tyrosine. Both studies found the C192Y mutation caused loss of α3 integrin binding. [171, 173] Collectively, these studies suggest the third disulfide bond in CD151 is structurally as well as functionally important.

5.3.1. TrxA-His$_6$-CD151

In wild-type *E. coli* cells disulfide bond formation occurs in the periplasm. The cytoplasm of *E. coli* is not favourable for the formation of disulfide bonds due to the presence of reductases and reducing agents such as glutathione (Grx1, Grx2, Grx3) and thioredoxin (Trx1, Trx2). [339] Conversely, in the periplasm a set of Dsb (disulfide bond isomerase) proteins catalyse disulfide bond formation. DsbA is a potent oxidase which catalyses the formation of disulfide bonds. [340] Once DsbA has donated its disulfide bond to the folding protein it becomes reduced and must be re-oxidised back to its active state by DsbB. [341] DsbA is promiscuous and tends to oxidise cysteines in a manner which can lead to proteins becoming misfolded. [342] The isomerase DsbC refolds proteins to their native disulfide-bonded state. [343] DsbD is a transmembrane protein responsible for maintaining DsbC in its reduced state. DsbD owes its reducing power to transfer of electrons from cytoplasmic thioredoxin which, in turn, receives electrons from the cytoplasmic pool of reduced nicotinamide adenine dinucleotide phosphate (NADPH) (Figure 5-46). [344]

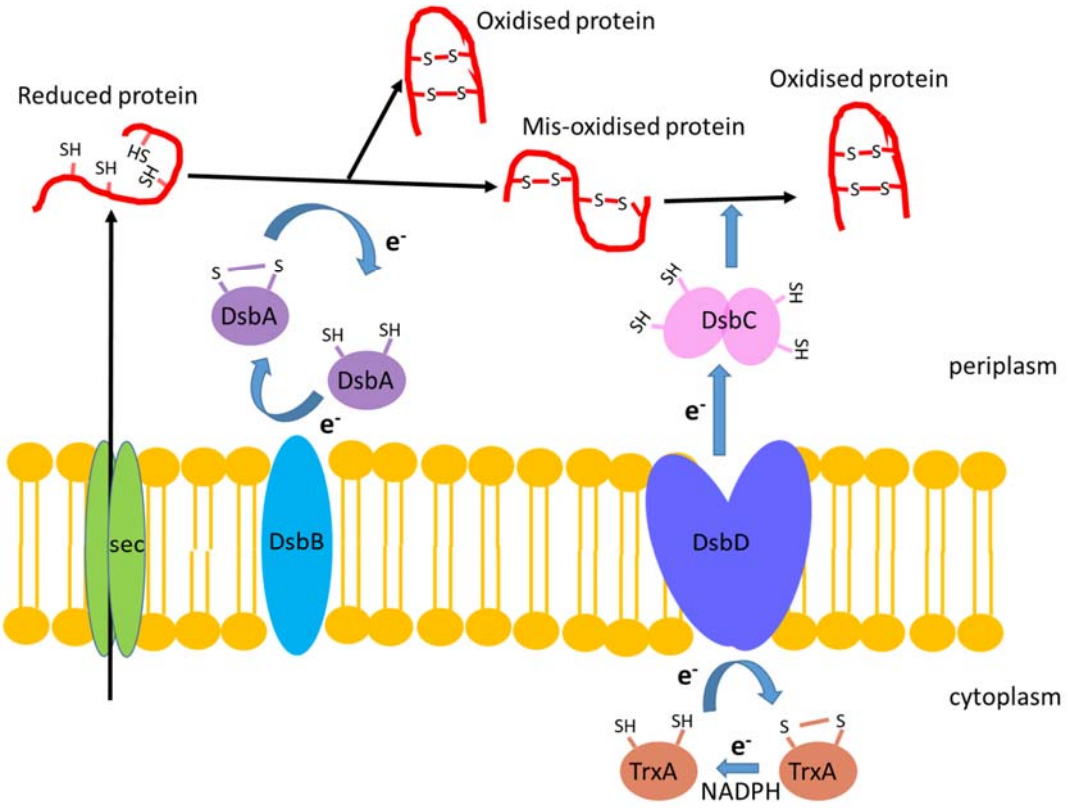


Figure 5-46 Periplasmic disulfide bond formation in gram negative bacteria. The reduced protein (red) is transported to the periplasm, usually via the sec pathway (green), where DsbA (purple) oxidises cysteine residues. DsbA is re-oxidised by the transmembrane protein DsbB (light blue). Mis-oxidised protein is isomerised to its native state by DsbC (pink) which is maintained in its reduced state by DsbD (blue). DsbD receives electrons from cytoplasmic thioredoxin (TrxA, brown), which receives electrons from cytoplasmic NADPH.

Cytoplasmic disulfide bond formation in SHuffle cells utilises a different pathway to periplasmic disulfide bond formation. SHuffle cells are mutant *E. coli* cells carrying deletions of thioredoxin reductase (*trxB*) and glutathione reductase (*gor*); these mutations are nonviable as ribonucleotide reductase (RNR), an essential protein, cannot be reduced to its active state. [345] The strain has been further modified to include mutant peroxidase AhpC, which has the ability to reduce Grx1 and restore some reducing power to the cell. [346] The thioredoxins remain in their oxidised state and can oxidise the folding protein's disulfide bonds. [347] Thioredoxins, like DsbA, form disulfide bonds indiscriminately. The SHuffle strain is still further engineered to express DsbC in the cytoplasm which should isomerise mis-oxidised proteins to their native state. [348]

In the periplasm electrons are transported from cytoplasmic thioredoxin via DsbD to periplasmic DsbC (Figure 5-46). In SHuffle cells cytoplasmic DsbC receives electrons directly from the oxidised thioredoxins. As I have observed mis-folded protein produced by the SHuffle cells I hypothesised that the limited reducing power of the SHuffle cytoplasm may be inhibiting the redox dependant isomerase activity of DsbC. I considered that including thioredoxin as a fusion partner with the CD151 LEL may be beneficial for the folding of the LEL. Thioredoxin is highly soluble and is known to confer solubility to otherwise insoluble proteins, especially when fused to the N-terminus. [349, 350] To exploit this property the CD151 LEL cDNA was cloned into the MscI and NotI sites of pET32a+ vector (Novagen) which encodes an N-terminal thioredoxin tag (TrxA) followed by a His_6 tag to facilitate purification, and a TEV cleavage site for tag removal; resulting in a fusion protein of 27,080 Da.

TrxA-His_6-CD151 was transformed into SHuffle T7 Express cells and a small scale 200 ml culture was grown for a trial protein expression as described previously. Harvested cells were processed as described previously, and the soluble fraction of the cell lysate purified over a 5ml HisTrap™. Samples taken during expression and purification were analysed by SDS-PAGE (Figure 5-47).

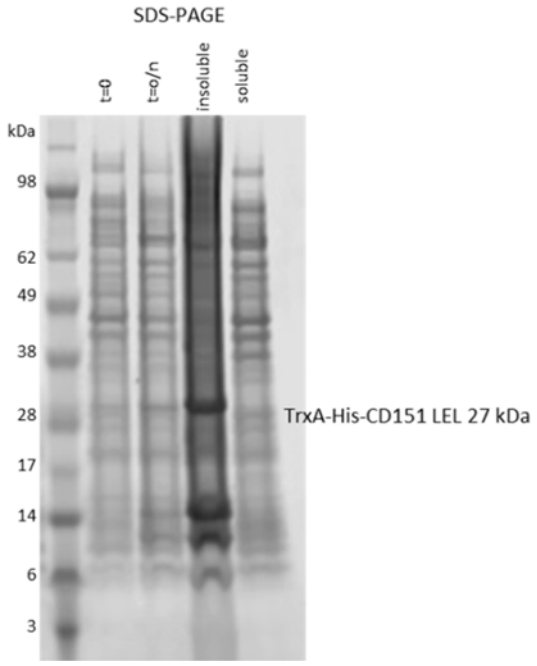


Figure 5-47 SDS-PAGE analysis of TrxA-His_6-CD151 LEL expression in SHuffle T7 Express cells.

A protein band running slightly above the 28 kDa marker may be the post-induction protein band but there was also a faint band at the same size in the pre-induction material. There was a band appearing in the post-induction sample at approximately 14 kDa, which is much smaller than the expected 27 kDa, however it is known that TrxA forms a very compact structure [351], as does CD151, so it is possible that the fusion protein migrated faster than anticipated on the gel. The post-induction band at 14 kDa corresponded to a large band of material in the insoluble fraction but there was no comparable protein visible in the soluble fraction. The only distinct protein bands occurring in the soluble material were above 38 kDa, suggesting that if the protein was TrxA-His_6-CD151 it was misfolded and forming oligomers.

5.3.2. His_6-GFP-CD151

Green fluorescent protein (GFP) from *Aequorea victoria* is a 27 kDa protein that exhibits green fluorescence when exposed to blue light. The most common usage of GFP tags is to visualise and monitor protein expression. The GFP protein structure consists of a tightly packed, eleven β-sheet containing barrel shape, with the chromophore in the centre of the barrel, shielded from quenching by aqueous solvent. [352] Correct folding and structure of the barrel are therefore required to maintain fluorescence. As such GFP can be considered to be a folding reporter and it has been demonstrated that GFP fluorescence can indicate proper folding of the GFP's fusion partner. [353] I considered that using a GFP tag may be advantageous when expressing recombinant CD151LEL to simplify the monitoring of protein folding and to streamline the purification process by eliminating incorrectly folded material earlier in the process.

The coding sequence for residues 1–230 of yEGFP (GenBank accession No. U73901)[354], a GFP with S65G and S72A mutations to enhance fluorescence (excitation maximum at 490 nm and emission maximum at 517 nm) and codon-optimised for bacterial expression, was cloned into the NdeI–BamHI sites of pET28a (Novagen) which encodes an N-terminal His_6-thrombin purification tag for expression in *E coli*. [355, 356] Subcloning of CD151 LEL cDNA using the BamHI restriction site resulted in a two residue linker (Gly-Ser) between yEGFP(1–230) and CD151 for a final Mw of 40,267 Da.

His_6-GFP-CD151 cDNA in pET28a vector was transformed into SHuffle T7 Express cells and small scale 200 ml culture was grown for a trial protein expression as previously described. The harvested cells were processed as previously described, and the soluble fraction of the cell lysate purified over a 5 ml HisTrap™.

Samples taken during expression and purification were analysed by SDS-PAGE. (Figure 5-48)

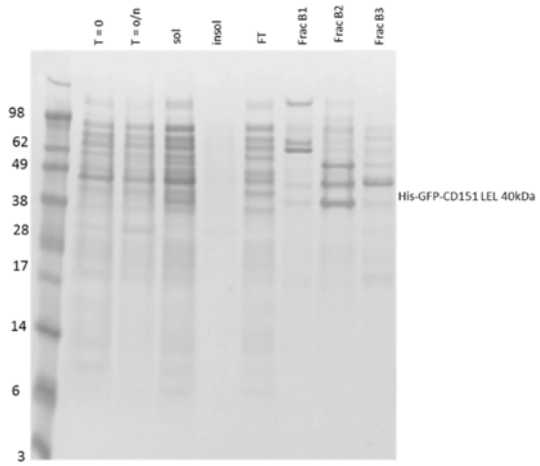


Figure 5-48 SDS-PAGE analysis of His_6-GFP-CD151 expression and purification.

Although there wasn't a clear protein band post-induction, there did seem to be a faint band at the right size in the soluble material which also appeared in the material eluted from the HisTrap (fraction B2). This fraction was concentrated and purified by gel filtration using a Superdex™ 200 10 300 column. As correctly folded GFP absorbs light at 488 nm, by monitoring the gel filtration purification at A280 and A488 it was possible to identify eluted material that contained the correctly folded fusion protein (Figure 5-49).

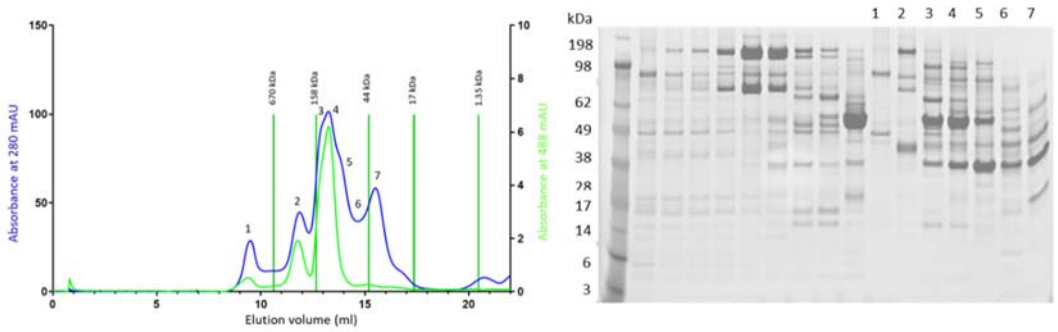


Figure 5-49 Gel filtration purification of His_6-GFP-CD151 fusion protein on a Superdex 200 10 300 column. (left) Protein and GFP was measured by absorbance at 280 (blue line, left Y-axis) and 488 (green line, right Y-axis) mAU, respectively. Fractions from each numbered peak were analysed by SDS-PAGE (right), lanes on the gel corresponding to fractions of the gel filtration elution are indicated by numbering. (Unlabelled lanes are from an unrelated experiment) The column was previously calibrated with Mw standards, the elution times and Mws are indicated by the dark green lines.

Fractions 6 and 7 appeared to contain the correct size protein by SDS-PAGE analysis (Figure 5-49, right); however, there was no corresponding peak of absorbance at 488 mAU in the chromatogram (Figure 5-49, left). GFP will not absorb light if the protein is not correctly folded, suggesting the monomer in factions 6 and 7 is misfolded. Contrary to this, the greatest absorption at 488 mAU was in a protein peak corresponding to a much higher Mw than expected. This suggests that the CD151 LEL was misfolded and oligomerising through the cysteines, however, the GFP was folded correctly and sensitive to light at 488 nm. Thus, this method of protein production proved to be unsatisfactory.

5.3.3. Periplasmic expression

As discussed above, disulfide bond formation in *E. coli* usually occurs in the periplasm (section 5.3.1). Translocation to the periplasm is dependent on fusing a suitable leader peptide that directs the unfolded protein into the periplasm via the Sec (post-translationally) or the SRP (co-translationally) systems. [357, 358] CD151 LEL cDNA with an N-terminal His_6-TEV tag was cloned into a pET-22b(+) resulting in a fusion protein with an N-terminal *pelB* signal sequence to direct the protein through the Sec pathway.

The vector was transformed into Lemo21(DE3) and BL21(DE3) cells. In Lemo21(DE3) cells T7 RNA polymerase activity can be modulated by T7 lysozyme, which is expressed from the rhamnose promotor. [359] Translocation of the unfolded protein through the Sec pathway to the periplasm is potentially a rate limiting step and can result in the

accumulation of protein in the cytoplasm. Lemo21(DE3) cells allow the expression level to be tuned so that protein expression can be optimised by adding inhibiting L-rhamnose to the expression culture.

A clone of each cell type was grown overnight in LB with ampicillin selection at 30°C. Eight 30 ml flasks of Lemo21(DE3) cells were grown at 37°C in media containing either 0, 250, 750 or 2000 µM of L-rhamnose, until OD_{600} nm reached 0.6 and expression was induced with 0.4 mM IPTG. Cells in four of the flasks were grown at 37°C for an additional 4 hours then harvested; the remaining four flasks were incubated at 16°C overnight before harvesting. A 30 ml flask of the BL21(DE3) was incubated at either 37°C or 16°C as above for comparison.

Cells were harvested by spinning at 3000 x g for 20 minutes, the supernatant was removed, and the cells were gently resuspended in a hypertonic solution of 200 mM Tris, 500 mM sucrose, 1 mM EDTA pH 8 and incubated on ice for 30 minutes. In hypertonic solution the cell contracts, osmotically sensitising the inner membrane and separating it from the cell wall. [360] The cells were then centrifuged at 16000 x g for 20 minutes. The cell pellet was resuspended in a hypotonic solution of 5 mM $MgSO_4$ to osmotically shock the cells, incubated on ice for 30 minutes then centrifuged at 16000 x g for 20 minutes. This supernatant is the hypotonic extract and should contain the periplasmic proteins. The remaining pelleted material was solubilised in 8 M urea. Samples from each fraction were analysed by SDS-PAGE (Figure 5-50, Figure 5-51).

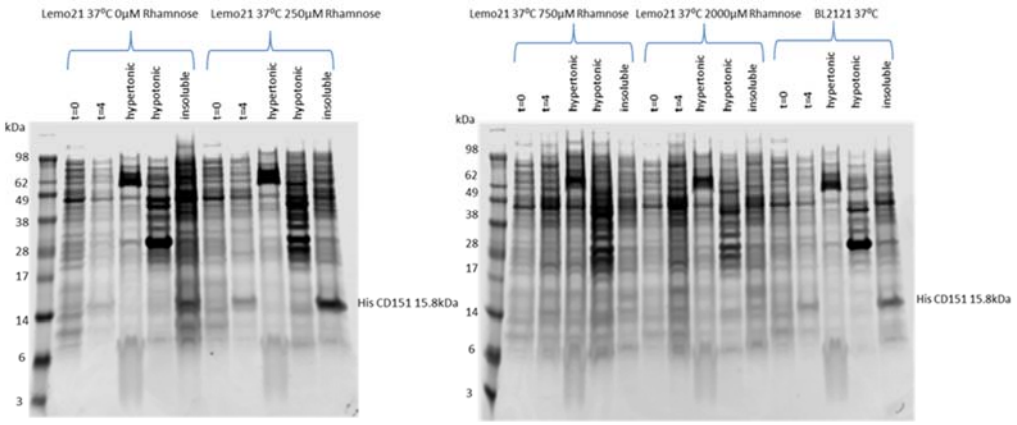


Figure 5-50 Periplasmic expression of His_6-CD151 at 37°C in Lemo21(DE3) cells with L-rhamnose tuning 0 and 250 µM (left) and 750 and 2000 µM (right). BL21(DE3) (far right) included for comparison.

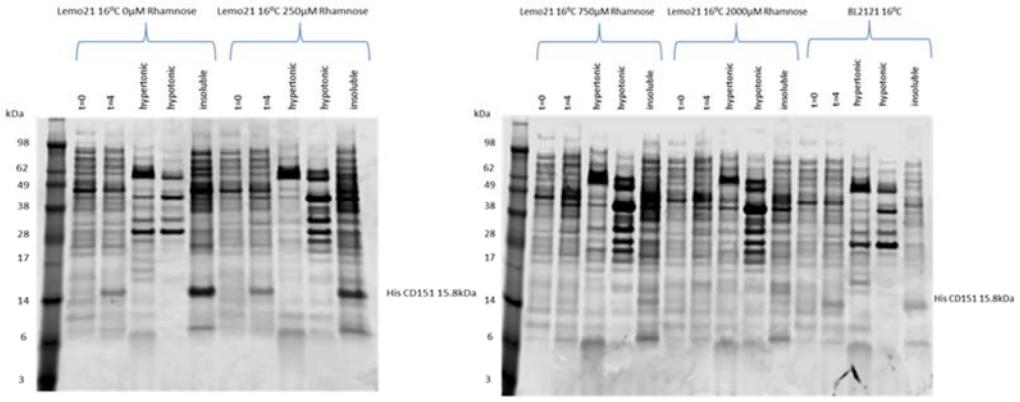


Figure 5-51 Periplasmic expression of His_6-CD151 at 16°C in Lemo21(DE3) cells with L-rhamnose tuning 0 and 250 µM (left) and 750 and 2000 µM (right). BL21(DE3) (far right) included for comparison.

SDS-PAGE analysis reveals protein bands at the correct size post-induction in Lemo21(DE3) cells grown at 37°C at all concentrations of L-rhamnose, with decreasing band intensity as L-rhamnose concentration increases, as expected. BL21(DE3) at 37°C also showed a clear post-induction band of protein; however in all cases protein was only present in the insoluble 8 M urea fraction (Figure 5-50).

At 16°C there were distinct bands of protein post-induction at 0 and 200 µM L-rhamnose and in the BL21(DE3) cells and only faint bands at 0.75 and 2 M L-rhamnose (Figure 5-51).

Western blot analysis of the hypertonic and hypotonic fractions of the Lemo21(DE3) expressions probed with anti-CD151 Ab and anti-His Ab revealed distinct bands of reactivity to the anti-His Ab in the hypertonic fractions of the 16°C expression with 0 and 250 µM L-rhamnose (Figure 5-52). Faint bands were also visible in the 37°C expression hypertonic fractions at 250 and 750 µM L-rhamnose. Banding was also present at higher Mws suggesting misfolded protein was forming dimers and trimers. Probing with the anti-CD151 Ab did not reveal any reactivity. The lack of reactivity to the CD151 conformational Ab and the presence of the higher Mw products was discouraging as it suggested that the protein was not correctly folding and non-specifically aggregating to form oligomers. The presence of a large amount of the fusion protein in the insoluble fractions (Figure 5-51) suggested that the protein was not efficiently translocated from the cytoplasm, possibly due to overwhelming the sec pathway machinery with large quantities of heterologously expressed protein. These results indicated that periplasmic export of the CD151 LEL, in order to promote formation of structurally indispensable disulfide bonds, was not a viable alternative for production of correctly folded CD151 LEL protein.

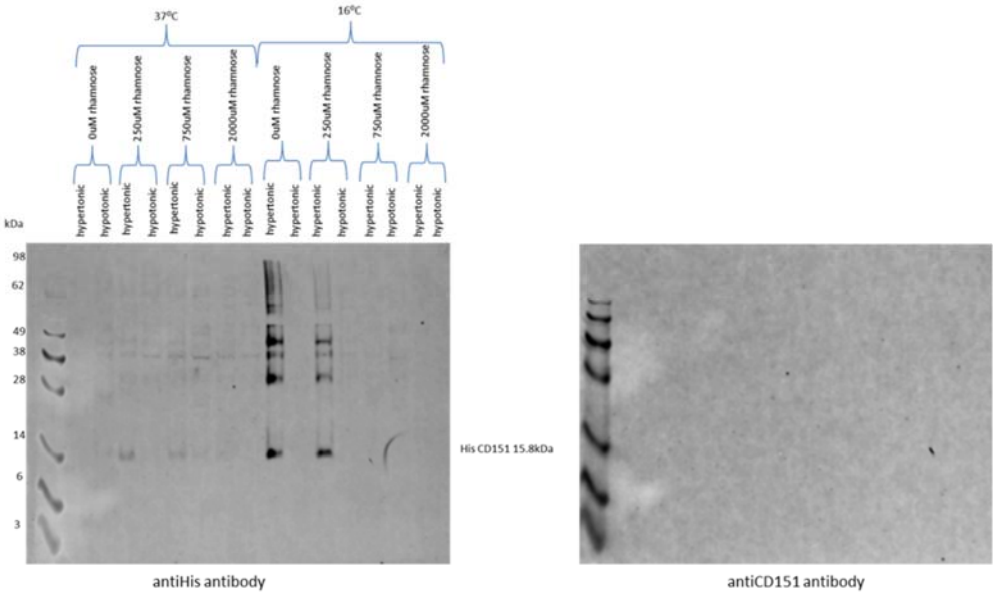


Figure 5-52 Western blot analysis of hypertonic and hypotonic supernatant fractions derived from Lemo21 (DE3) periplasmic expression of His_6-CD151 LEL at 37°C and 16°C. Blot was probed with anti-CD151 (right) and anti-His Abs (left).

5.3.4. His_8-MBP-TEV-CD151 LEL

5.3.4.1. Expression in BL21(DE3) cells

The *E. coli* maltose-binding protein (MBP) is known to enhance the solubility and folding of its fusion partners and is a popular choice for the production of recombinant proteins in prokaryotic expression systems. [361-363] MBP fusion also facilitates affinity purification on amylose resin to a high degree of purity. [364] For this reason I decided to examine the possibility of purifying an N-terminal MBP fusion of CD151 LEL; however, amylose resin can have variable binding efficiencies and co-purification of endogenous *E. coli* MBP can be problematic. [365] To pre-empt possible purification problems the construct was designed with a His_8 tag for purification and a TEV cleavage site for cleavage of the His_8-MBP tag.

The pET30a vector containing His_8-MBP-CD151 LEL cDNA was transformed into BL21(DE3) cells and grown in LB media with kanamycin selection. Glucose was added to a final concentration of 0.2% to suppress amylase expression. [366] A trial 2 L expression was carried out at 37°C: cells induced with 0.1 mM IPTG at OD_{600} nm = 0.7 and then grown for 4 hours before harvesting and processing the cellular lysate as described

previously. Samples of the expression and lysate were analysed by SDS-PAGE. (Figure 5-53)

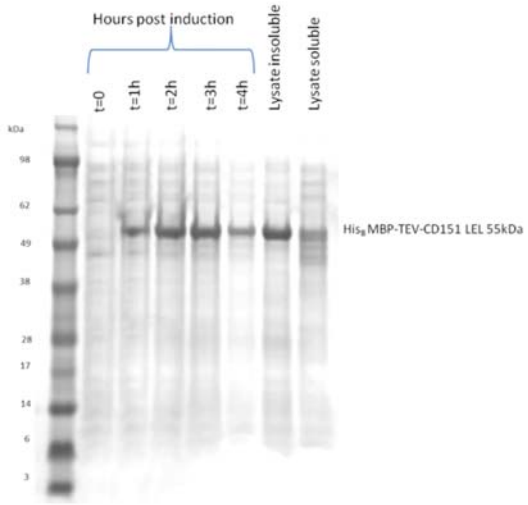


Figure 5-53 Expression of His_8-MBP-TEV-CD151 LEL in BL21 cells at 37^{o}C.

Although there was a large amount of insoluble protein at the predicted size for the fusion protein, there was also a protein band in the soluble material. (Figure 5-53) The cell lysate was purified over a 5 ml HisTrap™ column and the eluted fractions analysed by SDS-PAGE. (Figure 5-54)

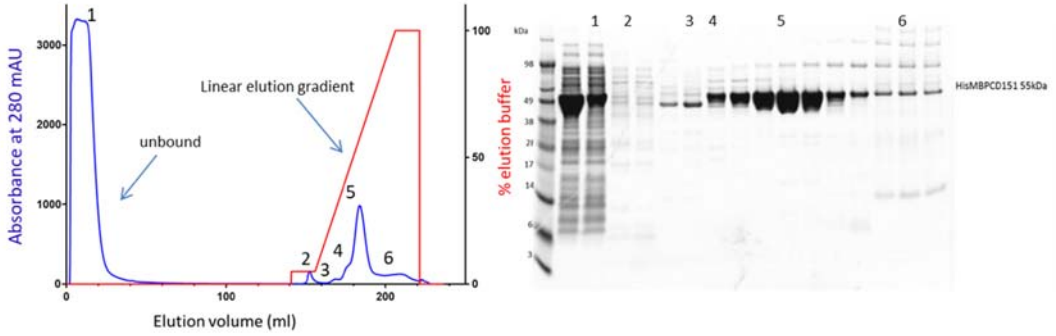


Figure 5-54 Purification of His_8-MBP-CD151 LEL over a HisTrap column (left) and analysis of the fractions by SDS-PAGE. The chromatogram shows the protein absorbance at A280 nm in blue and the percentage of elution buffer in red (left panel). Numbered peaks correspond to the numbered lanes in the SDS-PAGE gel (right panel).

The elution contained a lower Mw protein which appeared slightly earlier than the main protein peak, fraction 3, which may be endogenous MBP (42.5 kDa) as it contains three histidine residues which are likely deprotonated under the purification conditions and possibly interact with the Ni2+ media. Fraction 5 was analysed by gel filtration using a Superdex 200 10 300 column (GE). Eluted fractions were analysed by SDS-PAGE (Figure 5-55).

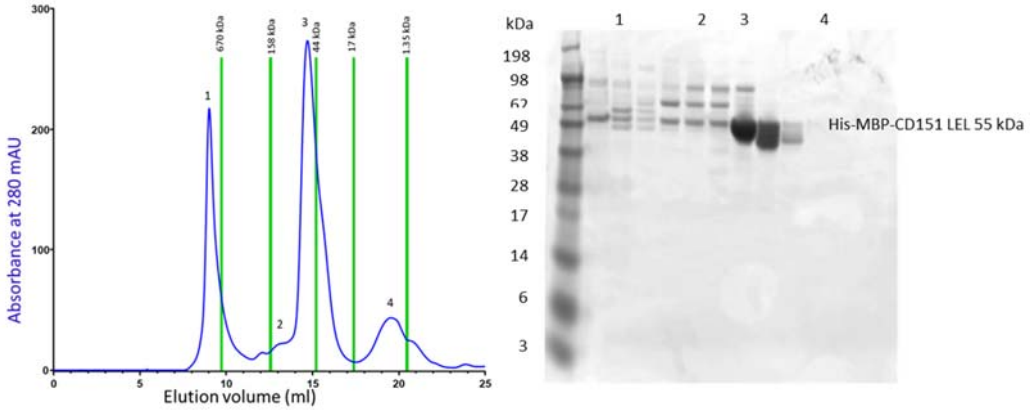


Figure 5-55 Gel filtration chromatogram of HisTrap purified His_8-MBP-CD151 and analysis of the eluted fractions by SDS-PAGE. Protein was purified using a Superdex 200 10 300 column which was pre-calibrated with molecular standards (green lines) (left panel). SDS-PAGE analysis shows the majority of the His_8-MBP-CD151 LEL is in peak 3 (right panel).

The protein standard ovalbumin (44 kDa) elutes at 15.9 ml on the Superdex 200 10 300 column. Extrapolating from this, elution of His_8-MBP-CD151 (55 kDa) should precede ovalbumin elution, suggesting that the protein peak eluting at 14.71 ml is the MBP fusion protein (Figure 5-55). However, by SDS-PAGE analysis every eluted fraction contained protein of the correct size. Most of the protein seemed to be in the fractions corresponding to the 14.71 ml peak with some higher Mw contaminant. The presence of the fusion protein in earlier eluted peaks suggests that the protein forms aggregates in solution which are disrupted by the effect of detergent in the SDS sample buffer.

5.3.4.1.1. TEV cleavage of the His_8-MBP tag

Tobacco etch virus (TEV) is a 27 kDa protease which specifically targets the amino acid sequence ENLYFQG/S and cleaves between the Q and G or S. For the purpose of these experiments I expressed and purified a double mutant of TEV (L56V, S135G), which has enhanced stability and solubility, and contains a His_6 tag for purification and removal of the TEV protein from experimental samples, by affinity chromatography, post cleavage.

[367] TEV is routinely used at a 1:100 ratio of OD_{280} nm TEV:protein for cleavage. [368] A trial cleavage was carried out at RT and 4°C with samples taken hourly to four hours then overnight. Samples were analysed by SDS-PAGE (Figure 5-56).

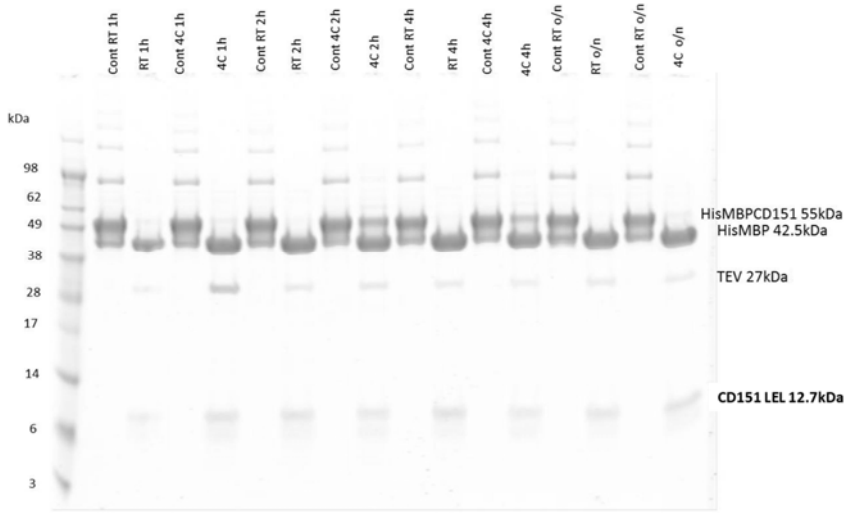


Figure 5-56 Time course of TEV cleaving the His_8-MBP from CD151 LEL at RT or 4°C comparing with control samples (cont) that did not contain TEV.

Cleavage for one hour at RT appeared to be complete. At 4°C there was still some intact protein present at 4 hours, but cleavage seemed to be mostly complete as there was little change after incubation overnight, with only a very faint band of intact protein remaining. TEV is maximally active at 34°C but only three-fold less active at 4°C. [369] In order to maintain the CD151 protein in a favourable environment, future cleavages were performed at 4°C for one hour.

A larger scale TEV cleavage of the HisTrap™ purified His_8-MBP-CD151 protein was carried out. 1 ml of protein at 9.5 mg/ml was incubated with 83 µl TEV (1:100 OD_{600}) and incubated at 4°C for one hour. The protein was purified over a 5 ml HisTrap™ column and fractions analysed by SDS-PAGE (Figure 5-57).

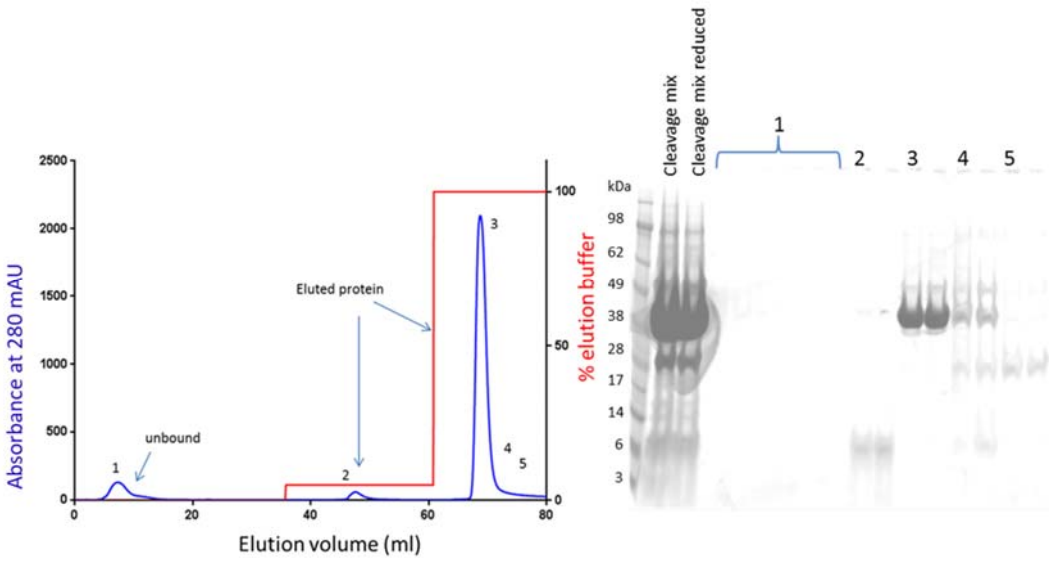


Figure 5-57 Separation of the cleaved CD151 LEL from the His_8-MBP tag over a HisTrap column. (left) All samples on SDS-PAGE were run as either oxidised (left lane) and reduced (right lane).

CD151 LEL should not bind to the HisTrap column, while the cleaved His-tagged MBP and TEV proteins should be retained. Analysis by SDS-PAGE (Figure 5-57, right) reveals that the unbound peak of protein is not CD151 LEL, which appears to be in peak 2.

The band of CD151 LEL visible on SDS-PAGE (Figure 5-57) is diffuse and may indicate protein degradation or multiple conformations. Samples were analysed by liquid chromatography time-of-flight mass spectrometry (LC-TOF) (Figure 5-58).

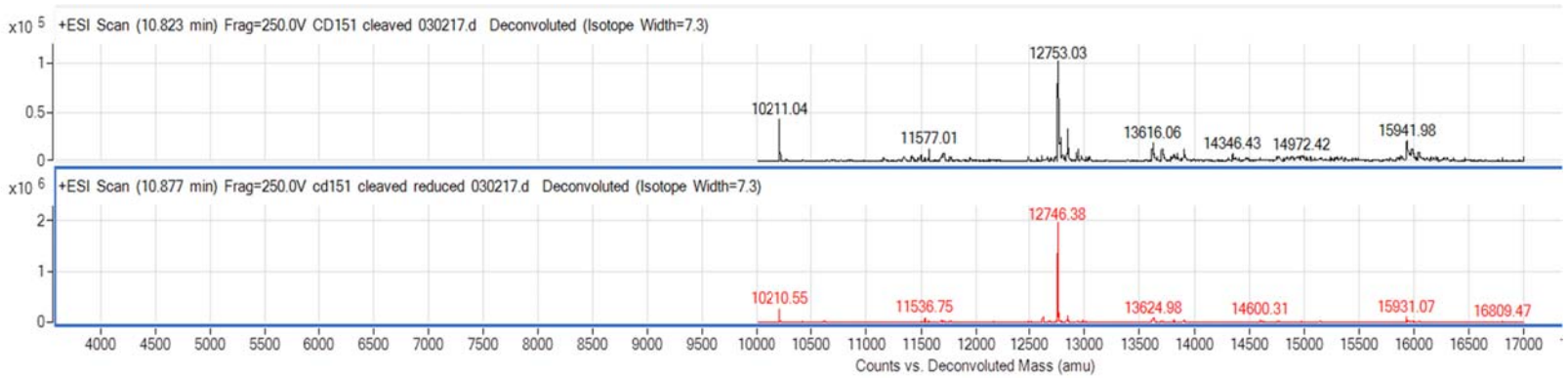


Figure 5-58 LC-TOF analysis of cleaved CD151 LEL reduced and oxidised.

The expected Mw is 12755.08 Da when reduced.LC-TOF analysis shows a protein peak at 12753 Da oxidised (top) and 12746.4 Da when reduced (bottom).

LC-TOF analysis revealed a 6 Da difference in size between reduced and oxidised protein species, suggesting the presence of three disulfides. (However, the mass decreased when the protein was reduced when the addition of 6 H^+ should have resulted in a mass increase. The most logical explanation for this is that I transposed the tubes in the sample rack. I was unable to repeat the experiment as protein and access to the equipment was limited.) These data also indicated that the sample was heterogeneous, with only a small amount of contamination of what is potentially break down product. The presence of the three disulfides was encouraging; however, it appeared that the protein was incorrectly folded, and aggregated once cleaved from the MBP tag.

5.3.4.2. Expression in SHuffle cells

The MBP tag was moderately successful in solubilising the CD151 LEL protein however once cleaved the CD151 became unstable suggesting it was not folding correctly. SHuffle cells are known to promote disulfide bond formation, but also assist with folding even in the absence of cysteines. [370] The His_8-MBP-TEV-CD151 LEL construct was transformed into SHuffle T7 Express cells and grown in a 200 ml culture in RM with 0.2% glucose under kanamycin selection. The culture was grown to OD_{600} of 0.6 and induced with 0.1 mM IPTG and grown overnight at 16^oC. The cells were harvested, processed as previously described and samples were analysed by SDS-PAGE (Figure 5-59).

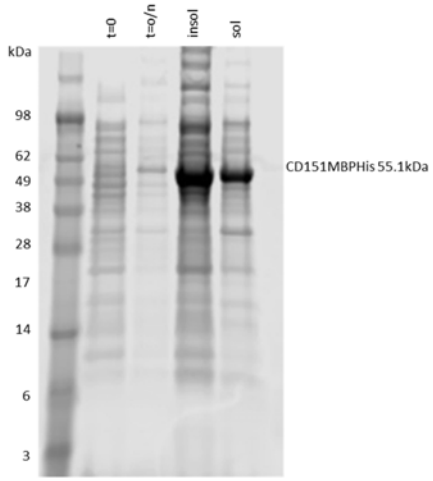


Figure 5-59 SDS-PAGE analysis of the expression of His_8-MBP-CD151 LEL in SHuffle cells.

The soluble fraction of the expression showed a strong band of protein at the approximate size of the MBP construct. The lysate was purified over a 5 ml HisTrap™ column as previously described (Figure 5-60).

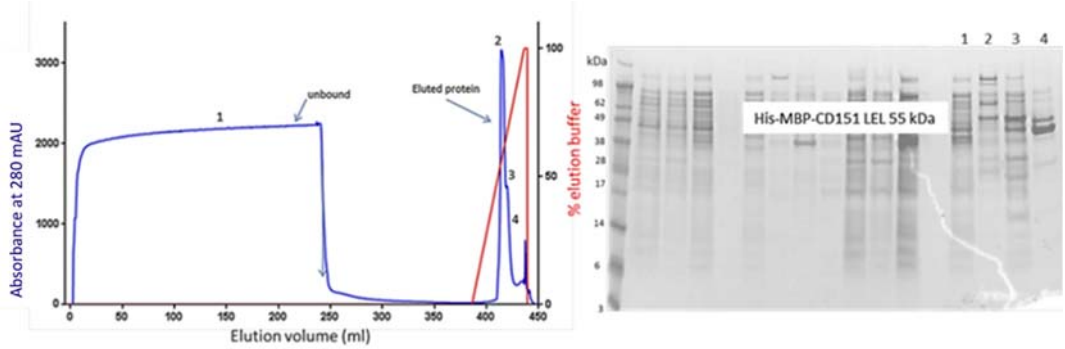


Figure 5-60 Elution profile of the soluble fraction of the cell lysate from His_8-MBP-CD151 LEL expression in SHuffle cells purified over a HisTrap™ column. (left panel). Fractions were analysed by SDS-PAGE (right panel), numbered fractions (right) correspond to numbered lanes on the gel. Unlabelled lanes on the gel are from an unrelated experiment.

All eluted fractions containing protein at the right size, as assayed by SDS-PAGE, were pooled, concentrated to 110 ml using a 10 kDa cut-off spin filter and desalted into 70 mM HEPES 100 mM NaCl 5% glycerol pH7.5 using a HiLoad desalt 20 10 column (GE). The protein was then further purified over a Superdex 200 26 60 gel filtration column (Figure 5-61).

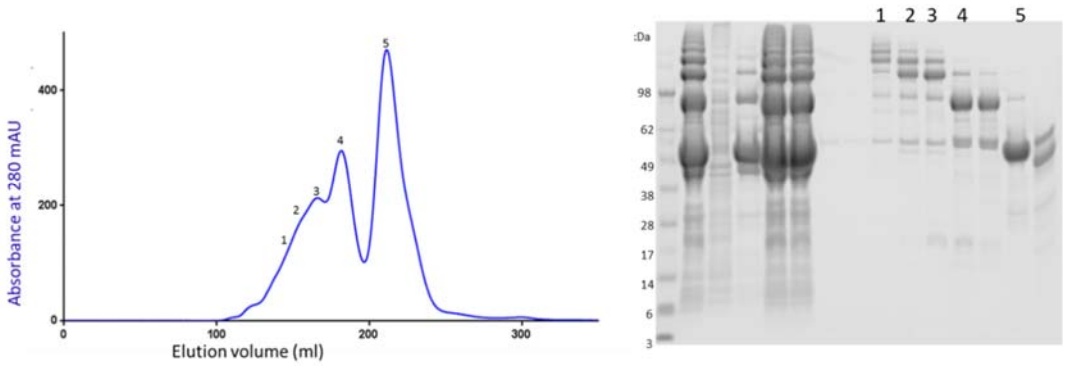


Figure 5-61 Elution profile of His_8-MBP-CD151 LEL on a Superdex 200 26 60 gel filtration column. (left) The blue line represents the absorbance at 280 nm (mAU). Fractions were analysed by SDS-PAGE (right). The numbers on the elution profile correspond to the numbers on the SDS-PAGE gel.

Fraction 5 from gel filtration was analysed by mass spectrometry (Table 14).

Table 14 Summary of the LC-TOF mass spectrometry analysis of His_8-MBP-CD151 LEL.

	His_8-MBP-CD151 M_w (Da)	ΔM_w ($M_w^{(obs)}$ -$M_w^{(predicted)}$) (Da)
Predicted	55127	-
Observed M_w	54977.4	-149.6
Observed M_w (reduced)	54982.7	-144.3

Table 14 shows the summary of the LC-TOF mass spectrometry analysis of the purified His_8-MBP-CD151 LEL. The predicted monoisotopic mass is shown in the top row. The second row contains the mass as determined by LC-TOF. This weight is ~150 Da less than the expected weight, which corresponds to the weight of the initiating methionine. N-terminal methionine excision is a common form of PTM that usually only occurs when the second amino acid is Gly, Ala, Ser, Thr, Cys, Pro or Val. [371] Although the second amino acid in this case is His, a conformation which is not favourable to methionine excision, it has been documented as occurring in immature *E. coli* sequences. [372] When reduced the mass increased by over 5 Da, which corresponds to the breakage of the three disulfide bonds (Agilent LC-TOF has an error of ± 0.5 Da).

This protein was used in crystallisation trials.

5.3.5. MBP-CD151 LEL Surface Entropy Reduced mutants

As cleavage of the CD151 LEL from the MBP tag proved problematic, it seemed that the best approach to making recombinant protein for structural studies was to include a tag that could facilitate protein crystallisation. Moon et. al. generated five MBP tags containing surface entropy reducing mutations designed to encourage protein crystallisation, largely by decreasing the energetic favourability of protein interactions with water molecules. [373] The surface entropy reduction (SER) mutations present in each MBP tag are listed in Table 15.

Table 15 SER mutations present in the MBP tags

Vector	SER mutation	Mw when fused to CD151 LEL (Da)
pMALX(A)	D82A/K83A	53013.9
pMALX(B)	E172A/N173A	52882.8
pMALX(C)	D82A/K83A/K239A	52825.7
pMALX(D)	E172A/N173A/K239A	52825.7
pMALX(E)	D82A/K83A/E172A/N173A/K239A	52767.6

CD151 LEL cDNA was cloned into each of the five vectors including a second pMALX(A) clone (A2) then transformed into BL21(DE3), SHuffle T7 Express and Rosetta2(DE3) plyss *E. coli* strains. Colonies for pMALX(C) and (E) in Rosetta2(DE3) plyss failed to grow. Two colonies from each successful transformation were grown in small scale 10 ml cultures with LB, 0.2% glucose with ampicillin selection, induced with 0.1 mM IPTG at OD_{600} 0.6 and allowed to grow for a further three hours before harvesting. Cell pellets were resuspended and sonicated to rupture the cells then centrifuged 15000 x g to pellet the insoluble material. Soluble lysates were analysed by SDS-PAGE (Figure 5-62).

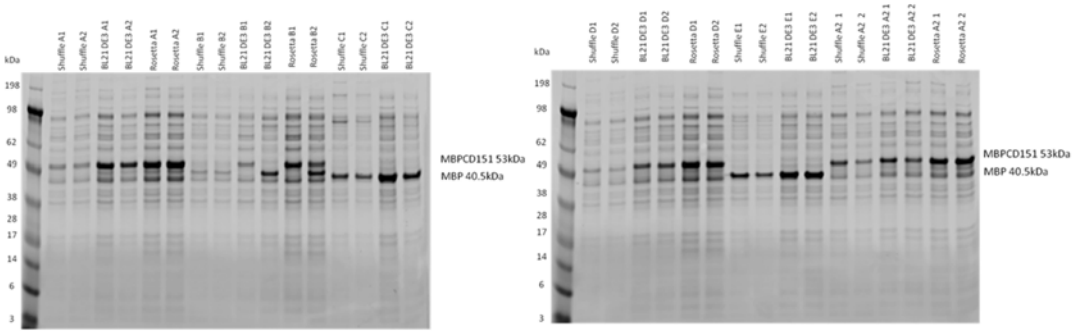


Figure 5-62 SDS-PAGE analysis of soluble fraction of the five different MBP SER mutants CD151 LEL fusion protein expressed in BL21(DE3), SHuffle T7 Express and Rosetta2(DE3) plyss cells.

All clones of pMALX(A) and (D) showed a protein band at the correct size. Clones BL21(DE3) B1 and Rosetta2(DE3) plyss B1 and B2 also showed protein at the correct size. Clones Rosetta A2, B1 and D1, SHuffle A2.1 and BL21 A1 were chosen as the "best" clones as they gave the best band density at the correct size with the least amount of contaminating proteins present. The lysate samples of these clones were analysed again by SDS-PAGE and Western blot, including the pre-induction sample to confirm the 55 kDa band is the expressed protein (Figure 5-63).

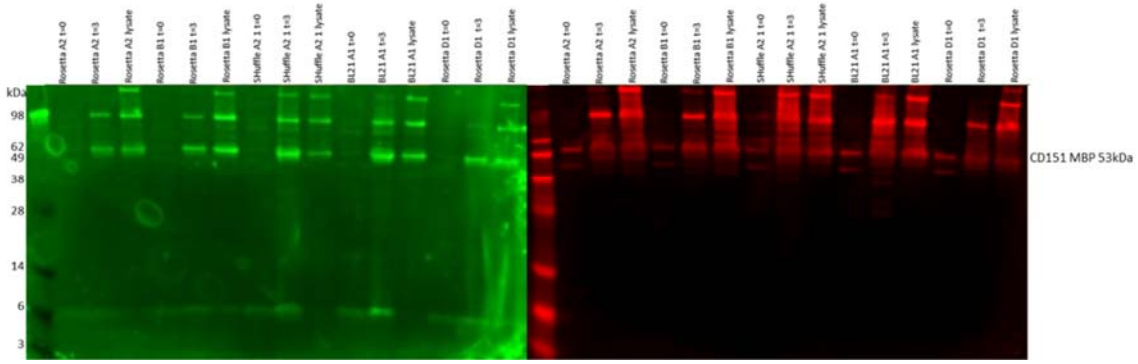


Figure 5-63 Western blot analysis of the expression of the SER mutant CD151 LEL fusion proteins. The same blot was probed with multiple Abs: first with anti-CD151 Ab Abcam 125363 rabbit polyclonal and anti-MBP mouse monoclonal; then with secondary Abs anti-rabbit IR800 (green, left panel) and anti-mouse IR680 (red, right panel).

The Western blot revealed that the pre-induction band of protein visible in SDS-PAGE was reactive to the anti-MBP Ab (Figure 5-63, right) suggesting the presence of

endogenously produced MBP. The anti-CD151 Ab was reactive to some low Mw bands, suggesting that the CD151 LEL is breaking down. The higher Mw species were reactive to both CD151 and MBP Abs suggesting that they were oligomers of misfolded protein. Clones SHuffle A2.1 and Rosetta A2 were chosen as the "best" clones as they had the least amount of breakdown product and the highest expression level. The Shuffle clone was chosen for further analysis as it likely had the best chance of producing correctly folded protein. It is interesting to note that the best clones were both produced by the MBP mutant with the least number of mutations.

200 ml cultures of each of the two clones were grown and processed as previously and the soluble lysate was purified on an MBPTrap™ (GE). Fractions were analysed by SDS-PAGE (Figure 5-64).

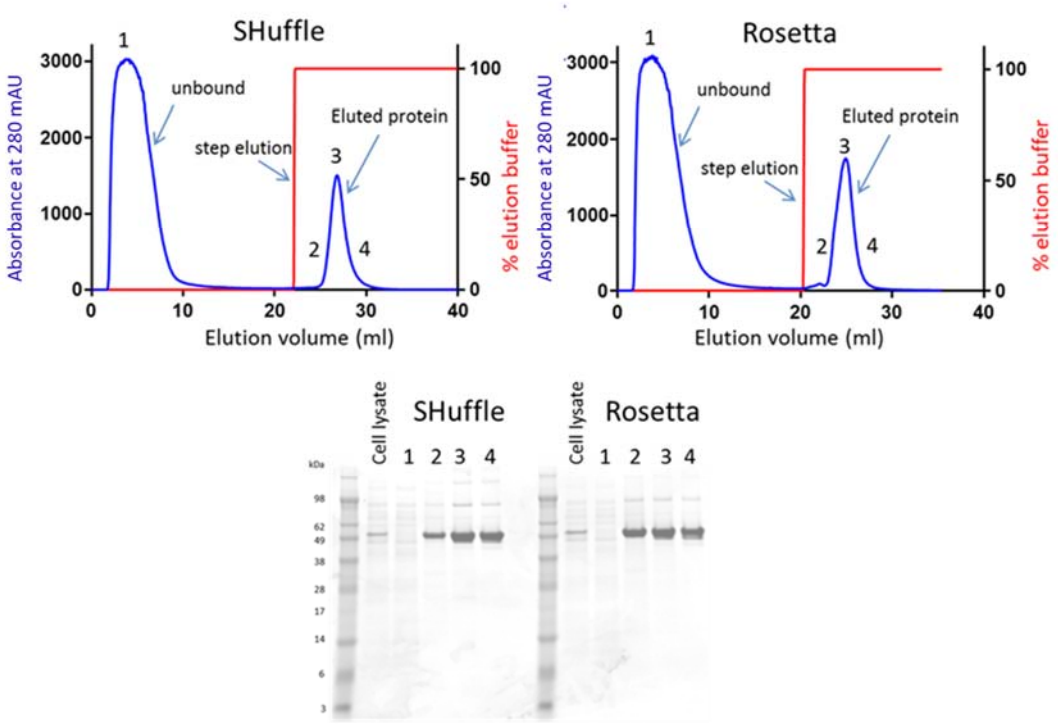


Figure 5-64 Chromatogram of the elution profile of the soluble fraction of the expression of pMALX(A)-CD151 LEL in SHuffle (top left) and Rosetta cells (top right) eluting from an MBPTrap™ Fractions were analysed by SDS-PAGE. (bottom)

Approximately 9 mg of protein was purified from the SHuffle expression and 12 mg from the Rosetta expression. The samples were concentrated to 500 µl using a 10 kDa cut-off

spin concentrator and further purified on a Superdex 75 10 300 (GE) gel filtration column and eluted fractions were analysed by SDS-PAGE (Figure 5-65).

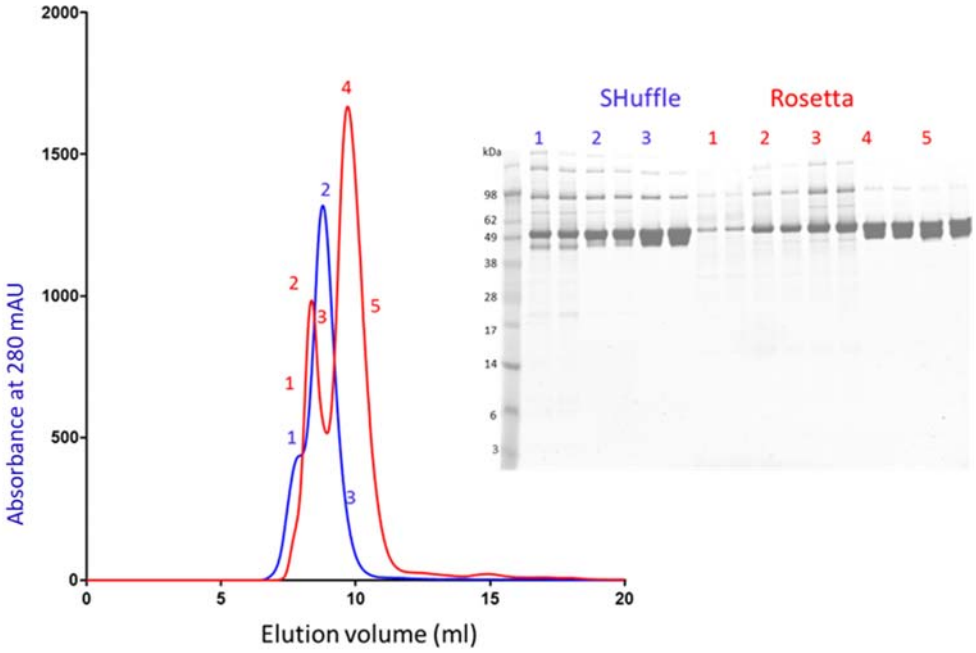


Figure 5-65 Chromatogram of the elution profile from the gel filtration purification of pMALX(A)-CD151 expressed in SHuffle cells (blue) and Rosetta cells (red). (right panel) Fractions were analysed by SDS-PAGE (left panel) and numbered fractions on the chromatogram correspond to the numbered samples on SDS-PAGE.

The gel filtration profiles showed two protein peaks for each sample: the Rosetta expressed protein had clearly separated peaks while the SHuffle produced material was less defined. Samples analysed by SDS-PAGE were run as oxidised or reduced with 100 mM DTT. As there was no discernible difference between the oxidised and reduced samples it seemed that reduction was unable to separate oligomeric proteins, suggesting that promiscuous formation of disulfide bonds may not be driving the formation of high order oligomers. The high Mw oligomers eluted at the same time as the 53 kDa protein as the Superdex 75 column does not have the resolving power to separate proteins greater than 70 kDa.

LC-TOF mass spectrometry analysis of the samples reveals that the SHuffle material contained 1 – 2 disulfides and the Rosetta material did not contain any (Table 16). The pMALX(A)-CD151 LEL fusion protein produced in SHuffle cells was used in crystallisation trials (refer section 6.3.2).

Table 16 LC-TOF mass spectrometry analysis of pMALX(A)-CD151 LEL fusion protein purified from either SHuffle or Rosetta cell expressions. Samples were reduced with 100 mM TCEP.

Expression host	Mw oxidised (Da)	Mw reduced (Da)
SHuffle T7 Express	53009.2	53012.1
Rosetta2(DE3) plyss	53015.8	53016.2

5.3.6. His_6-SUMO-CD151

The small ubiquitin-like modifier (SUMO) proteins are post translationally attached to other proteins and are involved with numerous cellular processes such as protein targeting and folding. [374] A His_6-SUMO purification and expression tag fused to the N-terminus of the target protein has been shown to facilitate protein folding and increase solubility. An additional benefit of the SUMO tag is that the SUMO protease, ULP1, is highly efficient and cleaves at the C-terminus of SUMO leaving the heterologous protein with a native N-terminus. [375]

The pET-30a(+) vector containing the His_6-SUMO-CD151 LEL cDNA was transformed into SHuffle T7 Express cells and grown and processed as previously. Fractions from the expression were analysed by SDS-PAGE (Figure 5-66).

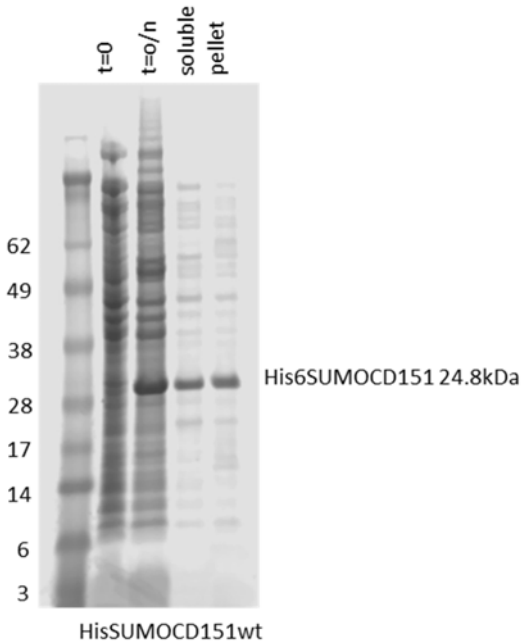


Figure 5-66 His_6-SUMO-CD151 LEL expression in SHuffle T7 Express cells. Samples were taken at induction (t=0), after overnight growth at 16°C (t=o/n) and of the soluble and insoluble (pellet) fractions of the cell lysate. A clear band of material is present in the soluble fraction.

The soluble fraction of the cell lysate was purified over a HisTrap™ column. As observed in all previous purifications, the misfolded CD151 LEL formed oligomers. For this construct each additional protein unit forming the oligomer carries the His6 tag, increasing the affinity for the nickel resin. Theoretically, the monomeric species, containing only one His tag, should elute at a lower percentage of imidazole. As such, the protein was eluted with a step-wise gradient rather than the usual linear gradient (Figure 5-67).

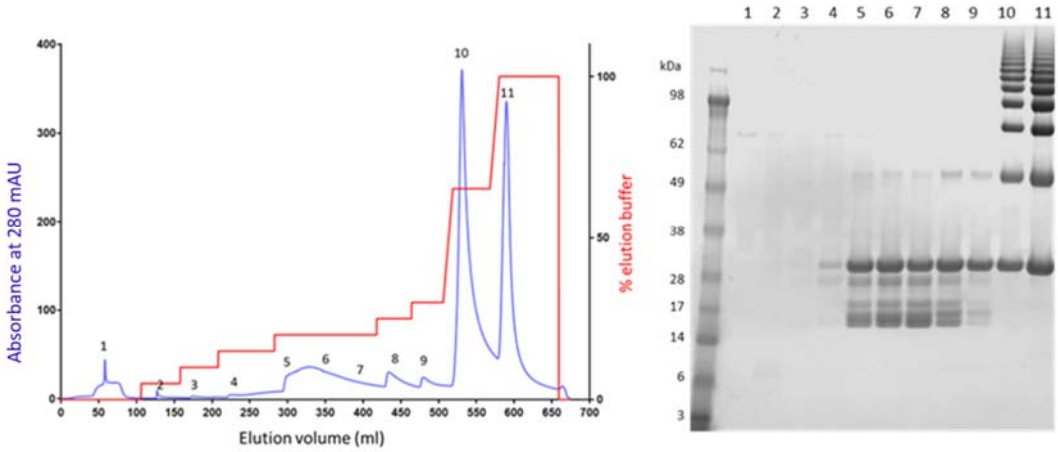


Figure 5-67 Chromatogram of the step-wise elution of His_6-SUMO-CD151 LEL from a HisTrap column. (left) The red line represents the concentration of the elution buffer (right ordinate axis). Samples from each step were analysed by SDS-PAGE (right panel). Numbered fractions on the chromatogram correspond to numbered samples on SDS-PAGE.

Analysis of the elution fractions by SDS-PAGE revealed that fractionation was occurring with low Mw protein eluting in the earlier fractions as anticipated. (Figure 5-67) The later fraction still appeared to contain monomer, suggesting the aggregated protein was disrupted to monomer by the SDS sample buffer and may not be forming covalently bonded oligomers.

The protein was concentrated using a 30 kDa cut off spin concentrator to try to separate the 24 kDa monomer from the oligomers. SDS-PAGE analysis showed that while the filtrate did contain monomer, there was also oligomer present. The filtrate and the

retained material were almost identical under oxidising conditions (Figure 5-68). As the oligomer could not have passed through the filter this indicates that the oligomers must be forming post filtration and that monomeric and oligomeric species exist in an equilibrium. When the samples were reduced the oligomers in the filtrate mostly reduced to monomer while the oligomers in the retained sample were more resistant to reduction. It is possible that the bands of protein at 50 kDa and 80 kDa were not CD151 LEL.

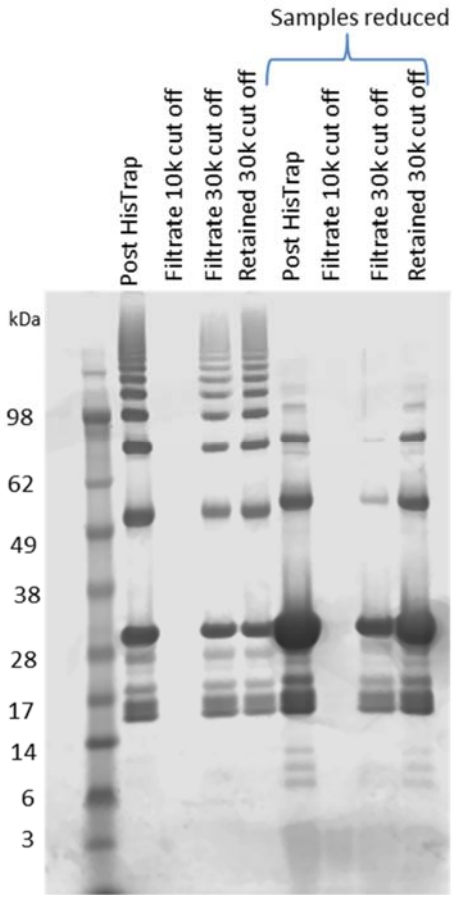


Figure 5-68 SDS-PAGE analysis of His_6-SUMO-CD151 LEL purified by affinity chromatography and filtered through a 10 kDa and 30 kDa cut off spin filter. Samples are shown oxidised and reduced. The "filtrate" is the material that has passed through the filter and the "retained" material is sample that is retained in the filter unit.

For disulfide bonds to form the redox active thiolate anions must be deprotonated ($Cys\text{-}S^-$). Environmental pH has an influence on thiolate reactivity and activity can be quenched by acidifying the environment which protonates the free thiol from $Cys\text{-}S^-$ to Cys-SH. [376, 377] Constitutive disulfide bonds are generally stable and not redox-regulated, however non-native disulfides are less stable and can rearrange. One

possibility is that filtered monomeric protein may spontaneously form and break transient disulfide bonds, resulting in the presence of higher order oligomeric protein in the filtered sample. To minimise this effect, I searched for a low pH buffer that could protonate the thiol groups to reduce reactivity but was also conducive to protein stability. A thermal melt assay with a broad pH range was performed. The assay revealed that the protein was less stable at pH below 6.5, with 50 mM MES pH 6.5 200 mM NaCl being a suitable buffer for protein stability.

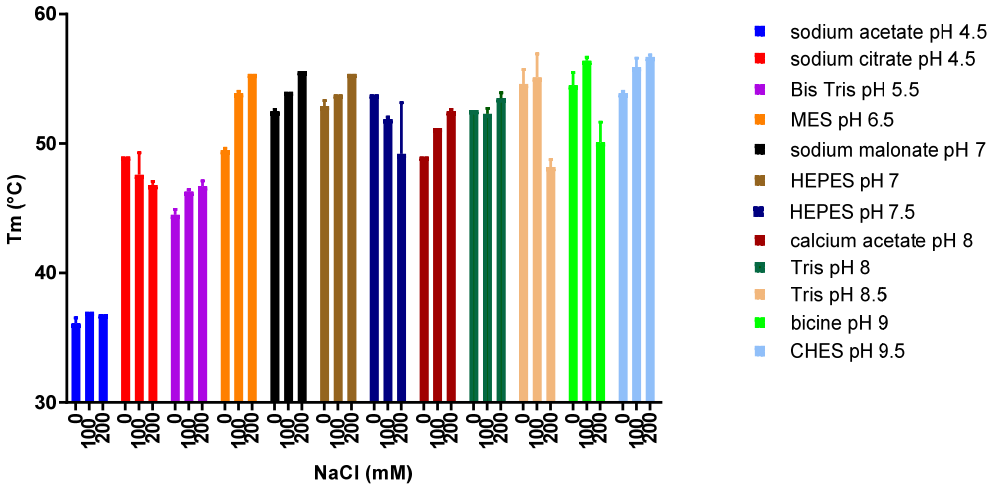


Figure 5-69 Thermal melt assay of His_6-SUMO-CD151 LEL in various pH and salt conditions.

A sample of the protein was buffer exchanged into 50 mM MES 200 mM NaCl pH 6.5 and separated into oligomer and monomer via gel filtration. The fractionated sample was analysed by SDS-PAGE, incubated over night at 4^{o}C, then reanalysed. No shift in oligomerisation occurred; however high Mw oligomers were present in the later fractions. This suggested that the misfolded monomers formed oligomers post elution. A sample was also fractionated into 70 mM HEPES 100 mM NaCl pH7.5 as a control. This sample, when analysed by SDS-PAGE, showed a greater oligomer content which did not shift post elution. The monomer was stable after 24 hours. These experiments indicated that the lower pH environment did not stabilise disulfide bond formation in the native fold of the protein, or alternatively, it is possible that pH 6.5 was not sufficiently acidic to result in widespread thiol protonation.

The protein was concentrated and purified over Superdex 200 16 60 column (GE) and a fraction containing mostly monomer was isolated.

5.3.6.1. Cleavage with ULP-1

ULP-1 is a very active cysteinyl protease which cleaves the SUMO protein in a highly specific manner, recognising the tertiary structure of the SUMO rather than an amino acid sequence. The catalytic core of yeast ULP-1 was expressed using standard bacterial approach and purified by Ni^{2+} ion affinity chromatography. The purified protease contained His_6 tags at both N- and C-termini to facilitate its removal from the protein sample following digestion.

A 60 µM solution of His_6-SUMO-CD151 (2.25 mg in 1.5 ml) in 70 mM HEPES 100 mM NaCl 5% glycerol pH 7.5 was incubated with 1 µl of 65 µM ULP-1 at RT with samples taken every half hour. Samples were analysed by SDS-PAGE (Figure 5-70).

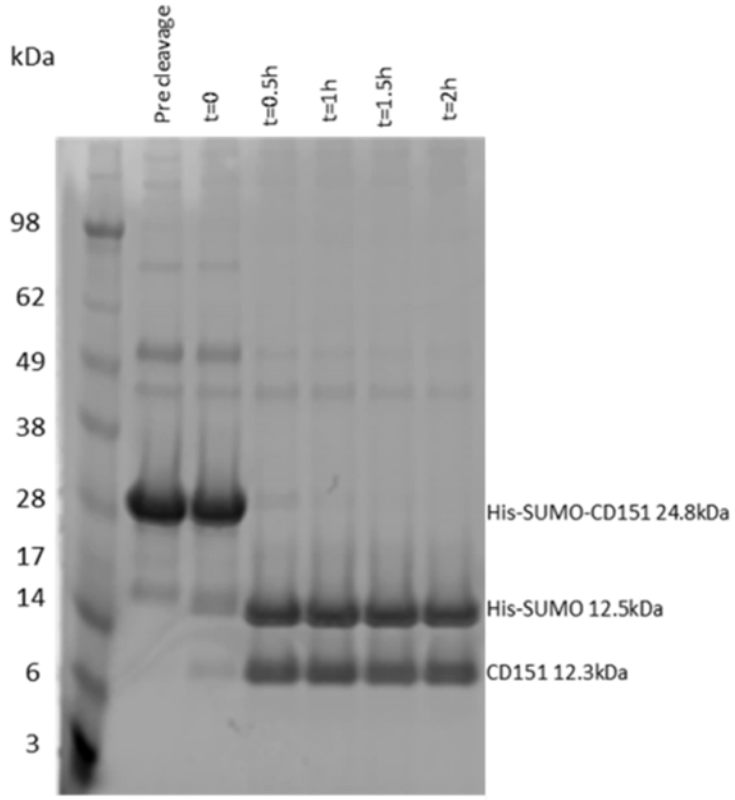


Figure 5-70 Time course of the cleavage of SUMO from CD151 LEL by ULP1 protease.

Cleavage of the SUMO tag from the CD151 LEL was almost complete after only 30 minutes, after 2 hours only a very faint band of uncleaved protein was visible in the gel (Figure 5-70). This was an encouraging result as ULP-1 protease exclusively recognises folded SUMO protein, suggesting the protein is correctly folded.

A larger scale cleavage of 14.5 mg of His6-SUMO-CD151 was performed, as above, incubated for 2 hours at RT. The reaction mix containing the cleaved His6-SUMO, CD151 LEL and the His-tagged ULP1 protease was passed over a HisTrap column (Figure 5-71).

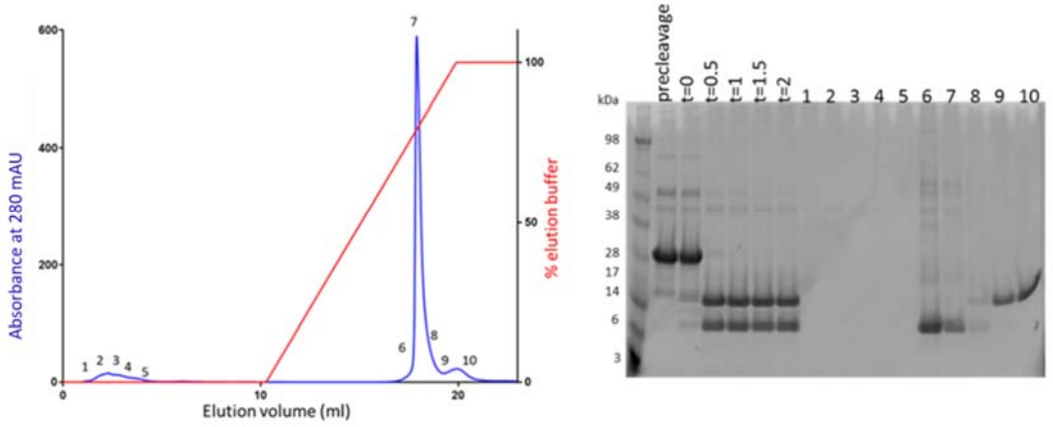


Figure 5-71 Chromatogram of the elution profile of His_6-SUMO-CD151 protein cleaved by ULP1 The protein was purified over a HisTrap column, retained protein was eluted with a 0 - 100% gradient with 500 mM imidazole. (left) Fractions were analysed by SDS-PAGE (right) and numbered fractions correspond to the samples on the SDS-PAGE gel. The gel also shows the protein cleavage reaction at 0.5 hour intervals to 2 hours.

The cleaved CD151 LEL should pass through the HisTrap column, as such the protein should be in first peak in the sample flow through. Fractions containing protein were analysed by SDS-PAGE (Figure 5-71).

The protein concentration of the first peak was too low to detect on SDS-PAGE, the second peak seemed to contain CD151 LEL and the third peak the His-SUMO. This elution profile was curious, as it suggested that CD151 LEL was retained on the column, possibly by virtue of native histidine residues coordinating the Ni^{2+} capture resin. The CD151 LEL contains five histidine residues, at positions 23, 27, 42, 85 and 107. It is possible that several of these residues may adopt a conformation in folded CD151 LEL that is favourable to coordinating a nickel ion, and as such, elution requires competition from imidazole in the elution buffer. Even though the histidine residues are in close proximity it seems unlikely that they would be binding to the nickel resin with such high affinity as to require approximately 65% elution buffer to dissociate it from the resin (Figure 5-72).

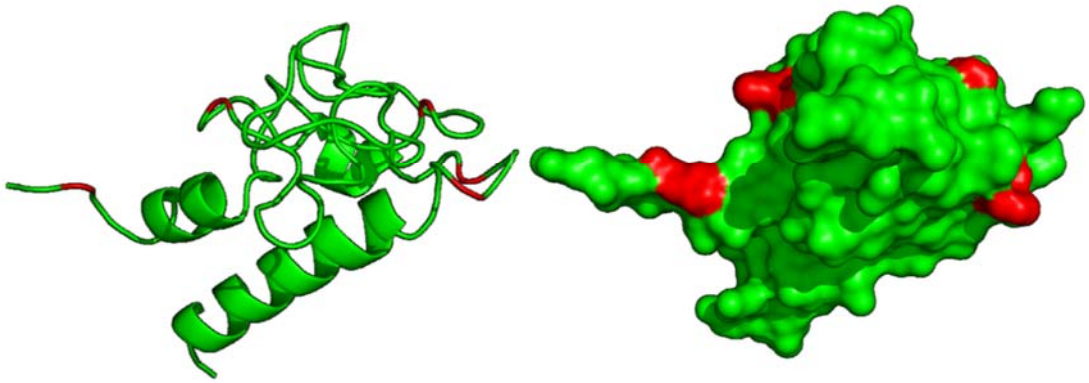


Figure 5-72 Cartoon structure of my CD151 LEL model (left) showing the five histidine residues in red. With the surface overlayed (right) showing surface available histidine residues that coordinate the nickel ion on the NTA resin.

A sample of the eluted protein, presumed to be CD151 LEL, was analysed by mass spectrometry (Figure 5-73).

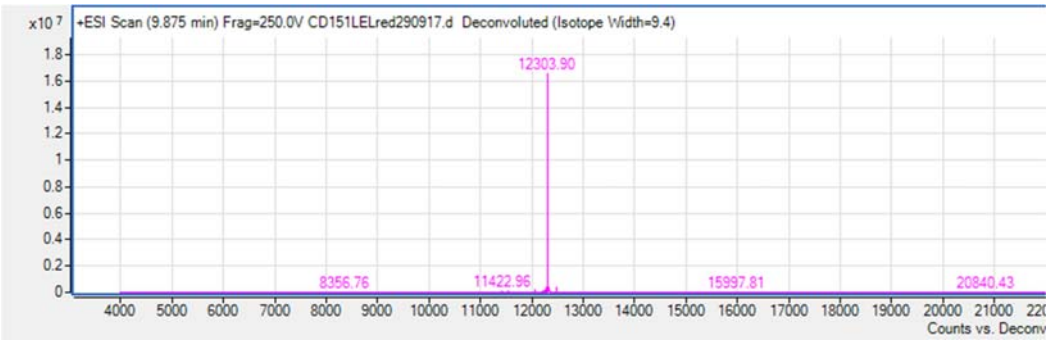


Figure 5-73 LC-TOF mass spectrometry analysis of the purified CD151 LEL reduced with 100 mM TCEP reveals a protein with Mw 12303.9 Da.

The predicted monoisotopic mass of CD151 LEL is 12303.6 Da, and the Agilent LC-TOF is accurate to within 0.5 Da, indicating the CD151 LEL is the correct size. The predicted mass of the cleaved His-SUMO protein is 12496.9 Da and this mass does not appear in the sample. The mass spectra confirms that the peak eluted from the HisTrap column (Figure 5-71) is the CD151 LEL.

A sample of the CD151 LEL was buffer exchanged into 50 mM NaF and analysed by CD using the Contin-LL method and reference data set SP175. [378] The spectra showed

that the protein was all helical (Table 17), strongly suggesting that the purified protein was folded (Figure 5-74).

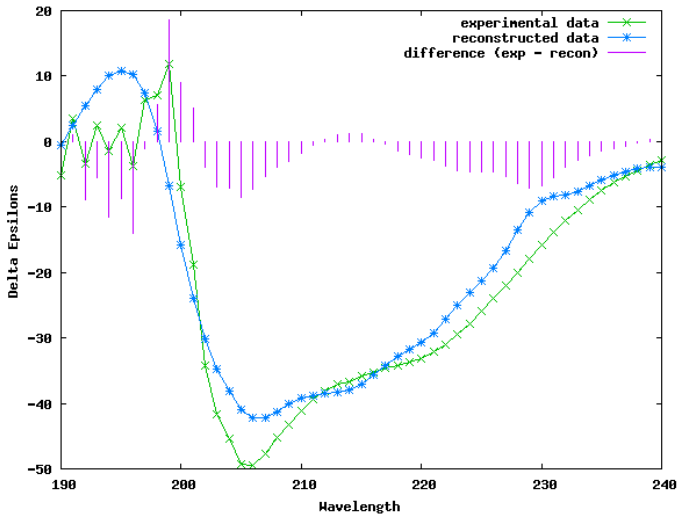


Figure 5-74 CD spectra of the CD151 LEL. Experimental data are plotted in green; the calculated spectrum derived from the calculated output secondary structure is plotted in blue and the difference spectra is depicted in vertical lines in pink. The plot shows a reasonable fit to the calculated data, with some deviation in the lower wavelength.

Table 17 The content of secondary structure predicted from the CD spectra.

Result	Helix1	Helix2	Strand1	Strand2	Turns	Unordered	Total
1	0.755	0.245	0.000	0.000	0.000	0.000	1
2	0.866	0.134	0.000	0.000	0.000	0.000	1

5.3.7. Disulfide bond mapping

CD151 LEL was digested with LysC (0.1 µg/µl) at a 1:100 molar ratio, incubated for 12 hours at 37°C. Sequence analysis by ExPASy PeptideCutter [379] identified 6 LysC cleavage sites in the CD151 LEL (Figure 5-75).

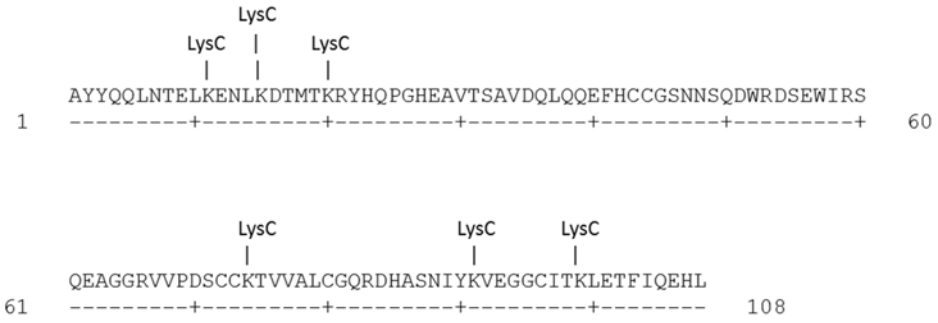


Figure 5-75 Sequence of CD151 LEL indicating the LysC cleavage sites, cleavage occurs to the right of the lysine.

Proteolytic cleavage using LysC was employed as a strategy to examine folding of the purified CD151 LEL protein. The rationale for these experiments is as follows: if two to three of the disulfide bonds that characterise the correct folding of the CD151 LEL are present in the purified protein then proteolytic cleavage by LysC should result in a disulfide-linked protein consisting of three discrete fragments, summing to a total Mw of 8.8 kDa (Figure 5-76).

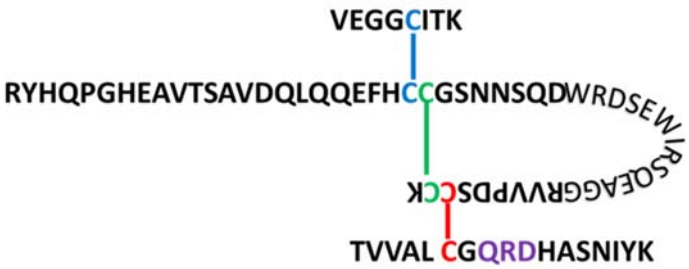


Figure 5-76 Schematic of the fragments of CD151 LEL cleaved by LysC held into an 8.8 kDa protein by disulfide bonds.

The protein fragment arrangement shown in Figure 5-76 is the predicted disulfide bond formation based on the conserved cysteines in the tetraspanin protein family. The digested CD151 LEL protein was examined using TOF mass spectrometry (MS-TOF). The change in mass when the purified LEL is reduced indicates that three disulfides are

forming (Figure 5-73, Figure 5-78). Using the Network Protein Sequence Analysis calculator for finding the number of possibilities for SS bridges in proteins, a protein with six cysteines, forming three disulfide bonds with no free SH groups, contains 15 different possible conformations. [380]

The mass/charge spectra were inspected and masses from the different charge series calculated. Calculation of the mass from the charge series identified in (Figure 5-77) using the equation (Mw = (m/z*charge)-charge) results in a Mw of 8806 Da. For example, the charge peak circled in red on the far right of the spectra has an m/z of 1468.68 and a charge of +6, using the formula Mw = (1468.68*6)-6 = 8806 Da. This result confirmed the presence of the 3 peptide species predicted (Figure 5-76), however it also confirmed that many other peptide combinations were present.

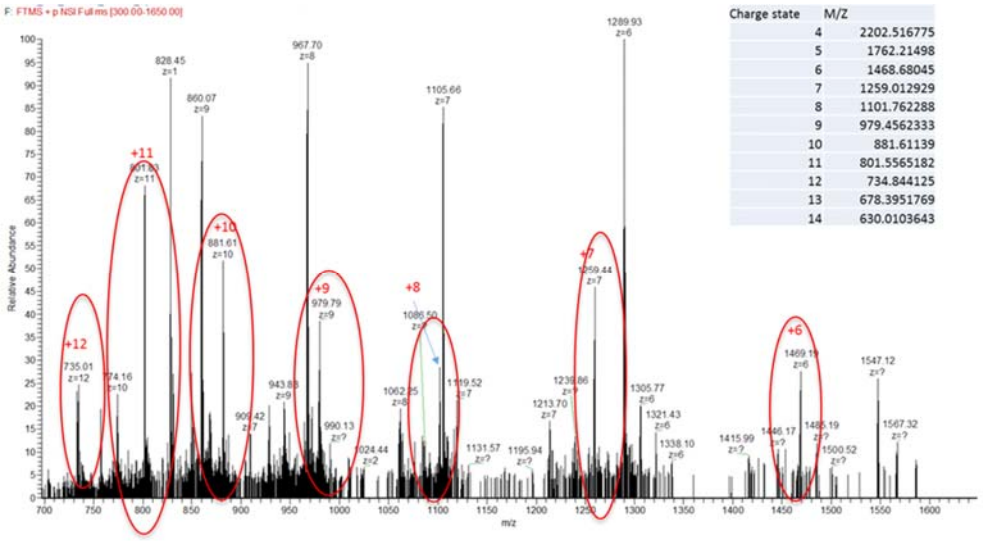


Figure 5-77 Mass/charge spectra of the MS-TOF analysis of the LysC digestion fragments of CD151 LEL. The charge series circled in red and summarised in the inset table, calculate to a protein with a mass of 8.8 kDa indicating the predicted disulfide linked peptides are present.

Analysis of the m/z spectra (Figure 5-77) by the protein metrics software Byonic™, which identifies disulfide bond crosslinked peptides, resulted in the identification of several peptide conformations. The results are summarised in Table 18.

Table 18 Summary of the disulfide bonded peptides observed in LysC digested CD151 LEL using the Byonic™ protein metrics software. The cysteines forming disulfide bonds between peptide 1 and 2 are indicated in red. Where Peptide 2 is absent, Peptide 1 is forming intra-peptide disulfide bonds.

Peptide 1	**Peptide 2**
RYHQPGHEAVTSAVDQLQQEFHCCGSNNSQDWRDSEWIRSQEAGGRVVPDSCCKTVVALCG	TVVALCGQRDHASNIYK
RYHQPGHEAVTSAVDQLQQEFHCCGSNNSQDWRDSEWIRSQEAGGRVVPDSCCK	
TVVALCGQRDHASNIYK	VEGGCITK
LCGQRDHASNIYK	VEGGCITK
RYHQPGHEAVTSAVDQLQQEFHCC	
GGRVVPDSCCK	
NSQDWRDSEWIRSQEAGGRVVPDSCCK	

These data suggested that the protein formed non-native disulfide bonds and may explain the difficulties with protein production and crystallisation (see CHAPTER 6).

The Water's Vion IMS QTof Ion Mobility Quadrupole Mass Spectrometer integrates mass spectrometry with ion mobility and enables sample separation in three dimensions: mass to charge, intensity, and drift time. This allows the separation of heterogeneous complexes with similar mass but different drift times due to variations in the shape and topology of the protein. [381] To further examine the molecular species present in the purified CD151 LEL the sample was analysed by the Vion IMS QTof (Figure 5-78).

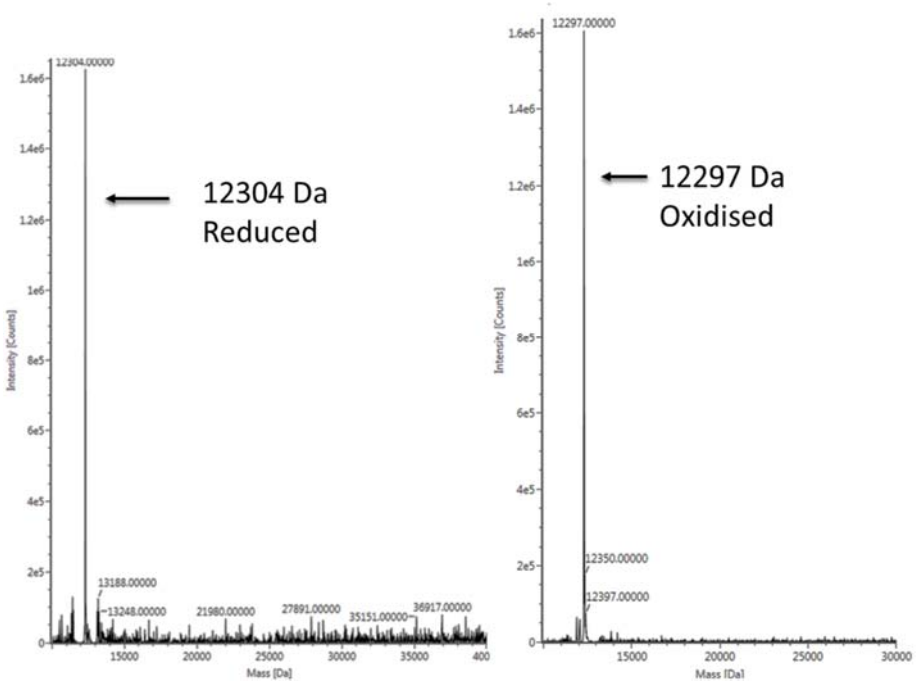


Figure 5-78 Deconvoluted mass spectrometry spectra of reduced (left) and oxidised (right) CD151 LEL. The expected reduced mass is 12303.6 Da as calculated by ExPASy [325], which is confirmed by the reduced spectra (left) and the sample mass decreased by 7 Da when oxidised (right), indicating that three disulfides are present.

The CD151 LEL was analysed for drift time versus m/z in both reduced and oxidised states (Figure 5-79, Figure 5-80).

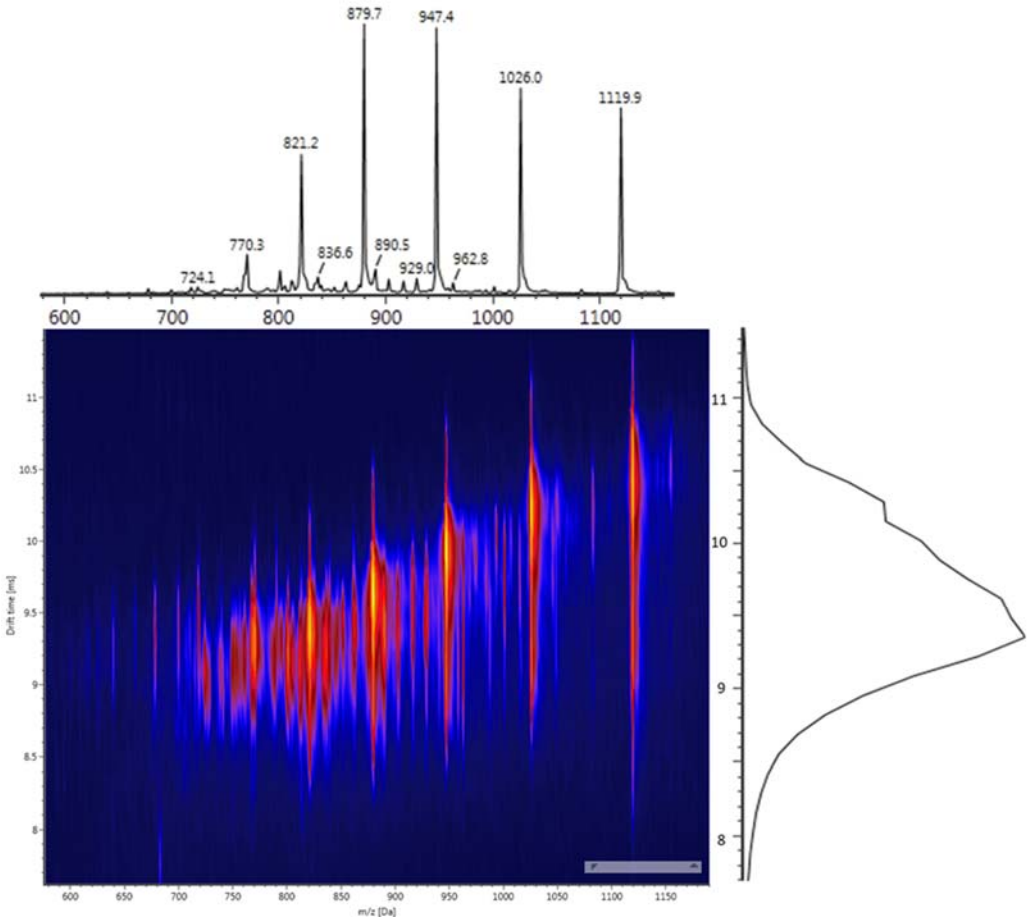


Figure 5-79 Plot of drift time versus m/z for oxidised CD151 LEL. (Central figure) The mass spectrum (projected on top) shows a charge series which is reflected in the drift time data. The drift time profile (projected on the side) indicates the spread of the drift times. The longer drift time and multiple charge groups indicate that the protein is present in different forms which are subject to different drift times.

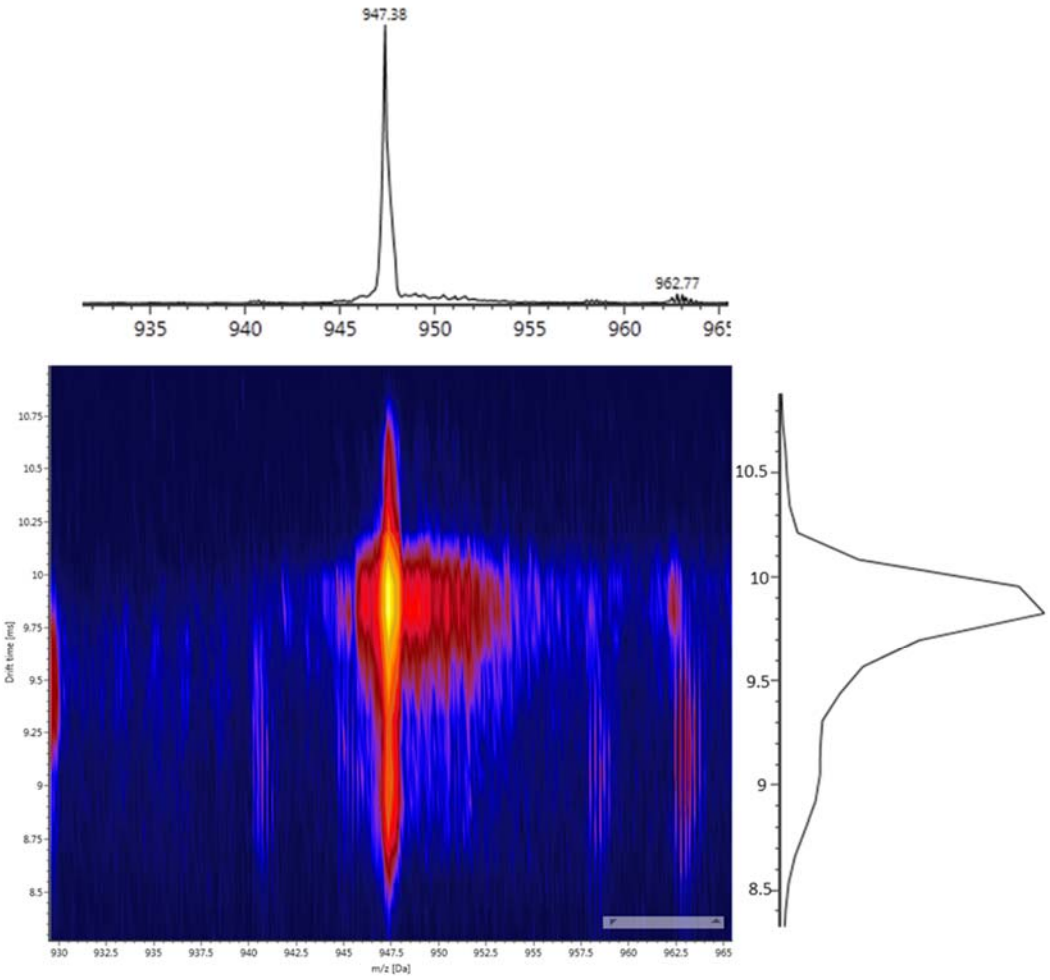


Figure 5-80 Plot of drift time versus m/z for reduced CD151 LEL. (Central figure) The mass spectrum (projected on top) shows a single charge series which is reflected in the drift time data. The drift time profile (projected on the side) indicates the narrow spread of the drift. The shorter drift time and single charge groups indicate that the protein is present in one form.

When analysing the protein by Mw a single species was present in both the reduced and oxidised states (Figure 5-78). However, when measuring the ion mobility, it was apparent that the protein is forming different shapes (Figure 5-79, Figure 5-80). Three disulfide bonds were formed, as indicated by the change in mass when reduced, so the different shaped protein species most likely arise from alternate disulfide configurations. These results indicate that the purified protein formed multiple soluble, non-native protein species arising from promiscuous formation of disulfide bonds.

5.3.8. His_6-SUMO-CD151 LEL PROSS mutation

Protein Repair One-Stop Shop (PROSS) is a website operated by the Fleishman Lab at the Weizmann Institute of Science, Israel (https://pross.weizmann.ac.il). After inputting a protein sequence and structure the algorithm provides several mutated sequences that are predicted to be more stable. [382] The only structure available for CD151 LEL is the homology model based on Sm-Tsp-2 which has a 19% amino acid sequence similarity. The PROSS website states that any homology model with less than 40% sequence similarity will not be accurate enough to generate stabilising mutants. I decided to go ahead with the process, despite the low homology similarity, for if the mutations were successful they could provide a useful strategy for protein purification and, indirectly, lend support for the veracity of the homology model.

Seven models were furnished by the PROSS web site, varying from minimal substitutions (model 1) to 10 amino acid substitutions (model 7) that the PROSS algorithm suggested would be most likely to stabilise the protein (Figure 5-81).

As the chances of success with this method were low I decided to only proceed with model 7 which contained all ten substitutions (Figure 5-82).

Figure 5-81 Multiple sequence alignment of the seven PROSS mutation models compared to the wild-type CD151 LEL sequence.

```
CD151LEL wild-type AYYQQLNTELKENLKDTMTKRYHQPGHEAVTSAVDQLQQEFHCCGSNNSQDWRDSEWIRSQEAGGRVVPDSCCKTVVALCGQRDHASNIYKVEGGCITKLETFIQEHL
CD151LEL mutant 7  VYYQQLNEELKKNMRDTMTKRYHQPGAEEVTSAVDRLQQEFHCCGSNNSQDWRDSEWIRSQEAGGRVVPDSCCKTVVANCGQRDHASNIYKVEGGCITKLEEFIQEHLR
```

Figure 5-82 CD151 LEL PROSS model 7 compared with the wild-type LEL, substitutions shown in red.

A His_6-SUMO-CD151 LEL PROSS model 7 gene cloned into a pET-30a(+) expression vector was transformed into SHuffle T7 Express cells and grown and processed as described previously. Fractions from the expression were analysed by SDS-PAGE (Figure 5-83).

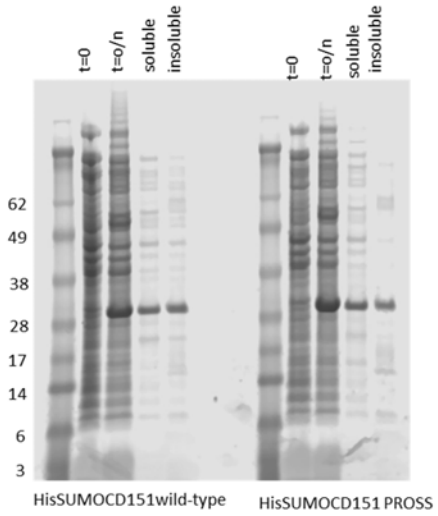


Figure 5-83 SDS-PAGE analysis of expression samples of His_6-SUMO-CD151 LEL wild-type (left) for comparison and with the 10 PROSS model 7 mutations (right). There did not appear to be any improvement in yield or solubility with the PROSS model 7 mutant.

The soluble fraction of the cell lysate was purified by affinity chromatography and the eluted fractions analysed by SDS-PAGE (Figure 5-84).

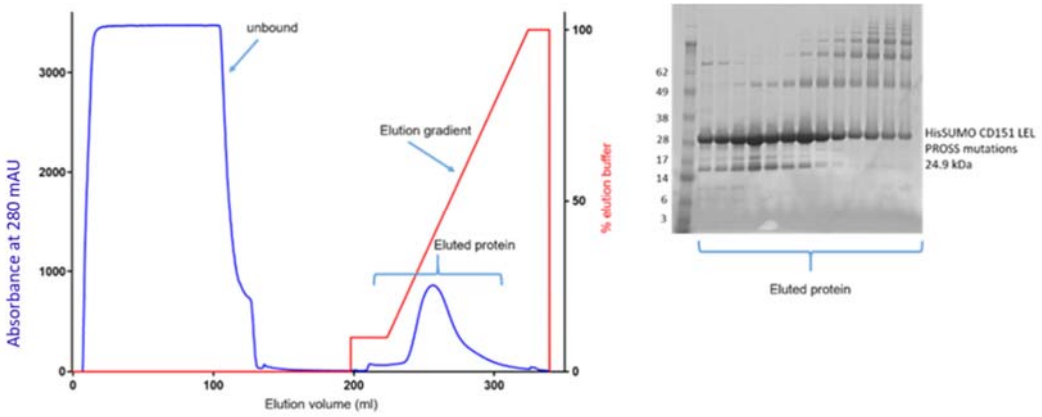


Figure 5-84 Chromatogram of the purification of His_6-SUMO-CD151 LEL PROSS fusion protein from cell lysate by HisTrap (left) and analysis of the eluted fractions by SDS-PAGE (right).

The SDS-PAGE analysis of affinity purified His_6-SUMO-CD151 LEL PROSS model 7 mutant protein gave the same profile as the His_6-SUMO-CD151 wild-type. Lack of improvement in solubility or reduction in the higher Mw banding present in the SDS-PAGE analysis suggested that the mutations had little effect on CD151 LEL protein expression. The PROSS algorithm appears to involve promoting hydrophobic interactions in the core of the folded protein while, at the same time, increasing the potential for surface residues to interact favourably with solvating water molecules; thus, decreasing the overall free energy of the protein in solution and, presumably, leading to greater protein stability. The failure of the mutations to increase protein stability suggests two likely alternatives: first, that the homology model used to discern between buried and solvent accessible side-chains is too inaccurate, and/or second, that the mutations, of themselves, were ineffective in driving an energetic minimum towards the stably folded protein species. In the interest of practicality these possibilities were not further investigated, and purification of this protein construct was not further pursued.

5.4. *INSECT CELL EXPRESSION*

Insect cell expression of heterologous protein can produce protein with PTMs similar to mammalian cells. One advantage of baculovirus mediated insect cell protein expression is that transduction of insect cells occurs by a recombinant baculovirus that can be propagated in the laboratory at little expense, compared to the large reagent costs for transient transfection of mammalian cell cultures. Another advantage of insect cell

culture is the advantage of growing to higher density and with much simpler conditions; e.g. there is no requirement for a CO_2 enriched atmosphere to buffer proton concentration in common, inexpensive, insect cell media formulations. [383-385]

CD151 LEL cDNA with an N-terminal Honey Bee Mellitin (HBM)-His_8-TEV tag was custom synthesised by Genscript® with codon optimisation for insect cells and cloned into a pFastBac1 vector using Not1-Xbal cloning sites. This resulted in a predicted fusion protein of 16862.9 Da. HBM is a signal peptide that directs the protein to the ER and through the secretory pathway. Once in the ER the signal peptide is cleaved, and the final excreted protein has a predicted molecular mass of 14252.6 Da.

Bacmid DNA was generated as per Materials and Methods (2.2.3.3) and the DNA was analysed by agarose gel electrophoresis (Figure 5-85).

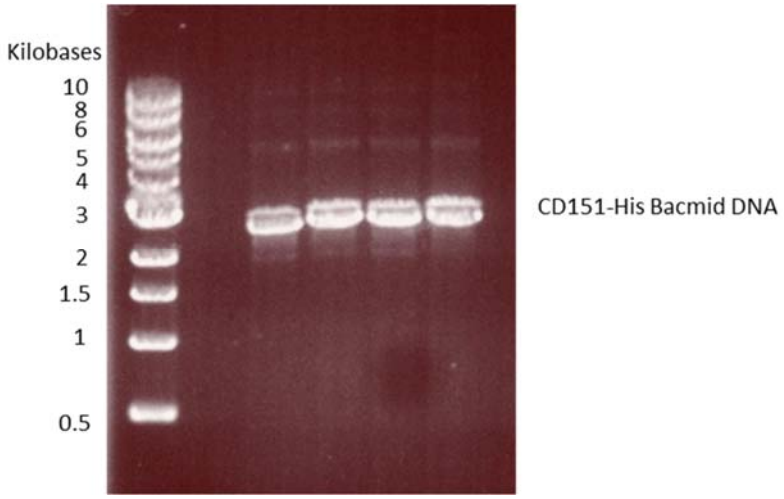


Figure 5-85 Agarose gel electrophoresis of DNA purified from DH10a colonies.

Baculovirus was grown and amplified as per method (section 2.2.3.5) and protein was expressed in *Sf*21 cells as per method (section 2.2.3.6).

Harvested media was dialysed against 20 mM Tris 100 mM NaCl 5% glycerol pH 7.5 then affinity purified using a HisTrap 5 ml column and fractions analysed by SDS-PAGE (Figure 5-86).

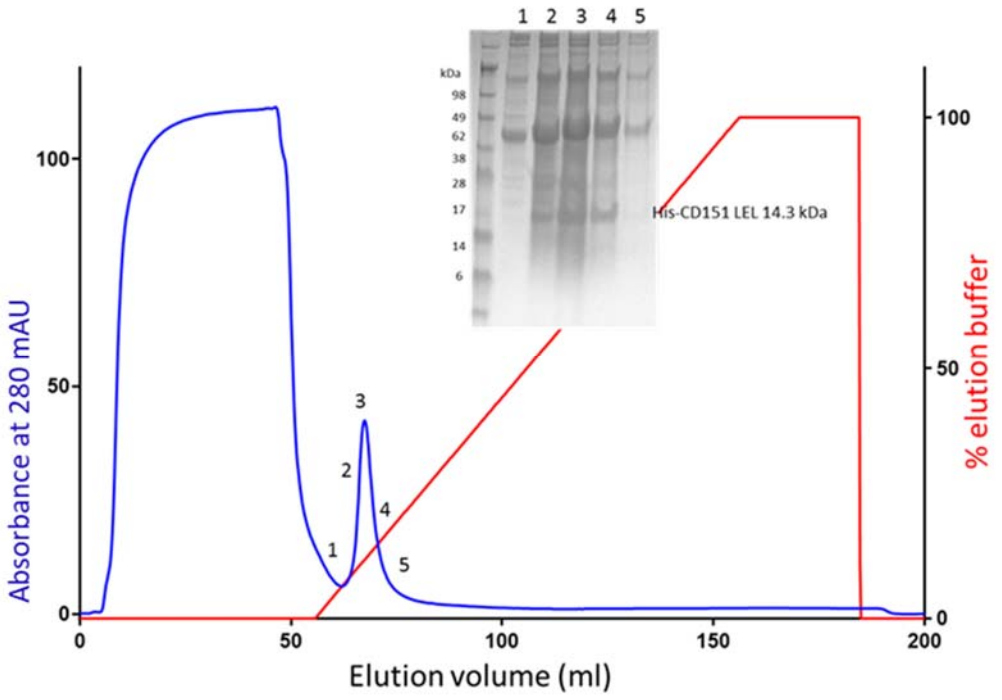


Figure 5-86 Chromatogram of the purification of media from *Sf*21 cells infected with baculovirus on a HisTrap. Fractions were analysed by SDS-PAGE (inset) the fractions numbered on the chromatogram correspond to the numbered samples on SDS-PAGE.

Although the majority of the protein seemed to be high Mw there was a band that possibly corresponded to the His_6-CD151 LEL at 14.3 kDa. 0.5ml of fraction 3 from the HisTrap purification was analysed on a Superdex 75 10 300 column (Figure 5-87).

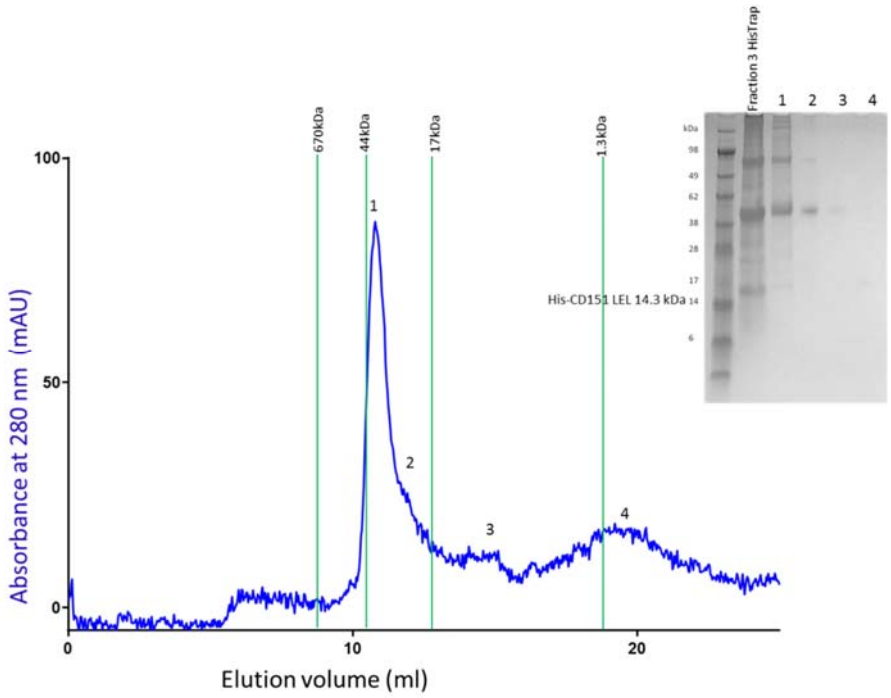


Figure 5-87 Chromatogram of the purification of His_6-CD151 LEL produced in *Sf*21 cells. The column was precalibrated with Mw standards, elution times shown in green. Fractions were analysed by SDS-PAGE (insert) and numbered fractions correspond to the numbered samples on SDS-PAGE.

The lower Mw protein from the affinity purification seemed to elute with the higher Mw protein in fraction 1. This suggested that the protein aggregated in solution and only became monomeric when disrupted by the SDS in the SDS-PAGE sample buffer.

Protein expression was repeated using *Sf*9 cells. *Sf*21 cells are a clonal derivative of *Sf*9 cells and there is evidence that *Sf*9 cells may support varying levels of expression and differential glycosylation to *Sf*21. [386]

*Sf*9 cells were grown and transduced with recombinant baculovirus as described previously. Media was harvested four days post transduction and dialysed against 50 mM HEPES, 100 mM NaCl, 10% glycerol pH 7.5. Media was purified by affinity purification using a HisTrap 5ml column and fractions analysed by SDS-PAGE (Figure 5-88).

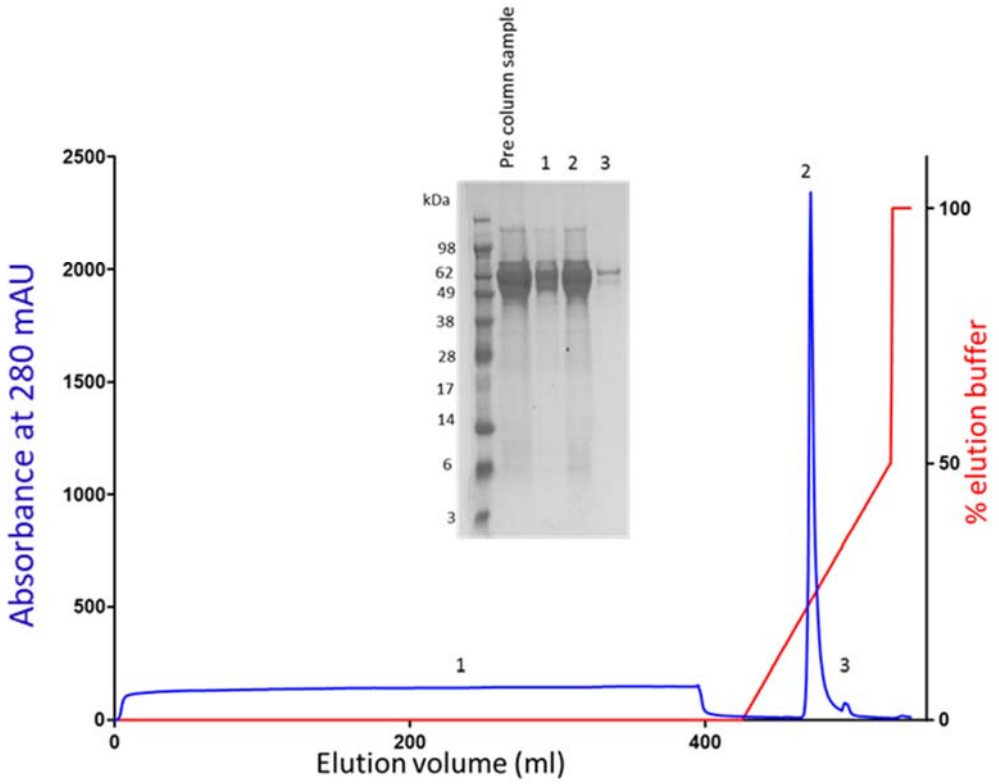


Figure 5-88 Chromatogram of the affinity purification of His_6-CD151 LEL from *Sf*9 expression. Fractions were analysed by SDS-PAGE (insert) and numbered fractions correspond to numbered samples on SDS-PAGE, a sample of the dialysed media prior to purification was also included.

There did not appear to be any material of the correct size, at 14.3 kDa, present in the sample. A sample of the most predominant peak, fraction 2, was analysed by mass spectrometry (Figure 5-89).

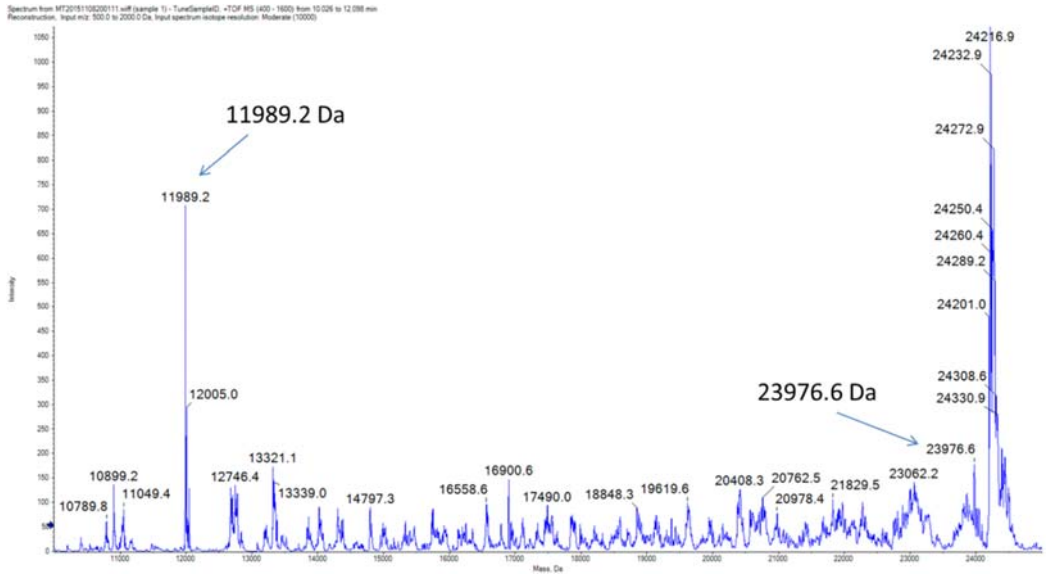


Figure 5-89 Deconvoluted spectra from MS-TOF analysis of fraction 2 from affinity purification of His_6-CD151 LEL produced in *Sf*9 cells showing a main protein peak of 11989.2 Da.

Mass spectrometry analysis revealed a main protein peak at 11.9 kDa. This is much smaller than the expected 14.3 kDa for His_6-CD151 LEL. It also does not correlate with any of the protein bands visible with SDS-PAGE (Figure 5-87, Figure 5-88). Carboxyl and cysteine proteases are known to be produced in virus infected insect cells and it is possible that the protein was digested during expression. [387] It is most probable that the truncation occurred from the C-terminus since the protein bound to the nickel affinity purification resin, indicating that the N-terminal His tag was still intact. If the protein was truncated from amino acid 200 the resulting fragment would be 11.9 kDa. This truncation removes the cysteine at 208 leaving an odd number of cysteines so there is the potential to form an intramolecular bond and result in the 23.9 kDa protein also observed in the MS-TOF deconvoluted spectra (Figure 5-89). His_6-CD151 LEL sequence analysed using PROSPER (Protease specificity prediction server), identified likely metallopeptidase-2 and -9 sites at that position. [388] Cleavage by endogenous proteases at these sites would result in a truncated protein of the size observed by mass spectrometry. EDTA and PMSF were routinely added to the harvested media. EDTA chelates metal ions required for metallopeptidase activity and PMSF is a serine protease inhibitor. However, addition of these protease inhibitors did not improve protein yield or prevent protein truncation, suggesting that the proteolytic activity occurred during protein expression. Viral transduction of the cells interferes with integrity of the cell membrane and cell viability is progressively reduced post transduction. It is likely that, owing to disintegration of cells subsequent to baculovirus transduction, cytoplasmic

proteases are constantly being released into the culture media during expression, leading to degradation of existing CD151 LEL protein.

Protein production in insect cells was not further pursued; however, further expression condition optimisation including the investigation of the addition of protease inhibitors such as E-64 to the culture media during expression and alternative protease substrates, such as foetal bovine serum, to the culture media may be advantageous. [389]

5.5. *MAMMALIAN EXPRESSION*

Prokaryotic expression systems often fail to generate correctly folded forms of eukaryotic proteins, with PTMs such as disulfide bonding and glycosylation being either very difficult or impossible to re-produce or mimic in bacterial cells. [390] While I experienced some success with producing disulfide bonded protein in *E. coli*, CD151 LEL contains one N-linked glycosylation site at position 159 which may be important for protein folding, conformation, stability and solubility. Mammalian cell expression systems are an obvious choice for producing mammalian proteins as they enable these crucial PTMs. The drawbacks with mammalian expression systems include expensive reagents, specialised media and facilities, consumables and the time required to establish and grow the cells and to express the recombinant protein. For these reasons mammalian cell expression was not the first choice for protein expression for this project; however, due to seemingly insurmountable difficulties encountered using bacterial expression systems, protein expression in mammalian cells appeared necessary to further the aims of this project.

A new CD151 LEL construct was synthesised encoding an Ig Kappa signal sequence, the CD151 LEL and a C-terminal His_6 tag for purification. The N-linked glycosylation process occurs in the lumen of the ER. For the nascent protein to be translocated from the cytosol to the ER the N-terminal of the protein must be recognised by the signal recognition particle while the protein is still being synthesised on the ribosome. The signal recognition particle then delivers the protein complex to the ER. [391] The murine Ig kappa chain leader sequence for protein secretion is a commonly employed signal peptide that is recognised by the signal recognition particle for effective translocation to the ER where the signal peptide is cleaved and was chosen for this construct. [392] Typically, CD151 would be anchored to the cell surface by virtue of its TM domain; however, for the purpose of purifying the LEL domain, the isolated domain was treated as a secreted protein, with the expectation that replacing the leader peptide with that of a constitutively secreted protein (Ig Kappa) would facilitate secretion and subsequent purification from the extracellular milieu. The cDNA encoding Ig Kappa-CD151 LEL-His_6 was cloned into the mammalian expression vector pcDNA3.1(+) using the cloning sites

Nhe1-Not1, resulting in a predicted recombinant protein Mw, once cleaved from the Ig Kappa signal peptide, of 13,692 Da.

HEK293F cells were grown in suspension to an OD_{600} nm of 1 and transiently transfected as described. Transfected cells were grown in two flasks, one harvested at day three and the second at day four. Media was centrifuged at 800 x g for 5 minutes to pellet the cell and the supernatant was concentrated in a 3 kDa cut-off spin concentrator to 10 ml. Samples from the expression were analysed by Western blot (Figure 5-90).

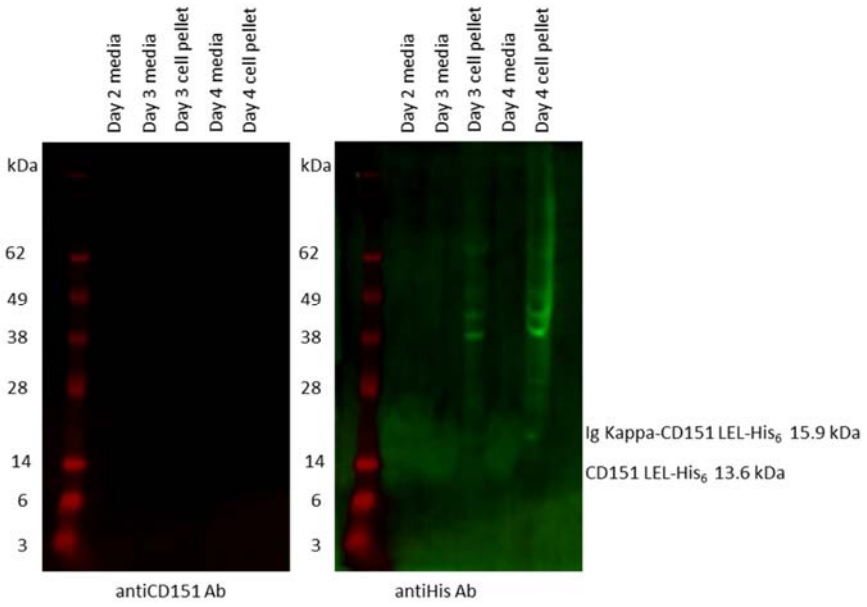


Figure 5-90 Western blot analysis of the media and cell pellets of HEK293 cells expressing CD151 LEL. The same blot was probed with multiple Abs: anti-CD151 Ab NOVUS 210127 mouse mAb conjugated to Alexa Fluor IR 700 (left) and anti-His mouse mAb conjugated with DyLight 680 (right).

There was no reactivity to the anti-CD151 Ab. His-tagged protein was detected by the anti-His Ab but not at the correct size for the secreted protein and only in the insoluble fraction. This indicates that the protein was expressed but was retained in the cytosol, likely owing to misfolding. (Figure 5-90). Mammalian expression of CD151 LEL was not further investigated.

5.6. *CONCLUSIONS AND FUTURE DIRECTIONS*

The principal aims of this project were to establish a reproducible methodology for the production, and purification to homogeneity, of the LEL domain of CD151 for structural biology and biophysical analysis; ultimately, providing the foundation for rationally developing small molecules to target CD151 function as a basis for developing drugs to treat cancers such as prostate cancer. Production of correctly folded recombinant CD151 LEL in quantities amenable to subsequent empirical experimentation proved to be a challenging task; principally owing to the propensity of cysteine residues in the CD151 LEL to promiscuously form non-native disulfide bonds. Several divergent approaches proved unsuccessful in promoting correct formation of the native disulfide bonds, and therefore in promulgating correct folding of recombinantly expressed CD151 LEL.

The most promising expression construct was the SUMO-tagged fusion protein. This was the only construct that produced stable protein when the fusion tag was removed by proteolytic cleavage. The arrangement of disulfide bonds in recombinant CD151 LEL produced using this method was, however, heterogeneous. There was evidence that some of the LEL formed the correct disulfide arrangement, which was encouraging. It is possible that further refinement of expression and purification conditions for this construct may yield a method for reproducibly expressing substantial quantities of correctly folded CD151 LEL protein that are amenable to structural and biophysical characterisation.

The preceding results beg the question; how have other research groups published studies of isolated, purified, CD151 LEL where the present experiments were unsuccessful in isolating high-quality recombinant protein? At face value this question is perplexing; however, on closer interrogation the authors of these studies frequently admit difficulties in expressing and purifying CD151 LEL, and 'quality control' assays to determine the fidelity of heterologously expressed CD151 LEL are rarely available in published data. Examples from the existing literature are as follows:

- Barreiro et al. produced CD151-LEL fused to GFP but stated "..a low rate of proper folding in solution which precluded its use in functional studies."[393]
- Scheltz et al. expressed CD151 in *D. discoideum* as a GFP fusion protein with a Kozak consensus sequence, an enhancer of translation [394], and adapted codon usage. Their discussion states "Yet, CD151 expression was poor." [395]
- Tarry et al. produced CD151 as a GFP fusion protein in *E. coli* and found that "…Tspan24-GFP possessed Mws of approximately …290 kDa.." [297] This is several times the expected Mw and suggests the protein is misfolded and forming oligomers.

- Ho et al. produced CD151 LEL as a GST fusion protein in *E. coli* and although they found the protein inhibited virus infection of adherent monocyte-derived macrophages, they also state that it was "...typically 5- to 10-fold less effective than CD63..". They confirmed correct conformation of the proteins by Western blotting using conformation-specific Ab's. [315] As my previous experiments demonstrate, CD151 LEL protein that reacts with conformation dependent Ab still contains multiple aberrantly folded protein species. It is possible that the results in this publication reflect the activity of a small, correctly folded, proportion of the overall purified sample of CD151 LEL.
- Skaar et al. attempted to express CD151 as a GFP fusion protein in *S. cerevisiae* using fluorescence measurement to detect protein production. They found that CD151 was "..not possible to produce in yeast using the conditions tested..".[396]
- Yang et al. were able to produce full-length CD151 using stable mammalian expression in NIH 3T3 cells as a N-terminal myc labelled fusion protein. Western blot analysis of their expression products shows a diffuse band at the expected 28 kDa but also a band at over 200 kDa which they describe as "protein multimerisation".[298] The diffuse band visible in Western Blot analysis suggests a protein with multiple conformations and is reflected in results I obtained (Figure 5-17,Figure 5-70,Figure 5-71).

Homology modelling of the CD151 LEL places the disulfide bonds in an area of protein that may be intrinsically disordered (Figure 4-6). I have also demonstrated that CD151 LEL promiscuously forms non-native disulfide bonds (Table 18); thus, further increasing the number of energetically favourable conformational states of the protein. Chaperone-mediated folding occurs when the chaperones bind to protein via exposed hydrophobic regions that would typically be buried in a folded, or lipid embedded, protein. [397] These hydrophobic regions cause instability when exposed to water molecules and are the main thermodynamic driving force for protein folding. [398]

Producing CD151 LEL in SHuffle cells as a SUMO fusion protein was successful enough to result in stable, soluble protein. The protein remained in solution and seemed to have secondary structure when analysed by CD (Figure 5-74); however, analysis by ion mobility showed the protein was present in multiple conformations.

Analysis of the CD151 LEL sequence using ExPASy ProtScale [399] and selecting the Kyte & Doolittle hydropathy scale produces a plot of the hydrophobicity score vs amino acid (Figure 5-91). [284]

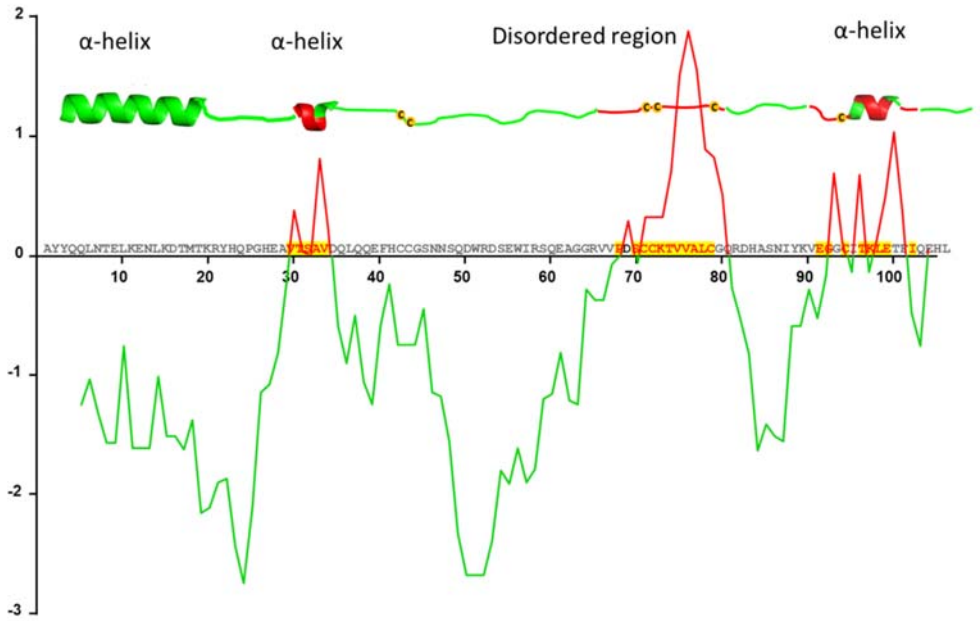


Figure 5-91 The Kyte and Doolittle hydrophobicity plot overlayed with the protein sequence and predicted secondary structure of the CD151 LEL. Hydrophobic regions with a value >0 are highlighted in red, the location of the cysteine residues on the structure map are highlighted by orange circles.

Examining the position of the hydrophobic amino acids on the homology model and overlaying the proposed secondary structure with the hydropathy plot reveals that all of the cysteines are located in regions of the protein that are predicted to be disordered and four are also in hydrophobic regions. Of the disulfide bonded peptides of CD151 LEL, produced by LysC digestion and identified by the protein metrics Byonic™ software, four of the five pairs were of peptides from the hydrophobic regions (Figure 5-76). It is possible that, energetically favourable, promiscuous interactions between hydrophobic regions of the protein, enabled by the inherent flexibility of the disordered sequences in these regions, allows formation of energetically stable non-native protein species.

Comparison of the hydrophobicity, the helical segments and the location of the disulfide bonds in the CD151 LEL with the published structures of the CD81 and sm-TSP-2 LELs reveals that although CD81 contains more hydrophobic amino acids, they are mostly located in the defined helical areas (Figure 5-92). Sm-TSP-2 contains fewer hydrophobic amino acids and is a mostly helical protein. The comparison suggests that the CD81 and sm-TSP-2 LELs have much less flexibility in their conformational arrangement and are less prone to misfolding than the CD151 LEL.

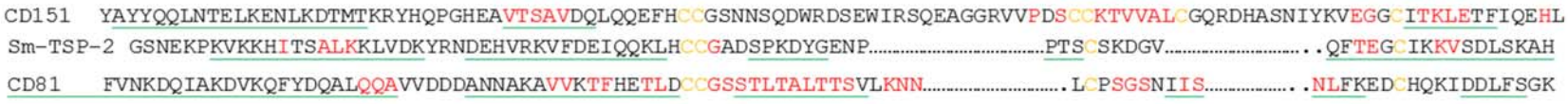


Figure 5-92 Comparison of the hydrophobic (red) and helical (green underline) content and the location of the cysteines (orange) of the CD151, sm-TSP-2 and CD81 LELs.

Searching the PDB for CD81 structures returns sixteen results: the original LEL structure from 2001 (PDB ID: 1G8Q) and the full-length structure from 2016 (PDB ID: 5TCX) being the top two search results. Of the remaining fourteen structures, one is in complex with magnesium ions (PDB ID: 3X0E); one is a dimeric form of the original LEL structure (PDB ID: 1IV5), which has since been discounted as a crystallographic anomaly. [309] Two CD81 LEL structures are non-human (PDB ID: 3X0E, 3X0G); five are in complex with an Ab or Ab fragment (PDB ID: 5DFV, 5DFW, 6EJG, 6EJM, 6EK2) and the remaining five structures are examinations of different crystallographic packing arrangements of the same protein construct (PDB ID: 5M2C, 5M33, 5M3D, 5M3T, 5M4R). In the 1G8Q structure two of the helices form a cleft-like motif within a region identified as the HCV E2 glycoprotein binding site. [200] Structure 1IV5 is considered to be a 'closed' form of the CD81 structure as the cleft is absent. [200] However, molecular dynamics studies by Neugebauer et al. suggest that the cleft observed in the open 1G8Q conformation is a crystallographic artefact and not the physiologically relevant conformation. [400]

Overlaying structures 5TCX, 1IV5 and 3X0E, as the most physiologically relevant forms, shows some variation in the third and fourth helices that contain the most hydrophobic amino acids (Figure 5-93).

Figure 5-93 Overlay of CD81 LEL crystal structures 5TCX (green), 1IV5 (pink), 3X0E (blue) with disulfide bonds depicted (orange). The sequence of the CD151 LEL (top) indicates the regions of helix (green underline), hydrophobic amino acids (red) and cysteines (orange). The blue lines indicate the protein sequence that corresponds to the segment of structure with the highest structural variability. 5TCX is the structure of full length CD81 but has been truncated to the LEL for this image.

A recent study by Cunha et al. used crystallographic molecular dynamics to present six different conformations of the CD81 LEL demonstrating inherent plasticity in this molecule in the third and fourth helices (Figure 5-93). [401] The authors postulate that the disulfide bond at C157-C175 is disrupted when CD81 binds to the E2 domain of HCV allowing fusion of the virus-receptor complex. Conformational flexibility of CD81 and potential rearrangement of disulfide bonds is in keeping with the formation of heterologous protein species upon the expression and purification of the CD151 LEL, and suggests that a propensity to populate a wide conformational space may be a characteristic feature of tetraspanin family proteins.

Separating conformationally correct, heterologously expressed, CD151 LEL protein from the complex environment of incorrectly folded protein seems a complicated endeavour.

It may be possible to separate protein species using hydrophobic interaction chromatography, however very fine tuning of the conditions would be required to separate such similar proteins. Affinity chromatography using a conformation-specific Ab immobilised on the resin may also separate the proteins. However, this method would not be cost effective for producing high quality folded protein on a large scale. Further exploration of expression technologies that are amenable to producing conformationally correct CD151 are likely the best approach to producing protein for downstream structural biology and biophysical experiments, as such a method would negate many of the inefficient and costly laboratory handling techniques involved in protein purification.

Expression of the CD151 LEL in mammalian cells was briefly explored but due to time and budgetary constraints was not thoroughly investigated. Using mammalian expression was unsuccessful and all the CD151 LEL produced was insoluble in the cell lysate, indicating the protein was misfolding and aggregating in the ER. The construct used for mammalian expression contained an Ig kappa signal leader sequence to facilitate translocation of the nascent fusion protein into the ER. As the full-length protein, including the Ig kappa leader sequence, was detected by Western blot it seems the protein was not being translocated. Ig kappa is a commonly used leader sequence from the murine immunoglobulin kappa light chain; testing other signal sequences may be beneficial and improve the result. Often the best choice for signal sequence is the proteins native signal peptide. However, analysis of the sequence of CD151 by ExPASy SignalP 4.1[402] server did not identify a clear signal sequence (Figure 5-94).

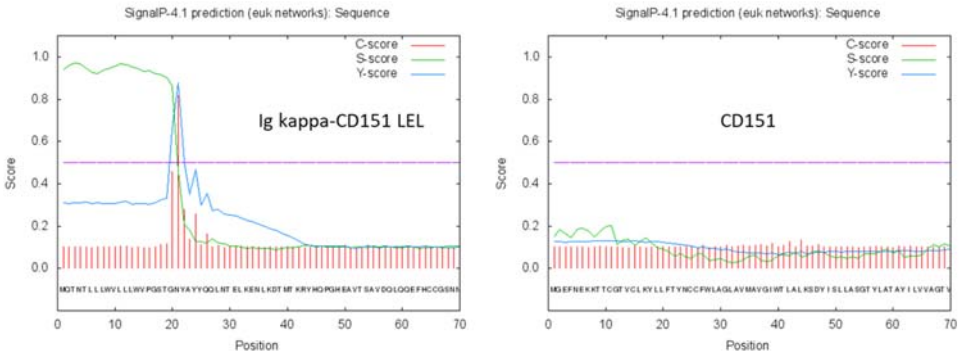


Figure 5-94 Analysis of the signal peptide and cleavage site of Ig kappa-CD151 LEL (left) and the full length CD151.

Figure 5-94 shows the C-score, cleavage site score, (red) is high at the position immediately after the cleavage site, the S-score, signal peptide score, (green) distinguishes signal peptides from mature protein and the Y-score, combined cleavage site score, (blue) is the geometric average of the C-score and the slope of the S-score

and is a more accurate prediction of the true cleavage site. A non-secretory protein will have low scores close to the negative target value of 0.1.

The signal peptide analysis shows that the Ig kappa signal peptide used should have been functional. For multiple membrane spanning proteins that have an intracellular N-terminus, the signal peptide is usually incorporated in the polypeptide signal-start sequence that commences the transfer of the protein through the membrane. Subsequent domains typically contain either another start-transfer signal sequence or a stop-transfer sequence that anchors the protein in the membrane. [403] As such, I expected to find a signal peptide at the N-terminus of CD151 and although the S-score for this region is slightly above the rest of the protein it is not adequate to distinguish a signal peptide. (Figure 5-94) It is possible that mammalian expression of the CD151 LEL would be more successful incorporating the native N-terminal sequence as opposed to an artificially introduced signal peptide, if one could be identified.

Further exploration of mammalian expression is also warranted due to glycosylation of the CD151 LEL. It is possible that glycosylation contributes to the overall stability and solubility of CD151 LEL. While, owing to the heterogeneous nature of glycosylation, glycosylated protein is not ideal for crystallographic studies, mammalian cell strains such as HEK GnTi (ATTC: CRL-3022) cells limit glycosylation and may therefore be a viable vehicle for CD151 LEL expression. Possible future strategies for expressing and purifying CD151 LEL from mammalian cell culture would benefit from a systematic approach to identifying protein constructs that produce correctly folded, soluble protein in small-scale transient transfections, before progressing to large scale expression cultures.

Recombinant CD151 is available commercially from Novus Biologicals as a GST fusion protein produced in an *in vitro* wheat germ expression system, however the datasheet states that conformation and functionality are not validated. [404] *In vitro* protein synthesis, or cell-free protein synthesis, utilises the biological machinery of protein synthesis such as ribosomes, aminoacyl-tRNA synthetases, translation initiation and elongation factors etc., in a cell extract that is free of cell walls, genomic DNA and other cellular debris. [405] Additives such as buffers and proteins designed to correctly fold proteins can increase the yield of soluble and active protein. This expression method was not explored during this project but may be an alternative method for producing conformationally correct CD151 LEL.

The results presented in this chapter provide an extensive study to recombinantly producing human CD151 LEL protein in prokaryotic cells, and show that expressing and purifying CD151 LEL from bacterial cells is possible, if not without complication.

CHAPTER 6. DETERMINING THE 3D ATOMIC STRUCTURE OF CD151 LEL

6.1. *INTRODUCTION*

Structure-based drug design (SBDD) utilises structural information to facilitate virtual screening of large databases of commercially available compounds to identify potential hits that dock into the selected region of the 3D structure. The compounds are then scored and ranked based on their predicted interactions and fit with the target site. The best hits are purchased and tested using various biochemical screening methods, such as NMR, SPR and ITC.

If an experimentally determined structure is not available a homology model can be used for drug design. It has been shown that the most successfully predicted compound-protein complexes are modelled on templates with a sequence identity higher than 30%. [406] For CD151 LEL the only structures available for homology modelling share less than 20% sequence identity, so homology modelling and results for virtual screening of the model would likely prove unreliable.

X-ray crystallography is the most established method for determining the 3D structure of proteins, with over 90% of the structures deposited in the Protein Data Bank (PDB) in 2017 being solved using this technique. [407] For this method the soluble protein is purified to a high level and coaxed in to forming protein crystals. The crystals are then probed using an intense beam of monochromatic X-rays which results in a light diffraction pattern that reflects the 3D arrangement of atoms in the protein crystal. This pattern can then be used to calculate an electron density map; essentially, a recreation of the location of atomic-level features of the protein crystal. In turn, this map can be used to infer the position of individual atoms, and the bonds that connect them, to generate a 3D, atomic-scale, model of the protein structure. [408, 409]

Establishing suitable conditions to coax proteins to form high quality protein crystals is, fundamentally, an empirical process that often proves to be difficult and time-consuming. Even when homogenous soluble protein is available, and trials result in formation of protein crystals, the crystals may not be of sufficient quality to produce a high-resolution diffraction pattern. Fundamentally, crystallisation of a protein from solution requires the purified protein to be in a supersaturated state. Supersaturation refers to the concentration of protein that exceeds the capacity of the solvating solution to maintain soluble protein, at which protein molecules precipitate from solution as either disordered aggregate or form an ordered crystalline lattice (Figure 6-1). [410] Crystallisation conditions are typically probed with respect to pH, temperature, protein concentration and buffer composition to identify the conditions that are favourable for promoting crystallisation of a protein of interest. Under suitable conditions individual

protein molecules arrange into a non-covalently packed, repeating array that propagates to form macroscopic crystals. Empirical testing of many conditions is usually required before an appropriate crystallisation condition for the protein of interest is found. The range of conditions that influence crystal formation is vast and there are no set rules to produce high quality crystals; however, broad screening for promising conditions, followed by optimising crystal growth by fine-tuning conditions can lead to the production of diffracting crystals. [411]

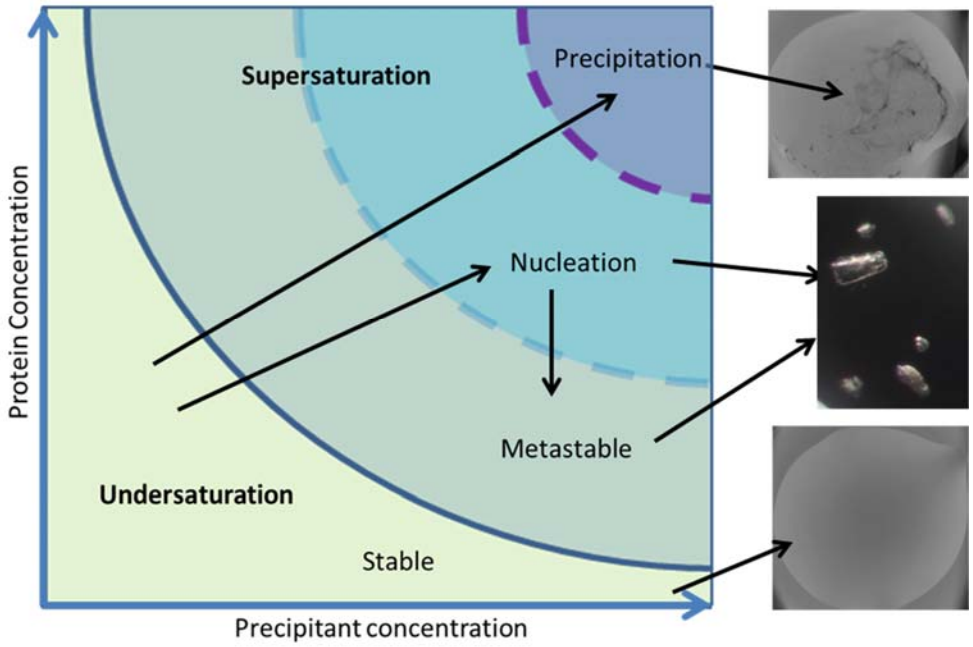


Figure 6-1 Phase diagram for the crystallisation of proteins. The undersaturation region denotes the concentration of protein and precipitate conditions where the protein is stable and stays in solution (below dark blue line).

As the concentration of protein and precipitate increases and the condition becomes supersaturated (above dark blue line) the protein stability decreases until nucleation occurs and crystals form (above blue dashed line). As crystals form the concentration of protein in solution decreases and the condition reverts to the metastable phase, where no new crystals are nucleated, but the existing crystals grow larger. If the concentration of protein and precipitate increases beyond the nucleation zone, the protein will precipitate as amorphous aggregate (above the dashed purple line). Pictures on the right show examples of CD151 LEL in each condition: (Top) amorphous precipitate, (Middle) a nucleation event that has led to crystal formation and the metastable condition that allowed crystal growth, (Bottom) protein and precipitate in the stable, undersaturation state i.e. soluble. This figure represents a theoretical reduction of the crystallisation process; other factors including temperature, pH, the addition of different salts and

precipitating agents strongly influence crystal formation by altering the nature of the phase diagram.

There are many techniques available for producing protein crystals. The most widespread, and effective, technique for protein crystallisation is vapour diffusion in either hanging-drop or sitting-drop formats (Figure 6-2). Both techniques work with very small amounts of protein and give comparable results. The sitting-drop format is widely used owing to the ease with which it can be adapted for high-throughput, robotic automation of establishing experiments to determine appropriate crystallisation conditions. These experiments can then be coupled to robotic imaging systems to further streamline the process of identifying conditions suitable for protein crystallisation.

The practical considerations of protein crystallisation using vapour diffusion are relatively straight forward, and are explained as follows: a small volume of buffer containing soluble protein and precipitant are sealed in a closed chamber that also contains a larger reservoir containing buffer and precipitants in higher concentrations. As water-vapour equilibration between the protein solution and the osmotically more concentrated reservoir solution occurs, progressive concentration of protein and buffer constituents in the protein solution may lead to conditions where the protein forms crystals. (Figure 6-2). [412]

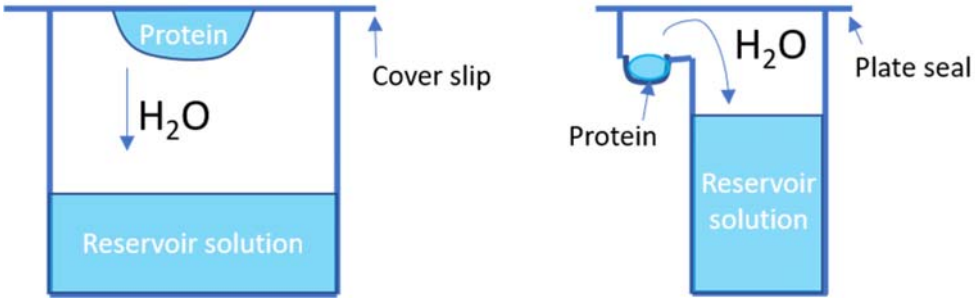


Figure 6-2 Schematic of two vapour diffusion techniques. Hanging-drop (left) where a drop of protein is placed on a cover slip and inverted over a reservoir of buffer and precipitant. Sitting-drop (right) where the drop is placed on a pedestal separate from the reservoir. Both methods require a sealed system, isolated from the external environment, to permit equilibration between the drop and the reservoir.

There are, to date, no reports of CD151 crystal structures. At the onset of this project the only published human tetraspanin structures were of CD81 (see 4.2.1), the LEL of

which is smaller and contains only four cysteines so is an inadequate model for CD151 LEL. Virtual screening is a crucial step in the search for small molecule inhibitors to refine and focus the compounds selected for screening. This step is futile without reliable structural data. Solving the 3D structure of CD151 LEL is a fundamental requirement in the drug development process.

6.2. *METHODS*

6.2.1. Pre-Crystallisation Test

The Pre-Crystallisation Test (PCT™) (Hampton Research) is used to determine the protein concentration most likely to result in crystal formation. [413, 414] The test provides a way to evaluate the proteins' sensitivity to salt and precipitant concentrations and if the protein is at a concentration likely to be successful for subsequent crystallisation screening.

PCT Formulation:

A1: 0.1 M Tris-HCl pH 8.5, 2.0 M Ammonium sulfate
B1: 0.1 M Tris-HCl pH 8.5, 1.0 M Ammonium sulfate
A2: 0.1 M Tris-HCl pH 8.5, 0.2 M Magnesium chloride, 30% w/v Polyethylene glycol 4,000
B2: 0.1 M Tris-HCl pH 8.5, 0.2 M Magnesium chloride, 15% w/v Polyethylene glycol 4,000

Table 19 Possible PCT results and recommended action as per the Hampton Research method. [310]

PCT Reagent A1/B1 results	PCT Reagent A2/B2 Results	Recommended Action
Heavy amorphous precipitate	Heavy amorphous precipitate	Dilute sample 1:1, repeat test
Clear	Clear	Concentrate sample to half the volume, repeat test
Light granular precipitate	Clear	Perform Screen
Clear	Light granular precipitate	Perform Screen
Heavy amorphous precipitate	Light granular precipitate	Perform Screen
Heavy amorphous precipitate	Clear	Perform PCT with B1 & B2/perform diagnostic testing
Clear	Heavy amorphous precipitate	Perform PCT with B1 & B2/perform diagnostic testing

6.2.2. Crystallisation trials

6.2.2.1. Surface Entropy Reduced Mutant

Almost 400 structures of MBP fusion proteins solved by X-ray crystallography have been deposited in the PDB to date. The structures comprise a range of conformations of the fusion protein-linker-MBP arrangement and in some cases the use of a surface entropy reduced mutant MBP seems to have facilitated the formation of crystals. In the case of the structure of APO MBP-MCL1 (PDB ID: 4WMS) the protein had been successfully expressed as a fusion with SUMO, TRX and MBP yet only the MBP yielded crystals. [415] This approach seemed promising as a means to facilitate the crystallisation of the CD151 LEL.

Of the five MBP-SER mutants [373] fused with CD151 LEL trialled, (page 173) only MBP(A) CD151 LEL expressed and purified to a quality suitable for crystallisation. The protein remained stable at concentrations up to 29.5 mg/ml. 96-well sitting-drop vapour diffusion crystallisation plates with a wide variety of conditions were set up as per Table 20.

Table 20 CD151 LEL crystallisation conditions trialled in 96-well sitting drop vapour diffusion plates. All crystallisation plates were housed at 22°C and duplicates of the Anatrace plates were housed at 4°

Protein	mg/ml	Protein Buffer	Qiagen Screen	Molecular Dimensions Screen	Anatrace Screens	Jena Bioscience	Inhouse and C3
MBP(A) CD151 LEL	17	20 mM Na Citrate pH 5	The NH_4SO_4 Suite		Microlytic MCSG1		
	20	PBS			Microlytic MCSG1		
	29.5	70 mM HEPES, 100 mM NaCl, 5% glycerol pH 7.5	The PEGs Suite	MemStart and MemSys HT-96	Microlytic MCSG1	JBS Screen JCSG++	Custom
His_6-SUMO-CD151 LEL	10	70 mM HEPES, 100 mM NaCl, 5% glycerol pH 7.5	The NH_4SO_4 Suite		Microlytic MCSG1		
Refolded CD151 LEL	1.8	70 mM HEPES, 100 mM NaCl, 5% glycerol pH 7.5	The NH_4SO_4 Suite		Microlytic MCSG1		
CD151 LEL cleaved from His_6-SUMO	3.8	70 mM HEPES, 100 mM NaCl, 5% glycerol pH 7.5	The NH_4SO_4 Suite	MemStart and MemSys HT-96	Microlytic MCSG1		Cubic_C3, Shotgun_C3

6.2.2.1. His_6-SUMO-CD151 LEL

The SUMO tag is popular for protein purification but is usually cleaved off prior to crystallisation. [416] A search of the PDB, however, reveals over 100 SUMO fusion protein structures. The His_6-SUMO-tagged CD151 LEL was highly soluble and remained soluble at concentrations up to 10 mg/ml. It is possible that a highly soluble tag such as SUMO, fused to an insoluble protein merely results in a level of solubility that allows expression and purification but that the fusion protein becomes insoluble once the tag is cleaved. To circumvent this potential problem a standard crystallisation screen was carried out to assess the viability of crystallising the protein with the tag in-place (Table 20).

6.2.2.2. CD151 LEL

Purified CD151 LEL, produced by either refolding or after the cleavage of the His_6-SUMO tag, was less stable than the tagged versions of the protein and concentrations of >5 mg/ml were difficult to attain. Evaluation of the protein by PCT (section 6.2.1) indicated

that a lower concentration may be viable for crystallisation. Several commercial crystallisation screens were trialled. Based on results from the commercial screens, further condition optimisation was trialled using 24-well VDX hanging drop plates (Table 20).

6.3. *RESULTS*

6.3.1. PCT

The PCT test was performed, as described above, using CD151 LEL produced by the refold method at 3.8 mg/ml.

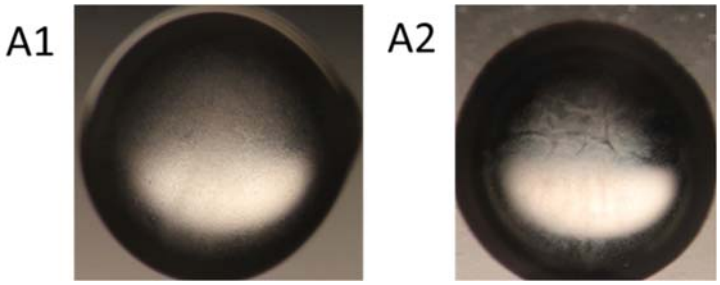


Figure 6-3 PCT test of CD151 LEL at 3.8 mg/ml. Protein was mixed with equal volume (1 µl) of reagent A1 (left) and A2 (right) on a coverslip and mounted onto a VDX plate with 1 ml of the same reagent in the well. The drop with A1 shows heavy precipitate and the drop with A2 shows heavy amorphous precipitate.

Using Table 19 to evaluate the drops shown in Figure 6-3 suggested that the protein was too concentrated. As the result for A1 was somewhat ambiguous I repeated the test using reagents B1 & B2. (Figure 6-4)

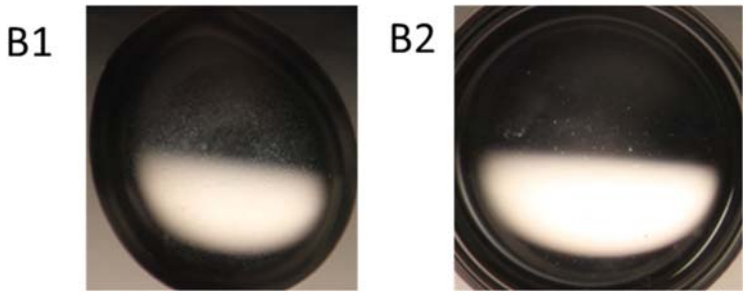


Figure 6-4 Pictures of PCT conditions of CD151 LEL at 3.8 mg/ml mixed with equal volume (1 µl) of reagent B1(left) and B2 (right) on a coverslip and mounted onto a VDX plate with 1ml of the same reagent in the well. The drop with B1 shows light precipitate and the drop with B2 shows light precipitate/clear.

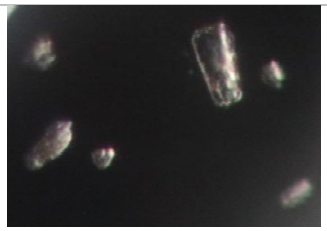
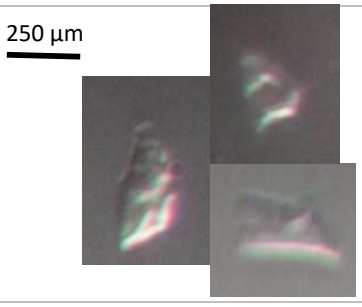
Referring to Table 19 again, the result suggested that the protein concentration was appropriate to perform crystallisation screening.

6.3.2. Crystallisation trials of MBP(A)-CD151 LEL and His_6-SUMO-CD151 LEL

Crystallisation screening of MBP(A)-CD151 LEL and SUMO-CD151 LEL was performed a using Crystal Gryphon robotic liquid handling system. 96-well plates with 0.2 µl vol. protein solution and 50 µl reservoir solution were prepared. Reservoir conditions were taken from 17 commercial protein crystallisation screens. Many of the drops containing MBP(A)-CD151 LEL formed precipitate and/or phase separation. Phase separation is a possible indicator of conditions that may be favourable to protein crystallisation, and a custom screen was designed to explore the chemical space around conditions under which phase separation occurred, with a higher concentration of protein to encourage nucleation events. This approach was unsuccessful; however, and did not result in protein crystals.

Only 3 of the 1632 conditions tested resulted in protein crystal formation. These were all with His_6-SUMO-CD151 LEL protein using the Qiagen NH_4SO_4 Suite. The details of the crystallisation conditions are summarised in Table 21.

Table 21 Conditions under which His_6-SUMO-CD151 LEL formed crystals using the Qiagen NH_4SO_4 Suite.

	Crystal	Condition
A		1 M lithium sulfate, 1.6 M ammonium sulfate
B		0.2 M NaCl, 1.6 M ammonium sulfate, 0.1 M sodium HEPES pH 7.5
C		1.8 M ammonium sulfate, 0.1 M sodium MES pH 6.5

From Table 21, crystals in group A formed from a light, granular precipitate with a plate like form and sharp, irregular facets. Crystals in group B were irregular and rounded, and the crystals in group C were large and irregular.

Five of the above crystals were transferred to a cryogenic solution comprising 20% ethylene glycol in their respective well solution before being frozen in liquid nitrogen. The frozen crystals were taken to the Australian Synchrotron for X-ray diffraction data collection on the MX2 beamline. None of the diffraction patterns collected from these crystals were suitable for further analysis.

6.4. *DISCUSSION*

The two most successful precipitants used in protein crystallisation are ammonium sulfate and polyethylene glycol (PEG) in the approximate concentration ranges 1.0 -2.5 M and 10 – 35 % respectively. The PCT contains ammonium sulfate and PEG at either a low or high concentration within these ranges. If the protein forms a heavy amorphous precipitate at the lower precipitant concentration ranges, then most of the results from the screening experiments would also give a heavy precipitate. This indicates that the protein concentration is too high. Likewise, if they were clear at the higher precipitant concentration ranges, then most of the results from the screening experiments would also give clear drops, indicating the protein concentration is too low. However, this is a simplified test and other factors such as pH, salt concentration and temperature can also influence protein crystallisation. The results of the PCT show that CD151 LEL forms heavy precipitate in the presence of ammonium sulfate. This was reflected in the conditions under which crystals formed, i.e. contained either 1.6 or 1.8 M ammonium sulfate.

As was observed during protein characterisation, the MBP(A)-CD151 LEL only contained one or two disulfide bonds and the CD151 LEL segment of the protein may not have been correctly folded. This would prevent crystallisation. The most promising results were from the His_6-SUMO-CD151 LEL, which at the time seemed to be correctly folded. However, recent data from my mass spectrometry studies indicate that non-native disulfides are forming which would increase heterogeneity in the protein folding and prevent crystallisation. Other sources of protein heterogeneity may be from protein degradation and aggregation. The absence of the N-linked glycosylation at N159 may also be hindering crystallisation. Evidence suggests that CD151 lacking this glycosylation is able to bind integrin indicating the protein is functionally correct, however, the glycosylation may be required for conformation and stability. [287] Despite screening over 2500 conditions with four different protein constructs as well as optimising crystallisation conditions manually in hanging drop trays, a favourable condition has not yet been identified. The results obtained are, however, encouraging and provide the basis for future crystallisation trials.

CHAPTER 7. CD151 LEL DIRECT BINDING ASSAY

7.1. *INTRODUCTION*

It is well established that the CD151 association with integrin α3β1 plays an integral role in cancer progression. [179, 180, 213, 214, 417, 418] Inhibition of this interaction by targeted Abs or gene deletion has been demonstrated to inhibit tumour progression. [175, 419, 420] The discovery of a synthetic compound that could specifically inhibit this interaction, yet allow normal physiological function of CD151, would have huge potential for the prevention of cancer metastasis.

Small molecule drugs have many advantages over biological drugs like Abs and Ab fragments due to their size, defined character, homogeneity, low immunogenicity, stability and low manufacturing costs. [421, 422] Typically, small molecule structure-based drug design (SBDD) targets specific protein receptors to modulate biological function. Discovering potentially active compounds requires knowledge of either the structure of the target or of a known ligand. This information is used to perform a "virtual" screen of known compound libraries, in which the structures of millions of drug-like molecules are assessed for their ability to bind the target protein or mimic the known ligand. Putative ligands from the virtual screen are then selected for experimental screening by various molecular interaction techniques such as SPR, DSF and NMR or functional assays where they are available. Binding compounds are chosen as hits and analogues of these compounds are developed to improve binding affinity and solubility. Promising hits, together with their analogues, are then developed into lead families of compounds. This method of SBDD has been the basis for many industrial and academic drug discovery projects since the 1990's. [423]

In the absence of relevant structural information on which to base a virtual screen, an alternative strategy is required. Fragment based drug discovery (FBDD) has become a successful technique for the identification of hit compounds on which to base further screens. Rather than targeting a known structure, FBDD uses a high-throughput binding assay, such as SPR, DSF or NMR, to identify 'fragments' – small molecules that represent typical substructures of drug-like molecules – as weak ligands for a target protein. Fragments that show binding are then elaborated to develop drug-like molecules with improved binding affinity and provide the basis for traditional drug development through medicinal chemistry. Fragment libraries are designed to both maximise chemical space coverage and enable rapid hit optimisation. [424, 425] Due to the difficulties associated with the structural determination of CD151 LEL an FBDD approach was employed to identify binding fragments as the basis for drug development.

7.1.1. High throughput screen using surface plasmon resonance

SPR is a phenomenon that occurs at the interface between media with two different refractive indices, to detect and measure analyte-ligand interactions. The ligand is immobilised on a thin inert metal film, usually gold, encased within a sensor chip. The opposite side of the chip acts as a prism with a higher refractive index; this side of the chip is illuminated by polarised light under conditions of total internal reflection. The light generates an evanescent wave field across the interface into the medium of lower refractive index - the liquid compartment. Any changes of mass due to the binding of soluble molecules to the immobilised ligand in the liquid compartment induces a change in the refractive index, which leads to a change in the angle of the reflected light. This results in absorption of energy via the evanescent wave field and a drop in the intensity of the reflected light is observed. This real-time measurement is recorded by an arbitrary resonance unit (RU) where 1 RU = 1pg of protein per mm^2 (Figure 7-1). [426, 427]

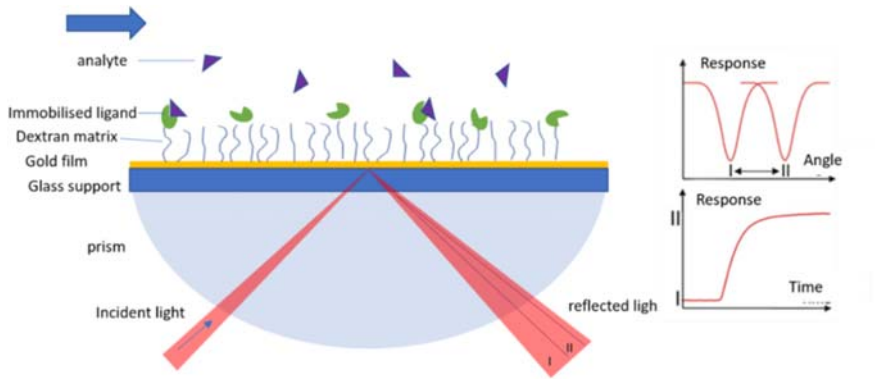


Figure 7-1 Schematic of the SPR phenomenon. The picture shows the gold film which supports the dextran matrix that the ligand (green) is immobilised on. The analyte (purple) flows across the chip surface and any binding to the ligand results in a change in the angle of the reflected light (red). The change in response is measured in response units and reflects the real-time association and dissociation of the analyte to the ligand. Figure adapted from Biacore™ Assay Handbook, GE Healthcare.

The use of SPR as the primary method for screening fragment libraries is common, as the technique allows automated high-throughput analysis of hundreds of fragments per day with very little protein consumption. SPR can be used to both screen and confirm binding specificity, as well as providing information on the affinity and kinetics of the binding interaction. [427-429] For FBDD by SPR the ligand bound to the sensor chip is typically the protein target of interest, and the analytes tested are the library of fragments.

7.1.2. Fragment library

The Monash Fragment Platform (MFP) at the Monash Institute of Pharmaceutical Sciences (Parkville, Victoria, Australia) provides collaborative and commercial researchers with access to a high-quality, well characterised fragment library. [430] The library has been designed to maximise chemical space coverage and rapid hit optimisation and has been extensively characterised by physiochemical methods to screen the fragments for size, solubility and purity. The fragments are also screened using the PAINS (Pan Assay Interference Compounds) filter to eliminate undesirable reactivity. [431] The use of such well-defined fragments alleviated the necessity to perform basic quality assessments such as clean screening (see section 3.3.8). As such, the screen development on CD151 LEL focused on protein specific issues such as immobilisation.

7.1.3. Protein immobilisation

There are many techniques for attaching the ligand to the chip surface. [257] Covalent coupling to the dextran matrix on the sensor chip surface is a common method for ligand attachment, which utilises free reactive groups on the protein surface such as amines or thiols, and uses chemical reactions to join to groups on the matrix surface. Although this method requires limited chemical modification of the ligand, most proteins can tolerate this and will not lose activity. Covalent coupling results in a stable surface that can be regenerated to remove bound analyte and reused many times. Heterogeneity in orientation of proteins immobilised by this method is likely, and if the available reactive groups are located near the binding site then steric hindrance of analyte binding becomes an issue.

Capturing ligands via a specific interaction, such as an Ab immobilised on the chip surface or a tag on the protein terminus will avoid steric effects on the binding site and give homogeneous orientation of the ligand. However, the affinity for the ligand and the capturing molecule needs to be high enough so that the ligand does not dissociate from the surface during the analysis. [257] Finding an appropriate immobilisation or capturing method is an important optimisation step for quality analysis of binding interactions. While my original intention was to capture the CD151 LEL on a streptavidin surface via a biotinylated N-terminal Avi tag, protein expression and purification difficulties made this option unviable. The proposed structure of CD151 LEL from my homology model and the available His purification tag allowed the examination of alternative immobilisation techniques.

7.1.3.1. Amine coupling

For amine coupling the surface of the chip is activated with a mixture of 0.4 M 1-ethyl-3-(3-dimethylaminopropyl)-carbodiimide (EDC) and 0.1 M N-hydroxysuccinimide (NHS)

in water to give reactive succinimide esters in the carboxymethylated dextran matrix. When ligand is passed over the chip surface the esters react with primary amines on the ligand surface to covalently link the ligand to the dextran matrix. Once the desired immobilisation level is reached any uncoupled succinimide esters are blocked with 1 M ethanolamine-HCl pH 8.5. [257]

The dextran matrix on the sensor chip surface is negatively charged in buffer conditions with pH above 3.5. By diluting the ligand in buffer with a pH between 3.5 and the isoelectric point (pI) of the protein, the protein becomes positively charged and the electrostatic attraction between the chip surface and the ligand pre-concentrates the ligand within the dextran matrix to increase the efficiency of covalent coupling. The optimal pH for ligand coupling will induce enough positive charges to preconcentrate the protein in the dextran matrix without compromising the stability of the protein. At low pH the covalent cross-linking efficiency can also be reduced, such that the simple approach of using a pH 3.5 buffer may be counter-productive.

The CD151 LEL contains six lysine residues that could potentially link the protein, via amine coupling, to the dextran matrix on the sensor chip surface. The homology model of the CD151 LEL shows all 6 lysines available on the surface of the protein with none in close proximity to the integrin binding site (Figure 7-2).

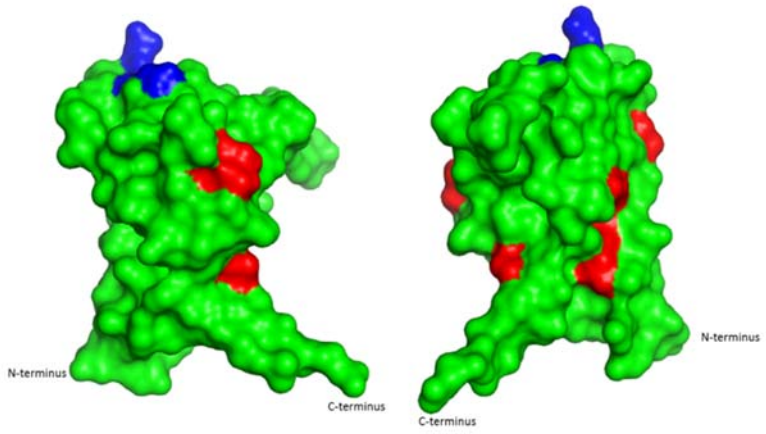


Figure 7-2 Front and rear view of the surface of the homology model of CD151 LEL with lysine residues highlighted in red and the integrin binding site shown in blue.

7.1.3.2. His capture

The patents on dextran surfaces used in the Biacore began to expire in 2010; since then there has been a huge growth in the manufacture of SPR sensor chips. [432] One of the manufacturers that has been very successful in developing alternative sensor chips is XanTec bioanalytics GmbH in Dusseldorf, Germany. The sensor chip matrix produced by

XanTec bioanalytics GmbH uses a linear, synthetic polycarboxylate which has improved signal-to-noise ratios and produces lower nonspecific interactions compared to the microbially produced branched dextrans used by Biacore manufacturer GE Healthcare. The XanTec chips are also produced with a range of defined matrix densities to allow multilayer ligand immobilisation, providing a basis for signal amplification. [433] This advance in chip technology has increased the potential for using capture immobilisation for small molecule screening.

Poly-histidine tags are commonly used to provide purification mechanisms for recombinant proteins [434] and also provide a mechanism for immobilising proteins on the surface of an SPR chip. Using the traditional GE NTA sensor chips for His capture is an excellent technique for protein-protein interactions (PPIs); the chips allow the capture and recapture of protein many times and the signal generated by proteins interacting is proportionally larger than the problematic signal to noise ratio. The chips are, however, prone to non-specific binding, which in PPI studies can be overcome by limiting the surface density of the protein and using bovine serum albumin (BSA) and increased salt and/or detergent in the running buffer. However, protein-small molecule screening requires high protein surface densities which increase surface instability, and BSA is known to bind to small molecules and will adversely affect the binding kinetics if included in the running buffer. [435, 436] The improved chemistry of the XanTec chips largely overcomes these limitations and allows the capture method to be employed for small molecule screening. This method of protein immobilisation is advantageous as it allows the protein to be stripped from the chip surface and fresh protein captured several times during the analysis. This allows longer analysis experiments to be performed without the risk of denaturing the protein on the chip.

7.1.4. Solvent correction

The solvent correction curve adjusts the response values for the effects of varying bulk refractive index effects. DMSO, which is routinely used to dissolve drug-like molecules and fragments, has a high refractive index. The reference surface will produce a larger bulk shift due to the higher concentration of DMSO near the chip surface, compared with the target surface where DMSO is excluded from the chip surface by the immobilised ligand (excluded volume). As the response is different in each flow cell it cannot simply be cancelled out after reference subtraction. Small differences in the concentration of DMSO in the sample and running buffer can also lead to changes and inconsistencies in response. A standard solvent correction uses eight injections of running buffer with a range of DMSO concentrations around the assumed concentration; e.g. for running buffer with 2% DMSO a concentration range from approximately 1 – 3% would be used. The principles for employing a solvent correction are illustrated in Figure 7-3; however once the step is included in the analysis method the evaluation software will apply these calculations automatically.

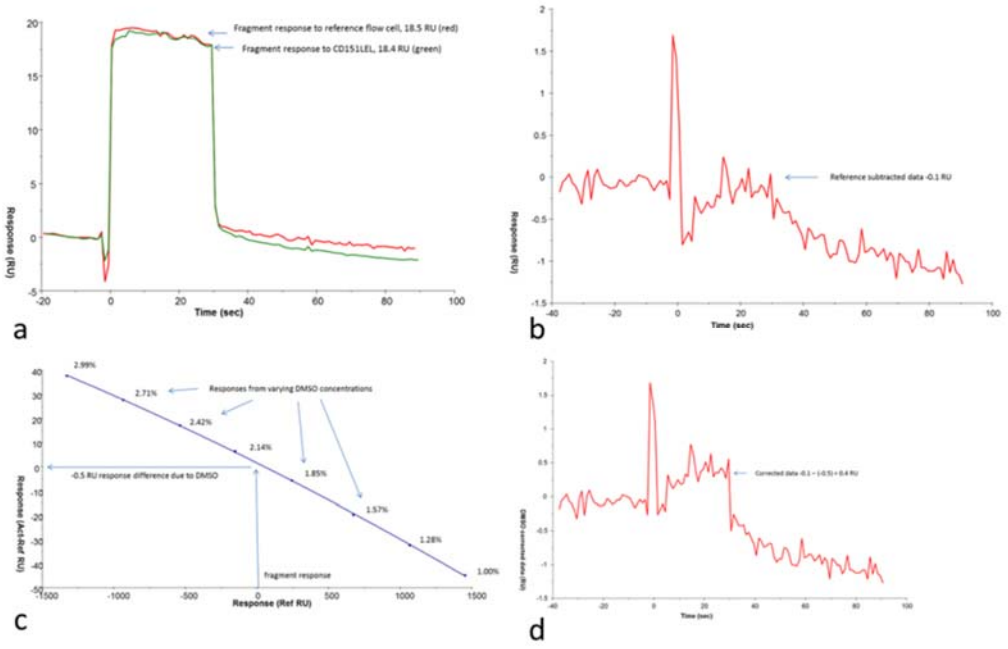


Figure 7-3 Solvent correction. (a) Injection of fragment over captured CD151 LEL (green) and the reference flow cell (red). (b) Response to the reference flow cell subtracted from the CD151 LEL data. (c) A series of eight different concentrations of DMSO in running buffer were run every 50 cycles during the screen. The response of the solutions from the reference surface ranged from approximately -1300 to 1500 RU. A calibration curve plotting the difference in response between the CD151 LEL and the reference flow cell (Act-Ref RU) versus the response from the reference flow cell (Ref RU) was used to correct for DMSO mismatch in the sample and running buffer. (d) Corrected fragment response.

7.2. *RESULTS AND DISCUSSION*

7.2.1. Protein immobilisation

All of the protein used for these experiments was thoroughly examined for purity by SDS-PAGE and size exclusion chromatography, assessed for correct size and the presence of disulfide bonds by LC-TOF and conformation examined by CD (see CHAPTER 5 for details).

7.2.1.1. Protein pre-concentration

To determine the optimal pH for pre-concentration of the CD151 LEL to the chip surface, CD151 LEL cleaved from the His-SUMO tag was diluted to 56 µg/ml in 10 mM sodium acetate buffer at three different pH values, pHs 4, 4.5 and 5, and each sample injected across a blank CM5 sensor chip for 180 seconds. The pI of the CD151 LEL, as calculated from the sequence using ExPASy Bioinformatics Resource Portal ProtParam [399] , is 5.67 indicating that the protein should exhibit a net positive charge at all three pH values. The amount of binding of the CD151 LEL to the blank surface at the three different pH was compared. (Figure 7-4)

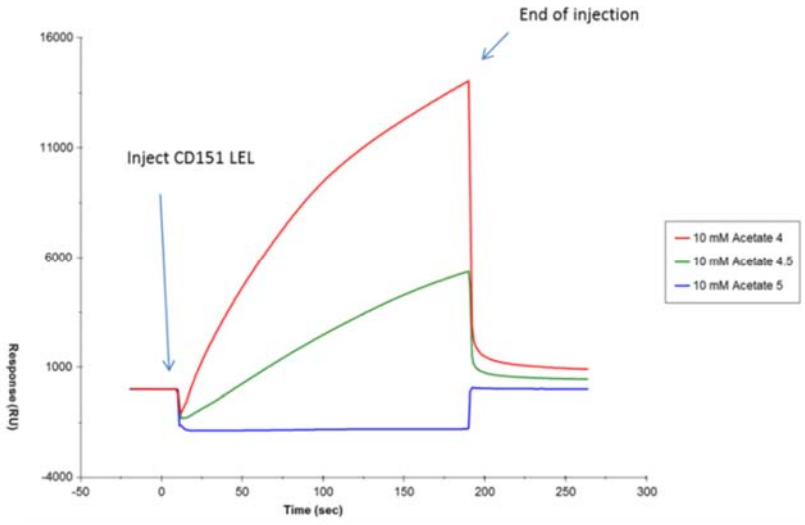


Figure 7-4 SPR response curves for CD151 LEL binding to a blank CM5 chip under three different pH conditions in 10 mM sodium acetate. Protein at 56 µg/ml was injected over the chip surface at 30 µl/minute for 180 seconds and the change in refractive index observed using a Biacore T200.

Figure 7-4 shows that at pH 5 the net positive charge on CD151 LEL is not sufficient to efficiently preconcentrate the protein at the dextran surface layer of the chip. At pH 4 the positive charge is clearly able to preconcentrate the protein, but a pH this low could be both detrimental to the protein's stability and reduce cross-linking efficiency. From this experiment it is evident that pH 4.5 is sufficient to preconcentrate the protein. Analysis by DSF indicates that the protein is stable at this pH (Figure 5-69).

7.2.1.2. **Amine coupling**

CD151 LEL was purified by gel filtration on a Superdex 200 10 300 SEC column into HBS-P (immobilisation buffer) to a final concentration of 200 nM and coupled via covalent amine coupling on a GE CM5 chip as per manufacturer's instructions (Figure 7-5). The final immobilisation level was approximately 700 RU.

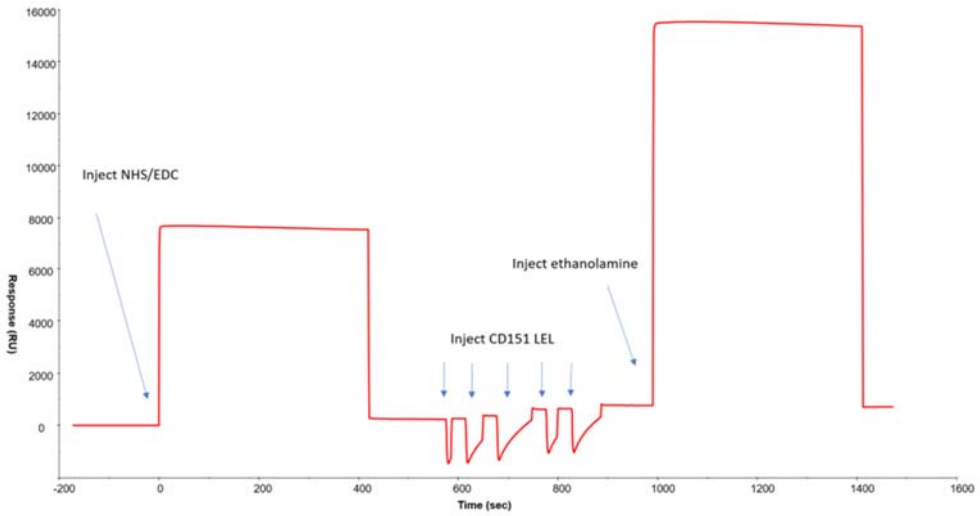


Figure 7-5 Sensorgram of the amine coupling of CD151 LEL on a CM5 sensor chip. The surface was activated with a 7-minute injection of NHS/EDC then CD151 LEL at 200 nM was injected in pulses until approximately 700 RU of protein was immobilised. The surface was then deactivated with a 7-minute injection of ethanolamine. Running buffer was HBS-P and the flow rate was 10 µl/minute.

7.2.1.3. **His capture**

His_6-CD151 LEL was purified by gel filtration on a Superdex 200 10 300 SEC column into HBS-P and captured at 100 nM on a XanTec NiHC 1500m sensor chip which had been pre-loaded with Ni^{2+} as per manufacturer's instructions (Figure 7-6). The final immobilisation level was approximately 8000 RU. This trial was to determine the efficiency of this capture method and protein was stripped from the surface using 350 mM EDTA and recaptured for each experiment.

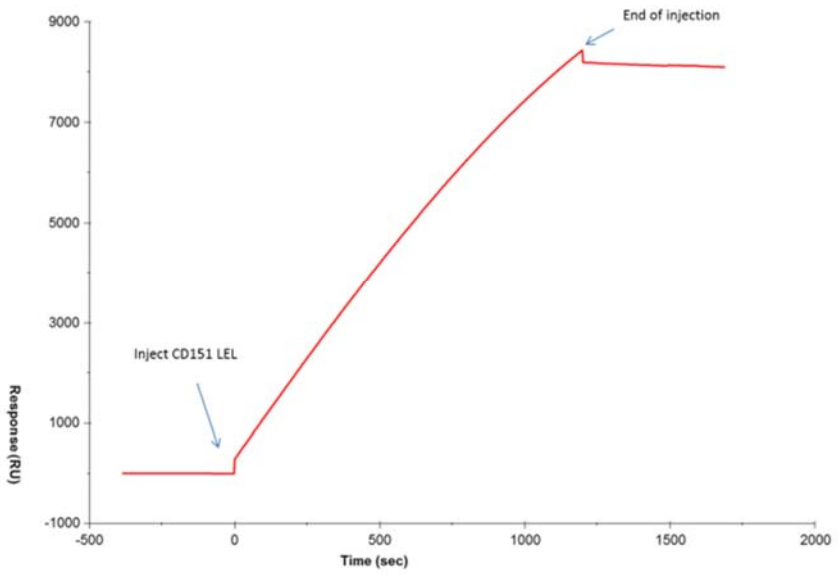


Figure 7-6 Capture of His_6-CD151-LEL on a XanTec NiHC 1500m sensor chip. CD151-LEL at 100 nM was injected for 20 minutes resulting in over 8000 RU of protein captured on the chip surface.

7.2.2. Assay validation

7.2.2.1. Amine coupled protein

Binding of the anti-CD151 Ab 11G5α (ab33315 Abcam Australia Pty Ltd) was tested using single cycle kinetics at a concentration range from 9 nM serially diluted threefold to 0.1 nM, running buffer was HBS-P+ (Figure 7-7).

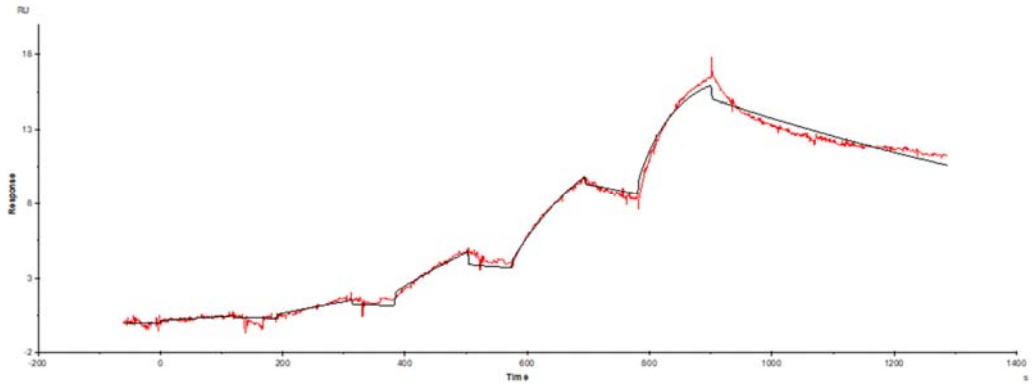


Figure 7-7 mAb 11G5a binding to CD151 LEL immobilised on a CM5 sensor chip via amine coupling. The Ab bound with a K_D = 0.5 nM and a Rmax of 16.9 RU. The red curve represents the raw data and the black curve is the fitted data using a 1:1 Langmuir binding model.

The Rmax value describes the binding capacity of the surface and is dependent on the amount of ligand immobilised on the chip surface. [252] Using Equation 2-2, the theoretical amount of IgG that could bind to the immobilised CD151 LEL is over 4200 RU. The actual Rmax, although typically lower than the theoretical Rmax, in this instance is over 250 times lower than expected suggesting that the CD151 LEL and/or the Ab has a low active concentration (i.e. functional protein vs total protein) or that the amine coupling is occluding the Ab binding site. This result indicates that the amine coupling method of immobilisation is not suitable for studying CD151 LEL.

7.2.2.2. **His captured protein**

His_6-CD151 LEL at 200 nM was immobilised to 3000 RU on a XanTec NiHC 1500m chip via His capture. Binding of the anti-CD151 Ab 11G5α (ab33315 Abcam Australia Pty Ltd) was tested using single cycle kinetics at a concentration range diluted threefold from 27 nM to 0.3 nM in HBS-P+ running buffer(Figure 7-8).

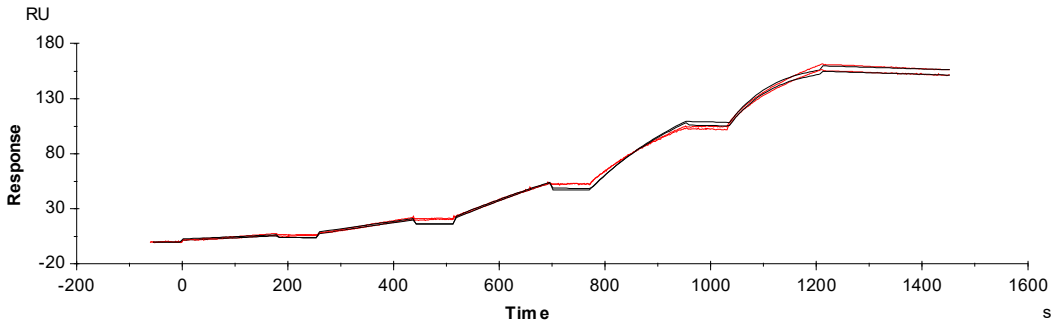


Figure 7-8 mAb 11G5a binding to His_6-CD151 LEL captured on a XanTec NiHC 1500m sensor chip via His capture. The Ab bound with a K_D = 0.2 nM and a Rmax of 167 RU. The red curve represents the raw data and the black curve is the fitted data using a 1:1 Langmuir binding model.

Using the Rmax formula the theoretical amount of binding of the Ab to the captured CD151 LEL is over 18200 RU. Although this is over 100 times the actual result, the fit of the Ab binding curve is much better than that observed with the amine coupled protein. This result, with the added advantage of being able to renew the protein on the surface as needed, makes the His capture approach a more suitable method for SPR studies of CD151 LEL.

Phosphate buffered saline (PBS) is the manufacturer's recommended running buffer for small molecule analysis. A comparison between the initial Ab check in HBS and an equivalent analysis using 11G5α Ab in PBS + 0.005% TWEEN®20 was therefore carried out to select appropriate buffer conditions for the fragment screening analysis (Figure 7-9). Although the kinetics of binding and apparent affinity of the Ab for CD151 LEL in PBS are similar to those obtained in HBS, the curves do not fit as well to the model and the replicates do not overlay as closely. On the basis of these results, HBS-P was selected as the running buffer for the fragment library screen.

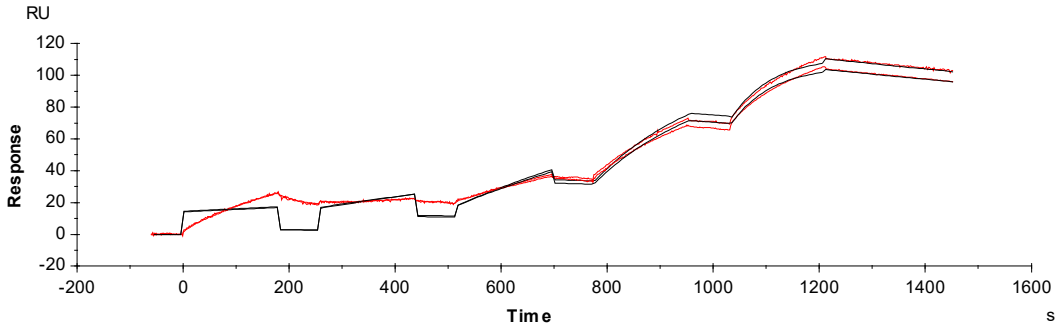


Figure 7-9 Single cycle kinetic evaluation of the mAb 11G5a binding to His_6-CD151 LEL captured on a XanTec NiHC 1500m sensor chip via His capture in PBS. The Ab bound with an affinity of K_D = 0.6 nM and a Rmax of 116 RU. The red curve represents the raw data and the black curve is the fitted data using a 1:1 Langmuir model.

7.2.3. Screen of a fragment library

261 fragments from a well-characterised fragment library purchased from the Monash Institute of Pharmaceutical Sciences

(https://www.monash.edu/research/infrastructure/platforms-pages/fragment) were screened against the His_6-CD151 LEL. His_6-CD151 LEL was captured on the chip surface to approximately 8000 RU at the beginning of the screen then stripped from the chip and fresh protein captured every 100 cycles. The fragments were diluted to 200 µM in HBS-P 2% DMSO and individually injected across the captured His_6-CD151 LEL at 100 µl/minute for 30 seconds and allowed to dissociate for 60 seconds. A solvent correction curve was run every 50 cycles and a blank cycle every 10. 11G5α mAb at 3 nM was injected at the end of each 100 cycle block to monitor the protein activity.

The binding response of the fragments was solvent corrected, double referenced against a blank injection and the reference surface and corrected for Mw. Of the 261 fragments screened, 140 bound preferentially to the reference surface and were excluded from analysis. Using Equation 2-2 to calculate the expected Rmax for the fragment binding gives a value of approximately 86 RU. Five fragments appeared to show promiscuous binding, in spite of binding below the Rmax. [437] Although these fragments are binding below theoretical Rmax, the shape of the curve indicates that the fragment is binding in a non-stoichiometric manner, most likely due to fragment aggregation (Figure 7-10).

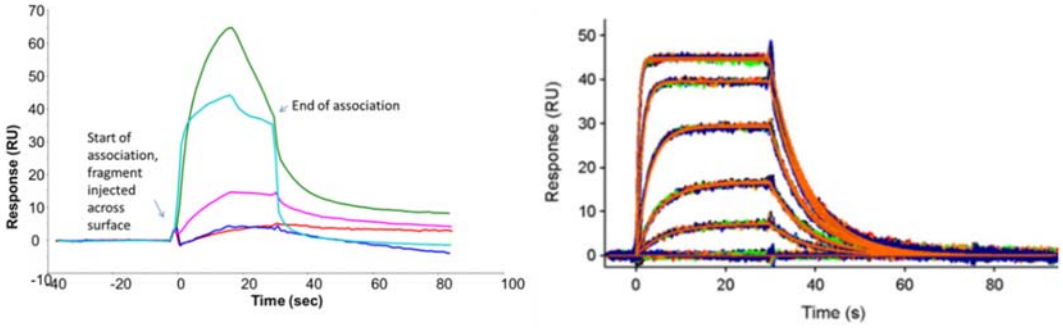


Figure 7-10 Sensorgrams of five promiscuous binders (left) and typical fragment responses (right).

Promiscuous binding, in this instance, is identified by the slow association rate, dissociation of the fragment during the association phase and slow dissociation or pseudo-irreversible binding to the protein. These types of sensorgram perturbations are typically seen when the fragments are accumulating non-specifically and forming aggregates on the protein surface. The sensorgram on the right, Figure 7-10, shows sulphanilamide binding to carbonic anhydrase II, a commonly used model system, taken from Rich et al. 2010. [438] The model sensorgrams show a steady association curve that reaches saturation with increased analyte concentration and a steady dissociation curve that returns to baseline.

Of the remaining 116 fragments the binding response ranged from 0.01 to 1.6 RU (Figure 7-11). This is much lower than anticipated for a binding event. The calculated Rmax for this interaction is over 80 RU.

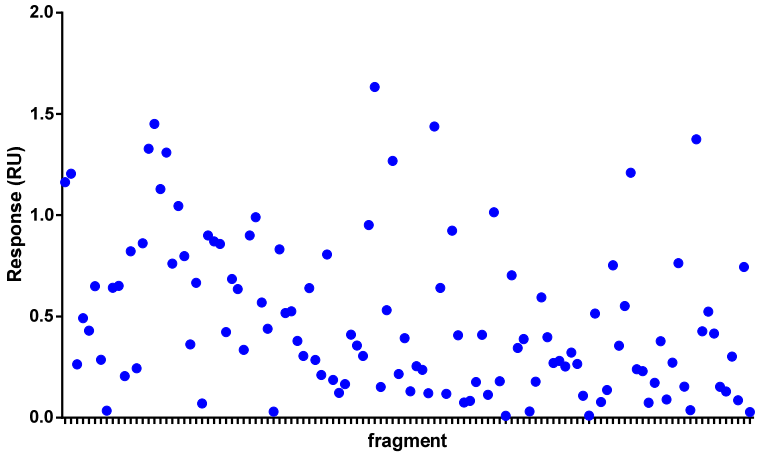


Figure 7-11 The binding response (RU) of 116 fragments to CD151 LEL. Data are double referenced and normalised for Mw.

As a positive control small molecule for the binding site was unavailable, the only method available for evaluating the surface activity during the fragment screen was to inject the conformationally specific Ab 11G5α. For a 150 kDa Ab and >7000 RU of CD151 LEL captured on the chip surface, the binding response as calculated by Equation 2-2

is well over 66,000 RU. The binding response observed for the Ab was approximately 50 RU (Figure 7-12).

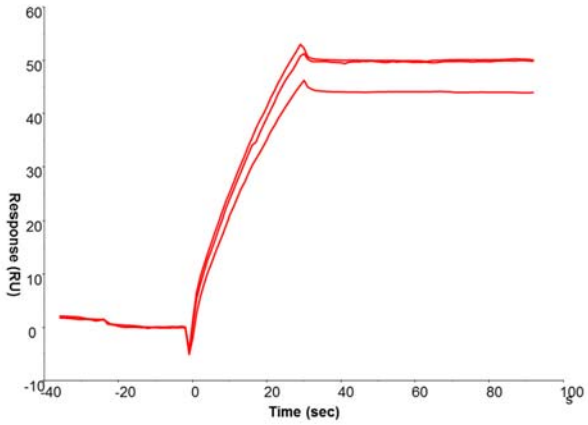


Figure 7-12 Sensorgram of CD151 Ab 11G5α binding to captured CD151 LEL.

The Ab binding response represents a very small fraction of the expected binding. The 11G5α Ab is a mouse mAb that recognises human CD151 cell surface antigen. [439] The crowding of the CD151 LEL on the chip surface may be preventing recognition by such a large analyte. Mass spectrometry and CD analysis of the protein (Figure 5-73, Figure 5-74) indicated that the protein was the correct size and tertiary structure, however the proteins low binding to the Ab and lack of hits from the fragment library suggested there was an unknown problem with the fidelity of the protein.

A true positive control would be a compound of similar size to the analytes being screened and bind specifically to the targeted binding site. When searching for novel binders this, by definition, is not possible. The “next best” approach would be to use an Ab fragment such as a Fab to monitor the protein conformation, stability and activity during analysis. The use of a whole Ab as a positive control for a fragment screen represents the “least best” option. However, in this instance, was the only control available. The reproducible binding of the Ab established protein activity and stability and, although far from ideal, was valuable for validating the assay conditions.

7.3. *CONCLUSION*

The advantages of low protein consumption, rapid assay development, HTS and kinetic validation of hits has established SPR based fragment screening as a common primary screening technique employed in drug discovery projects. [440, 441]

In attempting to apply the SPR technique to fragment screening against CD151 LEL, I was able to establish immobilisation methods and demonstrate protein stability and reproducibility when binding to a positive control Ab. The lack of any fragments binding

to the CD151 LEL was disappointing. The fragment library used is designed to maximise chemical space coverage and it is unusual not to see any binding hits. It is unlikely that the protein was degraded or misfolded as it was binding the 11G5α mAb with high affinity and any exposed hydrophobic regions would have resulted in more false positive hits from the fragments. The capture via the His tag would have resulted in a homogenous orientation of the protein on the chip surface presenting the fragments with unobstructed access to the surface of the protein. The most likely explanation for the lack of binding is that the CD151 LEL does not contain any structural cavities or pockets, rendering it a very difficult if not "undruggable" target for the FBDD approach. Without an appropriately sized positive control to validate the binding site it is impossible to asses if this is the case. My current modelling of the LEL suggests that there is a surface crevice across, and two small pockets adjacent to, the targeted QRD binding site. As the binding site is located on a flexible loop it is possible that a conformation exists that opens a surface channel connecting the two pockets. (Figure 7-13)

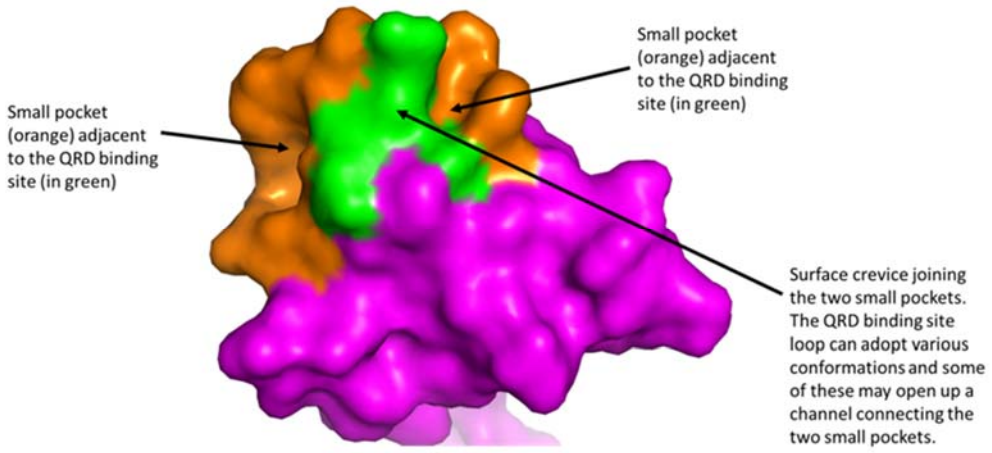


Figure 7-13 Surface model of the CD151 LEL showing potential binding pockets on the QRD binding site. The QRD binding site is shown in green; the binding pockets are shown in orange.

It is also possible that the loop forms a protrusion that docks into a pocket on integrin. Without accurate 3D structural information, it is impossible to confidently predict which of these scenarios is more likely. The lack of binding from the fragment screen strengthens the importance of solving the structure of CD151 LEL to facilitate the discovery of small molecule inhibitors by virtual screening approaches or developing any hits from a HTS approach.

It is likely that the problems that arose during the development of the fragment screen were due to the formation of non-native disulfide bonds as established in section 5.3.7. The experiments carried out using the Water's Vion IMS QToF mass spectrometer (Figure 5-79,Figure 5-80) revealed the disulfide mismatches, however this data was only available after the SPR fragment screen had been completed. If only a small percentage of the immobilised protein was correctly folded, then binding to the Ab would be reduced. If the aberrant disulfide bonds fold the LEL in such a way that protects the hydrophobic regions, the LEL may present a smooth surface that would not bind any fragments. In light of the mass spectrometry information about the disulfide formation, the fragment screen potentially worked perfectly.

CHAPTER 8. **FINAL COMMENTS**

The search for new treatments for cancer is yielding more effective drugs, treatment strategies and improving patient outcomes in the western world. In the USA the cancer death rate has fallen 26% since 1991. The decreasing death rate is likely due to a reduction in the use of tobacco and advances in early detection and treatment. [442] However, globally new cancer cases and deaths are increasing and it is estimated there was 18.1 million new cases and 9.6 million deaths in 2018. The increasing incidence and mortality rates worldwide are reflective of the growth and ageing of the population and other factors associated with socioeconomic development. Prostate cancer and leukaemia represented 3.8% and 3.2% respectively of all cancer deaths in 2018. [443] The need for new diagnostic and treatment strategies is ongoing.

The move from cytotoxic chemotherapy towards molecularly targeted therapies has increased the number of successful treatments available; however, progress is slow. Of the 53 new drugs approved by the U. S. FDA's Centre for Drug Evaluation and Research in 2018, only 11 were to treat cancer. [37] Currently cancer therapeutics have the lowest clinical trial success rate of all major diseases, which in 2015 was at around 3%. [444, 445] The National Cancer Institute (NCI), the principal federal agency for cancer research in the USA, spent over $8 billion on cancer research in 2018 alone, yet with only 11 new cancer drugs entering the marketplace this does not seem like a reasonable return. [446, 447]

Failure rates in drug development are most often attributed to toxicology issues during clinical development, with late-stage attrition for cancer drugs as high as 70% in phase II and 59% in phase III clinical trials. [448, 449] When considering the targets and compounds that are investigated and rejected at the initial laboratory level, the actual attrition rates are no doubt much higher. There are numerous scientific, technical, economic and personnel hurdles to overcome during the drug discovery and development process and as has been clearly demonstrated by this project, target selection and meticulous validation is crucial.

Both CD33 and CD151 have been thoroughly investigated as therapeutic targets. Multiple studies and reviews have validated their role in cancer progression. For CD33 this has translated into several ADCs being developed to treat AML and although clinical success has been limited, the potential remains promising. The compounds I identified that bind CD33 seemed to give convincing data in the initial screening; however, the attempt to develop higher affinity analogues was unsuccessful and the original compounds, although undergoing testing, have yet to be validated by complimentary methods. This finding does not invalidate the initial results but illustrates the necessity for meticulous analysis of results and careful identification and consideration of false positives. The progression from hit to lead compounds is fraught with difficulties.

The use of SPR as a primary screen for small molecule and fragment libraries is common as the instruments are sensitive enough to detect low affinity binders and with a high throughput format that can screen hundreds of compounds in a few days. SPR is also used to characterise and validate binding compounds to prioritise further investigation. Successful identification of hits by SPR requires detailed and specialised instrument preparation, assay development, compound handling, screening, conformation testing and data analysis. [450] But even the most fastidious user, in spite of their best efforts, can still be confronted with confusing and ambiguous data. Deciphering the good from the bad can involve multiple repeats of experiments, redesign of approach, optimisation of conditions and should involve analogous techniques to confirm the data.

From 1998 to 2009 Rich et al. published annual reviews of the optical biosensor literature. In 1998, when SPR was a relatively young technology, 384 publications were reviewed. [451] By 2009, the last year of their reviews, 1514 publications were examined. [452] While in the 1998 review they lamented that the majority of the published literature was poor, by 2009 they increased their approval to 20%. This is still a low figure and hopefully SPR data quality has increased in the proceeding decade, but the evidence remains that execution and analysis of SPR experiments is not a simple matter. As the technique becomes mainstream in drug discovery laboratories, the danger is that poor quality data leading to false positives will increase the attrition rate of novel compounds in drug development.

When choosing the binders for CD33 the priority was given to the higher affinity binders, which biased compound selection. In hindsight, a more reasonable approach would have been to target the lower affinity binders that gave better quality data, such as 2966, and focus on improving affinity with compound analogues. While the binding data for 2971 was better quality than that for 2952, it also bound the negative control. This is likely why the 2971 analogue compounds preferentially bound the R119A negative control protein, as the original compound may not have targeted the R119 sialic acid-binding site. The technique and assay used to screen CD33 is valid; problems arose from protein quality and data analysis. As such, the binding data obtained for CD33 during this project should only be considered preliminary, not definitive and so require further examination.

The recombinantly produced CD33 and CD151 LEL used for this project were of varying quality. Data quality was constantly negatively impacted by poor quality protein, as is evident by the low % functional ligand as calculated by Equation 2-2. For CD33, assay quality was further impacted by having access to only small amounts of the compounds which were not highly soluble and were prone to non-specific binding. Assay quality was optimised for CD33 by improving the protein construct to eliminate dimerisation through the third cysteine residue, changing the capture method to biotin-streptavidin,

using a higher grade of sensor chip and careful buffer optimisation. Which, even though improved binding and surface stability, did not ultimately aid detection of a reliable, high affinity binder. An ideal small molecule or fragment screen would include a positive control of the same size as the compounds being screened and an unrelated negative control. When searching for novel binding compounds, however, a positive control often isn't available. For the case of CD33 this was somewhat compensated by the inclusion of a negative control protein to assess specificity; and for CD151 an Ab was used to monitor protein activity. Having since established that the recombinant CD151 has a dynamic structure with some disorder, the specificity of the Ab used must be called into question. The positive CD33 control compound P22 that became available later in the project was useful for validating the cysteine mutation (CD33 C36S). However, P22 availability was minimal as it was expensive and complex to synthesise and thus the compound wasn't able to be used during the screening.

The use of analogous methods to confirm hit compounds is crucial as each method of testing reveals different information about the binding interaction and each have their own strengths and weaknesses. STD-NMR establishes the chemical authenticity of the compounds and fragments and indicates binding. SPR can measure binding affinity and specificity. A parallel screening study by Weilens et al. in 2013 compared the hits identified by NMR and SPR for the same protein target. [453] In their preliminary screen of 455 compounds, 62 compounds were identified as hits by NMR and 16 by SPR. There was no overlap of compounds from either group. The SPR hits were re-analysed in NMR and vice versa. Under NMR conditions three of the SPR hits were insoluble and one contained breakdown product. In SPR, two of the NMR hits gave no binding and five contained DMSO mismatch and gave a negative response. These results illustrate the complimentary nature of the two techniques as false positives can be identified with the different information available from each technique. Retesting and validation of hits by X-ray crystallography confirmed 15 of the NMR identified compounds and 6 of the SPR identified compounds were true binders. This led the authors to conclude that the main reason for the lack of overlap of the confirmed hits was due to the selection criteria for the classification of a hit by the two methods, rather than experimental conditions. The Weilens et al. study clearly demonstrates the subjective nature of the hit-to-lead method and the necessity of the use of complimentary methods to confirm hits.

Two of the contributing authors from the Weilens et al. paper later collaborated on a further comparison of fragment and compound screening data using native state electrospray ionisation mass spectrometry (ESI-MS), SPR and X-ray crystallography. [454] 70 compounds were screened using SPR and ESI-MS and 25 compounds were found to be positive binders by SPR and 38 by ESI-MS. All 25 SPR positive compounds were positive in ESI-MS. Of the 25 compounds positive in both ESI-MS and SPR, 13 were confirmed by X-ray crystallography, 10 were not tested due to lack of material and 2 did

not bind. Unfortunately, none of the compounds positive in ESI-MS and negative in SPR were further investigated so it remains unknown if they were false negatives by SPR or false positives by ESI-MS. Regardless of this omission, it is evident that the two techniques display a high consensus and would be an efficient means of screening compound and fragment libraries.

Rich et al. conducted a global benchmark study using SPR based biosensors which included 150 participants from 20 countries using 18 different biosensors. [437] Each of the participants were given the same protein samples and asked to determine the kinetic rate constants. Of the 258 data sets that were collected, 5 were disregarded as problematic and the remaining produced an affinity of K_D = 0.62 ± 0.98 nM. This demonstrates the reliability and reproducibility of SPR when applied by skilled independent researchers.

During my attempts to produce recombinant CD151 protein, I became increasingly frustrated with the lack of biophysical evidence presented in the various papers to back up their claim that they had successfully done so. The common use of Western blot as definitive proof of protein quality is spectacularly inadequate. Proteins that are aggregated in solution appear as monomer by SDS-PAGE as the SDS disrupts the aggregates. Transfer from the SDS-PAGE gel to the Western blot membrane can be incomplete or patchy. Abs used for probing Western blots are notoriously prone to non-specific binding, and Western blots themselves can be ambiguous with background noise masking bands and other random noise from poorly handled membranes or unfiltered buffer. They provide very little information regarding conformation or protein quality. Obviously within the constraints of publishing a journal article, not every technique used can be reported and much background information such as protein quality control is simply assumed to have been done as routine. Complicated protocols that have taken years to develop may be reduced to a few lines in a materials and methods section, and while there is no intention to mislead or misdirect, a method under these circumstances may become irreproducible.

When performing quality control experiments on the recombinantly expressed CD151, I routinely assessed protein quality by Western blot, DLS, CD and mass spectrometry. All of these techniques showed that I had good quality protein. This, however, was not reflected in the SPR data or crystallographic studies. It wasn't until the ion mobility analysis became available to me in the final stages of my project that I was able to ascertain that the protein was misfolding, which explained the difficulties I experienced. A general lab rule-of-thumb is that if a protein is misfolded it will be insoluble due to the exposed hydrophobic core. This may be true for many proteins but was not true for CD151. In fact, when examining the hydrophobic plot of CD151 (Figure 5-91) it is apparent that C155 and C156 are in a hydrophilic region yet are expected to covalently

bind to C208 and C185 respectively, which are both in a hydrophobic region. This is perhaps why the cysteines are simply bonding with their nearest neighbour with similar hydrophobicity rather than twisting into the proposed correct form. The results obtained in this project indicate that CD151 is not suitable for recombinant expression in prokaryotic systems, at least by the systems trialled, as the disulfide bonds are far too prone to mis-forming. It may be possible, with increased optimisation to improve expression in a eukaryotic system, with co-expression of integrin which has been shown to exhibit the same expression patterns, suggesting some co-operative relationship. [219] Another approach would be to express and determine the structure of the full-length protein as has been done for CD81. [201] Otherwise CD151 assays may be limited to cell based and *in vivo* models.

Since completion of this project there have been several interesting publications regarding tetraspanin structures. The full length structure of CD53 and CD9 have been solved by X-ray diffraction (PBD ID: 6WVG and 6K4J, respectively). [455, 456] Both of these tetraspanins, like CD81 have four cysteines in the LEL. CD53 was recombinantly produced in yeast (*Komagataella pastoris*) and CD9 and CD81 were produced in insect cells (*Spodoptera frugiperda*), presumably to facilitate post translational modifications such as disulfide bonding. In order to crystallise CD9 the authors truncated the LEL region by five amino acids (Thr175 – Lys179) reducing it from 84 amino acids to 79 along with a reduction in the length of the cytoplasmic tail from seven amino acids to five. These deletions were required for concurrent studies using the recombinant protein in an *in vitro* fertilisation assay, however, the authors do not comment on the effect the deletion may have had on the crystallisation of the protein. [456] CD53 was mutated to remove one glycosylation site from the LEL and to prevent palmitoylation in the transmembrane regions to improve crystal diffraction. [455]

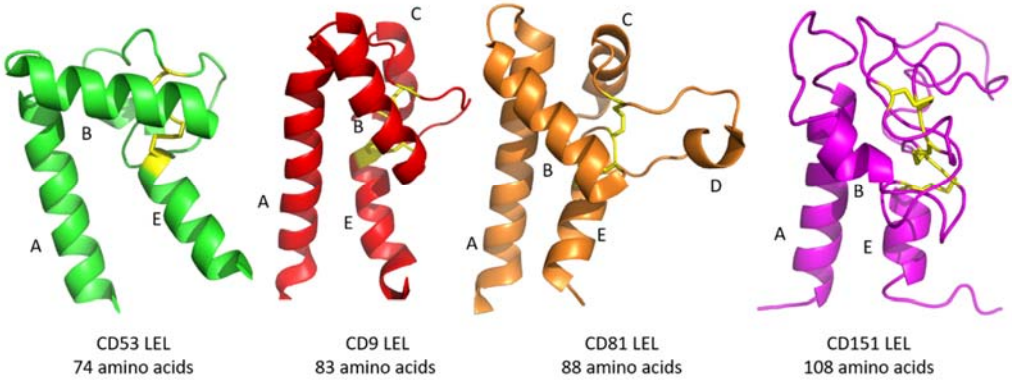


Figure 8-1 Cartoon representations of the crystal structures of CD53, CD9 and CD91 LELs and my model of CD151 LEL with disulfide bonds shown in yellow.

The cartoon representation of the three known LEL crystal structures structures and the model structure of CD151 LEL illustrates the increase in complexity with the CD151 (Figure 8-1). The tetraspanin LELs form a head region supported by two helical stalks, (labelled A and E in Figure 8-1), that connect to the transmembrane domains. CD81 contains two more helical regions, labelled C and D, which folds into two antiparallel loops, stabilised by the two conserved disulfides. [457] CD53 lacks the C and D helices found in CD81, CD9 lacks the D helix and it is not known if CD151 has any structure in this C – D region. A significant finding of recent publications is that CD81, CD53 and CD9 have a cholesterol dependant open and closed conformation wherein the head region rotates to expose the C – D region for interaction with a binding partner. [201, 456, 458]

Homology modelling of CD151 by Purushothaman & Thiruvenkatam using CD81 as the structural template includes helices in the C and D regions, although their modelling does not include disulfide bonds. [459] Furthermore molecular dynamic simulations by Purushothaman & Thiruvenkatam suggest that CD151 undergoes the same cholesterol dependant conformational changes as CD81 and CD53. They hypothesise that when cholesterol is bound CD151 adopts a stable, closed conformation (Figure 8-2).

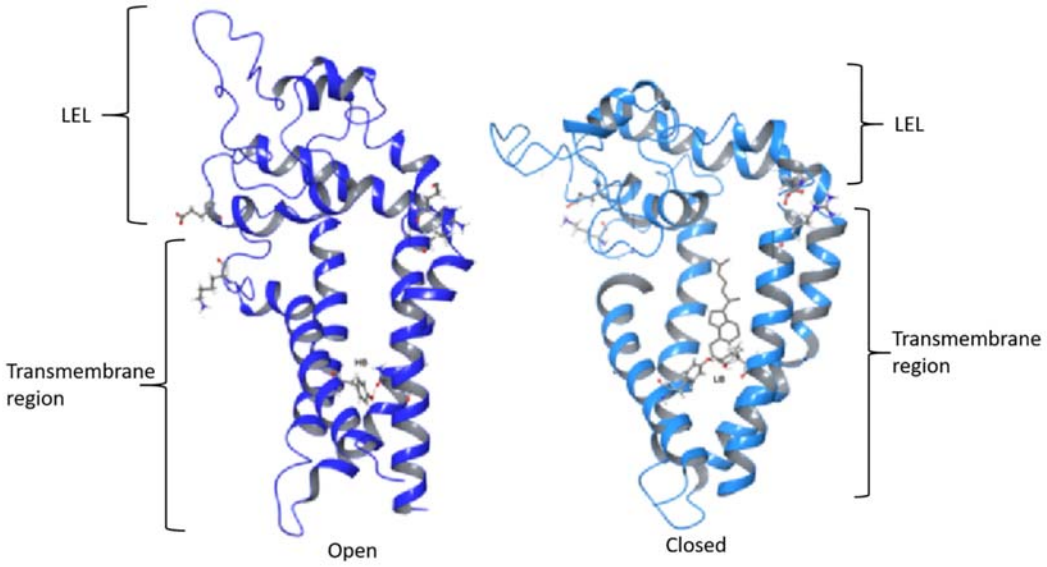


Figure 8-2 Homology model of CD151, based on the crystal structure of CD81, illustrating the open (left) and closed (right) conformations. Adapted from Purushothaman & Thiruvenkatam 2019.

The hypothesis that CD151 LEL is more stable in the closed conformation when bound to cholesterol aligns with my findings that the detergent CHAPS stabilises the LEL (see 5.3.2.3.2). The structural similarities of CHAPS and cholesterol are shown in Figure 8-3.

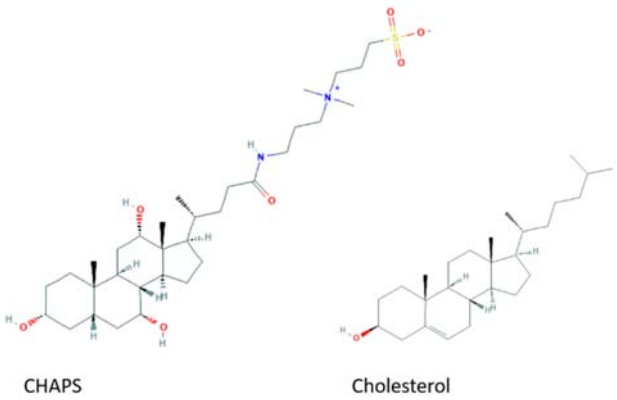


Figure 8-3 2D structure of CHAPS (left) and cholesterol (right).

Purushothaman & Thiruvenkatam also claim in their paper to have successfully produced recombinant CD151 LEL in *E. coli.* as an MBP fusion protein. Their protein characterisation includes SDS-PAGE, Western blotting and CD for the intact constructs. Although they state the MBP tag was cleaved and the protein was "in good confirmation [sic] with respect to the proper folding...". Their initial lysis buffer contains 10 mM β-mercaptoethanol, which would have prevented disulfide bonds forming once the protein was released from the reducing environment of the cell. Their text simply does not mention disulfides; I find this omission perplexing given the in-depth molecular dynamics simulations they report. [459]

This leads me back to the point raised on page 202 of how some research groups manage to express CD151 while others struggle. Is it simply a problem of not asking the right questions? If you don't look for disulfides, then you'll never know if they are problematic. This raises the question of how many other proteins may have been erroneously used in biochemical assays, without their conformation being accurately assessed, and may be one of the reasons for inhibitors or activity assays generating different results in *in vitro* and *in vivo* models.

In 2014 the NIH published a set of guidelines for reporting preclinical data in which they suggest that journals should have no or generous limits for method sections. [460] The guidelines were established to address the issue of reproducibility and rigor of research findings and as of 2017 approximately 80 journals had endorsed the guidelines. While initiatives such as this are helping to maintain scientific integrity, the very nature of grass

roots research in the current funding climate, which encourages a “publish or perish” culture, it is inevitable that shortcuts are taken. A collaboration between the Centre for Open Science and Science Exchange, the Reproducibility Project, aims to independently replicate results from high-profile papers in cancer biology. [461] The replication studies are yielding a mixed bag of results, falling into the categories of reproducible data, inconsistent data and irreproducible data. It is impossible to verify the reproducibility of every publication and there is limited scope for researchers to comment when methods don’t add up.

An open access online journal club called “PubPeer”, established in 2012, encourages researches to upload and expose journal articles they believe contain research misconduct. This may be anything from photo-shopped images to flawed methodology. There are currently 73164 papers in their database, which equates to a thousand articles uploaded every month. [462] The accused authors are free to respond, some justify their data and offer explanations, others humbly correct mistakes or even retract their papers. One court case was launched against the web site administrators but was dismissed on appeal. The majority of authors simply don’t comment. Some of the papers are from so-called prestigious journals such as *Nature* and *Cell*, indicating the standing of the journal is no protection to this behaviour.

It is cause for concern that the peer review process seems so flawed. With many different entities identifying poor quality publications, one wonders whether members of review panels are qualified to do their job or if work pressure leads to rushed reviews. I believe that is the crux of the flawed data problem – no one can know everything; review panels are comprised of humans after all. The complexity of modern scientific analysis and the sheer volume of instruments, techniques and methods available to researchers has us all applying ourselves to many fields. There seems to be a move away from dedicated technical specialists and a move towards DIY, which is diluting expertise and resulting in data generated by scientists who don’t quite know what they are doing. SPR data that looks fine to researchers and reviewers who don’t fully understand the nuances of the technique, is later identified as problematic by experts such as Rich et al. [452] Post-publishing review is therefore an important process that needs wider recognition and application. User groups and journal clubs can also be an invaluable resource for cross checking data and for guidance on interpretation and method design.

The drug discovery process from target identification through to lead optimisation can be done entirely without animal models, employing assays based on recombinant protein or cells. Often animal testing is only used to evaluate toxicology, not efficacy, of lead compounds prior to first-in-human testing. [463] Animal models are known to be inaccurate when used to evaluate efficacy and often fail to fully mimic human diseases and human toxicity has occurred when animals showed none. [464] Some of these

discrepancies can be sourced to errors in the lack of understanding of molecular disease mechanisms and the fundamental differences between mice and men. [465] Indeed mouse Siglecs 1, 2 and 4 are orthologous to the human, whilst the mouse and human Siglec 3 (CD33) have many differences including sialic acid recognition. [466] This precludes the use of a mouse model for CD33 assays, although CD33 knockout mice are available and there are efforts to knock-in the human CD33. [467]

The attrition rate for cancer drugs from concept to market is enormous. When dealing with proteins as badly behaved as recombinant CD151 it is easy to see why the translation rate from drug discovery to clinical trials is low. This thesis explores many options for expressing CD151 LEL as a recombinant protein and its use in FBDD. My discovery of the misfolding of CD151, and the discovery by others of the dynamic nature of the structure, provides an explanation of why many of the published articles mention difficulty working with this protein.

The small molecule-screening assay developed for CD33 was successful in identifying some hits. The compounds identified are proof-of-principal that the R119 binding site can be targeted, even though the initial analogue refinements were not successful, the compounds form a basis for further exploration.

The majority of the body of work presented here largely comprises negative results but would be an invaluable resource for anyone considering working on CD151 *in vitro*. It is unfortunate that negative results are not more widely and unashamedly shared in the scientific community, as knowing what not to do can provide vital guidance when knowing what to do is a mystery.

A quote frequently attributed to Albert Einstein “if we knew what we were doing, we wouldn’t call it research”, is a fair assessment of scientific research, although there is no evidence he actually said this.

REFERENCES

1. Naghavi, M., et al., *Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016.* The Lancet, 2017. **390**(10100): p. 1151-1210.
2. *Global Cancer Facts & Figures.* American Cancer Society, 2018.
3. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.
4. Vlaeminck-Guillem, V., *Extracellular Vesicles in Prostate Cancer Carcinogenesis, Diagnosis, and Management.* Front Oncol, 2018. **8**: p. 222.
5. Rosenquist, R., M. Esteller, and C. Plass, *Introduction: Epigenetics in cancer.* Semin Cancer Biol, 2018.
6. Samet, J., et al., *Choice of cancer therapy varies with age of patient.* JAMA, 1986. **255**(24): p. 3385-90.
7. *Surgery*. About Cancer 2019; Available from: https://www.cancer.org.au/about-cancer/treatment/surgery.html.
8. Martin Abeloff, J.A., John Niederhuber, Michael Kastan, w. McKenna, *Abeloff's Clinical Oncology*. Vol. 4th edition. 2008: Chruchill Livingstone. 2592.
9. Yarbro, C.H., D. Wujcik, and B. Holmes Gobel, *Cancer Nursing*. 2016, Sudbury, UNITED STATES: Jones & Bartlett Learning, LLC.
10. Airley, R., *Cancer Chemotherapy: Basic Science to the Clinic*. 2009: Wiley-Blackwell. 342.
11. Corrie, P.G., *Cytotoxic chemotherapy: clinical aspects.* Medicine, 2008. **36**(1): p. 24-28.
12. Oeffinger, K.C., et al., *Chronic health conditions in adult survivors of childhood cancer.* N Engl J Med, 2006. **355**(15): p. 1572-82.
13. Armstrong, G.T., et al., *Reduction in Late Mortality among 5-Year Survivors of Childhood Cancer.* N Engl J Med, 2016. **374**(9): p. 833-42.
14. Housman, G., et al., *Drug resistance in cancer: an overview.* Cancers (Basel), 2014. **6**(3): p. 1769-92.
15. Padma, V.V., *An overview of targeted cancer therapy.* Biomedicine (Taipei), 2015. **5**(4): p. 19.
16. Hanahan, D. and R.A. Weinberg, *The hallmarks of cancer.* Cell, 2000. **100**(1): p. 57-70.
17. Charlton, P.S., J., *Targeted therapy in cancer.* Medicine, 2016. **44**(1): p. 34-38.
18. Souho, T., et al., *Cancer hallmarks and malignancy features: Gateway for improved targeted drug delivery.* Biotechnol Adv, 2018. **36**(7): p. 1928-1945.
19. Hudziak, R.M., et al., *p185HER2 monoclonal antibody has antiproliferative effects in vitro and sensitizes human breast tumor cells to tumor necrosis factor.* Mol Cell Biol, 1989. **9**(3): p. 1165-72.
20. Cho, H.S., et al., *Structure of the extracellular region of HER2 alone and in complex with the Herceptin Fab.* Nature, 2003. **421**(6924): p. 756-60.
21. Huang, M., et al., *Molecularly targeted cancer therapy: some lessons from the past decade.* Trends Pharmacol Sci, 2014. **35**(1): p. 41-50.
22. Rowley, J.D., *Identificaton of a translocation with quinacrine fluorescence in a patient with acute leukemia.* Ann Genet, 1973. **16**(2): p. 109-12.
23. Lugo, T.G., et al., *Tyrosine kinase activity and transformation potency of bcr-abl oncogene products.* Science, 1990. **247**(4946): p. 1079-82.
24. Daley, G.Q., R.A. Van Etten, and D. Baltimore, *Induction of chronic myelogenous leukemia in mice by the P210bcr/abl gene of the Philadelphia chromosome.* Science, 1990. **247**(4944): p. 824-30.
25. Pray, L.A., *Gleevec: the Breakthrough in Cancer Treatment.* Nature Education, 2008. **1**(1).
26. Ropero, S., et al., *Trastuzumab plus tamoxifen: anti-proliferative and molecular interactions in breast carcinoma.* Breast Cancer Res Treat, 2004. **86**(2): p. 125-37.

27. Argiris, A., et al., *Synergistic interactions between tamoxifen and trastuzumab (Herceptin).* Clin Cancer Res, 2004. **10**(4): p. 1409-20.
28. Cooper, G.M., *Functions of Cell Surface Receptors*. 2 ed. The Cell: A Molecular Approach

ed. S. (MA). 2000: Sinauer Associates.
29. Grimm, D., et al., *Diagnostic and therapeutic use of membrane proteins in cancer cells.* Curr Med Chem, 2011. **18**(2): p. 176-90.
30. Ghirlanda, G., *Design of membrane proteins: toward functional systems.* Curr Opin Chem Biol, 2009. **13**(5-6): p. 643-51.
31. Allen, T.M., *Ligand-targeted therapeutics in anticancer therapy.* Nat Rev Cancer, 2002. **2**(10): p. 750-63.
32. Bareford, L.M. and P.W. Swaan, *Endocytic mechanisms for targeted drug delivery.* Adv Drug Deliv Rev, 2007. **59**(8): p. 748-58.
33. Almen, M.S., et al., *Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin.* BMC Biol, 2009. **7**: p. 50.
34. Li, Y.H., et al., *Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics.* Nucleic Acids Res, 2018. **46**(D1): p. D1121-D1127.
35. Cheever, M.A., et al., *The prioritization of cancer antigens: a national cancer institute pilot project for the acceleration of translational research.* Clin Cancer Res, 2009. **15**(17): p. 5323-37.
36. Zhu, X.R., et al., *A case report of targeted therapy with apatinib in a patient with advanced gastric cancer and high serum level of alpha-fetoprotein.* Medicine (Baltimore), 2016. **95**(37): p. e4610.
37. Hoelder, S., P.A. Clarke, and P. Workman, *Discovery of small molecule cancer drugs: successes, challenges and opportunities.* Mol Oncol, 2012. **6**(2): p. 155-76.
38. Zheng, H., et al., *The future of crystallography in drug discovery.* Expert Opin Drug Discov, 2014. **9**(2): p. 125-37.
39. Dias, D.M. and A. Ciulli, *NMR approaches in structure-based lead discovery: recent developments and new frontiers for targeting multi-protein complexes.* Prog Biophys Mol Biol, 2014. **116**(2-3): p. 101-12.
40. Renaud, J.P., et al., *Cryo-EM in drug discovery: achievements, limitations and prospects.* Nat Rev Drug Discov, 2018. **17**(7): p. 471-492.
41. Vyas, V.K., et al., *Homology modeling a fast tool for drug discovery: current perspectives.* Indian J Pharm Sci, 2012. **74**(1): p. 1-17.
42. Adams, D., *Life, the universe, and everything*. 1982, London: Pan Books.
43. De Las Rivas, J. and C. Fontanillo, *Protein-protein interactions essentials: key concepts to building and analyzing interactome networks.* PLoS Comput Biol, 2010. **6**(6): p. e1000807.
44. Chautard, E., N. Thierry-Mieg, and S. Ricard-Blum, *Interaction networks: from protein functions to drug discovery. A review.* Pathol Biol (Paris), 2009. **57**(4): p. 324-33.
45. Voet, D., J.G. Voet, and C.W. Pratt, *Fundamentals of biochemistry : life at the molecular level*. Vol. Fifth edition. 2016, United States: Hoboken, NJ : John Wiley & Sons.
46. Larsen, T.A., A.J. Olson, and D.S. Goodsell, *Morphology of protein-protein interfaces.* Structure, 1998. **6**(4): p. 421-7.
47. Jones, S. and J.M. Thornton, *Principles of protein-protein interactions.* Proc Natl Acad Sci U S A, 1996. **93**(1): p. 13-20.
48. Veselovsky, A.V., et al., *Protein-protein interactions: mechanisms and modification by drugs.* J Mol Recognit, 2002. **15**(6): p. 405-22.

49. Fry, D.C. and L.T. Vassilev, *Targeting protein-protein interactions for cancer therapy.* J Mol Med (Berl), 2005. **83**(12): p. 955-63.
50. Khan, S.H., et al., *Protein-protein interactions: principles, techniques, and their potential role in new drug development.* J Biomol Struct Dyn, 2011. **28**(6): p. 929-38.
51. Nero, T.L., et al., *Oncogenic protein interfaces: small molecules, big challenges.* Nat Rev Cancer, 2014. **14**(4): p. 248-62.
52. Clackson, T. and J.A. Wells, *A hot spot of binding energy in a hormone-receptor interface.* Science, 1995. **267**(5196): p. 383-6.
53. Hall, D.R., et al., *Lessons from Hot Spot Analysis for Fragment-Based Drug Discovery.* Trends Pharmacol Sci, 2015. **36**(11): p. 724-736.
54. London, N., B. Raveh, and O. Schueler-Furman, *Druggable protein-protein interactions--from hot spots to hot segments.* Curr Opin Chem Biol, 2013. **17**(6): p. 952-9.
55. Wells, J.A. and C.L. McClendon, *Reaching for high-hanging fruit in drug discovery at protein-protein interfaces.* Nature, 2007. **450**(7172): p. 1001-9.
56. Vassilev, L.T., et al., *In vivo activation of the p53 pathway by small-molecule antagonists of MDM2.* Science, 2004. **303**(5659): p. 844-8.
57. Yu, S., et al., *Potent and orally active small-molecule inhibitors of the MDM2-p53 interaction.* J Med Chem, 2009. **52**(24): p. 7970-3.
58. Baudino, T.A., *Targeted Cancer Therapy: The Next Generation of Cancer Treatment.* Curr Drug Discov Technol, 2015. **12**(1): p. 3-20.
59. Adams, G.P. and L.M. Weiner, *Monoclonal antibody therapy of cancer.* Nat Biotechnol, 2005. **23**(9): p. 1147-57.
60. Ross, J., et al., *Antibody-based therapeutics in oncology.* Expert Rev Anticancer Ther, 2003. **3**(1): p. 107-21.
61. Smith, S.L., *Ten years of Orthoclone OKT3 (muromonab-CD3): a review.* J Transpl Coord, 1996. **6**(3): p. 109-19; quiz 120-1.
62. Midtvedt, K., et al., *Individualized T cell monitored administration of ATG versus OKT3 in steroid-resistant kidney graft rejection.* Clin Transplant, 2003. **17**(1): p. 69-74.
63. Harding, F.A., et al., *The immunogenicity of humanized and fully human antibodies: residual immunogenicity resides in the CDR regions.* MAbs, 2010. **2**(3): p. 256-65.
64. Brekke, O.H. and G.A. Loset, *New technologies in therapeutic antibody development.* Curr Opin Pharmacol, 2003. **3**(5): p. 544-50.
65. Singh, S., et al., *Monoclonal Antibodies: A Review.* Curr Clin Pharmacol, 2018. **13**(2): p. 85-99.
66. Chadd, H.E. and S.M. Chamow, *Therapeutic antibody expression technology.* Curr Opin Biotechnol, 2001. **12**(2): p. 188-94.
67. Goldberg, E.P., et al., *Intratumoral cancer chemotherapy and immunotherapy: opportunities for nonsystemic preoperative drug delivery.* J Pharm Pharmacol, 2002. **54**(2): p. 159-80.
68. Drews, J., *Drug discovery: a historical perspective.* Science, 2000. **287**(5460): p. 1960-4.
69. Meldrum, N.U. and F.J. Roughton, *Carbonic anhydrase. Its preparation and properties.* J Physiol, 1933. **80**(2): p. 113-42.
70. Schwartz, W.B., *The effect of sulfanilamide on salt and water excretion in congestive heart failure.* N Engl J Med, 1949. **240**(5): p. 173-7.
71. Lavanya V, M.A.A.A., Neesar Ahmed, Arun K. Rishi, Shazia Jamal, *Small molecule inhibitors as emerging cancer therapeutics.* Integrative Cancer Science and Therapeutics, 2014. **1**(3): p. 39-46.
72. Coussens, N.P., et al., *Small-Molecule Screens: A Gateway to Cancer Therapeutic Agents with Case Studies of Food and Drug Administration–Approved Drugs.* Pharmacological Reviews, 2017. **69**: p. 479-496.

73. Anderson, A.C., *The process of structure-based drug design.* Chem Biol, 2003. **10**(9): p. 787-97.
74. Broach, J.R. and J. Thorner, *High-throughput screening for drug discovery.* Nature, 1996. **384**(6604 Suppl): p. 14-6.
75. Bleicher, K.H., et al., *Hit and lead generation: beyond high-throughput screening.* Nat Rev Drug Discov, 2003. **2**(5): p. 369-78.
76. Brazil, R. *Navigating Drug Discovery with High-Throughput Screening*. Technology Networks, 2018.
77. Hevener, K.E., et al., *Hit-to-Lead: Hit Validation and Assessment.* Methods Enzymol, 2018. **610**: p. 265-309.
78. Mayr, L.M. and P. Fuerst, *The future of high-throughput screening.* J Biomol Screen, 2008. **13**(6): p. 443-8.
79. Gribbon, P., et al., *Evaluating real-life high-throughput screening data.* J Biomol Screen, 2005. **10**(2): p. 99-107.
80. Zhang, J.H., T.D. Chung, and K.R. Oldenburg, *A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays.* J Biomol Screen, 1999. **4**(2): p. 67-73.
81. Shoichet, B.K., *Screening in a spirit haunted world.* Drug Discov Today, 2006. **11**(13-14): p. 607-15.
82. Ciulli, A., *Biophysical screening for the discovery of small-molecule ligands.* Methods Mol Biol, 2013. **1008**: p. 357-88.
83. *Journal of Leukemia*. 2018; Available from: https://www.omicsonline.org/leukemia.php.
84. Kato, M. and A. Manabe, *Treatment and biology of pediatric acute lymphoblastic leukemia.* Pediatr Int, 2018. **60**(1): p. 4-12.
85. Hallek, M., T.D. Shanafelt, and B. Eichhorst, *Chronic lymphocytic leukaemia.* Lancet, 2018. **391**(10129): p. 1524-1537.
86. Hunger, S.P. and C.G. Mullighan, *Acute Lymphoblastic Leukemia in Children.* N Engl J Med, 2015. **373**(16): p. 1541-52.
87. Dohner, H., D.J. Weisdorf, and C.D. Bloomfield, *Acute Myeloid Leukemia.* N Engl J Med, 2015. **373**(12): p. 1136-52.
88. Dohner, H., et al., *Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel.* Blood, 2017. **129**(4): p. 424-447.
89. *Leukemia - Acute Myeloid - AML: Statistics*. Leukemia - Acute Myeloid - AML 2018; Available from: https://www.cancer.net/cancer-types/leukemia-acute-myeloid-aml/statistics.
90. P C Nowell, D.A.H., *A minute chromosome in human Chronic Granulocytic leukemia.* Science, 1960. **132**.
91. Groffen, J., et al., *c-sis is translocated from chromosome 22 to chromosome 9 in chronic myelocytic leukemia.* J Exp Med, 1983. **158**(1): p. 9-15.
92. Pray, L.A., *Gleevec: the breakthrough in cancer treatment.* Nature Education, 2008. **1**(1).
93. Short, N.J., M.E. Rytting, and J.E. Cortes, *Acute myeloid leukaemia.* Lancet, 2018. **392**(10147): p. 593-606.
94. Ding, L., et al., *Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing.* Nature, 2012. **481**(7382): p. 506-10.
95. Nakano, Y., et al., *Molecular evolution of acute myeloid leukaemia in relapse: unstable N-ras and FLT3 genes compared with p53 gene.* Br J Haematol, 1999. **104**(4): p. 659-64.
96. Craddock, C., et al., *Biology and management of relapsed acute myeloid leukaemia.* Br J Haematol, 2005. **129**(1): p. 18-34.

97. Laing, A.A., et al., *Unlocking the potential of anti-CD33 therapy in adult and childhood acute myeloid leukemia.* Exp Hematol, 2017. **54**: p. 40-50.
98. Macauley, M.S., P.R. Crocker, and J.C. Paulson, *Siglec-mediated regulation of immune cell function in disease.* Nat Rev Immunol, 2014. **14**(10): p. 653-66.
99. Krupka, C., et al., *CD33 target validation and sustained depletion of AML blasts in long-term cultures by the bispecific T-cell-engaging antibody AMG 330.* Blood, 2014. **123**(3): p. 356-65.
100. Ehninger, A., et al., *Distribution and levels of cell surface expression of CD33 and CD123 in acute myeloid leukemia.* Blood Cancer J, 2014. **4**: p. e218.
101. Jandus, C., H.U. Simon, and S. von Gunten, *Targeting siglecs--a novel pharmacological strategy for immuno- and glycotherapy.* Biochem Pharmacol, 2011. **82**(4): p. 323-32.
102. Audran, R., et al., *Internalization of human macrophage surface antigens induced by monoclonal antibodies.* J Immunol Methods, 1995. **188**(1): p. 147-54.
103. Angata, T., C.M. Nycholat, and M.S. Macauley, *Therapeutic Targeting of Siglecs using Antibody- and Glycan-Based Approaches.* Trends Pharmacol Sci, 2015. **36**(10): p. 645-660.
104. Laszlo, G.S., E.H. Estey, and R.B. Walter, *The past and future of CD33 as therapeutic target in acute myeloid leukemia.* Blood Rev, 2014. **28**(4): p. 143-53.
105. Feldman, E.J., et al., *Phase III randomized multicenter study of a humanized anti-CD33 monoclonal antibody, lintuzumab, in combination with chemotherapy, versus chemotherapy alone in patients with refractory or first-relapsed acute myeloid leukemia.* J Clin Oncol, 2005. **23**(18): p. 4110-6.
106. Williams, B.A., et al., *Antibody Therapies for Acute Myeloid Leukemia: Unconjugated, Toxin-Conjugated, Radio-Conjugated and Multivalent Formats.* J Clin Med, 2019. **8**(8).
107. Jurcic, J.G., *Androgen Maintenance Therapy for Acute Myeloid Leukemia.* J Clin Oncol, 2017. **35**(4): p. 381-383.
108. Engel, P., et al., *CD Nomenclature 2015: Human Leukocyte Differentiation Antigen Workshops as a Driving Force in Immunology.* J Immunol, 2015. **195**(10): p. 4555-63.
109. Kelm, S. and R. Ravindran, *3.24 - Siglecs*, in *Comprehensive Glycoscience*, H. Kamerling, Editor. 2007, Elsevier: Oxford. p. 523-538.
110. Bornhofft, K.F., et al., *Siglecs: A journey through the evolution of sialic acid-binding immunoglobulin-type lectins.* Dev Comp Immunol, 2018. **86**: p. 219-231.
111. Zhou, J.Y., et al., *The Glycoscience of Immunity.* Trends Immunol, 2018. **39**(7): p. 523-535.
112. Varki, A., Schauer, R., *Sialic Acids*, in *Essentials of Glycobiology*, C.R. Varki A, Esko JD, et al., Editor. 2009, Cold Spring Harbour Laboratory Press: Cold Spring Harbour, NY.
113. Varki, N.M. and A. Varki, *Diversity in cell surface sialic acid presentations: implications for biology and disease.* Lab Invest, 2007. **87**(9): p. 851-7.
114. Cornish, A.L., et al., *Characterization of siglec-5, a novel glycoprotein expressed on myeloid cells related to CD33.* Blood, 1998. **92**(6): p. 2123-32.
115. Nicoll, G., et al., *Identification and characterization of a novel siglec, siglec-7, expressed by human natural killer cells and monocytes.* J Biol Chem, 1999. **274**(48): p. 34089-95.
116. Kikly, K.K., et al., *Identification of SAF-2, a novel siglec expressed on eosinophils, mast cells, and basophils.* J Allergy Clin Immunol, 2000. **105**(6 Pt 1): p. 1093-100.
117. Crocker, P.R., *Siglecs: sialic-acid-binding immunoglobulin-like lectins in cell-cell interactions and signalling.* Curr Opin Struct Biol, 2002. **12**(5): p. 609-15.
118. Varki, A. and T. Angata, *Siglecs--the major subfamily of I-type lectins.* Glycobiology, 2006. **16**(1): p. 1R-27R.
119. Orr, S.J., et al., *CD33 responses are blocked by SOCS3 through accelerated proteasomal-mediated turnover.* Blood, 2007. **109**(3): p. 1061-8.

120. Brinkman-Van der Linden, E.C., et al., *Human-specific expression of Siglec-6 in the placenta.* Glycobiology, 2007. **17**(9): p. 922-31.
121. Wang, X., et al., *Expression of Siglec-11 by human and chimpanzee ovarian stromal cells, with uniquely human ligands: implications for human ovarian physiology and pathology.* Glycobiology, 2011. **21**(8): p. 1038-48.
122. Schwardt, O., S. Kelm, and B. Ernst, *SIGLEC-4 (MAG) Antagonists: From the Natural Carbohydrate Epitope to Glycomimetics.* Top Curr Chem, 2015. **367**: p. 151-200.
123. Crocker, P.R., J.C. Paulson, and A. Varki, *Siglecs and their roles in the immune system.* Nat Rev Immunol, 2007. **7**(4): p. 255-66.
124. Walter, R.B., et al., *ITIM-dependent endocytosis of CD33-related Siglecs: role of intracellular domain, tyrosine phosphorylation, and the tyrosine phosphatases, Shp1 and Shp2.* J Leukoc Biol, 2008. **83**(1): p. 200-11.
125. Miles, L.A., et al., *Small Molecule Binding to Alzheimer Risk Factor CD33 Promotes Abeta Phagocytosis.* iScience, 2019. **19**: p. 110-118.
126. Sgroi, D., A. Nocks, and I. Stamenkovic, *A single N-linked glycosylation site is implicated in the regulation of ligand recognition by the I-type lectins CD22 and CD33.* J Biol Chem, 1996. **271**(31): p. 18803-9.
127. Freeman, S., et al., *A comparative study of the asparagine-linked oligosaccharides on siglec-5, siglec-7 and siglec-8, expressed in a CHO cell line, and their contribution to ligand recognition.* Eur J Biochem, 2001. **268**(5): p. 1228-37.
128. Mortland, L., et al., *Clinical significance of CD33 nonsynonymous single-nucleotide polymorphisms in pediatric patients with acute myeloid leukemia treated with gemtuzumab-ozogamicin-containing chemotherapy.* Clin Cancer Res, 2013. **19**(6): p. 1620-7.
129. Imai, K. and S. Mitaku, *Mechanisms of secondary structure breakers in soluble proteins.* Biophysics (Nagoya-shi), 2005. **1**: p. 55-65.
130. Coutsias, E.A., C. Seok, and K.A. Dill, *Using quaternions to calculate RMSD.* J Comput Chem, 2004. **25**(15): p. 1849-57.
131. Gonzalgo, M.L. and W.B. Isaacs, *Molecular pathways to prostate cancer.* J Urol, 2003. **170**(6 Pt 1): p. 2444-52.
132. Jemal, A., et al., *Cancer statistics, 2006.* CA Cancer J Clin, 2006. **56**(2): p. 106-30.
133. Schulz, W.A., M. Burchardt, and M.V. Cronauer, *Molecular biology of prostate cancer.* Mol Hum Reprod, 2003. **9**(8): p. 437-48.
134. *Cancer compendium: information and trends by cancer type*. 2018 2018 [cited 2018; Available from: https://www.aihw.gov.au/reports/cancer/cancer-compendium-information-trends-by-cancer/report-contents/prostate-cancer.
135. Australia, C. *Prostate Cancer*. 2020; Available from: https://canceraustralia.gov.au/affected-cancer/cancer-types/prostate-cancer/statistics.
136. Frankel, S., et al., *Screening for prostate cancer.* Lancet, 2003. **361**(9363): p. 1122-8.
137. Bax, M., et al., *Dendritic cell maturation results in pronounced changes in glycan expression affecting recognition by siglecs and galectins.* J Immunol, 2007. **179**(12): p. 8216-24.
138. Pokorny, M., et al., *MRI-guided in-bore biopsy for prostate cancer: what does the evidence say? A case series of 554 patients and a review of the current literature.* World J Urol, 2018.
139. Sohail, A., et al., *Role of key players in paradigm shifts of prostate cancer bone metastasis.* Cancer Manag Res, 2018. **10**: p. 1619-1626.

140. Bianchini, D., et al., *Antitumour activity of enzalutamide (MDV3100) in patients with metastatic castration-resistant prostate cancer (CRPC) pre-treated with docetaxel and abiraterone.* Eur J Cancer, 2014. **50**(1): p. 78-84.
141. Ferraldeschi, R., et al., *PTEN protein loss and clinical outcome from castration-resistant prostate cancer treated with abiraterone acetate.* Eur Urol, 2015. **67**(4): p. 795-802.
142. Efesoy, O., et al., *Complications of transrectal ultrasound-guided 12-core prostate biopsy: a single center experience with 2049 patients.* Turk J Urol, 2013. **39**(1): p. 6-11.
143. Silver, D.A., et al., *Prostate-specific membrane antigen expression in normal and malignant human tissues.* Clin Cancer Res, 1997. **3**(1): p. 81-5.
144. Raff, A.B., A. Gray, and W.M. Kast, *Prostate stem cell antigen: a prospective therapeutic and diagnostic target.* Cancer Lett, 2009. **277**(2): p. 126-32.
145. Hinkle, G.H., et al., *Prostate cancer abdominal metastases detected with indium-111 capromab pendetide.* J Nucl Med, 1998. **39**(4): p. 650-2.
146. Slovin, S.F., *Targeting novel antigens for prostate cancer treatment: focus on prostate-specific membrane antigen.* Expert Opin Ther Targets, 2005. **9**(3): p. 561-70.
147. Sharifi, N., et al., *HER2 gene amplification in patients with prostate cancer: Evaluating a CISH-based method.* Oncol Lett, 2016. **12**(6): p. 4651-4658.
148. Ziada, A., et al., *The use of trastuzumab in the treatment of hormone refractory prostate cancer; phase II trial.* Prostate, 2004. **60**(4): p. 332-7.
149. Lara, P.N., Jr., et al., *Trastuzumab plus docetaxel in HER-2/neu-positive prostate carcinoma: final results from the California Cancer Consortium Screening and Phase II Trial.* Cancer, 2004. **100**(10): p. 2125-31.
150. Sandvig, K. and A. Llorente, *Proteomic analysis of microvesicles released by the human prostate cancer cell line PC-3.* Mol Cell Proteomics, 2012. **11**(7): p. M111 012914.
151. Ang, J., et al., *CD151 protein expression predicts the clinical outcome of low-grade primary prostate cancer better than histologic grading: a new prognostic indicator?* Cancer Epidemiol Biomarkers Prev, 2004. **13**(11 Pt 1): p. 1717-21.
152. Copeland, B.T., M.J. Bowman, and L.K. Ashman, *Genetic ablation of the tetraspanin CD151 reduces spontaneous metastatic spread of prostate cancer in the TRAMP model.* Mol Cancer Res, 2013. **11**(1): p. 95-105.
153. Goel, H.L., et al., *Integrin signaling aberrations in prostate cancer.* Am J Transl Res, 2009. **1**(3): p. 211-20.
154. Goel, H.L., et al., *Integrins in prostate cancer progression.* Endocr Relat Cancer, 2008. **15**(3): p. 657-64.
155. Knudsen, B.S. and C.K. Miranti, *The impact of cell adhesion changes on proliferation and survival during prostate cancer development and progression.* J Cell Biochem, 2006. **99**(2): p. 345-61.
156. Sroka, I.C., et al., *The laminin binding integrin alpha6beta1 in prostate cancer perineural invasion.* J Cell Physiol, 2010. **224**(2): p. 283-8.
157. Hunter, K. and A.F. Chambers, *Introduction to special issue on 'epigenetic and physiological regulation of metastasis'.* Clin Exp Metastasis, 2009. **26**(1): p. 3-4.
158. van der Flier, A. and A. Sonnenberg, *Function and interactions of integrins.* Cell Tissue Res, 2001. **305**(3): p. 285-98.
159. Davis, T.L., et al., *Unique expression pattern of the alpha6beta4 integrin and laminin-5 in human prostate carcinoma.* Prostate, 2001. **46**(3): p. 240-8.
160. Bonkhoff, H., U. Stein, and K. Remberger, *Differential expression of alpha 6 and alpha 2 very late antigen integrins in the normal, hyperplastic, and neoplastic prostate: simultaneous demonstration of cell surface receptors and their extracellular ligands.* Hum Pathol, 1993. **24**(3): p. 243-8.

161. Perrais, M., et al., *E-cadherin homophilic ligation inhibits cell growth and epidermal growth factor receptor signaling independently of other cell interactions.* Mol Biol Cell, 2007. **18**(6): p. 2013-25.
162. Nagle, R.B., et al., *Adhesion molecules, extracellular matrix, and proteases in prostate carcinoma.* J Cell Biochem Suppl, 1994. **19**: p. 232-7.
163. Nishiuchi, R., et al., *Ligand-binding specificities of laminin-binding integrins: a comprehensive survey of laminin-integrin interactions using recombinant alpha3beta1, alpha6beta1, alpha7beta1 and alpha6beta4 integrins.* Matrix Biol, 2006. **25**(3): p. 189-97.
164. Ramovs, V., L. Te Molder, and A. Sonnenberg, *The opposing roles of laminin-binding integrins in cancer.* Matrix Biol, 2017. **57-58**: p. 213-243.
165. Hao, J., et al., *Investigation into the mechanism of the loss of laminin 5 (alpha3beta3gamma2) expression in prostate cancer.* Am J Pathol, 2001. **158**(3): p. 1129-35.
166. Schmelz, M., et al., *Different phenotypes in human prostate cancer: alpha6 or alpha3 integrin in cell-extracellular adhesion sites.* Neoplasia, 2002. **4**(3): p. 243-54.
167. Pouliot, N. and N. Kusuma, *Laminin-511: a multi-functional adhesion protein regulating cell migration, tumor invasion and metastasis.* Cell Adh Migr, 2013. **7**(1): p. 142-9.
168. Vences-Catalan, F. and S. Levy, *Immune Targeting of Tetraspanins Involved in Cell Invasion and Metastasis.* Front Immunol, 2018. **9**: p. 1277.
169. Hemler, M.E., *Integrin associated proteins.* Curr Opin Cell Biol, 1998. **10**(5): p. 578-85.
170. Yauch, R.L., et al., *Highly stoichiometric, stable, and specific association of integrin alpha3beta1 with CD151 provides a major link to phosphatidylinositol 4-kinase, and may regulate cell migration.* Mol Biol Cell, 1998. **9**(10): p. 2751-65.
171. Yauch, R.L., et al., *Direct extracellular contact between integrin alpha(3)beta(1) and TM4SF protein CD151.* J Biol Chem, 2000. **275**(13): p. 9230-8.
172. Berditchevski, F., et al., *Analysis of the CD151-alpha3beta1 integrin and CD151-tetraspanin interactions by mutagenesis.* J Biol Chem, 2001. **276**(44): p. 41165-74.
173. Kazarov, A.R., et al., *An extracellular site on tetraspanin CD151 determines alpha 3 and alpha 6 integrin-dependent cellular morphology.* J Cell Biol, 2002. **158**(7): p. 1299-309.
174. Zevian, S., N.E. Winterwood, and C.S. Stipp, *Structure-function analysis of tetraspanin CD151 reveals distinct requirements for tumor cell behaviors mediated by alpha3beta1 versus alpha6beta4 integrin.* J Biol Chem, 2011. **286**(9): p. 7496-506.
175. Ke, A.W., et al., *Generation and characterization of a tetraspanin CD151/integrin alpha6beta1-binding domain competitively binding monoclonal antibody for inhibition of tumor progression in HCC.* Oncotarget, 2016. **7**(5): p. 6314-22.
176. Kim, J.H., et al., *Anti-invasion and anti-migration effects of miR-199a-3p in hepatocellular carcinoma are due in part to targeting CD151.* Int J Oncol, 2016. **49**(5): p. 2037-2045.
177. Oren, R., et al., *TAPA-1, the target of an antiproliferative antibody, defines a new family of transmembrane proteins.* Mol Cell Biol, 1990. **10**(8): p. 4007-15.
178. Ashman, L.K., et al., *The murine monoclonal antibody, 14A2.H1, identifies a novel platelet surface antigen.* Br J Haematol, 1991. **79**(2): p. 263-70.
179. Testa, J.E., et al., *Eukaryotic expression cloning with an antimetastatic monoclonal antibody identifies a tetraspanin (PETA-3/CD151) as an effector of human tumor cell migration and metastasis.* Cancer Res, 1999. **59**(15): p. 3812-20.
180. Kohno, M., et al., *CD151 enhances cell motility and metastasis of cancer cells in the presence of focal adhesion kinase.* Int J Cancer, 2002. **97**(3): p. 336-43.

181. Zijlstra, A., et al., *The inhibition of tumor cell intravasation and subsequent metastasis via regulation of in vivo tumor cell motility by the tetraspanin CD151.* Cancer Cell, 2008. **13**(3): p. 221-34.
182. Nishiuchi, R., et al., *Potentiation of the ligand-binding activity of integrin alpha3beta1 via association with tetraspanin CD151.* Proc Natl Acad Sci U S A, 2005. **102**(6): p. 1939-44.
183. Fitter, S., et al., *Transmembrane 4 superfamily protein CD151 (PETA-3) associates with beta 1 and alpha IIb beta 3 integrins in haemopoietic cell lines and modulates cell-cell adhesion.* Biochem J, 1999. **338 (Pt 1)**: p. 61-70.
184. Sincock, P.M., G. Mayrhofer, and L.K. Ashman, *Localization of the transmembrane 4 superfamily (TM4SF) member PETA-3 (CD151) in normal human tissues: comparison with CD9, CD63, and alpha5beta1 integrin.* J Histochem Cytochem, 1997. **45**(4): p. 515-25.
185. Karamatic Crew, V., et al., *CD151, the first member of the tetraspanin (TM4) superfamily detected on erythrocytes, is essential for the correct assembly of human basement membranes in kidney and skin.* Blood, 2004. **104**(8): p. 2217-23.
186. Sachs, N., et al., *Kidney failure in mice lacking the tetraspanin CD151.* J Cell Biol, 2006. **175**(1): p. 33-9.
187. Baleato, R.M., et al., *Deletion of CD151 results in a strain-dependent glomerular disease due to severe alterations of the glomerular basement membrane.* Am J Pathol, 2008. **173**(4): p. 927-37.
188. Wright, M.D., et al., *Characterization of mice lacking the tetraspanin superfamily member CD151.* Mol Cell Biol, 2004. **24**(13): p. 5978-88.
189. TOISTER, E. and I. Shachar, *Compositions and methods for treating inflammation*. 2012, Google Patents.
190. Haeuw, J.-f.B., FR), Goetsch, Liliane (Ayze, FR), *Use of an anti-CD151 antibody in the treatment of cancer*. 2012, Pierre Fabre Medicament (Boulogne Billancourt, FR): United States.
191. Hopkins, A.L. and C.R. Groom, *The druggable genome.* Nat Rev Drug Discov, 2002. **1**(9): p. 727-30.
192. Russ, A.P. and S. Lampel, *The druggable genome: an update.* Drug Discov Today, 2005. **10**(23-24): p. 1607-10.
193. Garcia-Espana, A., et al., *Appearance of new tetraspanin genes during vertebrate evolution.* Genomics, 2008. **91**(4): p. 326-34.
194. Metzelaar, M.J., et al., *CD63 antigen. A novel lysosomal membrane glycoprotein, cloned by a screening procedure for intracellular antigens in eukaryotic cells.* J Biol Chem, 1991. **266**(5): p. 3239-45.
195. Stipp, C.S., T.V. Kolesnikova, and M.E. Hemler, *Functional domains in tetraspanin proteins.* Trends Biochem Sci, 2003. **28**(2): p. 106-12.
196. Seigneuret, M., et al., *Structure of the tetraspanin main extracellular domain. A partially conserved fold with a structurally variable domain insertion.* J Biol Chem, 2001. **276**(43): p. 40055-64.
197. Boucheix, C. and E. Rubinstein, *Tetraspanins.* Cell Mol Life Sci, 2001. **58**(9): p. 1189-205.
198. Boucheix, C., et al., *Tetraspanins and malignancy.* Expert Rev Mol Med, 2001. **2001**: p. 1-17.
199. Shoham, T., et al., *Building of the tetraspanin web: distinct structural domains of CD81 function in different cellular compartments.* Mol Cell Biol, 2006. **26**(4): p. 1373-85.
200. Kitadokoro, K., et al., *CD81 extracellular domain 3D structure: insight into the tetraspanin superfamily structural motifs.* EMBO J, 2001. **20**(1-2): p. 12-8.
201. Zimmerman, B., et al., *Crystal Structure of a Full-Length Human Tetraspanin Reveals a Cholesterol-Binding Pocket.* Cell, 2016. **167**(4): p. 1041-1051 e11.

202. Levy, S. and T. Shoham, *Protein-protein interactions in the tetraspanin web.* Physiology (Bethesda), 2005. **20**: p. 218-24.
203. Higginbottom, A., et al., *Identification of amino acid residues in CD81 critical for interaction with hepatitis C virus envelope glycoprotein E2.* J Virol, 2000. **74**(8): p. 3642-9.
204. Ryu, F., et al., *Domain analysis of the tetraspanins: studies of CD9/CD63 chimeric molecules on subcellular localization and upregulation activity for diphtheria toxin binding.* Cell Struct Funct, 2000. **25**(5): p. 317-27.
205. Charrin, S., et al., *A physical and functional link between cholesterol and tetraspanins.* Eur J Immunol, 2003. **33**(9): p. 2479-89.
206. Silvie, O., et al., *Cholesterol contributes to the organization of tetraspanin-enriched microdomains and to CD81-dependent infection by malaria sporozoites.* J Cell Sci, 2006. **119**(Pt 10): p. 1992-2002.
207. van Deventer, S.J., V.E. Dunlock, and A.B. van Spriel, *Molecular interactions shaping the tetraspanin web.* Biochem Soc Trans, 2017. **45**(3): p. 741-750.
208. Rubinstein, E., et al., *CD9, CD63, CD81, and CD82 are components of a surface tetraspan network connected to HLA-DR and VLA integrins.* Eur J Immunol, 1996. **26**(11): p. 2657-65.
209. Zhu, G.Z., et al., *Residues SFQ (173-175) in the large extracellular loop of CD9 are required for gamete fusion.* Development, 2002. **129**(8): p. 1995-2002.
210. Kawakami, Y., et al., *Tetraspanin CD9 is a "proteolipid," and its interaction with alpha 3 integrin in microdomain is promoted by GM3 ganglioside, leading to inhibition of laminin-5-dependent cell motility.* J Biol Chem, 2002. **277**(37): p. 34349-58.
211. Berditchevski, F., et al., *A novel link between integrins, transmembrane-4 superfamily proteins (CD63 and CD81), and phosphatidylinositol 4-kinase.* J Biol Chem, 1997. **272**(5): p. 2595-8.
212. Zhang, X.A., A.L. Bontrager, and M.E. Hemler, *Transmembrane-4 superfamily proteins associate with activated protein kinase C (PKC) and link PKC to specific beta(1) integrins.* J Biol Chem, 2001. **276**(27): p. 25005-13.
213. Sugiura, T. and F. Berditchevski, *Function of alpha3beta1-tetraspanin protein complexes in tumor cell invasion. Evidence for the role of the complexes in production of matrix metalloproteinase 2 (MMP-2).* J Cell Biol, 1999. **146**(6): p. 1375-89.
214. Chien, C.W., et al., *Regulation of CD151 by hypoxia controls cell adhesion and metastasis in colorectal cancer.* Clin Cancer Res, 2008. **14**(24): p. 8043-51.
215. Yoo, S.H., et al., *CD151 expression can predict cancer progression in clear cell renal cell carcinoma.* Histopathology, 2011. **58**(2): p. 191-7.
216. Zoller, M., *Tetraspanins: push and pull in suppressing and promoting metastasis.* Nat Rev Cancer, 2009. **9**(1): p. 40-55.
217. Wang, H.X., et al., *Tetraspanin protein contributions to cancer.* Biochem Soc Trans, 2011. **39**(2): p. 547-52.
218. Hasegawa, M., et al., *CD151 dynamics in carcinoma-stroma interaction: integrin expression, adhesion strength and proteolytic activity.* Lab Invest, 2007. **87**(9): p. 882-92.
219. Sterk, L.M., et al., *Association of the tetraspanin CD151 with the laminin-binding integrins alpha3beta1, alpha6beta1, alpha6beta4 and alpha7beta1 in cells in culture and in vivo.* J Cell Sci, 2002. **115**(Pt 6): p. 1161-73.
220. Serru, V., et al., *Selective tetraspan-integrin complexes (CD81/alpha4beta1, CD151/alpha3beta1, CD151/alpha6beta1) under conditions disrupting tetraspan interactions.* Biochem J, 1999. **340 (Pt 1)**: p. 103-11.

221. Sterk, L.M., et al., *The tetraspan molecule CD151, a novel constituent of hemidesmosomes, associates with the integrin alpha6beta4 and may regulate the spatial organization of hemidesmosomes.* J Cell Biol, 2000. **149**(4): p. 969-82.
222. Hynes, R.O., *Integrins: versatility, modulation, and signaling in cell adhesion.* Cell, 1992. **69**(1): p. 11-25.
223. Schwartz, M.A., M.D. Schaller, and M.H. Ginsberg, *Integrins: emerging paradigms of signal transduction.* Annu Rev Cell Dev Biol, 1995. **11**: p. 549-99.
224. Timpl, R., et al., *Laminin--a glycoprotein from basement membranes.* J Biol Chem, 1979. **254**(19): p. 9933-7.
225. Winterwood, N.E., et al., *A critical role for tetraspanin CD151 in alpha3beta1 and alpha6beta4 integrin-dependent tumor cell functions on laminin-5.* Mol Biol Cell, 2006. **17**(6): p. 2707-21.
226. Yamada, M., et al., *The tetraspanin CD151 regulates cell morphology and intracellular signaling on laminin-511.* FEBS J, 2008. **275**(13): p. 3335-51.
227. Yanez-Mo, M., et al., *Regulation of endothelial cell motility by complexes of tetraspan molecules CD81/TAPA-1 and CD151/PETA-3 with alpha3 beta1 integrin localized at endothelial lateral junctions.* J Cell Biol, 1998. **141**(3): p. 791-804.
228. Sincock, P.M., et al., *PETA-3/CD151, a member of the transmembrane 4 superfamily, is localised to the plasma membrane and endocytic system of endothelial cells, associates with multiple integrins and modulates cell function.* J Cell Sci, 1999. **112 (Pt 6)**: p. 833-44.
229. Janes, S.M. and F.M. Watt, *New roles for integrins in squamous-cell carcinoma.* Nat Rev Cancer, 2006. **6**(3): p. 175-83.
230. Tokuhara, T., et al., *Clinical significance of CD151 gene expression in non-small cell lung cancer.* Clin Cancer Res, 2001. **7**(12): p. 4109-14.
231. Yang, X.H., et al., *CD151 accelerates breast cancer by regulating alpha 6 integrin function, signaling, and molecular organization.* Cancer Res, 2008. **68**(9): p. 3204-13.
232. Sadej, R., et al., *Tetraspanin CD151 regulates transforming growth factor beta signaling: implication in tumor metastasis.* Cancer Res, 2010. **70**(14): p. 6059-70.
233. Kwon, M.J., et al., *Clinical significance of CD151 overexpression in subtypes of invasive breast cancer.* Br J Cancer, 2012. **106**(5): p. 923-30.
234. Sadej, R., et al., *CD151 regulates tumorigenesis by modulating the communication between tumor cells and endothelium.* Mol Cancer Res, 2009. **7**(6): p. 787-98.
235. Hashida, H., et al., *Clinical significance of transmembrane 4 superfamily in colon cancer.* Br J Cancer, 2003. **89**(1): p. 158-67.
236. Zhu, G.H., et al., *Expression and prognostic significance of CD151, c-Met, and integrin alpha3/alpha6 in pancreatic ductal adenocarcinoma.* Dig Dis Sci, 2011. **56**(4): p. 1090-8.
237. Ke, A.W., et al., *Role of overexpression of CD151 and/or c-Met in predicting prognosis of hepatocellular carcinoma.* Hepatology, 2009. **49**(2): p. 491-503.
238. Suzuki, S., et al., *Prognostic significance of CD151 expression in esophageal squamous cell carcinoma with aggressive cell proliferation and invasiveness.* Ann Surg Oncol, 2011. **18**(3): p. 888-93.
239. Yauch, R.L. and M.E. Hemler, *Specific interactions among transmembrane 4 superfamily (TM4SF) proteins and phosphoinositide 4-kinase.* Biochem J, 2000. **351 Pt 3**: p. 629-37.
240. Klosek, S.K., et al., *CD151 forms a functional complex with c-Met in human salivary gland cancer cells.* Biochem Biophys Res Commun, 2005. **336**(2): p. 408-16.
241. Franco, M., et al., *The tetraspanin CD151 is required for Met-dependent signaling and tumor cell growth.* J Biol Chem, 2010. **285**(50): p. 38756-64.

242. Shi, G.M., et al., *CD151 modulates expression of matrix metalloproteinase 9 and promotes neoangiogenesis and progression of hepatocellular carcinoma.* Hepatology, 2010. **52**(1): p. 183-96.
243. Shiomi, T., et al., *Pericellular activation of proMMP-7 (promatrilysin-1) through interaction with CD151.* Lab Invest, 2005. **85**(12): p. 1489-506.
244. Gustafson-Wagner, E. and C.S. Stipp, *The CD9/CD81 tetraspanin complex and tetraspanin CD151 regulate alpha3beta1 integrin-dependent tumor cell behaviors by overlapping but distinct mechanisms.* PLoS One, 2013. **8**(4): p. e61834.
245. Whitmore, L. and B.A. Wallace, *Protein secondary structure analyses from circular dichroism spectroscopy: methods and reference databases.* Biopolymers, 2008. **89**(5): p. 392-400.
246. van Stokkum, I.H., et al., *Estimation of protein secondary structure and error analysis from circular dichroism spectra.* Anal Biochem, 1990. **191**(1): p. 110-8.
247. Society, A.C. *About Acute Myeloid Leukemia*. Cancer 2018; Available from: https://www.cancer.org/cancer/acute-myeloid-leukemia/about/key-statistics.html.
248. Wan, H., *An overall comparison of small molecules and large biologics in ADME testing.* ADMET & DMPK, 2016. **4**(1): p. 1-22.
249. Bornhorst, J.A. and J.J. Falke, *Purification of proteins using polyhistidine affinity tags.* Methods Enzymol, 2000. **326**: p. 245-54.
250. Smyth, D.R., et al., *Crystal structures of fusion proteins with large-affinity tags.* Protein Sci, 2003. **12**(7): p. 1313-22.
251. Waugh, D.S., *An overview of enzymatic reagents for the removal of affinity tags.* Protein Expr Purif, 2011. **80**(2): p. 283-93.
252. Healthcare, G., *Biacore Assay Handbook*, ed. G. Lifesciences. 2012.
253. Marquart, A. *SPR-Pages*. 2018; Available from: https://www.sprpages.nl/.
254. Brinkman-Van der Linden, E.C. and A. Varki, *New aspects of siglec binding specificities, including the significance of fucosylation and of the sialyl-Tn epitope. Sialic acid-binding immunoglobulin superfamily lectins.* J Biol Chem, 2000. **275**(12): p. 8625-32.
255. Crocker, P.R., et al., *Siglecs: a family of sialic-acid binding lectins.* Glycobiology, 1998. **8**(2): p. v.
256. Zhuravleva, M.A., K. Trandem, and P.D. Sun, *Structural implications of Siglec-5-mediated sialoglycan recognition.* J Mol Biol, 2008. **375**(2): p. 437-47.
257. Healthcare, G., *Biacore Sensor Surface Handbook*. 2005.
258. Nieba, L., et al., *BIACORE analysis of histidine-tagged proteins using a chelating NTA sensor chip.* Anal Biochem, 1997. **252**(2): p. 217-28.
259. O'Shannessy, D.J., et al., *Detection and quantitation of hexa-histidine-tagged recombinant proteins on western blots and by a surface plasmon resonance biosensor technique.* Anal Biochem, 1995. **229**(1): p. 119-24.
260. Gershon, P.D. and S. Khilko, *Stable chelating linkage for reversible immobilization of oligohistidine tagged proteins in the BIAcore surface plasmon resonance detector.* J Immunol Methods, 1995. **183**(1): p. 65-76.
261. Kimple, A.J., et al., *A capture coupling method for the covalent immobilization of hexahistidine tagged proteins for surface plasmon resonance.* Methods Mol Biol, 2010. **627**: p. 91-100.
262. Attrill, H., et al., *Siglec-7 undergoes a major conformational change when complexed with the alpha(2,8)-disialylganglioside GT1b.* J Biol Chem, 2006. **281**(43): p. 32774-83.
263. Sreerama, N. and R.W. Woody, *Estimation of protein secondary structure from circular dichroism spectra: comparison of CONTIN, SELCON, and CDSSTR methods with an expanded reference set.* Anal Biochem, 2000. **287**(2): p. 252-60.

264. Scientific, T., *Instructions EZ-Link Sulfo-NHS-LC-LC-Biotin*. 2016, Thermo Fischer Scientific Inc: USA.
265. Glaser, R.W., *Antigen-antibody binding and mass transport by convection and diffusion to a surface: a two-dimensional computer model of binding and dissociation kinetics.* Anal Biochem, 1993. **213**(1): p. 152-61.
266. Lifesciences, G., *Biacore T200.* 2013. **Data File 28-9794-15 AB**: p. 7.
267. Rillahan, C.D., et al., *Disubstituted Sialic Acid Ligands Targeting Siglecs CD33 and CD22 Associated with Myeloid Leukaemias and B Cell Lymphomas.* Chem Sci, 2014. **5**(6): p. 2398-2406.
268. Aykul, S. and E. Martinez-Hackert, *Determination of half-maximal inhibitory concentration using biosensor-based protein interaction analysis.* Anal Biochem, 2016. **508**: p. 97-103.
269. Fu, H., *Protein-Protein Interactions: Methods and Applications*. 2004: Springer Science & Business Media. 532.
270. Baldi, P. and R. Nasr, *When is chemical similarity significant? The statistical distribution of chemical similarity scores and its extreme values.* J Chem Inf Model, 2010. **50**(7): p. 1205-22.
271. Willett, P., *Similarity-based virtual screening using 2D fingerprints.* Drug Discov Today, 2006. **11**(23-24): p. 1046-53.
272. Mohabatkar, H., P. Rabiei, and M. Alamdaran, *New Achievements in Bioinformatics Prediction of Post Translational Modification of Proteins.* Curr Top Med Chem, 2017. **17**(21): p. 2381-2392.
273. Jenkins, N., L. Murphy, and R. Tyther, *Post-translational modifications of recombinant proteins: significance for biopharmaceuticals.* Mol Biotechnol, 2008. **39**(2): p. 113-8.
274. Huang, S., et al., *The phylogenetic analysis of tetraspanins projects the evolution of cell-cell interactions from unicellular to multicellular organisms.* Genomics, 2005. **86**(6): p. 674-84.
275. Benz, C., et al., *Viral escape and T-cell immunity during ganciclovir treatment of cytomegalovirus infection: case report of a pancreatico-renal transplant recipient.* Transplantation, 2003. **75**(5): p. 724-7.
276. Daniel, E., et al., *Validated System for Centralized Grading of Retinopathy of Prematurity: Telemedicine Approaches to Evaluating Acute-Phase Retinopathy of Prematurity (e-ROP) Study.* JAMA Ophthalmol, 2015. **133**(6): p. 675-82.
277. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
278. Aschauer, P., et al., *Crystal structure of the Saccharomyces cerevisiae monoglyceride lipase Yju3p.* Biochim Biophys Acta, 2016. **1861**(5): p. 462-70.
279. Sievers, F., et al., *Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.* Mol Syst Biol, 2011. **7**: p. 539.
280. The UniProt, C., *UniProt: the universal protein knowledgebase.* Nucleic Acids Res, 2017. **45**(D1): p. D158-D169.
281. *Transmembrane*. 2018; Available from: https://www.uniprot.org/help/transmem.
282. Puls, K.L., et al., *The molecular characterisation of a novel tetraspanin protein, TM4-B(1).* Biochim Biophys Acta, 1999. **1447**(1): p. 93-9.
283. Wistow, G., et al., *Expressed sequence tag analysis of human RPE/choroid for the NEIBank Project: over 6000 non-redundant transcripts, novel genes and splice variants.* Mol Vis, 2002. **8**: p. 205-20.
284. Kyte, J. and R.F. Doolittle, *A simple method for displaying the hydropathic character of a protein.* J Mol Biol, 1982. **157**(1): p. 105-32.

285. Helenius, A. and M. Aebi, *Intracellular functions of N-linked glycans.* Science, 2001. **291**(5512): p. 2364-9.
286. Fitter, S., et al., *Molecular cloning of cDNA encoding a novel platelet-endothelial cell tetra-span antigen, PETA-3.* Blood, 1995. **86**(4): p. 1348-55.
287. Baldwin, G., et al., *Tetraspanin CD151 regulates glycosylation of (alpha)3(beta)1 integrin.* J Biol Chem, 2008. **283**(51): p. 35445-54.
288. Matuszewski, L., et al., *Assessment of bone marrow angiogenesis in patients with acute myeloid leukemia by using contrast-enhanced MR imaging with clinically approved iron oxides: initial experience.* Radiology, 2007. **242**(1): p. 217-24.
289. Kovalenko, O.V., et al., *Structural organization and interactions of transmembrane domains in tetraspanin proteins.* BMC Struct Biol, 2005. **5**: p. 11.
290. Stuck, M.W., S.M. Conley, and M.I. Naash, *PRPH2/RDS and ROM-1: Historical context, current views and future considerations.* Prog Retin Eye Res, 2016. **52**: p. 47-63.
291. Hemler, M.E., *Tetraspanin proteins mediate cellular penetration, invasion, and fusion events and define a novel type of membrane microdomain.* Annu Rev Cell Dev Biol, 2003. **19**: p. 397-422.
292. Hemler, M.E., *Tetraspanin functions and associated microdomains.* Nat Rev Mol Cell Biol, 2005. **6**(10): p. 801-11.
293. Yang, X., et al., *Palmitoylation supports assembly and function of integrin-tetraspanin complexes.* J Cell Biol, 2004. **167**(6): p. 1231-40.
294. Kitadokoro, K., *[Structural biology of human CD81, a receptor for hepatitis C virus].* Uirusu, 2004. **54**(1): p. 39-47.
295. Parthasarathy, V., et al., *Distinct roles for tetraspanins CD9, CD63 and CD81 in the formation of multinucleated giant cells.* Immunology, 2009. **127**(2): p. 237-48.
296. Bavik, C., et al., *The gene expression program of prostate fibroblast senescence modulates neoplastic epithelial cell proliferation through paracrine mechanisms.* Cancer Res, 2006. **66**(2): p. 794-802.
297. Tarry, M., et al., *Production of human tetraspanin proteins in Escherichia coli.* Protein Expr Purif, 2012. **82**(2): p. 373-9.
298. Yang, X., et al., *Palmitoylation of tetraspanin proteins: modulation of CD151 lateral interactions, subcellular distribution, and integrin-dependent cell morphology.* Mol Biol Cell, 2002. **13**(3): p. 767-81.
299. Basu, J., *Protein palmitoylation and dynamic modulation of protein function.* Current Science, 2004. **87**(2).
300. Ko, P.J. and S.J. Dixon, *Protein palmitoylation and cancer.* EMBO Rep, 2018. **19**(10).
301. Smotrys, J.E. and M.E. Linder, *Palmitoylation of intracellular signaling proteins: regulation and function.* Annu Rev Biochem, 2004. **73**: p. 559-87.
302. Farazi, T.A., G. Waksman, and J.I. Gordon, *The biology and enzymology of protein N-myristoylation.* J Biol Chem, 2001. **276**(43): p. 39501-4.
303. Linder, M.E. and R.J. Deschenes, *Palmitoylation: policing protein stability and traffic.* Nat Rev Mol Cell Biol, 2007. **8**(1): p. 74-84.
304. Sobocinska, J., et al., *Protein Palmitoylation and Its Role in Bacterial and Viral Infections.* Front Immunol, 2017. **8**: p. 2003.
305. Dunphy, J.T. and M.E. Linder, *Signalling functions of protein palmitoylation.* Biochim Biophys Acta, 1998. **1436**(1-2): p. 245-61.
306. Hancock, J.F., et al., *All ras proteins are polyisoprenylated but only some are palmitoylated.* Cell, 1989. **57**(7): p. 1167-77.
307. Naumenko, V.S. and E. Ponimaskin, *Palmitoylation as a Functional Regulator of Neurotransmitter Receptors.* Neural Plast, 2018. **2018**: p. 5701348.

308. Kitadokoro, K., et al., *Crystallization and preliminary crystallographic studies on the large extracellular domain of human CD81, a tetraspanin receptor for hepatitis C virus.* Acta Crystallogr D Biol Crystallogr, 2001. **57**(Pt 1): p. 156-8.
309. Zimmerman, B., et al., *Crystal Structure of a Full-Length Human Tetraspanin Reveals a Cholesterol-Binding Pocket.* Cell, 2016. **167**(4): p. 1041-1051 e11.
310. Jia, X., et al., *Solution structure, membrane interactions, and protein binding partners of the tetraspanin Sm-TSP-2, a vaccine antigen from the human blood fluke Schistosoma mansoni.* J Biol Chem, 2014. **289**(10): p. 7151-63.
311. Chung, S.Y. and S. Subbiah, *A structural explanation for the twilight zone of protein sequence homology.* Structure, 1996. **4**(10): p. 1123-7.
312. Kelley, L.A., et al., *The Phyre2 web portal for protein modeling, prediction and analysis.* Nat Protoc, 2015. **10**(6): p. 845-58.
313. Abraham, R.J., et al., *Charge calculations in molecular mechanics. Part 8. Partial atomic charges from classical calculations.* J Comput Aided Mol Des, 1991. **5**(1): p. 21-39.
314. Laskowski, R.A., et al., *PROCHECK: a program to check the stereochemical quality of protein structures.* Journal of Applied Crystallography, 1993. **26**(2): p. 283-291.
315. Ho, S.H., et al., *Recombinant extracellular domains of tetraspanin proteins are potent inhibitors of the infection of macrophages by human immunodeficiency virus type 1.* J Virol, 2006. **80**(13): p. 6487-96.
316. Shiau, A. and J.M. Smith, *Improved cat gene cassette for promoter analysis and genetic constructions.* Gene, 1988. **67**(2): p. 295-9.
317. Bair, A.E., D. Smith, and L. Lichty, *Intubation confirmation techniques associated with unrecognized non-tracheal intubations by pre-hospital providers.* J Emerg Med, 2005. **28**(4): p. 403-7.
318. Leong, L.E., P.A. Walker, and A.G. Porter, *Human rhinovirus-14 protease 3C (3Cpro) binds specifically to the 5'-noncoding region of the viral RNA. Evidence that 3Cpro has different domains for the RNA binding and proteolytic activities.* J Biol Chem, 1993. **268**(34): p. 25735-9.
319. Scientific, T. *GST tagged proteins - Production and Purification*. Pierce Protein Methods n.d. [cited 2014 16th September]; Available from: https://www.thermofisher.com/au/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/gst-tagged-proteins-production-purification.html#/legacy=www.piercenet.com.
320. Burt, S.M., T.J. Carter, and L.J. Kricka, *Thermal characteristics of microtitre plates used in immunological assays.* J Immunol Methods, 1979. **31**(3-4): p. 231-6.
321. Oliver, D.G., et al., *Thermal gradients in microtitration plates. Effects on enzyme-linked immunoassay.* J Immunol Methods, 1981. **42**(2): p. 195-201.
322. Bolen, D.W. and I.V. Baskakov, *The osmophobic effect: natural selection of a thermodynamic force in protein folding.* J Mol Biol, 2001. **310**(5): p. 955-63.
323. Forneris, F., et al., *ThermoFAD, a Thermofluor-adapted flavin ad hoc detection system for protein folding and ligand binding.* FEBS J, 2009. **276**(10): p. 2833-40.
324. Huynh, T.L., et al., *Fusion and Thermal Degradation Behavior of Symmetric Sulfur-Containing Quaternary Ammonium Bromides.* J Phys Chem B, 2016. **120**(7): p. 1330-5.
325. Artimo P, J.M., Arnold K, Baratin D, Csardi G, de Castro E, Duvaud S, Flegel V, Fortier A, Gasteiger E, Grosdidier A, Hernandez C, Ioannidis V, Kuznetsov D, Liechti R, Moretti S, Mostaguir K, Redaschi N, Rossier G, Xenarios I, and Stockinger H., *ExPASy: SIB bioinformatics resource portal.* Nucleic Acids Res, 2012. **40**(1): p. 597-603.
326. Nagy, P., *Kinetics and Mechanisms of Thiol–Disulfide Exchange Covering Direct Substitution and Thiol Oxidation-Mediated Pathways.* Antioxidants & Redox Signaling, 2013. **18**(13): p. 1623-1641.

327. Gilbert, H.F., *[2] Thiol/disulfide exchange equilibria and disulfidebond stability*, in *Methods in Enzymology*. 1995, Academic Press. p. 8-28.
328. Kramer, Ryan M., et al., *Toward a Molecular Understanding of Protein Solubility: Increased Negative Surface Charge Correlates with Increased Solubility.* Biophysical Journal, 2012. **102**(8): p. 1907-1915.
329. Ferguson, W.J., et al., *Hydrogen ion buffers for biological research.* Analytical Biochemistry, 1980. **104**(2): p. 300-310.
330. Gleason, D., *Tightening the loopholes in newborn identification.* Hospitals, 1966. **40**(15): p. 60-3.
331. Privé, G.G., *Detergents for the stabilization and crystallization of membrane proteins.* Methods, 2007. **41**(4): p. 388-397.
332. Hjelmeland, L.M., *A nondenaturing zwitterionic detergent for membrane biochemistry: design and synthesis.* Proc Natl Acad Sci U S A, 1980. **77**(11): p. 6368-70.
333. Zheng, H., et al., *Data mining of metal ion environments present in protein structures.* Journal of inorganic biochemistry, 2008. **102**(9): p. 1765-1776.
334. Chattopadhyay, A. and K.G. Harikumar, *Dependence of critical micelle concentration of a zwitterionic detergent on ionic strength: implications in receptor solubilization.* FEBS Lett, 1996. **391**(1-2): p. 199-202.
335. Okumura, M., et al., *Acceleration of disulfide-coupled protein folding using glutathione derivatives.* FEBS J, 2011. **278**(7): p. 1137-44.
336. Yamaguchi, H. and M. Miyazaki, *Refolding techniques for recovering biologically active recombinant proteins from inclusion bodies.* Biomolecules, 2014. **4**(1): p. 235-51.
337. Healthcare, G., *Rapid and efficient purification and refolding of a (histidine)6 -tagged recombinant protein produced in E. coli as inclusion bodies.* 2007.
338. Kitadokoro, K., et al., *Subunit association and conformational flexibility in the head subdomain of human CD81 large extracellular loop.* Biol Chem, 2002. **383**(9): p. 1447-52.
339. Prinz, W.A., et al., *The role of the thioredoxin and glutaredoxin pathways in reducing protein disulfide bonds in the Escherichia coli cytoplasm.* J Biol Chem, 1997. **272**(25): p. 15661-7.
340. Pan, J.L. and J.C. Bardwell, *The origami of thioredoxin-like folds.* Protein Sci, 2006. **15**(10): p. 2217-27.
341. Jander, G., N.L. Martin, and J. Beckwith, *Two cysteines in each periplasmic domain of the membrane protein DsbB are required for its function in protein disulfide bond formation.* EMBO J, 1994. **13**(21): p. 5121-7.
342. Kadokura, H., et al., *Snapshots of DsbA in action: detection of proteins in the process of oxidative folding.* Science, 2004. **303**(5657): p. 534-7.
343. McCarthy, A.A., et al., *Crystal structure of the protein disulfide bond isomerase, DsbC, from Escherichia coli.* Nat Struct Biol, 2000. **7**(3): p. 196-9.
344. Stewart, E.J., F. Katzen, and J. Beckwith, *Six conserved cysteines of the membrane protein DsbD are required for the transfer of electrons from the cytoplasm to the periplasm of Escherichia coli.* EMBO J, 1999. **18**(21): p. 5963-71.
345. Ortenberg, R., et al., *Interactions of glutaredoxins, ribonucleotide reductase, and components of the DNA replication system of Escherichia coli.* Proc Natl Acad Sci U S A, 2004. **101**(19): p. 7439-44.
346. Ritz, D., et al., *Conversion of a peroxiredoxin into a disulfide reductase by a triplet repeat expansion.* Science, 2001. **294**(5540): p. 158-60.
347. Stewart, E.J., F. Aslund, and J. Beckwith, *Disulfide bond formation in the Escherichia coli cytoplasm: an in vivo role reversal for the thioredoxins.* EMBO J, 1998. **17**(19): p. 5543-50.

348. Lobstein, J., et al., *SHuffle, a novel Escherichia coli protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm.* Microb Cell Fact, 2012. **11**: p. 56.
349. Lunn, C.A., et al., *Amplification and purification of plasmid-encoded thioredoxin from Escherichia coli K12.* J Biol Chem, 1984. **259**(16): p. 10469-74.
350. LaVallie, E.R., et al., *A thioredoxin gene fusion expression system that circumvents inclusion body formation in the E. coli cytoplasm.* Biotechnology (N Y), 1993. **11**(2): p. 187-93.
351. Ye, J., et al., *Crystal structure of an unusual thioredoxin protein with a zinc finger domain.* J Biol Chem, 2007. **282**(48): p. 34945-51.
352. Heim, R., D.C. Prasher, and R.Y. Tsien, *Wavelength mutations and posttranslational autoxidation of green fluorescent protein.* Proc Natl Acad Sci U S A, 1994. **91**(26): p. 12501-4.
353. Waldo, G.S., et al., *Rapid protein-folding assay using green fluorescent protein.* Nat Biotechnol, 1999. **17**(7): p. 691-5.
354. Cormack, B.P., et al., *Yeast-enhanced green fluorescent protein (yEGFP): a reporter of gene expression in Candida albicans.* Microbiology, 1997. **143 (Pt 2)**: p. 303-11.
355. Suzuki, N., et al., *Crystallization of small proteins assisted by green fluorescent protein.* Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 10): p. 1059-66.
356. Czabotar, P.E., et al., *Bax crystal structures reveal how BH3 domains activate Bax and nucleate its oligomerization to induce apoptosis.* Cell, 2013. **152**(3): p. 519-31.
357. den Blaauwen, T. and A.J. Driessen, *Sec-dependent preprotein translocation in bacteria.* Arch Microbiol, 1996. **165**(1): p. 1-8.
358. Luirink, J. and I. Sinning, *SRP-mediated protein targeting: structure and function revisited.* Biochim Biophys Acta, 2004. **1694**(1-3): p. 17-35.
359. Wagner, S., et al., *Tuning Escherichia coli for membrane protein overexpression.* Proc Natl Acad Sci U S A, 2008. **105**(38): p. 14371-6.
360. Wingfield, P.T., *Overview of the purification of recombinant proteins.* Curr Protoc Protein Sci, 2015. **80**: p. 6 1 1-35.
361. Kapust, R.B. and D.S. Waugh, *Escherichia coli maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused.* Protein Sci, 1999. **8**(8): p. 1668-74.
362. Fox, J.D., et al., *Maltodextrin-binding proteins from diverse bacteria and archaea are potent solubility enhancers.* FEBS Lett, 2003. **537**(1-3): p. 53-7.
363. Fox, J.D. and D.S. Waugh, *Maltose-binding protein as a solubility enhancer.* Methods Mol Biol, 2003. **205**: p. 99-117.
364. Lebendiker, M. and T. Danieli, *Purification of proteins fused to maltose-binding protein.* Methods Mol Biol, 2011. **681**: p. 281-93.
365. Duplay, P., et al., *Sequences of the malE gene and of its product, the maltose-binding protein of Escherichia coli K12.* J Biol Chem, 1984. **259**(16): p. 10606-13.
366. Biolabs, N.E. *pMAL™ protein fusion and purification system*. 2017; Available from: https://www.neb.com/products/e8200-pmal-protein-fusion-and-purification-system#Product%20Information.
367. Cabrita, L.D., et al., *Enhancing the stability and solubility of TEV protease using in silico design.* Protein Sci, 2007. **16**(11): p. 2360-7.
368. Waugh, D.S. *TEV Protease FAQ*. September 2010 [cited 2014; Available from: https://mcl1.ncifcrf.gov/waugh_tech/faq/tev.pdf#search=TEV.
369. Nallamsetty, S., et al., *Efficient site-specific processing of fusion proteins by tobacco vein mottling virus protease in vivo and in vitro.* Protein Expr Purif, 2004. **38**(1): p. 108-15.
370. Chen, J., et al., *Chaperone activity of DsbC.* J Biol Chem, 1999. **274**(28): p. 19601-5.

371. Giglione, C., A. Boularot, and T. Meinnel, *Protein N-terminal methionine excision.* Cell Mol Life Sci, 2004. **61**(12): p. 1455-74.
372. Hirel, P.H., et al., *Extent of N-terminal methionine excision from Escherichia coli proteins is governed by the side-chain length of the penultimate amino acid.* Proc Natl Acad Sci U S A, 1989. **86**(21): p. 8247-51.
373. Moon, A.F., et al., *A synergistic approach to protein crystallization: combination of a fixed-arm carrier with surface entropy reduction.* Protein Sci, 2010. **19**(5): p. 901-13.
374. Dohmen, R.J., *SUMO protein modification.* Biochim Biophys Acta, 2004. **1695**(1-3): p. 113-31.
375. Malakhov, M.P., et al., *SUMO fusions and SUMO-specific protease for efficient expression and purification of proteins.* J Struct Funct Genomics, 2004. **5**(1-2): p. 75-86.
376. Houk, J., R. Singh, and G.M. Whitesides, *Measurement of thiol-disulfide interchange reactions and thiol pKa values.* Methods Enzymol, 1987. **143**: p. 129-40.
377. Held, J.M. and B.W. Gibson, *Regulatory control or oxidative damage? Proteomic approaches to interrogate the role of cysteine oxidation status in biological processes.* Mol Cell Proteomics, 2012. **11**(4): p. R111 013037.
378. Lees, J.G., et al., *A reference database for circular dichroism spectroscopy covering fold and secondary structure space.* Bioinformatics, 2006. **22**(16): p. 1955-62.
379. Bairoch, A., et al., *The Universal Protein Resource (UniProt).* Nucleic Acids Res, 2005. **33**(Database issue): p. D154-9.
380. Combet, C., et al., *NPS@: network protein sequence analysis.* Trends Biochem Sci, 2000. **25**(3): p. 147-50.
381. Michaelevski, I., N. Kirshenbaum, and M. Sharon, *T-wave ion mobility-mass spectrometry: basic experimental procedures for protein complex analysis.* J Vis Exp, 2010(41).
382. Goldenzweig, A., et al., *Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability.* Mol Cell, 2016. **63**(2): p. 337-346.
383. Cole, C.R. and C.A. Smith, *Glycoprotein biochemistry (structure and function) — a vehicle for teaching many aspects of biochemistry and molecular biology.* Biochemical Education, 1989. **17**(4): p. 179-189.
384. Shi, X. and D.L. Jarvis, *Protein N-glycosylation in the baculovirus-insect cell system.* Curr Drug Targets, 2007. **8**(10): p. 1116-25.
385. Druzinec, D., et al., *Optimization of insect cell based protein production processes - online monitoring, expression systems, scale up.* Adv Biochem Eng Biotechnol, 2013. **136**: p. 65-100.
386. Hink, W.F., et al., *Expression of three recombinant proteins using baculovirus vectors in 23 insect cell lines.* Biotechnol Prog, 1991. **7**(1): p. 9-14.
387. Gotoh, T., et al., *Proteolytic activity and recombinant protein production in virus-infected Sf-9 insect cell cultures supplemented with carboxyl and cysteine protease inhibitors.* J Biosci Bioeng, 2001. **92**(3): p. 248-55.
388. Song, J., et al., *PROSPER: an integrated feature-based tool for predicting protease substrate cleavage sites.* PLoS One, 2012. **7**(11): p. e50300.
389. Ikonomou, L., Y.J. Schneider, and S.N. Agathos, *Insect cell culture for industrial production of recombinant proteins.* Appl Microbiol Biotechnol, 2003. **62**(1): p. 1-20.
390. Kwaks, T.H. and A.P. Otte, *Employing epigenetics to augment the expression of therapeutic proteins in mammalian cells.* Trends Biotechnol, 2006. **24**(3): p. 137-42.
391. Haryadi, R., et al., *Optimization of heavy chain and light chain signal peptides for high level expression of therapeutic antibodies in CHO cells.* PLoS One, 2015. **10**(2): p. e0116878.

392. Egea, P.F., R.M. Stroud, and P. Walter, *Targeting proteins to membranes: structure of the signal recognition particle.* Curr Opin Struct Biol, 2005. **15**(2): p. 213-20.
393. Barreiro, O., et al., *Endothelial tetraspanin microdomains regulate leukocyte firm adhesion during extravasation.* Blood, 2005. **105**(7): p. 2852-61.
394. Kozak, M., *An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs.* Nucleic Acids Res, 1987. **15**(20): p. 8125-48.
395. Scheltz, T., J. von Bulow, and E. Beitz, *Reducing isoform complexity of human tetraspanins by optimized expression in Dictyostelium discoideum enables high-throughput functional read-out.* Protein Expr Purif, 2017. **135**: p. 8-15.
396. Skaar, K., et al., *Expression and Subcellular Distribution of GFP-Tagged Human Tetraspanin Proteins in Saccharomyces cerevisiae.* PLoS One, 2015. **10**(7): p. e0134041.
397. Buchner, J., *Supervising the fold: functional principles of molecular chaperones.* FASEB J, 1996. **10**(1): p. 10-9.
398. Dyson, H.J., P.E. Wright, and H.A. Scheraga, *The role of hydrophobic interactions in initiation and propagation of protein folding.* Proc Natl Acad Sci U S A, 2006. **103**(35): p. 13057-61.
399. Gasteiger E., H.C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A, *Protein Identification and Analysis Tools on the ExPASy Server*, in *The Proteomics Protocols Handbook*. 2005, Humana Press. p. 571-607.
400. Neugebauer, A., C.D. Klein, and R.W. Hartmann, *Protein-dynamics of the putative HCV receptor CD81 large extracellular loop.* Bioorg Med Chem Lett, 2004. **14**(7): p. 1765-9.
401. Cunha, E.S., et al., *Mechanism of Structural Tuning of the Hepatitis C Virus Human Cellular Receptor CD81 Large Extracellular Loop.* Structure, 2017. **25**(1): p. 53-65.
402. Petersen, T.N., et al., *SignalP 4.0: discriminating signal peptides from transmembrane regions.* Nat Methods, 2011. **8**(10): p. 785-6.
403. White, S.H., *Translocons, thermodynamics, and the folding of membrane proteins.* FEBS Lett, 2003. **555**(1): p. 116-21.
404. *Recombinant Human CD151 protein Product datasheet*, N. Biologicals, Editor. 2018, Novus Biologicals.
405. Roos, C., et al., *High-Level Cell-Free Production of Membrane Proteins with Nanodiscs*, in *Cell-Free Protein Synthesis: Methods and Protocols*, K. Alexandrov and W.A. Johnston, Editors. 2014, Humana Press: Totowa, NJ. p. 109-130.
406. Beuming, T. and W. Sherman, *Current assessment of docking into GPCR crystal structures and homology models: successes, challenges, and guidelines.* J Chem Inf Model, 2012. **52**(12): p. 3263-77.
407. *PDB Statistics: Overall Growth of Released Structures Per Year*. 2018; Available from: https://www.rcsb.org/stats/growth/overall.
408. Ivaska, J., et al., *Integrin-protein kinase C relationships.* Biochem Soc Trans, 2003. **31**(Pt 1): p. 90-3.
409. Ryu, W.-S., *Molecular Virology of Human Pathogenic Viruses*. 2017: Academic Press. 440.
410. Parker, M.W., *Protein structure from x-ray diffraction.* J Biol Phys, 2003. **29**(4): p. 341-62.
411. Chayen, N.E. and E. Saridakis, *Protein crystallization: from purified protein to diffraction-quality crystal.* Nat Methods, 2008. **5**(2): p. 147-53.
412. McPherson, A. and J.A. Gavira, *Introduction to protein crystallization.* Acta Crystallogr F Struct Biol Commun, 2014. **70**(Pt 1): p. 2-20.
413. *PCT™ Pre-Crystallization Test*. 2018; Available from: https://www.hamptonresearch.com/product_detail.aspx?cid=1&sid=29&pid=10.

414. Watson, A.A. and C.A. O'Callaghan, *Crystallization and X-ray diffraction analysis of human CLEC-2.* Acta Crystallogr Sect F Struct Biol Cryst Commun, 2005. **61**(Pt 12): p. 1094-6.
415. Clifton, M.C., et al., *A Maltose-Binding Protein Fusion Construct Yields a Robust Crystallography Platform for MCL1.* PLoS One, 2015. **10**(4): p. e0125010.
416. Butt, T.R., et al., *SUMO fusion technology for difficult-to-express proteins.* Protein Expr Purif, 2005. **43**(1): p. 1-9.
417. Brzozowski, J.S.A.-O.h.o.o., et al., *Extracellular vesicles with altered tetraspanin CD9 and CD151 levels confer increased prostate cell motility and invasion.* (2045-2322 (Electronic)).
418. Detchokul, S., et al., *CD151 is associated with prostate cancer cell invasion and lymphangiogenesis in vivo.* Oncol Rep, 2014. **31**(1): p. 241-7.
419. Roselli, S., et al., *Deletion of Cd151 reduces mammary tumorigenesis in the MMTV/PyMT mouse model.* BMC Cancer, 2014. **14**: p. 509.
420. Palmer, T.D., et al., *Integrin-free tetraspanin CD151 can inhibit tumor cell motility upon clustering and is a clinical indicator of prostate cancer progression.* Cancer Res, 2014. **74**(1): p. 173-87.
421. PJ, D., *Biologicals and biosimilars: a review of the science and its implications.* Generics and Biosimilars Initiative Journal, 2012. **1**(1): p. 13-16.
422. Blackstone, E.A. and P.F. Joseph, *The economics of biosimilars.* Am Health Drug Benefits, 2013. **6**(8): p. 469-78.
423. Kunathikom, S., B.L. Anderson, and S.A. Collins, Jr., *Design of delay elements in a binary optical true-time-delay device that uses a White cell.* Appl Opt, 2003. **42**(35): p. 6984-94.
424. Bielavitz, S., J. Wisdom, and D.A. Pollack, *Effective mental health consumer education: a preliminary exploration.* J Behav Health Serv Res, 2011. **38**(1): p. 105-13.
425. Doak, B.C., R.S. Norton, and M.J. Scanlon, *The ways and means of fragment-based drug design.* Pharmacol Ther, 2016. **167**: p. 28-37.
426. Sciences, G.H.L., *Biacore T200 Instrument Handbook*. 2013.
427. Chavanieu, A. and M. Pugniere, *Developments in SPR Fragment Screening.* Expert Opin Drug Discov, 2016. **11**(5): p. 489-99.
428. Shepherd, C.A., A.L. Hopkins, and I. Navratilova, *Fragment screening by SPR and advanced application to GPCRs.* Prog Biophys Mol Biol, 2014. **116**(2-3): p. 113-23.
429. Olaru, A., et al., *Surface plasmon resonance (SPR) biosensors in pharmaceutical analysis.* Crit Rev Anal Chem, 2015. **45**(2): p. 97-105.
430. *Monash Fragment Platform*. 2018; Available from: https://www.monash.edu/research/infrastructure/mfp.
431. Baell, J.B. and G.A. Holloway, *New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays.* J Med Chem, 2010. **53**(7): p. 2719-40.
432. *Label-Free technologies For Drug Discovery*. 2011, United Kingdom: Wiley.
433. GmbH, X.b. *Products SPR Sensorchips*. 2018; Available from: https://www.xantec.com/products/spr_sensorchips.php.
434. Hochuli, E., H. Dobeli, and A. Schacher, *New metal chelate adsorbent selective for proteins and peptides containing neighbouring histidine residues.* J Chromatogr, 1987. **411**: p. 177-84.
435. Gui, B., et al., *Assessment of the Albumin-Bilirubin (ALBI) Grade as a Prognostic Indicator for Hepatocellular Carcinoma Patients Treated With Radioembolization.* Am J Clin Oncol, 2017.

436. Zhang, L., et al., *Interactions of Bovine Serum Albumin with Anti-Cancer Compounds Using a ProteOn XPR36 Array Biosensor and Molecular Docking.* Molecules, 2016. **21**(12).
437. Rigotti, N.A., et al., *An international survey of training programs for treating tobacco dependence.* Addiction, 2009. **104**(2): p. 288-96.
438. Rich, R.L., et al., *Biosensor-based fragment screening using FastStep injections.* Anal Biochem, 2010. **407**(2): p. 270-7.
439. Abcam, *Anti-CD151 antibody [11G5a] ab33315*. 2018.
440. Neumann, T., et al., *SPR-based fragment screening: advantages and applications.* Curr Top Med Chem, 2007. **7**(16): p. 1630-42.
441. Navratilova, I. and A.L. Hopkins, *Fragment screening by surface plasmon resonance.* ACS Med Chem Lett, 2010. **1**(1): p. 44-8.
442. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2018.* CA Cancer J Clin, 2018. **68**(1): p. 7-30.
443. Bray, F., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA Cancer J Clin, 2018. **68**(6): p. 394-424.
444. Cagan, R. and P. Meyer, *Rethinking cancer: current challenges and opportunities in cancer research.* Dis Model Mech, 2017. **10**(4): p. 349-352.
445. Wong, C.H., K.W. Siah, and A.W. Lo, *Estimation of clinical trial success rates and related parameters.* Biostatistics, 2018: p. kxx069-kxx069.
446. *Novel Drug Approvals for 2018*. Drug Innovation 2018; Available from: https://www.fda.gov/drugs/developmentapprovalprocess/druginnovation/ucm592464.htm.
447. *NCI Budget and Appropriations*. NCI Office of Budget and Finance 2018; Available from: https://www.cancer.gov/about-nci/budget#current-year.
448. Kola, I. and J. Landis, *Can the pharmaceutical industry reduce attrition rates?* Nat Rev Drug Discov, 2004. **3**(8): p. 711-5.
449. DiMasi, J.A. and H.G. Grabowski, *Economics of new oncology drug development.* J Clin Oncol, 2007. **25**(2): p. 209-16.
450. Giannetti, A.M., *Chapter Eight - From Experimental Design to Validated Hits: A Comprehensive Walk-Through of Fragment Lead Identification Using Surface Plasmon Resonance*, in *Methods in Enzymology*, L.C. Kuo, Editor. 2011, Academic Press. p. 169-218.
451. Myszka, D.G., *Survey of the 1998 optical biosensor literature.* J Mol Recognit, 1999. **12**(6): p. 390-408.
452. Rich, R.L. and D.G. Myszka, *Survey of the 2009 commercial optical biosensor literature.* J Mol Recognit, 2011. **24**(6): p. 892-914.
453. Wielens, J., et al., *Parallel screening of low molecular weight fragment libraries: do differences in methodology affect hit identification?* J Biomol Screen, 2013. **18**(2): p. 147-59.
454. Woods, L.A., et al., *Native State Mass Spectrometry, Surface Plasmon Resonance, and X-ray Crystallography Correlate Strongly as a Fragment Screening Combination.* J Med Chem, 2016. **59**(5): p. 2192-204.
455. Yang, Y., et al., *Open conformation of tetraspanins shapes interaction partner networks on cell membranes.* EMBO J, 2020. **39**(18): p. e105246.
456. Umeda, R., et al., *Structural insights into tetraspanin CD9 function.* Nat Commun, 2020. **11**(1): p. 1606.
457. Susa, K.J., et al., *A dynamic interaction between CD19 and the tetraspanin CD81 controls B cell co-receptor trafficking.* Elife, 2020. **9**.

458. Greenberg, Z.J., et al., *The Tetraspanin CD53 Regulates Early B Cell Development by Promoting IL-7R Signaling.* J Immunol, 2020. **204**(1): p. 58-67.
459. Purushothaman, G. and V. Thiruvenkatam, *High Yield Expression of Recombinant CD151 in E. coli and a Structural Insight into Cholesterol Binding Domain.* Mol Biotechnol, 2019. **61**(12): p. 905-915.
460. *Principles and Guidelines for Reporting Preclinical Research*. 2017; Available from: https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research.
461. *Reproducibility Project: Cancer Biology*. 2018; Available from: https://elifesciences.org/collections/9b1e83d1/reproducibility-project-cancer-biology.
462. *PUBPEER The online journal club*. 2018; Available from: https://blog.pubpeer.com.
463. *Improving and Accelerating Therapeutic Development for Nervous System Disorders: Workshop Summary.* Forum on Neuroscience and Nervous System Disorders. Vol. 2. 2014, Washingto (DC): National Academies Press (US).
464. Seok, J., et al., *Genomic responses in mouse models poorly mimic human inflammatory diseases.* Proc Natl Acad Sci U S A, 2013. **110**(9): p. 3507-12.
465. Mestas, J. and C.C.W. Hughes, *Of Mice and Not Men: Differences between Mouse and Human Immunology.* The Journal of Immunology, 2004. **172**(5): p. 2731.
466. Angata, T., et al., *Cloning and characterization of a novel mouse Siglec, mSiglec-F: differential evolution of the mouse and human (CD33) Siglec-3-related gene clusters.* J Biol Chem, 2001. **276**(48): p. 45128-36.
467. Brinkman-Van der Linden, E.C., et al., *CD33/Siglec-3 binding specificity, expression pattern, and consequences of gene deletion in mice.* Mol Cell Biol, 2003. **23**(12): p. 4199-206.